Submitted to the Annals of Applied Statistics

FOCUSED MODEL SELECTION FOR LINEAR MIXED MODELS, WITH AN APPLICATION TO WHALE ECOLOGY

By Céline Cunen[*], Lars Walløe[†] and Nils Lid Hjort[*]

University of Oslo

A central point of disagreement, in certain long-standing discussions about a particular whaling dataset in the Scientific Committee of the International Whaling Commission, has directly involved model selection issues for linear mixed effect models. The biological question under discussion is associated with a clearly defined parameter of primary interest, i.e. a focus parameter, which makes model selection with the Focused Information Criterion (FIC) more appropriate than other selection methods. Since the existing FIC methodology has not covered the case of linear mixed effects models, this article sets up the required framework and develops the necessary formulae for the relevant FIC. Our new criterion requires the asymptotic distribution of estimators derived for a given candidate linear mixed model, but with behaviour examined under a wider linear mixed model. These results, needed here to build our FIC, also have independent interest.

1. Introduction.

Linear mixed effect (LME) models have become a standard modelling tool for many problems in ecology and evolution (Bolker et al., 2009), as well as in many other fields where clustered or longitudinal data appear. A simple net search yields many papers and books where LME models are used in social and physical sciences, in biology, demography, etc. Ecological data typically exhibit dependencies, due for example to repeated sampling of observations within the same time-unit or within the same space (Grueber et al., 2011). LME models provide a framework for taking such dependencies into account.

As motivation for the methodological work presented in this article, we will study the Antarctic minke whale (*Balaenoptera bonaerensis*) data from the Japanese Whale Research Program under Special Permit in the Antarctic (the so-called JARPA 1, hereafter JARPA, see Government of Japan (1987)). This dataset contains many potential sources of dependencies, a

[*] Department of Mathematics.

[†] Institute of Basic Medical Sciences.

Keywords and phrases: Antarctic minke whale, focused information criterion, FIC plot, linear mixed effects models, model selection

typical characteristics of many ecological data. Crucially, the dataset has found itself at the centre of some long-standing discussions in the Scientific Committee of the International Whaling Commission (IWC-SC), with some of the central questions concerning model selection with LME models. Although few ecological analyses have been subject to critical scrutiny at this level, the challenge of model selection remains of high importance in many biological applications.

The JARPA minke whale dataset contains measurements on a large number of Antarctic minke whales caught over 18 consecutive years, from 1987/88 to 2004/05. One of the main research questions has been whether the body condition of the whales has decreased during the study years. The fat weight of each dissected whale is taken as a proxy for its body condition, and is used as the response variables in the analyses presented here. Several covariates of potential relevance were also recorded, including the year of capture, the date within each year, the sex, the body length, and different spatial covariates. The parameter describing the yearly change in body condition, hereafter referred to as the yearly decline, is the parameter of main interest (which we will call the *focus parameter*; note that in general the focus parameter can be a function of several of the model parameters).

The data were first analysed in Konishi et al. (2008) using linear regression models. These analyses faced criticism in the IWC-SC, mainly concerning the choice of model and the potential sources of dependencies left unaccounted for. In the following years, many amendments and other analyses have been proposed in the IWC, for instance Konishi and Walløe (2015), de la Mare, McKinlay and Welsh (2017), and McKinlay, de la Mare and Welsh (2017), including our own contribution in Cunen, Walløe and Hjort (2017). One of the central points of disagreement has been the choice of model selection criteria for LME models. The number of potential covariates in the dataset leads to a large number of candidate models to choose between. The choice of model can influence the estimate and the standard errors of the parameter of main interest. In addition, one is interested in choosing between models containing the focus parameter or not (an implicit test of the size of the yearly decline). There are many possible choices of model selection criteria for LME models (Müller, Scealy and Welsh, 2013), and the choice between them depends on the type of data and models at hand, and on the goal of the model selection.

Practitioners may be interested in model selection for different, overlapping reasons. On one hand the goal might be to select the candidate model which in a relevant sense is the closest to the true data generating mechanism. Criteria based on model fit and some penalisation for complexity aim at this goal; for LME models there exists a large number of criteria of this type; see references in Section 3. On the other hand, practitioners often seek a small model offering precise estimates of the quantities they are interested in. These two goals are related, but methods tailored for the first goal, may in some instances be sub-optimal or irrelevant for the second. Model selection procedures based on the focused information criteria (FIC) aim at the second goal: selecting models giving the most precise estimates of the quantity that we are interested in, the focus parameter (Claeskens and Hjort, 2003, 2008a). Since the existing FIC methodology did not cover the case of LME models, we have developed a FIC for these models, which is derived and discussed in this article. Crucially, our FIC framework is perfectly suited for addressing one of the main questions in the JARPA dataset: finding a model that estimates the yearly decline with the best precision.

The purpose of FIC is to select a model which minimises the estimated risk associated with the focus parameter. For example, let the focus parameter under scrutiny be the parameter representing a linear relationship between body condition and year, $\beta_{\rm year}$. Each candidate model M leads to an associated estimate $\widehat{\beta}_{M,\rm year}$. These carry mean squared errors, or risks, say

$$\text{mse}_M = \mathrm{E}\,(\widehat{\beta}_{M,\rm year} - \beta_{\rm year})^2 = \mathrm{Var}\,\widehat{\beta}_{M,\rm year} + (\mathrm{E}\,\widehat{\beta}_{M,\rm year} - \beta_{\rm year})^2. \tag{1.1}$$

Here, the expectations and variances are taken with respected to the assumed true data generating mechanism. The FIC scheme is then to estimate each of these risk measures from data,

$$\text{FIC}(M) = \widehat{\text{mse}}_M.$$

This necessitates working out good approximations to biases, variances and covariances, and then constructing estimators for these again. Importantly, the apparatus we develop can of course be applied more generally, to any focus parameter besides the $\beta_{\rm year}$ of the JARPA study.

Our article is structured as follows. In Section 2 we start with an illustration of the use of FIC in a very simple setting. That section provides the necessary intuition to readers not already familiar with FIC. Section 3 sets the basic framework with LME formulation and notation, along with a brief description of existing model selection criteria. Then in Section 4 we examine the behaviour of estimators used for one candidate LME model, but under the assumption that the data generating mechanism is a different LME model. There are several aspects of interest, regarding such consequences of LME model misspecification. The main purpose here, however, is to use these results and insights to develop the FIC approach, in Section 5. In Section 6 we report on relevant simulation experiments. The application of the model selection machinery to the JARPA data, concerned indeed with assessing the mse_M of (1.1) for a list of relevant candidate models, is discussed in Section 7. Finally, Section 8 offers further relevant discussion points and a list of concluding remarks.

2. Simple illustration of FIC. We begin with an illustration of the motivation underlying FIC in a very simplified setting. Here we will consider ordinary linear regression models (which of course are a special case of an LME model). Readers already familiar with FIC may skip this section.

Assume that we have $n = 20$ measurements of for example body condition (y) over time (x), say years, and that we are particularly interested in obtaining a good estimate of the body condition at a particular time point, for instance after 9 years. This quantity is our focus parameter $\mu = \mathrm{E}(Y|x_0 = 9)$. Further, say we have good biological reasons to assume that the following model is the true data generating mechanism, i.e. the wide model,

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i \tag{2.1}$$

with $i = 1, ..., n$ and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. Our focus parameter can then be expressed as $\mu_{\mathrm{wide}} = \mathrm{E}_{\mathrm{wide}}(Y \,|\, x_0 = 9) = \theta^{\mathrm{t}} x_f$, with column vectors $\theta = (\alpha, \beta, \gamma)^{\mathrm{t}}$ and $x_f = (1, x_0, x_0^2)^{\mathrm{t}}$. A natural estimator is $\widehat{\mu}_{\mathrm{wide}} = \widehat{\theta}^{\mathrm{t}} x_f = \widehat{\alpha} + 9\widehat{\beta} + 9^2\widehat{\gamma}$, using ordinary least squares. The idea underlying FIC, and which we hope to illustrate here, is that even though the wide model in (2.1) is the true model, a different model may provide more precise estimates of μ, at least in parts of the parameter space.

Let the candidate model be a smaller model, $y_i = \alpha_M + \beta_M x_i + \epsilon_i$ and $\epsilon_i \sim \mathrm{N}(0, \sigma_M^2)$. Here the natural estimator of the focus parameter is $\widehat{\mu}_M = \widehat{\theta}_M^{\mathrm{t}} x_{f,M} = \widehat{\alpha}_M + 9\widehat{\beta}_M$, with $\theta_M = (\alpha_M, \beta_M)^{\mathrm{t}}$ and $x_{f,M} = (1, x_0)^{\mathrm{t}}$. Clearly, this quantity is in general different from the $\widehat{\mu}_{\mathrm{wide}}$ from the wide model. The smaller model can still serve as an approximation to the truth, producing estimates of μ, which typically will be biased, but can have lower variance.

For the $n = 20$ datapoints in Figure 1 we obtain $\widehat{\mu}_{\mathrm{wide}} = 38.9$ and $\widehat{\mu}_M = 32.3$. Which of these two estimates should we prefer, i.e. should we trust the most? Throughout this paper we will seek to select the estimator, and by extension the model, which estimates the focus parameter with the best precision. We evaluate the precision of the two estimators by considering their risk in terms of mean squared errors,

$$\begin{aligned}
\mathrm{mse}(\widehat{\mu}_{\mathrm{wide}}; \theta, \sigma) &= (\mathrm{E}_{\mathrm{wide}}\, \widehat{\mu}_{\mathrm{wide}} - \mu)^2 + \mathrm{Var}_{\mathrm{wide}}\, \widehat{\mu}_{\mathrm{wide}}, \\
\mathrm{mse}(\widehat{\mu}_M; \theta, \sigma) &= (\mathrm{E}_{\mathrm{wide}}\, \widehat{\mu}_M - \mu)^2 + \mathrm{Var}_{\mathrm{wide}}\, \widehat{\mu}_M.
\end{aligned}$$

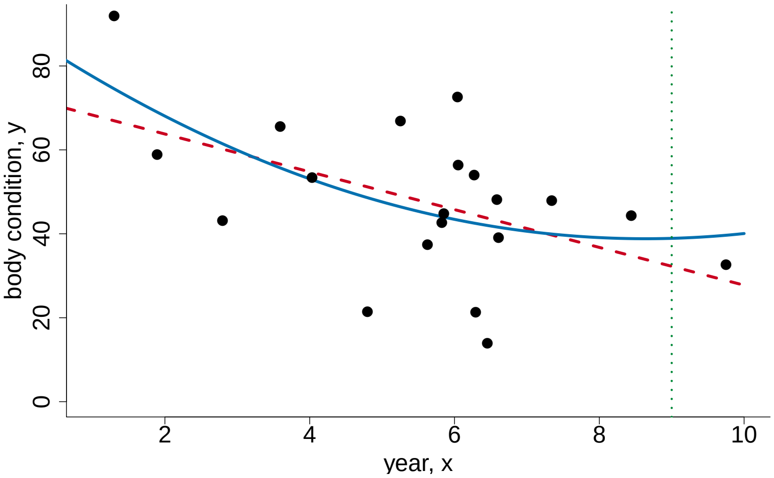

Fig 1. *The $n = 20$ datapoints and two estimated regression lines, for the wide model in blue and for the candidate model in dashed red. The vertical green line marks $x_0 = 9$, the point where we want to evaluate the expected body condition.*

These quantities are functions of θ and σ, the true parameters in the wide model, and also of the sample size n, the observed covariate matrix X and the particular year of interest x_0. The subscript in $\mathrm{E}_{\mathrm{wide}}$ and $\mathrm{Var}_{\mathrm{wide}}$ is meant to emphasise that the expectations and variances are taken with respect to *the true, wide model.* The simplicity of the linear regression models allows us to work out exact risk formulas. Let x be the column vectors of observed years, $X = [1, x, x^2]$ be the $n \times 3$ design matrix in the wide model and $X_M = [1, x]$ be the $n \times 2$ design matrix in the candidate model. Then

$$\begin{aligned}\mathrm{mse}(\widehat{\mu}_{\mathrm{wide}}; \theta, \sigma) &= 0 + \sigma^2 x_f^{\mathrm{t}} (X^{\mathrm{t}} X)^{-1} x_f,\\ \mathrm{mse}(\widehat{\mu}_M; \theta, \sigma) &= (x_{f,M}^{\mathrm{t}} (X_M^{\mathrm{t}} X_M)^{-1} X_M^{\mathrm{t}} X\theta - x_f^{\mathrm{t}}\theta)^2 \\ &\qquad + \sigma^2 x_{f,M}^{\mathrm{t}} (X_M^{\mathrm{t}} X_M)^{-1} x_{f,M}.\end{aligned} \tag{2.2}$$

The estimator from the wide model has zero bias by definition. For the particular dataset shown in Figure 1, where we have drawn a dataset from the wide model with $\alpha = 80, \beta = -10, \gamma = 0.5, \sigma = 15$, the true root mse values are then

$$\sqrt{\mathrm{mse}(\widehat{\mu}_{\mathrm{wide}}; \theta, \sigma)} = 9.00 \quad \text{and} \quad \sqrt{\mathrm{mse}(\widehat{\mu}_M; \theta, \sigma)} = 8.38.$$

So in this case, the estimator from the smaller model is more precise, even though it has some bias (around 5.02).

Since the mse formulas in (2.2) are functions of unknown parameters one needs to estimate their values when faced with a real dataset. In this case, this would essentially amount to plugging in the maximum likelihood estimates of the necessary parameters into (2.2). These estimated mean squared errors are the FIC scores of the two models. For the data in Figure 1 we obtain

$$\sqrt{\mathrm{fic}_{\mathrm{wide}}} = 9.90 \quad \text{and} \quad \sqrt{\mathrm{fic}_M} = 7.44.$$

Note that we usually prefer to present root-fic scores, and root mse, since these quantities are on the scale of the response variable. In this case, the FIC scores are reasonably good estimates of the true mse values and correctly identify the smaller model as the best model for estimating the focus parameter. In simple cases like this one, the mse formulas and their estimates are immediately available from basic statistical knowledge. In general situations, exact formulas are not available and one needs to settle for approximations of the true risk functions, usually obtained through large-sample considerations. The FIC framework described in the following sections offer a way to obtain estimated risks in the case of LME models.

3. Setup and existing approaches.

In this section we provide a short overview of the LME model, along with the necessary notation, and a review of some existing model selection approaches. In the next section we go into how estimators constructed using one LME candidate model behave when the data generating mechanism is a different and perhaps wider LME model.

3.1. *Linear mixed effect models.*

Suppose there are n natural groups in our data, with m_i potentially dependent observations within group i. The groups often correspond to observations collected in the same space or time, and are often referred to as *subjects* or individuals in the statistical literature. In the whaling example, the groups are defined by the year of sampling (more on this in Section 7). The general LME model takes the form

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n. \tag{3.1}$$

with y_i an $m_i \times 1$ vector of responses for the ith group, X_i a known $m_i \times p$ design matrix of covariates, and Z_i a known $m_i \times k$ design matrix corresponding to the random effects. Let β be the $p \times 1$ vector of fixed effect coefficients, and b_i the group-specific $k \times 1$ vector of random effects, with $b_i \sim \mathrm{N}_k(0, \sigma^2 D)$ and assumed independence across groups. The errors ϵ_i are independently distributed $\mathrm{N}_{m_i}(0, \sigma^2 I)$. In applications it is common to choose Z_i equal to a subset of the columns in X_i, but the following developments are valid for any choice of Z_i. The full parameter of the model is $\theta = (\beta, \sigma, D)$, with p fixed effects, k random effects, and $k(k+1)/2$ separate parameters for the symmetric positive definite $k \times k$ matrix D, yielding a total of $d = p + 1 + k(k+1)/2$ unknown parameters.

The model can also be written in a marginal form, as

$$y_i \sim \mathrm{N}_{m_i}\big(X_i\beta, \sigma^2(I + Z_i D Z_i^{\mathrm{t}})\big). \tag{3.2}$$

The corresponding marginal log-likelihood for one group is

$$\begin{aligned} \ell_i(\theta) = -\tfrac{1}{2}[&m_i \log(\sigma^2) + \log(|I + Z_i D Z_i^{\mathrm{t}}|) \\ &+ \sigma^{-2}(y_i - X_i\beta)^{\mathrm{t}}(I + Z_i D Z_i^{\mathrm{t}})^{-1}(y_i - X_i\beta)]. \end{aligned} \tag{3.3}$$

Assuming that the model is correct, we have classical results, in for instance Pinheiro and Bates (2000) and Verbeke and Lesaffre (1997), stating that the maximum likelihood (ML) estimates $\widehat{\theta}$ are consistent and asymptotically normal with asymptotic covariance matrix equal to the inverse Fisher information matrix. The information matrix (normalised by sample size) has the following block-diagonal form, for instance from Demidenko (2013),

$$J_n = n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \sigma^{-2}X_i^{\mathrm{t}}V_i^{-1}X_i & 0 & 0 \\ 0 & 2m_i\sigma^{-2} & \sigma^{-1}\mathrm{vec}(R_i)^{\mathrm{t}}W_k \\ 0 & \sigma^{-1}W_k^{\mathrm{t}}\mathrm{vec}(R_i) & \tfrac{1}{2}W_k^{\mathrm{t}}(R_i \otimes R_i)W_k \end{bmatrix}. \tag{3.4}$$

Here $\otimes$ is the Kronecker product; the vec vectorisation operation stacks the columns of the input matrix into a long vector;

$$V_i = I + Z_i D Z_i^{\mathrm{t}}, \quad R_i = Z_i^{\mathrm{t}} V_i^{-1} Z_i;$$

and W_k is the so-called duplication matrix, of size $k^2 \times k(k+1)/2$. These mathematical linear algebra tools, used here to reach accurate descriptions and then algorithms for variance matrices, are treated in Appendix B.

We note that the block-diagonal structure here implies that the ML estimator of β becomes asymptotically independent of the estimators of the variance-covariance parameters. This relationship holds when the model is correctly specified, but not when it is misspecified as we will see in the next section.

3.2. *Existing model selection approaches for LME models.* Müller, Scealy and Welsh (2013) offer a comprehensive review of most existing frameworks for model selection in LME models. In fact, they treat a somewhat wider class of models than we do here. Similarly to most of the literature on LME models, we limit ourselves to what Müller, Scealy and Welsh (2013) call the 'independent cluster model' where each observation belongs to only one natural group and all groups are independent of each other, giving a block diagonal covariance matrix for y_i. The authors describe and compare four classes of model selection methods; information criteria, shrinkage methods, fence methods, and a few Bayesian methods (for example Chen and Dunson (2003)). A large number of AIC- and BIC-like criteria exist in the literature, differing in their loss functions and in their penalty terms. The loss functions are usually some form of log-likelihood, either the marginal log-likelihood related to (3.3); the reduced or restricted log-likelihood associated with the REML methods, see e.g. Gumedze and Dunne (2011) and Demidenko (2013, Ch. 2); or the conditional log-likelihood. The most commonly used information criterion in this class is the so-called marginal AIC, which used the marginal log-likelihood and the straightforward penalty $d = p + q$ (where q is the number of variance-covariance parameters; in our formulation we have $q = 1 + k(k+1)/2$). Alternatively, there exists a large number of variants of the so-called conditional AIC, see for instance Vaida and Blanchard (2005). Some of the information criteria methods are designed for only selecting among the regression coefficients, like the recently proposed meanAIC by Craiu and Duchesne (2018) (but which is aimed at the whole Generalised linear mixed model (GLMM) class).

When the number of candidate models under consideration is large, shrinkage methods may have an advantage over information criteria, partly because it may be computationally infeasible to evaluate all 2^{p+q} candidate models. Shrinkage methods for LME models choose a candidate model by solving an optimisation problem with a LASSO-type criterion consisting of two terms: a measure of model fit and a penalty function ensuring that both the estimated coefficients and variance-covariance parameters may be shrunk to zero. In Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) the model fit is evaluated via the marginal log-likelihood, while in Peng and Lu (2012) a least squares criterion is used. These three methods also differ in their choice of penalty functions. Recently, Hui, Müller and Welsh (2017) proposed a method combining the penalised quasi-likelihood as a measure of model fit with adaptive lasso penalty functions.

Other model selection approaches include the fence method, originally proposed in Jiang et al. (2008). The method consists of estimating the loss of each candidate model (e.g. as the negative log-likelihood), finding the model with the minimal loss and constructing a fence around this model containing all the candidate models with a loss sufficiently close to the minimal loss.

The small set of models within the fence can then be investigated more carefully, for example selecting the least complex one.

The model selection methods briefly described here all assume, directly or indirectly, that the user wishes to identify a model which maximises the fit to the data, with a trade-off against complexity. This is also apparent through the choice of criteria for evaluating the model selection procedures used in the simulation studies of the aforementioned articles; most concern the probability of selecting the true (or correct) model, or maximising some measure of fit. The FIC approach, introduced in Section 2, is different in aim and spirit: there the goal of model selection is to identify the model which provides the most precise estimate of the focus parameter.

4. Behaviour of LME estimators under misspecification.

In order to develop our FIC machinery in the next section we need results regarding the behaviour of estimators constructed for one LME model, but examined when the real data generating mechanism is a different LME model. What we derive below leads to precise answers to the pertinent questions, in any situation where the true mechanism is some LME model A, but where LME model B is used to generate estimators. Our use of these results and insights will be in a context where model A is a well-defined and well thought through 'wide model', assumed to hold, and where a list of candidate models is being considered for comparison and ranking, say $M_1, \ldots, M_r$. The choice and use of such a wide model is discussed further in Section 5. We emphasise that the candidate models, generically called M below, do not have to be submodels of the wide model.

4.1. *Behaviour of the estimators from the candidate model.*

Let the wide model be defined as in Section 3, with the true parameter vector

$$\theta_{\rm true} = (\beta_{\rm true}, \sigma_{\rm true}, D_{\rm true}) \tag{4.1}$$

governing LME data as in (3.1) and (3.2); we also assume that $D_{\rm true}$ is positive definite, with each diagonal element positive. In addition, we contemplate using a different LME model, the candidate model M. We assume that this model is defined with respect to the same n groups as in the (3.1) formulation, and we write

$$y_i \sim {\rm N}_{m_i}\bigl(X_{M,i}\beta_M, \sigma_M^2(I + Z_{M,i}D_M Z_{M,i}^{\rm t})\bigr). \tag{4.2}$$

This model has design matrices, $X_{M,i}$ and $Z_{M,i}$, potentially different from those of the wide model, and hence also a different set of parameters, say $\theta_M = (\beta_M, \sigma_M, D_M)$. Often, but not necessarily, the candidate model will involve a subset of the covariates (i.e. columns) included in X_i and Z_i, respectively. Let the covariate matrix $X_{M,i}$ have dimension $m_i \times p_M$, with $Z_{M,i}$ being $m_i \times k_M$, and hence D_M being $k_M \times k_M$. We denote by d_M the total number of parameters in the candidate model.

The ML estimates $\widehat{\theta}_M$ can be obtained in the usual fashion, i.e. numerically maximising the log-likelihood function associated with M. Under natural conditions, these will aim for the least false parameters

$$\theta_{M,0} = (\beta_{M,0}, \sigma_{M,0}, D_{M,0}), \tag{4.3}$$

minimising the Kullback–Leibler divergence from the wide model to the candidate model, say $\int f(y, \theta_{\rm true}) \log\{f(y, \theta_{\rm true})/f_M(y, \theta_M)\}\, {\rm d}y$ with $f_M(y, \theta_M)$ denoting the density of the full dataset, under model M. Here 'aiming for' will mean 'converging in probability to', with the right large-sample setup, with n growing, see below. In more practical terms this translates to $\widehat{\theta}_M$ having high probability of coming close to this $\theta_{M,0}$.

Minimising the divergence corresponds to solving the following equation for $\theta_{M,0}$:

$$\sum_{i=1}^n {\rm E}_{\rm wide}\, u_{M,i}(y, \theta_{M,0}) = 0, \quad \text{where} \quad u_{M,i}(y, \theta_M) = \partial \ell_{M,i}(\theta_M)/\partial \theta_M,$$

the score function of the candidate model, via $\ell_{M,i}$, the log-likelihood for the ith group of the candidate model. The least false parameters $\theta_{M,0}$ of (4.3) depend on the true parameters $\theta_{\rm true}$ of the data generating mechanism, and also on the covariate matrices X_i and Z_i associated with the dataset. With some algebra, the three parts of the $\theta_{M,0}$ of (4.3) can be seen to be the solutions of the following three equations. First,

$$\begin{aligned} \beta_{M,0} &= \Big(\sum_{i=1}^n X_{M,i}^{\rm t} V_{M,0,i}^{-1} X_{M,i}\Big)^{-1} \sum_{i=1}^n X_{M,i}^{\rm t} V_{M,0,i}^{-1} X_i \beta_{\rm true}, \\ \sigma_{M,0}^2 &= \frac{1}{n_{\rm tot}} \sum_{i=1}^n \{\mu_{e,i}^{\rm t} V_{M,0,i}^{-1} \mu_{e,i} + \sigma_{\rm true}^2 \,{\rm Tr}(V_{M,0,i}^{-1} V_i)\}, \end{aligned} \tag{4.4}$$

with $V_{M,0,i} = I + Z_{M,i} D_{M,0} Z_{M,i}^{\rm t}$, $\mu_{e,i} = X_i \beta_{\rm true} - X_{M,i} \beta_{M,0}$ and $n_{\rm tot} = \sum_{i=1}^n m_i$. Next,

$$\begin{aligned} \sigma_{M,0}^2 n^{-1} \sum_{i=1}^n Z_{M,i}^{\rm t} V_{M,0,i}^{-1} Z_{M,i} &= \sigma_{\rm true}^2 n^{-1} \sum_{i=1}^n Z_{M,i}^{\rm t} V_{M,0,i}^{-1} V_i V_{M,0,i}^{-1} Z_{M,i} \\ &\quad + n^{-1} \sum_{i=1}^n Z_{M,i}^{\rm t} V_{M,0,i}^{-1} \mu_{e,i} \mu_{e,i}^{\rm t} V_{M,0,i}^{-1} Z_{M,i}. \end{aligned} \tag{4.5}$$

Note here that if the candidate model has all the right X_i, but not necessarily the right Z_i, then $\beta_{M,0} = \beta_{\rm true}$, which means that the $\widehat\beta_M$ is essentially unbiased for the correct $\beta_{\rm true}$ of the wide model.

Via some general arguments, similar to those used in Claeskens and Hjort (2008a, Ch. 2) for deriving large-sample approximations of ML estimators outside model conditions, one can demonstrate that

$$\begin{pmatrix} \sqrt{n}(\widehat\theta - \theta_{\rm true}) \\ \sqrt{n}(\widehat\theta_M - \theta_{M,0}) \end{pmatrix} = \begin{pmatrix} J_n^{-1} U_n \\ J_{M,n}^{-1} U_{M,n} \end{pmatrix} + \begin{pmatrix} \delta_n \\ \delta_{M,n} \end{pmatrix}. \tag{4.6}$$

Here, J_n is as in (3.4), but evaluated at $\theta_{\rm true}$. We also have $J_{M,n} = -n^{-1}\sum_{i=1}^n {\rm E}_{\rm wide}\,\partial^2 \ell_{M,i}(\theta_{M,0})/\partial\theta_M\partial\theta_M^{\rm t}$; $U_n = n^{-1/2}\sum_{i=1}^n \partial\ell_i(\theta_{\rm true})/\partial\theta$; $U_{M,n} = n^{-1/2}\sum_{i=1}^n \partial\ell_{M,i}(\theta_{M,0})/\partial\theta_M$; and δ_n and $\delta_{M,n}$ are remainder terms becoming small in probability. Here

$$\begin{pmatrix} U_n \\ U_{M,n} \end{pmatrix} \approx_d {\rm N}_{d+d_M}\left(0, \begin{pmatrix} J_n & C_{M,n} \\ C_{M,n}^{\rm t} & K_{M,n} \end{pmatrix}\right), \tag{4.7}$$

in which

$$K_{M,n} = n^{-1}\sum_{i=1}^n {\rm Var}_{\rm wide}\, u_{M,i}(y, \theta_{M,0}),$$
$$C_{M,n} = n^{-1}\sum_{i=1}^n {\rm Cov}_{\rm wide}\left\{u_i(y,\theta_{\rm true}), u_{M,i}(y,\theta_{M,0})\right\}.$$

We provide explicit formulae for the matrices $J_{M,n}, K_{M,n}, C_{M,n}$ in Appendix B. Note that the expectations and variances are taken with respect to the wide model. The $J_{M,n}$ and $K_{M,n}$ matrices will typically be different from each other, and $K_{M,n}$ in particular will have a more complex form than for the J_n given in (3.4). In particular, the matrices will in general no longer be block-diagonal.

In (4.7), '$\approx_d$' means 'approximately distributed as', and a precise asymptotic statement is that $(U_n, U_{M,n})$ converges in distribution, under mild Lindeberg type regularity conditions, to a multivariate zero-mean normal (U, U_M) with covariance matrix having components J, C_M, K_M, the appropriate limits of $J_n, C_{M,n}, K_{M,n}$. Under yet further but still mild regularity assumptions, the remainder terms in (4.6) will tend to zero in probability, and the left hand side of (4.6) has its consequent clear limit distribution, namely

$$\begin{pmatrix} J^{-1}U \\ J_M^{-1}U_M \end{pmatrix} \sim {\rm N}_{d+d_M}\left(0, \begin{pmatrix} J^{-1} & J^{-1}C_M J_M^{-1} \\ J_M^{-1}C_M^{\rm t}J^{-1} & J_M^{-1}K_M J_M^{-1} \end{pmatrix}\right). \tag{4.8}$$

The practical translation of this precise limit theorem is the directly useful approximation

$$\begin{pmatrix} \sqrt{n}(\widehat\theta-\theta_{\rm true}) \\ \sqrt{n}(\widehat\theta_M-\theta_{M,0}) \end{pmatrix} \approx_d {\rm N}_{d+d_M}\left(0, \begin{pmatrix} J_n^{-1} & J_n^{-1}C_{M,n}J_{M,n}^{-1} \\ J_{M,n}^{-1}C_{M,n}^{\rm t}J_n^{-1} & J_{M,n}^{-1}K_{M,n}J_{M,n}^{-1} \end{pmatrix}\right). \tag{4.9}$$

We have reached precise results and formulae for how LME estimators from different candidate models M behave, assuming that there is a certain wide LME model which generates the data. These results ought to have independent interest, e.g. for examining robustness, consequences of model misspecifications, etc. We also learn that the fixed effects part $\widehat\beta_M$ and the variance parts $\widehat\sigma_M$ and $\widehat D_M$ of a candidate model may exhibit dependence. This clashes with how estimator behaviour results for LME models are typically expressed in the literature. The reason is the block diagonal structure for the model based information matrix (3.4), where we find that the $J_{M,n}^{-1}K_{M,n}J_{M,n}^{-1}$ matrix of (4.9) is in general not block diagonal. Our present mandate is using (4.9) not to assess model misspecification issues, per se, but to find the required ingredients for the FIC, in Section 5.

4.2. *The approximate distribution of a focus parameter estimator.* Consider a focus parameter, say μ, like 'the yearly decline' for the minke whale application sketched in Section 1, or the probability $p(Y_{\rm new} > y_0 \,|\, x_0, z_0)$ that a future or not observed $Y_{\rm new}$ with covariate vectors x_0, z_0 will be bigger than some threshold y_0. The user needs to provide a definition of μ for each candidate model; typically μ will have a clear statistical interpretation. For the wide model we assume $\mu = \mu(\theta)$ can be expressed as a smooth function of $\theta = (\beta, \sigma, D)$. The ensuing ML estimator using the wide model is $\widehat\mu = \mu(\widehat\theta)$, aiming at the true value $\mu_{\rm true} = \mu(\theta_{\rm true})$. For a candidate model M, the μ can be expressed in terms of that model's parameter vector, say $\mu_M = \mu_M(\theta_M)$, with $\theta_M = (\beta_M, \sigma_M, D_M)$, and with $\mu_M(\cdot)$ also a smooth function. The ML estimator using model M is $\widehat\mu_M = \mu_M(\widehat\theta_M)$, aiming for the corresponding least false parameter value $\mu_{0,M} = \mu_M(\theta_{M,0})$, with $\theta_{M,0} = (\beta_{M,0}, \sigma_{M,0}, D_{M,0})$ as in (4.3).

Now introduce

$$c = \partial\mu(\theta_{\rm true})/\partial\theta \quad \text{and} \quad c_M = \partial\mu_M(\theta_{M,0})/\partial\theta_M, \tag{4.10}$$

column vectors of the relevant lengths. Via delta method arguments, see e.g. Schweder and Hjort (2016, Appendix), applied to (4.9), we have the following joint approximate distribution,

$$\begin{pmatrix} \sqrt{n}(\widehat{\mu} - \mu_{\rm true}) \\ \sqrt{n}(\widehat{\mu}_M - \mu_{M,0}) \end{pmatrix} \approx {\rm N}_2 \left(0, \begin{pmatrix} \nu_{\rm wide} & \nu_{M,c} \\ \nu_{M,c} & \nu_M \end{pmatrix} \right), \tag{4.11}$$

with $\nu_{\rm wide} = c^{\rm t} J_n^{-1} c$, $\nu_{M,c} = c^{\rm t} J_n^{-1} C_{M,n} J_{M,n}^{-1} c_M$, $\nu_M = c_M^{\rm t} J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1} c_M$. Thus $\widehat{\mu}$ is approximately unbiased, whereas competing estimators $\widehat{\mu}_M$ will be biased, but potentially with smaller variances. Note that this result, along with a precise limit distribution version, as per (4.8), is valid for each candidate model M. We shall use this actively below, for constructing our FIC scores for candidate models.

5. The FIC approach.

We start with some general objectives and principles for the FIC approach to model selection, before applying these to classes of LME models.

5.1. *The general FIC scheme.*

For various applications, depending also on the context, precise estimates for one or more focus parameters are more important to the practitioner than identifying the 'true' model, or finding the model with the best overall fit. Contrary to other information criteria and model selection procedures, the FIC aims at finding the model giving the most precise estimates of the parameter of main interest (say μ, henceforth called the focus parameter). The FIC was proposed in Claeskens and Hjort (2003, 2008a) and has been successfully applied to a healthy variety of models and contexts; these include semiparametric Cox regression models (Hjort and Claeskens, 2006), additive risk models (Gandy and Hjort, 2013), classes of time series models (Claeskens, Croux and Van Kerckhoven, 2007), generalised additive linear models Zhang and Liang (2011), with application areas ranging from finance and economics (Brownlees and Gallo (2008) and Behl et al. (2012)) to fisheries (Hermansen, Hjort and Kjesbu, 2016). Another and related FIC approach has been developed in Jullum and Hjort (2017, 2019). These two perspectives on the FIC relate partly to the definition of the widest potential candidate model, and partly to the mathematical tools used to approximate and assess mean squared errors, cf. (1.1), as further commented upon below.

The general FIC procedure consists of estimating the risk associated with each candidate model's estimate of the focus parameter, and then choosing the model with the smallest estimated risk. The most common risk measure is the mean squared error (mse), due to its natural interpretation and convenient separation into variance and squared-bias parts,

$$\text{mse}(\widehat{\mu}) = \text{Var}\,\widehat{\mu} + b^2 = \text{Var}\,\widehat{\mu} + (\text{E}\,\widehat{\mu} - \mu)^2.$$

Again, the expectations and variances are taken with respect to the assumed true data generating mechanism, which we call the wide model. In the next sections we will estimate the variance and squared bias parts separately. For a candidate model M we will provide FIC formulae of the form

$$\mathrm{fic}_M = \widehat{\mathrm{mse}}(\widehat{\mu}_M) = \widehat{\mathrm{bsq}}_M + \widehat{v}_M/n, \tag{5.1}$$

where $\widehat{\mathrm{bsq}}_M$ is an estimate of the squared bias and $\widehat{v}_M/n$ an estimate of the variance of $\widehat{\mu}_M$, the estimated focus parameter from model M. As before, n is the number of groups in the data.

Originally, as presented in Claeskens and Hjort (2003, 2008a), Hjort and Claeskens (2003), and as successfully followed for a range of later FIC constructions and contributions in the literature, the mean squared error is estimated via large-sample approximations derived inside a certain local misspecification framework. This in particular means working with asymptotic approximations coming out of a machinery where candidate models are within $O(1/\sqrt{n})$ of each other, where n is sample size (or, in the present framework, the number of natural groups in the data). Such a framework leads to clear FIC formulae in many situations, but is not necessarily well suited for all classes of models.

When it comes to LME models, there is a gap in the FIC literature. The $O(1/\sqrt{n})$ setup alluded to above may be worked with also for LME models, but with certain added issues and complexities, partly due to the fact that zero is not an inner point in the parameter spaces for the variances implied by the random effects. We shall instead work with a *fixed wide model*, which the user has to specify; cf. our setup for Section 4 above. Here, 'fixed' refers to the fact that the wide model does not change with n. This wide model should be sensible and flexible, i.e. rich enough in its parametrisation to encompass plausible submodels, and also biologically well motivated. Such a wide model will often be too complex to be the model actually selected in the end, since its implied standard errors for model parameters might easily be unnecessarily large. Its role, conceptually and operationally, is partly to secure means, biases, variances, covariances clear definitions, also for all candidate models, needed for (5.1).

Our FIC procedure will guide the user in choosing among a set of candidate models, often selecting a smaller model with less variable estimates, but potentially introducing some bias. The set of candidate models ought to be chosen with some care. Reducing the set of candidate models saves computational resources, enhances performance, and guards against certain issues and problems with post-selection inference.

To work our FIC scores of the type (5.1) for LME candidate models, we need mse expressions for the different $\widehat{\mu}_M$. We get such from the joint distribution of (4.11). The canonical mse expressions are

$$\text{mse}(\widehat{\mu}_{\text{wide}}) = 0^2 + \nu_{\text{wide}}/n \quad \text{and} \quad \text{mse}(\widehat{\mu}_M) = b_M^2 + \nu_M/n, \tag{5.2}$$

for estimators from the wide model and candidate model, respectively. The wide model estimator entails no bias, i.e. $\widehat{\mu}$ aims for μ_{true}, whereas the bias related to the candidate model is $b_M = \mu_{M,0} - \mu_{\text{true}}$. When it comes to model selection, we need to estimate b_M^2 and ν_M/n, for the appropriate list of candidate models.

5.2. *Estimation of the necessary quantities.* FIC formulae are worked out below, in the spirit of (5.1), as natural estimators of the mse expressions of (5.2). We need good estimators of the quantities appearing there. For the variances ν_{wide} and ν_M we have two general options; we discuss these first, before coming to the squared bias.

The first consists in using the formula for J_n in (3.4) and the formulae for $J_{M,n}$, $K_{M,n}$, $C_{M,n}$ in the appendix, and then plugging in the ML estimate $\widehat{\theta}$ for θ_{true} from the wide model, and the ML estimate $\widehat{\theta}_M$ for the candidate model. Upon noting that e.g. $J_{M,n}$ can be expressed as $J_{M,n}(\theta_{\text{true}}, \theta_{M,0})$, and similarly for the others, we may write

$$\widehat{J}_n = J_n(\widehat{\theta}),\ \widehat{J}_{M,n} = J_{M,n}(\widehat{\theta}, \widehat{\theta}_M),\ \widehat{K}_{M,n} = K_{M,n}(\widehat{\theta}, \widehat{\theta}_M),\ \widehat{C}_{M,n} = C_{M,n}(\widehat{\theta}, \widehat{\theta}_M).$$

Actually, $\widehat{C}_{M,n}$ is not required for the variances, only for the squared bias, see below. We can use either ML estimates or REML estimates, since these are large-sample equivalent (see for instance Demidenko (2013, Ch. 3)).

An alternative is to use the Hessian matrices from the optimisation routines as estimates of J_n and $J_{M,n}$. The estimates for the remaining matrices can be computed as $\widehat{K}_{M,n} = n^{-1}\sum_{i=1}^n u_{M,i}(y, \widehat{\theta}_M) u_{M,i}(y, \widehat{\theta}_M)^{\text{t}}$ and $\widehat{C}_{M,n} = n^{-1}\sum_{i=1}^n u_i(y, \widehat{\theta}) u_{M,i}(y, \widehat{\theta}_M)^{\text{t}}$. For our applications in this article we have used the first option, with plug-in for θ_{true} and $\theta_{M,0}$. We use

$$\widehat{c} = \partial\mu(\widehat{\theta})/\partial\theta \quad \text{and} \quad \widehat{c}_M = \partial\mu_M(\widehat{\theta}_{M,0})/\partial\theta_M$$

for (4.10), and these are straightforward to compute numerically in cases where explicit expressions are unavailable.

There are also different options for the estimation of the squared bias, i.e. b_M^2, with $b_M = \mu_{M,0} - \mu_{\text{true}}$. We begin from the natural $\widehat{b}_M = \widehat{\mu}_M - \widehat{\mu}$. A naive start estimator would be $\widehat{b}_M^2$. However, this estimator for b_M^2 overestimates the squared bias, in that $\text{E}\,\widehat{b}_M^2 = (\text{E}\,\widehat{b}_M)^2 + \text{Var}\,\widehat{b}_M$. An estimator repairing for this overshooting is therefore

$$\widehat{\mathrm{bsq}}_M = (\widehat{\mu}_M - \widehat{\mu})^2 - \widehat{\mathrm{Var}}\,\widehat{b}_M, \tag{5.3}$$

the latter term being an estimate of the variance. From (4.11) we have that $\mathrm{Var}\,\widehat{b}_M \approx n^{-1}(\nu_{\mathrm{wide}} + \nu_M - 2\nu_{M,c})$. We have estimates of all the necessary terms from the arguments in the previous paragraph, and obtain the following final estimator,

$$\widehat{\mathrm{bsq}}_M = (\widehat{\mu}_M - \widehat{\mu})^2 - n^{-1}(\widehat{\nu}_{\mathrm{wide}} + \widehat{\nu}_M - 2\widehat{\nu}_{M,c}). \tag{5.4}$$

Note that in some cases we can get a negative estimate of the bias squared. In some FIC schemes, it is common to truncate the bias squared estimate to zero, rather than allowing negative estimates (see for instance Claeskens and Hjort (2008a)). For the LME case, our simulation studies indicate that the FIC version with the untruncated bias squared estimate from (5.4) has better practical performance than the version where we truncate negative estimates to zero. When we do not truncate, the FIC scores are approximately unbiased estimates of the mse.

Instead of using $\widehat{C}_{M,n}$ and the other matrices for estimating $\mathrm{Var}\,\widehat{b}_M$, we can obtain an estimate by parametric bootstrapping from the estimated wide model. We generate new observations from the estimated wide model, we fit both the wide and the candidate model, and obtain a new estimate of the bias $\widehat{b}_{M,l} = \widehat{\mu}_{M,l} - \widehat{\mu}_l$. Here the subscript l indicates that this is the bias associated with the lth bootstrap sample. The procedure is repeated a number of times, and $\mathrm{Var}\,\widehat{b}_M$ is estimated by the empirical variance of the $\widehat{b}_{M,l}$. In the simulations and application we have used the first option for estimating the squared bias, relying on the formula for $\mathrm{Var}\,\widehat{b}_M$.

5.3. *Computing FIC scores.* We now have everything we need to compute the FIC scores of both the wide model and all candidate models. It starts from having a well-defined focus parameter μ, with a clear mathematical definition in all candidate models. Using the FIC formulae in practice then involves the following steps.

(a) Decide on the wide model, with parameter $\theta = (\beta, \sigma, D)$, and a list of candidate models M, with parameters $\theta_M = (\beta_M, \sigma_M, D_M)$. The focus parameter then needs to be expressed as $\mu(\theta)$ and $\mu_M(\theta_M)$ in the wide model and the candidate model, respectively.

(b) Estimate the parameters in the wide model, yielding $\widehat{\theta}$ (using either ML or REML, for example with R packages like `lme4` (Bates et al., 2014)), and find $\widehat{\mu} = \mu(\widehat{\theta})$. Similarly, estimate the parameters for each candidate model M, giving $\widehat{\theta}_M$, and compute $\widehat{\mu}_M = \mu_M(\widehat{\theta}_M)$.

(c) Differentiate μ with respect to θ for the wide model (with a formula, or numerically), at $\widehat\theta$, to find $\widehat c$; and similarly differentiate $\mu_M(\theta_M)$ with respect to θ_M for candidate model M, at position $\widehat\theta_M$, to find $\widehat c_M$; cf. (4.10).

(d) Compute $\widehat J_n = J_n(\widehat\theta)$ using the formula in (3.4). The FIC score for the wide model is then

$$\mathrm{fic}_{\mathrm{wide}} = n^{-1}\widehat\nu_{\mathrm{wide}} = n^{-1}\widehat c^{\mathrm{t}}\widehat J_n^{-1}\widehat c. \tag{5.5}$$

(e) For each candidate model M, compute $\widehat J_{M,n} = J_{M,n}(\widehat\theta,\widehat\theta_M)$, $\widehat K_{M,n} = K_{M,n}(\widehat\theta,\widehat\theta_M)$, $\widehat C_{M,n} = C_{M,n}(\widehat\theta,\widehat\theta_M)$, using the formulae in Appendix B. Then compute $\widehat\nu_{M,c} = \widehat c^{\mathrm{t}}\widehat J_n^{-1}\widehat C_{M,n}\widehat J_{M,n}^{-1}\widehat c_M$ and $\widehat\nu_M = \widehat c_M^{\mathrm{t}}\widehat J_{M,n}^{-1}\widehat K_{M,n}$ $\widehat J_{M,n}^{-1}\widehat c_M$.

(f) For each candidate model M, estimate the squared bias of the associated $\widehat\mu_M$, with $\widehat{\mathrm{bsq}}_M$ from (5.4). The FIC score for the candidate model is then

$$\mathrm{fic}_M = n^{-1}\widehat\nu_M + \widehat{\mathrm{bsq}}_M. \tag{5.6}$$

Note that with the bias squared estimator from (5.4) and no truncation to zero, some of the terms in the FIC score cancel each other out. The FIC score for the candidate model can actually be written as

$$\mathrm{fic}_M = 2n^{-1}\widehat\nu_{M,c} - n^{-1}\widehat\nu_{\mathrm{wide}} + (\widehat\mu_M - \widehat\mu)^2. \tag{5.7}$$

Thus, we do not actually need to estimate ν_M. Still, we often prefer to calculate, and evaluate, the variance and bias squared parts separately.

The computed FIC scores provide a ranking of the $r+1$ models under consideration, say, with r candidate models in addition to the wide model itself. It is also very useful to produce a *FIC plot*, consisting of the points

$$(\mathrm{fic}_M^{1/2}, \widehat\mu_M) \tag{5.8}$$

for all models. For this plotting purpose we prefer the root-FIC scores on the x axis, as they are on the scale of the estimates themselves, the Pythagoras square root of the squared standard deviation plus the squared bias. The farther to the left in the FIC plot, the better are the estimates. Such plots are given in Figures 6 and 7 below, pertaining to focus parameters for the whale dataset.

5.4. *A special case.* The FIC formulae presented above simplify significantly in the special case where the wide model and all the candidate models all have $D = D_M = 0$. This corresponds to there being no random effects and the models are reduced to normal linear models (where we can consider each observation as a member of a group of size one). Notably, when the focus parameter is a function of the regression coefficients alone, the variance part becomes

$$\frac{\nu_M}{n} = \sigma^2 c_M^{\mathrm{t}} \Big(\sum_{i=1}^{n} X_{M,i}^{\mathrm{t}} X_{M,i} \Big)^{-1} c_M.$$

This expression is equal to the exact variance of a function $\mu(\widehat{\beta}_M)$ of the ML estimator $\widehat{\beta}_M = (\sum_{i=1}^{n} X_{M,i}^{\mathrm{t}} X_{M,i})^{-1} \sum_{i=1}^{n} X_{M,i}^{\mathrm{t}} y_i$. Also, the expression for $\mathrm{Var}\, \widehat{b}_M$, in the formula for the squared bias in Section 5, turns out to be equal to the corresponding exact quantity. Thus, despite stemming from large-sample approximations, the FIC formulae presented in this article are exact (i.e. not approximate, and valid for any n) in the case of normal linear models with a focus parameter being a function of the regression coefficients. Incidentally, in this case, our FIC approach with a fixed wide model also coincides with the FIC formulae coming out of the local misspecification framework in Claeskens and Hjort (2008a, Ch. 6). The FIC formulae will also be exact in the related case where the covariance matrix of the random effects, D and D_M, are assumed to be known, for both the wide and the candidate model.

6. Simulations.

We illustrate our FIC procedure with a short simulation study. We simulate data with $n = 20$ groups, $m = 15$ observations in each, with $p = 4$ potential fixed effect covariates $X = [X_0, X_1, X_2, X_3]$ and $k = 4$ potential random effects $Z = [Z_0, Z_1, Z_2, Z_3]$. We set $X_0 = Z_0 = 1$ and also let $Z = X$, i.e. each group is allowed to have a potentially different intercept, as well as potentially different slopes corresponding to each fixed covariate. This is a common choice in many applications. The non-intercept covariates are drawn from a multivariate normal distribution with zero means and relatively high correlations ($\mathrm{corr}(X_1, X_2) = 0.45$, $\mathrm{corr}(X_1, X_3) = 0.7$, $\mathrm{corr}(X_2, X_3) = 0.95$). Correlated covariates are typical for many ecological applications. We present four different simulation experiments below, differing only in their choice of focus parameters. The true model is

$$y_{i,j} = \beta_0 + b_{0,i} + \beta_1 x_{1,i,j} + \beta_2 x_{2,i,j} + b_{1,i} z_{1,i,j} + b_{2,i} z_{2,i,j} + \epsilon_{i,j}$$

with $\epsilon_{i,j} \sim \mathrm{N}(0, \sigma^2)$, $\sigma = 1$, $\beta = (1, 1, 1)^{\mathrm{t}}$, $b_i \sim \mathrm{N}_3(0, \sigma^2 D)$, and

$$D = (1/\sigma^2)\begin{bmatrix} 9 & 4 & 0.1 \\ 4 & 4 & 0 \\ 0.1 & 0 & 0.1 \end{bmatrix}.$$

We consider the wide model M_0, along with four candidate models described in the following table.

	Description	p	k	d
M_0	X_0, X_1, X_2, X_3 and Z_0, Z_1, Z_2, Z_3	4	4	15
M_1	X_0, X_1, X_2 and Z_0, Z_1, Z_2	3	3	10
M_2	X_0, X_1, X_2 and Z_0, Z_1	3	2	7
M_3	X_0, X_1, X_2 and Z_0	3	1	5
M_4	X_0, X_1 and Z_0, Z_1	2	2	6

Thus, the wide model M_0 includes one unnecessary fixed effect covariate (X_3), one unnecessary random effect (Z_3) and one random effect that has very small influence (Z_2). The true model is M_1 (but M_0 is of course a *correct* model, but with more parameters than necessary). The models were fitted with REML.

Table 1

Simulation results for $\mu_1 = \beta_1$. For each model, we give the true root mean squared error for $\widehat{\beta}_1$, the average FIC score and the percentage of rounds where the model has the lowest FIC score (i.e. the winning model).

	M_0	M_1	M_2	M_3	M_4
true $\sqrt{\mathrm{mse}}$	0.558	0.459	0.458	0.498	0.670
$\sqrt{\mathrm{fic}}$	0.559	0.398	0.398	0.441	0.634
winning %	4	18	32	32	14

In the first simulation experiment, our focus parameter is simply one of the fixed effect coefficients $\mu_1 = \beta_1$. The results are shown in Figure 2 and also in Table 1. In the Figure, we have the variance part in the left panel, the squared bias part in the middle panel and the total FIC score in the right panel. The red lines are the ‘true’ variance and squared bias values (determined by averaging over 1000 datasets) while the grey crosses are the variance and squared bias parts of the FIC scores evaluated in 100 (different) datasets. The black lines are the average $\widehat{v}_M$ and $\widehat{b}^2_M$ from these 100 datasets. The true root mean squared errors for $\widehat{\beta}_1$ are given in the table. Thus, the

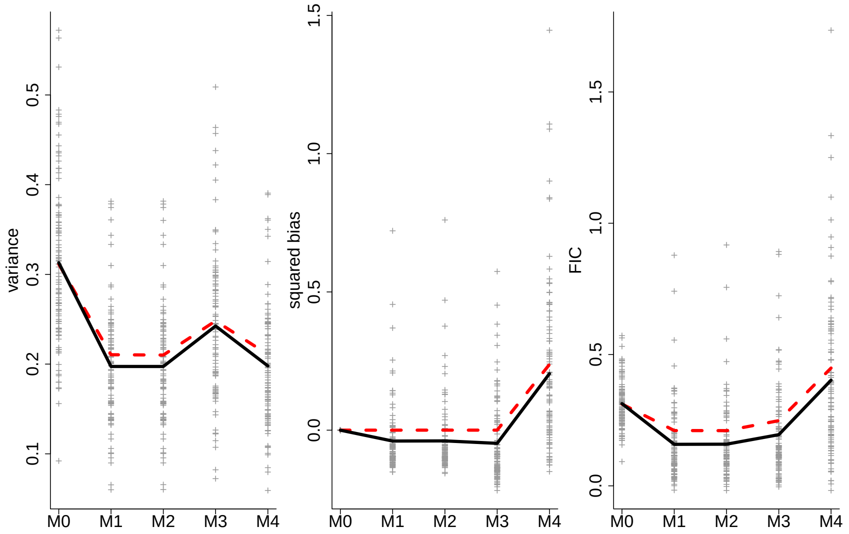

Fig 2. *Simulation results for $\mu_1 = \beta_1$. The variance in the left panel, the squared bias in the middle panel and the total FIC score in the right panel. The dashed red lines are the true values, the grey crosses are the variance and squared bias parts of the FIC scores on 100 simulated datasets and the full black lines are the averages of these.*

smaller model M_2 is the best model for estimating β_1 in this setting(i.e. with the smallest mse), but M_1 is almost as good. The wide model, with some unnecessary parameters for both the fixed and random effects, estimates β_1 with larger variance than the best models. While the model omitting one of the important fixed effects, M_4, has β_1 estimates with large bias. For 50% of the rounds, M_2 or M_1 were given the lowest FIC score, while M_3 was the winning model in 32% of the rounds. Thus, for many of the runs, M_3 was incorrectly considered better than M_2. The estimators from these two models only had a small difference in true mse, however.

Table 2

Simulation results for μ_2. For each model, we give the true root mean squared error for $\widehat{\mu}_2$, the average FIC score and the percentage of rounds where the model has the lowest FIC score (i.e. the winning model).

	M_0	M_1	M_2	M_3	M_4
true $\sqrt{\text{mse}} \times 10^2$	7.201	7.181	7.167	8.289	9.264
$\sqrt{\text{fic}} \times 10^2$	6.308	6.309	6.311	7.544	8.335
% best	32	17	33	18	0

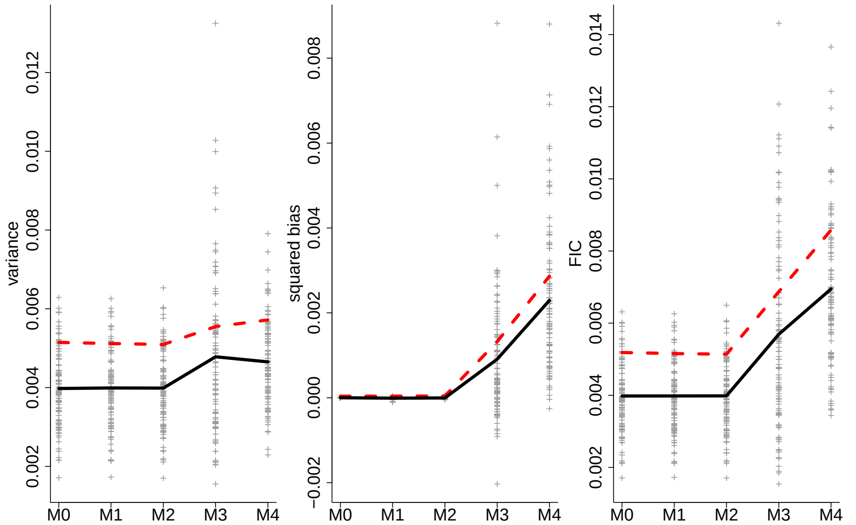

Fig 3. *Simulation results for μ_2. The variance in the left panel, the squared bias in the middle panel and the total FIC score in the right panel. The dashed red lines are the true values, the grey crosses are the variance and squared bias parts of the FIC scores on 100 simulated datasets and the full black lines are the averages of these.*

In our second experiment, the parameter of interest is $\mu_2 = n^{-1}\sum_{i=1}^{n}$ corr $(y_{i,j}, y_{i,k})$, with $j \neq k$, the average correlation between two observations belonging to the same group. The quantity depends on D and σ and is straightforward to evaluate given the design matrix Z. The true value of this focus parameter was 0.70 (given the covariates). The results are shown in Figure 3 and also in Table 2. The model M_2 gave μ_2 estimates with the smallest mean squared error, but M_0 and M_1 had almost similar performance. Our procedure selected one of these models in 82% of the runs. The worse model M_4 was never selected by FIC.

Table 3

Simulation results for μ_3. For each model, we give the true root mean squared error for $\widehat{\mu}_3$, the average FIC score and the percentage of rounds where the model has the lowest FIC score (i.e. the winning model).

	M_0	M_1	M_2	M_3	M_4
true $\sqrt{\text{mse}}$	0.046	0.045	0.059	1.250	0.375
$\sqrt{\text{fic}}$	0.047	0.045	0.063	1.199	0.388
% best	10	76	14	0	0

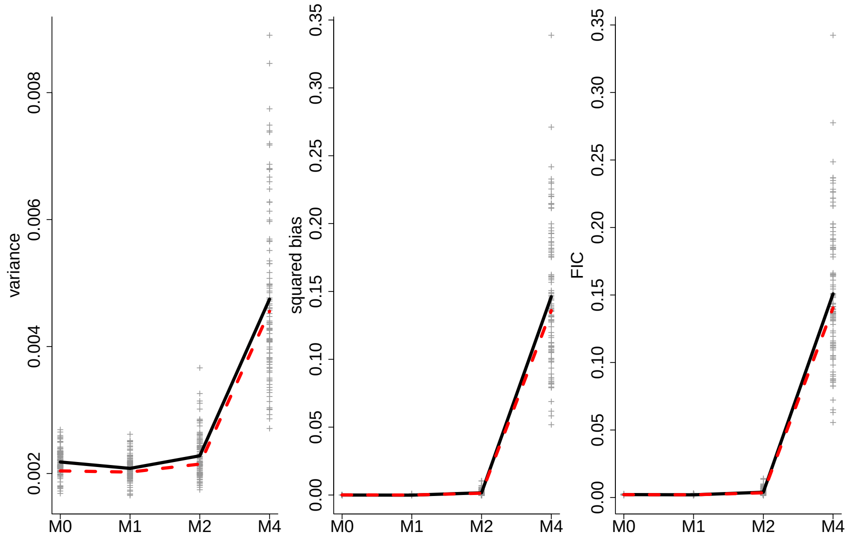

FIG 4. *Simulation results for $\mu_3 = \sigma$. The variance in the left panel, the squared bias in the middle panel and the total FIC score in the right panel. The dashed red lines are the true values, the grey crosses are the variance and squared bias parts of the FIC scores on 100 simulated datasets and the full black lines are the averages of these. Here we do not show the results for M_3 since this model had much higher values than the other models, both for the variance and squared bias.*

In our third experiment, the parameter of interest was $\mu_3 = \sigma$, the residual standard deviation. The results are shown in Figure 4 and also in Table 3. The true root mean squared errors for $\widehat{\mu}_3$ indicate that M_1 gave the most precise estimates of the standard deviation, while M_3 and M_4 were far worst. These two models were never selected by FIC.

TABLE 4

Simulation results for μ_4. For each model, we give the true root mean squared error for $\widehat{\mu}_4$, the average FIC score and the percentage of rounds where the model has the lowest FIC score (i.e. the winning model).

	M_0	M_1	M_2	M_3	M_4
true $\sqrt{\text{mse}}$	0.663	0.568	0.567	0.590	0.877
$\sqrt{\text{fic}}$	0.559	0.391	0.392	0.461	0.628
% best	0.03	0.37	0.32	0.13	0.15

The fourth experiment has $\mu_4 = \mathrm{E}(Y|x_1 = -0.5, x_2 = 0.5, x_3 = -0.1) = 1$ as focus parameter. This is the expected value of the response for some specific value of the covariate vector. This focus parameter is a function

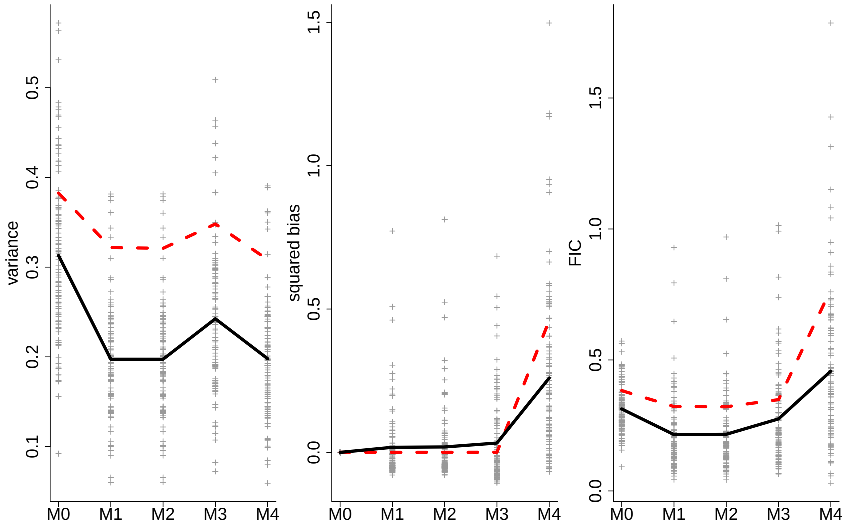

Fig 5. *Simulation results for μ_4. The variance in the left panel, the squared bias in the middle panel and the total FIC score in the right panel. The dashed red lines are the true values, the grey crosses are the variance and squared bias parts of the FIC scores on 100 simulated datasets and the full black lines are the averages of these.*

of the fixed effects coefficients only, but the inclusion or non-inclusion of random effects still influences its estimation. Figure 5 and Table 4 display the results. The truly best model, in terms of root mean squared errors, was M_2, with M_1 very close. One of these two models was selected by FIC in 69% of the rounds.

The simulations here illustrate two important facts. First, the quality of a model, in terms of the (true) mse of the corresponding estimator, varies depending on the choice of focus parameter. The true model is identical in all three experiments, but the model best suited to estimate the focus parameter varies, as does the true ranking of the different models. We see that for some foci, M_1 yields the most precise estimates, while for others M_2 is better. The quality of the largest model M_0 is also very different: for μ_1 the unnecessary parameters in the model clearly inflate the variance, while this effect is not apparent for the two other foci. Note that the true ranking of the different models also depends on the observed covariate matrices X and Z. For ease of comparisons they are the same in all three experiments, but different draws of the X and Z matrices can alter the true ranking of the models to some degree. The correlations between the covariates also influence the ranking of the models.

Second, the simulations demonstrate that our FIC approach successfully estimates the mean squared error associated with different focus parameters. The approach works both for foci that are functions of the fixed effect coefficient only, and for foci that depend on the variance-covariance parameters. These and similar experiments we have conducted, indicate that the precision of the mse estimates depend on n, m and also on the type of focus parameter; with seemingly less accurate estimates when the focus depends more on the covariance matrix D rather than on the regression coefficients only. However, even if the mse estimates are not always highly accurate for finite n, the relative ranking of the different candidate models can still be correct. When faced with a specific dataset, the FIC approach often identifies the model which estimates the focus with the best precision. When the wrong model is favoured it is very commonly a model with almost similar performance as the best one. The strength and benefit of the FIC approach manifest themselves in the ability to select different models for different purposes.

7. Application: Energy storage in Antarctic minke whales. As mentioned in the introduction, the minke whale dataset from the JARPA programme has been the object of considerable interest, discussion, and controversy. Our analyses concern the body condition of the whales and whether it has decreased during the 18 years of the JARPA period (1987/88 to 2004/5). Interest lies in the potential changes in body condition of minke whales, because these could herald deeper transformations in the Antarctic ecosystem; see Konishi et al. (2008); Konishi and Walløe (2015); Cunen, Walløe and Hjort (2020) for more information about the data and for discussion of the findings. Precise estimates of the evolution of body condition are also required for the construction of ecological models, as in Mori and Butterworth (2006), see also Laws (1977). Several proxies for body condition were studied in our report (Cunen, Walløe and Hjort, 2017), but here we limit ourselves to the response `fatweight`, the dissected fat reserves of each whale in kilograms. We have measurements of the fat weight of 683 different whales. Further, for each whale, we have measurements of seven independent variables that are considered relevant for explaining differences in fat weight. The independent variables are the year of capture, the date within each year, the sex, the body length, the age, the factor `region` which denotes one of three different longitudinal areas where the whales were caught (west, east and Ross sea), and finally, the binary indicator `diatom` which denotes whether the whales had little or substantial diatom coverage (diatom coverage is assumed to be an indicator of time spent in cold waters). The whales are caught during the austral summer, from the end of November to March; 1 denoting a whale caught on the 1st of December.

Our primary interest lies in the potential *change* in fat weight, so our focus parameter will be a function of the parameters related to `year`. However, in order to obtain a correct estimate of the yearly decline it is crucial that the rest of the model is well-specified. In Cunen, Walløe and Hjort (2017, 2020) we used considerable efforts to motivate our choice of wide model, but these arguments are outside the scope of the current article. The full wide model is quite large, with several interactions. According to prior biological knowledge, `date` is assumed to be one of the most important variables influencing the fat weight. The whales are in the Antarctic to gain weight so the coefficient related to date is expected to be large and positive. Also, the relationship between body condition and date is expected to be different from year to year, possibly due to random fluctuations in krill production. Hence, we include each year as a random effect influencing the effect of `date`. Further, it is assumed that the fat weight may be influenced by many other random processes with yearly variations. We therefore include the different years as a random effect influencing the intercept too.

The wide model we will use here is similar to the one we used in the report, with a few alterations, partly due to discussions in the IWC Scientific Committee meetings in 2017 and 2018. We have the following model specification (in an R-type notation):

$$\begin{aligned}\texttt{fatweight} \sim\ & \texttt{year} + \texttt{year}^2 + \texttt{bodylength} + \texttt{sex} + \texttt{diatom} + \texttt{date} + \texttt{date}^2 + \\ & \texttt{age} + \texttt{sex} * \texttt{diatom} + \texttt{diatom} * \texttt{date} + \texttt{diatom} * \texttt{date}^2 + \\ & \texttt{bodylength} * \texttt{sex} + \texttt{bodylength} * \texttt{date} + \\ & \texttt{bodylength} * \texttt{date}^2 + \texttt{sex} * \texttt{date} + \texttt{sex} * \texttt{date}^2 + \\ & \texttt{bodylength} * \texttt{sex} * \texttt{date} + \texttt{bodylength} * \texttt{sex} * \texttt{date}^2 + \\ & \texttt{age} * \texttt{sex} + \texttt{age} * \texttt{date} + \texttt{age} * \texttt{date}^2 + \texttt{age} * \texttt{sex} * \texttt{date} + \\ & \texttt{age} * \texttt{sex} * \texttt{date}^2 + \texttt{year} * \texttt{sex} + \texttt{year}^2 * \texttt{sex} + \texttt{region} + \\ & \texttt{year} * \texttt{region} + \texttt{year}^2 * \texttt{region} + \texttt{sex} * \texttt{region} + \\ & \texttt{diatom} * \texttt{region} + \texttt{region} * \texttt{date} + \texttt{region} * \texttt{date}^2 + \\ & (1 + \texttt{date} + \texttt{date}^2 \,|\, \texttt{year}).\end{aligned}$$

The model defined above has $p = 40$ fixed effect coefficients. The notation $(1 + \texttt{date} + \texttt{date}^2 \,|\, \texttt{year})$ specifies the random effect structure; the groups are defined by a categorical version of the year variable (so $n = 18$), and the Z_i matrix has 3 columns (a column of ones for the intercept, date and date squared). We thus have $k = 3$, giving a total of 47 parameters to estimate.

As explained in the introduction, the parameter of main interest in the IWC discussions was the yearly decline in the `fatweight` outcome variable. Since we have a quadratic year term in our wide model, with that part taking the form $\beta_{\rm year}x + \beta_{\rm year2}x^2$ for year x, a natural definition of the yearly decline is $\mu = \beta_{\rm year} + 2\beta_{\rm year2}x_0$, with x_0 the mean year in the dataset. The focus parameter corresponds to the derivative of the mean response, with respect to year, and evaluated in this mean year time point. This focus parameter can also be interpreted as the overall slope, the mean curve evaluated at the end point minus its value at the start point, divided by the length of time. For candidate models with only a linear relationship between body condition and year the term simplifies to $\beta_{\rm year}$ only. Furthermore, for those submodels where the fixed effect of year is not included, we have $\beta_{\rm year} = 0$, a parameter value which then is estimated with zero variance but with potentially big bias. Note also that one should take care to let the interactions between factor variables, in our case `region` and `sex`, and the `year` terms be defined as sum-to-zero contrasts. This ensures that $\beta_{\rm year}$ and $\beta_{\rm year2}$ can be interpreted as the parameters governing the *overall* yearly decline, and not the yearly decline for say males in some particular region.

The fitted wide model reveals some interesting features, see Cunen, Walløe and Hjort (2020). Most of the main effect estimates are relatively large (also compared to their standard errors), for instance the coefficients related to `age`, `bodylength` and `date`. This also concerns the terms related to `year`, which are discussed in the next paragraphs. Some of the interaction terms seem important as well. Still, the model contains a large number of parameters to estimate and the standard errors are therefore likely to be inflated. After estimation, it is crucial to evaluate whether the wide model adequately fits the data. This is particularly important in the case of model selection with FIC – since all the mse estimates rely on the wide model. In Cunen, Walløe and Hjort (2020) we investigate the quality of the wide model using different techniques. We have studied a number of diagnostic plots as recommended in Pinheiro and Bates (2000), and we have also carried out predictive simulations according to the recommendations in `lme4` (Bates et al., 2014). Incidentally, these authors use the term ‘posterior predictive simulations’, which is slightly misleading in the present frequentist setting.

In this illustration of the FIC methodology, we have limited ourselves to investigating five candidate models only, see Table 5. The full model specifications are given in the appendix. All the candidate models have a smaller number of fixed effects than the wide model. Note that the first candidate model M_1 has a more complex random effect structure than the wide model itself (with $k = 6$ giving a total of 21 random effect parame-

TABLE 5
Brief description and number of parameters in the wide model and five candidate models.

	Description	p	k	d
M_0	wide model	40	3	47
M_1	less interactions, quadratic year term	9	6	31
M_2	very simple, linear year term	5	2	9
M_3	very simple, linear year term	5	1	7
M_4	only linear year term	2	1	4
M_5	like the wide, but without year term	32	3	39

ters). This choice is meant to illustrate that there is nothing in the formulae hindering us from having candidate models with more random effects (or also more fixed effects) than the wide model. When it comes to interpreting the results, it is usually more natural to choose the wide model to be the largest possible plausible model, however. The models M_2 and M_3 are very simple (with few fixed effects), and differ only in the their random effects. Model M_4 includes only the linear year term in addition to a single random effect in the intercept. The last model, M_5, is the model without any year term, so $\mu_{M_5} = 0$. With the current focus parameter, the FIC score of such a model will have zero variance and a bias which only depends on the estimated focus parameter in the wide model and its estimated variance, $\mathrm{fic}_{M_5} = (0 - \widehat{\mu})^2 - n^{-1}\widehat{\nu}_{\mathrm{wide}}$. For this focus it is therefore not necessary to specify M_5; it includes all possible LME models without any fixed effect of year. In general, a full specification of M_5 will be necessary, here we use the same model as M_0, but without the linear and quadratic year terms.

The results from the model selection are given in the form of a FIC-plot in Figure 6. A FIC plot is a convenient graphical summary for model selection with FIC; it displays both the FIC scores and the estimated focus parameters for all the models under consideration. We see that M_2 gets the lowest FIC score, and that it has $\widehat{\mu} = -7.76$. The model without any fixed effect of year had a considerably larger FIC score than any of the other models. The winning model M_2 is very simple,

$$\texttt{fatweight} \sim \texttt{year} + \texttt{bodylength} + \texttt{sex} + \texttt{date} + (1 + \texttt{date}\,|\,\texttt{year}).$$

It assumes a linear relationship between body condition and year and contains only a few additional fixed effects. In the figure we have also included error bars representing the 95% confidence intervals for the focus parameter. Naturally, model M_5 with $\mu_{M_5} = 0$ has no uncertainty around its estimate. Note that the confidence intervals are all computed *under* the wide model, since this is assumed to be the true data generating mechanism. From the FIC plot we can conclude that our best estimate of the focus parameter is around 8 kilograms decline per year, or 80 kg loss of fat over a decade. Furthermore, assuming that the wide model holds, we may claim that the body condition decline has been negative and significant over the study period, since the confidence intervals associated with the best models all fall to the left of zero.

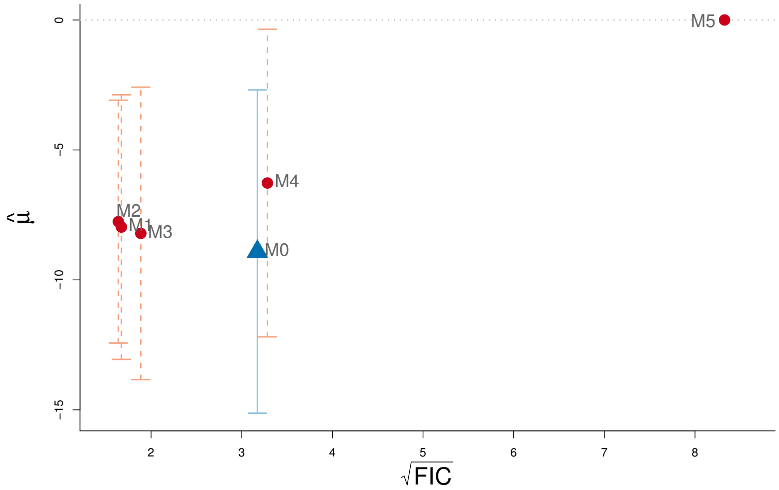

Fig 6. *Root-FIC scores and estimates of the yearly decline for the wide model (marked with a blue triangle) and the five candidate models. The scale of measurements is kilograms of fat. The bars represent 95% confidence intervals for the focus parameter, computed from each of the candidate models under the wide model.*

One might wonder about the uncertainty of the FIC scores. How stable are the MSE estimates? We can investigate this by parametric bootstrapping from the fitted wide model. Such investigations reveal that the ranking of the six models is reasonably stable, see Table 6. In more than half of the simulation runs, the models M_1 or M_2 are considered the best according to FIC. Model M_3 is selected quite seldom, but the wide model relatively often (around 18% of the time). The model without any fixed effect of year is chosen in about 5% of the runs.

As an illustration, we have investigated the same six models for another focus parameter, the probability of observing a whale with more than a certain amount of fat, say 1500 kilograms, given some covariate values, $\mu_2 = P(Y \geq 1500 \,|\, x_0, z_0)$. Here we chose to look at a 20 year old male whale, caught in the eastern region, of approximately mean length (8 metres), and

Table 6

Results from parametric bootstrap from the fitted wide model (1000 simulated dataset): the percentage of rounds where each model had the lowest FIC score (i.e. the winning model). Also, the AIC scores of the six models for the original dataset.

	M_0	M_1	M_2	M_3	M_4	M_5
% best	17.9	42.3	25.1	7.2	2.1	5.3
AIC	-349.97	-284.67	-304.93	-294.57	442.50	-349.84

which is caught towards the end of the season. Over the full dataset, the average fat weight of a whale is close to 1500 kilograms. First we consider a whale caught in year 1, then a whale caught in year 10.

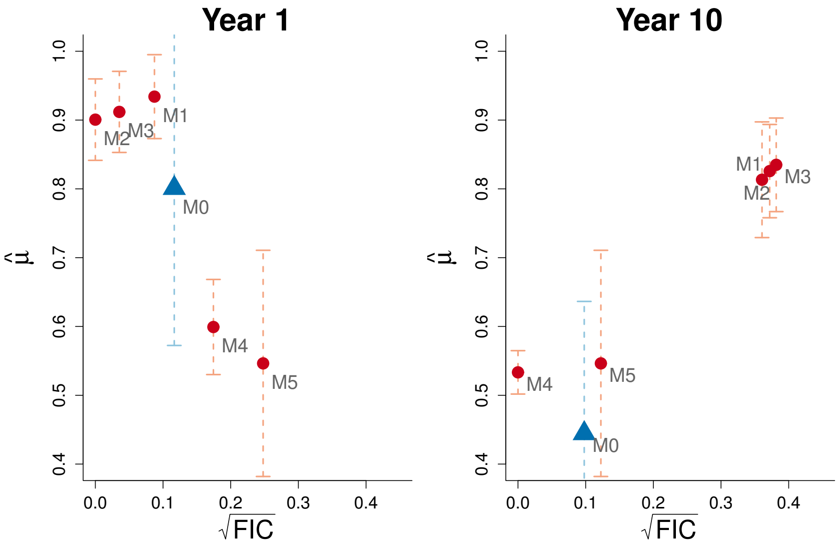

Fig 7. *Root-FIC scores and estimated probability of observing a whale with more than 1500 kilograms of fat. For year 1 the best model is M_2, while for year 10 the best model is M_4. The wide model is marked with a blue triangle. The bars represent 95% confidence intervals for the focus parameter, computed from each of the candidate models under the wide model.*

The FIC scores and estimates are given in Figure 7. For year 1, the estimates range from 0.50 to around 0.90. Naturally, the model without any fixed effect of year gives the same estimate for both years. Its ranking in terms of FIC is however very different; for year 1 the model without 'year' is considered the worst, while for year 10 M_5 is not far from the best, which reflects that year 10 is not far from the average year in the dataset. Note that some of the models are given a FIC score equal to zero. This is not a paradox, and only reflects that the true mse of these models is likely to be small.

Again, the ability of the FIC approach to select different models for different purposes is demonstrated by these examples. We have also computed the marginal AIC scores (see Müller, Scealy and Welsh (2013)) for the seven models under consideration, see Table 6. According to this criterion M_0, the wide model, is the best model, with M_5 very close. The extremely simple M_4 is clearly the worst. The disparity between the models preferred by AIC and by FIC constitutes no paradox and simply reflects that the two criteria have different aims. The models with very high AIC values, like M_4, are severely underspecified and fail to describe key aspects of the data. They should therefore not be used to answer inference questions in general, but they may still provide precise inference for *particular parameters of interest*, as seen in the FIC analyses.

8. Discussion and concluding remarks. In our paper we have developed and investigated a criterion for focused model selection in LME models. The Minke whale application presented above demonstrates that the methodology can be useful in practical situations where the main interest is the precise estimation of a well-defined focus parameter. The simulations indicate that the criterion estimates the risk associated with each candidate model with adequate precision, for a range of different foci.

Our framework offers ample flexibility in the choice of the wide model and the candidate models. The candidate models need not be nested within each other nor be inside the wide model, and may thus include covariates and random effects not present in the wide model. Also, the wide model need not necessarily be the most complex model, with the maximum number of parameters, although this may be natural in many cases.

Naturally, our FIC methodology also has some limitations. First, it is important to note that the derivation of our criterion relies on certain approximations, to biases and variances and covariances, and that these relate to large-sample approximations and asymptotics. Specifically, the asymptotic distributions are reached when the number of groups (n) increases to infinity; cf. Demidenko (2013). This applies particularly to focus parameters which are functions of the variance-covariance parameters. For functions of the linear mean parameters only, the normal approximations involved will still work well when the number of groups is small, but the total sample size $\sum_{i=1}^{n} m_i$ grows. The reliance on large-sample arguments is a characteristic shared with several other model selection criteria, like the Akaike and

Bayesian Information Criteria (i.e. the AIC and BIC), see Claeskens and Hjort (2008a, Chs. 2, 3). For certain foci, the accuracy of our FIC methodology may thus be limited for models with a small number of groups. However, in such situations it is common to assume that the group-specific effects, the b_i, are fixed rather than random. In that case, we simply have a normal linear model and our formulae are exact (see Section 5.4), i.e. regardless of n.

Users should also be aware that, in the limit, the bias part of the FIC score will always dominate the variance part. This is clear from the formulae in Section 5: the variance terms will disappear as the number of groups increase, but the squared bias will remain. The wide model, and potentially other candidate models without bias, will thus always be selected by our FIC procedure if the number of groups is very large. This property is quite natural considering the aim of FIC based model selection, i.e. the most precise estimates of the focus parameter. If one has enough data to estimate a big, plausible model without problems, there is not really anything to gain from model selection with FIC. In general, when the data volume is large, users are encouraged by the FIC scores to use more complex models.

In the FIC literature one often encounters a certain type of bias-variance trade-off where more complex models have estimates with large variances and small biases, while simpler models have more bias but smaller variances. This satisfying situation is not necessarily present with LME models. In some cases, our FIC formulae reveal that $\widehat{\mu}$ from a large and complex model both have small bias and small variance compared to smaller models. This occurs for instance when the models under consideration differ only in the random effects and the focus parameter is a function of the fixed effect coefficient only. Consider comparing the wide model and a candidate model in such a case. We have $X_{M,i} = X_i$, and then $\beta_{M,0} = \beta_{\rm true}$ (see Section 4.1). This leads to the matrices $J_{M,n}$, $K_{M,n}$, $C_{M,n}$ becoming block-diagonal, and allows us to simplify the mse formulae significantly. Since the focus parameter is a function of the fixed effect coefficient only, the non-zero elements in c and c_M are equal; we denote this non-zero part by c_β. Then, we get $\nu_{M,c} = \nu_{\rm wide}$ and the FIC formulae reduce to

$$\mathrm{fic}_{\rm wide} = \widehat{\sigma}^2 \widehat{c}_\beta^{\rm t} \Big(\sum_{i=1}^n X_i^{\rm t} \widehat{V}_i^{-1} X_i\Big)^{-1} \widehat{c}_\beta,$$

$$\mathrm{fic}_M = \widehat{\sigma}^2 \widehat{c}_\beta^{\rm t} \Big(\sum_{i=1}^n X_i^{\rm t} \widehat{V}_i^{-1} X_i\Big)^{-1} \widehat{c}_\beta + (\widehat{\mu}_{\rm wide} - \widehat{\mu}_M)^2.$$

The FIC scores will thus be equal only when $\widehat{\mu}_{\rm wide} = \widehat{\mu}_M$ and otherwise the wide model will have a smaller score than the simpler candidate model.

The two estimators have the same expectation (both are unbiased), but in practice they will only be exactly equal when $\widehat{V}_i = \widehat{V}_{M,i}$. The phenomenon described here originates from the asymptotic independence between regression coefficients and variance-covariance parameters in the wide model (see Section 3.1). It should influence how we interpret the FIC scores in situations where models only differ in their random effect structure. If the wide model has a much lower score than the candidate model it is clear that the random effect structure in the wide is necessary for the estimation of the focus. But when the two scores are *quite similar* it might be reasonable to select the smallest model, which often has advantages in increased numerical stability and ease of interpretation. When our focus parameter depends on the variance-covariance parameters as well as the regression coefficients we can obtain mse curves that look more like what we are used to in other FIC applications. The simple candidate model has clearly lower mse when the true data generating mechanism is within some distance of the candidate model, and outside that distance, the wide is preferred.

Our criterion depends on the choice of the wide model, and how well it represents the true data generating mechanism. In practice, the model building process will depend on the specifics of the sampling and the user's knowledge about the system under study. Depending on one's knowledge and preferences one may use techniques from exploratory data analysis or graphical tools from the causal inference literature (Greenland et al., 1999). In connection with the application described in Section 7 we have conducted some sensitivity checks and found that moderate changes to the wide model had little effect on the ranking of the different candidate models. Also, for the wide models we have investigated, the estimate of the focus parameter in the selected models was reasonably stable. More radical changes to the wide model should be expected to have greater effect, but we have not fully investigated this issue. There are different types of model misspecification of the wide model to consider. The true data generating model may contain unknown fixed or random effects, it may have random effects drawn from a different distribution than the normal, and it may be outside the linear mixed class altogether. Fully guarding against such misspecification of the wide model is unattainable, but extending our approach to even wider and more flexible wide models may lead to some improvements.

Finally, it is important to be aware of the problems of post-selection inference, see for instance Claeskens and Hjort (2008a, Ch. 7). Tests and confidence intervals computed after a model selection step will in general not be valid when one uses the same data for model selection and inference. Specifically, the confidence intervals might be too narrow, but the extent of undercoverage will depend on the specific dataset and model selection methodology. These problems are not specific to FIC, but concern all model selection methods. In Cunen, Walløe and Hjort (2020), we bypassed this problem by splitting the data into two parts, one for model selection and the second for inference. This is naturally a conservative solution, but there are more sophisticated approaches in the general model selection literature, see Berk et al. (2013); Tibshirani et al. (2016); Charkhi and Claeskens (2018). Naturally, and perhaps trivially, one can avoid the issues of post-selection, and the complexities of model selection in general, by choosing to use the wide model and not do model selection at all, see for instance Ver Hoef and Boveng (2015).

Modifying our FIC approach to other classes of wide models is one of several possible extensions of the methodology we have presented in this paper. Our FIC procedures assume that the random effects in both the wide model and the candidate models are normally distributed. Other distributional assumptions are possible and our FIC procedure can be adapted to such cases. Here one might make use of the contributions of Verbeke and Lesaffre (1997) and Heagerty and Kurland (2001) on the behaviour of estimators under misspecification of the random effects distribution in LME models. If one would like to avoid making distributional assumptions for the random effects, it is possible to estimate the wide model with the nonparametric maximum likelihood estimator of the random effect distribution (see Verbeke, Spiessens and Lesaffre (2001)). A FIC procedure using such a wide model might be more robust against potential misspecification of the random effects distribution

Another line of potential modifications consists of taking into account that the variance-covariance parameters in LME models are boundary parameters, meaning for instance that the covariance matrix D is required to be positive definite with each diagonal element positive. If some of the variance-covariance parameters are close to (or on) their boundary, the resulting asymptotic distribution of the ML estimators is not normal; see the theory exposited in Claeskens and Hjort (2008a, Ch. 10).

Our last remark is to point out that crucially, the methodology developed here can be extended to FIC model selection methods also for several other classes of candidate models, in different regression frameworks, see Claeskens, Cunen and Hjort (2019). As long as there is a fixed wide model, under which results for each candidate model corresponding to (4.9) and (4.11) can be reached, then only few more steps are required to reach a FIC in that framework. In particular, whereas Claeskens and Hjort (2008b) develop one type of mse approximations and FIC methods for generalised linear models, using local $O(1/\sqrt{n})$ asymptotics, the present approach actually leads to new and more versatile FIC formulae. These new FICs will have a different point of departure, namely the setting up of a fixed wide model, and can be derived as in the present paper, without any local asymptotics.

Acknowledgements. The authors thank Kenji Konishi and the other scientists at the Institute of Cetacean Research for obtaining the body condition data of Antarctic minke whale, and the IWC Scientific Committee's Data Availability Group (DAG) for facilitating the access to the data. We are also deeply grateful for the comments from the several anonymous reviewers.

References.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version* **1** 1–23.

Behl, P., Dette, H., Frondel, M. and Tauchmann, H. (2012). Choice is suffering: a focused information criterion for model selection. *Economic Modelling* **29** 817–822.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. et al. (2013). Valid post-selection inference. *Annals of Statistics* **41** 802–837.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24** 127–135.

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics* **66** 1069–1077.

Brownlees, C. T. and Gallo, G. M. (2008). On Variable Selection for Volatility Forecasting: The Role of Focused Selection Criteria. *Journal of Financial Econometrics* **6** 513–539.

Charkhi, A. and Claeskens, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* **105** 645–664.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2007). Prediction focussed model selection for autoregressive models. *Australian and New Zealand Journal of Statistics* **49** 359–379.

Claeskens, G., Cunen, C. and Hjort, N. L. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution* **7** 415.

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

Claeskens, G. and Hjort, N. L. (2008a). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

Claeskens, G. and Hjort, N. L. (2008b). Minimizing average risk in regression models. *Econometric Theory* **24** 493–527.

Craiu, R. V. and Duchesne, T. (2018). A scalable and efficient covariate selection criterion for mixed effects regression models with unknown random effects structure. *Computational Statistics and Data Analysis* **117** 154–161.

Cunen, C., Walløe, L. and Hjort, N. L. (2017). Decline in energy storage in Antarctic minke whales during the JARPA period: Assessment via the Focused Information Criterion (FIC). *IWC/SC/67A/EM04* 1–55.

Cunen, C., Walløe, L. and Hjort, N. L. (2020). Decreasing fat storage in Antarctic minke whales during the 1990ies. *Submitted for publication.*

de la Mare, W., McKinlay, J. and Welsh, A. (2017). Analyses of the JARPA Antarctic minke whale fat weight data set. *IWC/SC/67A/EM01* 1–57.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R.* John Wiley & Sons.

Gandy, A. and Hjort, N. L. (2013). Focused information criteria for semiparametric linear hazard regression Technical Report, Department of Mathematics, University of Oslo.

Greenland, S., Pearl, J., Robins, J. M. et al. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

Grueber, C. E., Nakagawa, S., Laws, R. J. and Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology* **24** 699–711.

Gumedze, F. N. and Dunne, T. T. (2011). Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications* **435** 1920–1944.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88** 973–985.

Hermansen, G. H., Hjort, N. L. and Kjesbu, O. S. (2016). Modern statistical methods applied on extensive historic data: Hjort liver quality time series 1859-2012 and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences* **73** 279–295.

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

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association* **101** 1449–1464.

Hui, F. K. C., Müller, S. and Welsh, A. H. (2017). Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association* **112** 1–11.

Ibrahim, J. G., Zhu, H., Garcia, R. I. and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503.

Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics* **36** 1669–1692.

Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* **27** 951–981.

Jullum, M. and Hjort, N. L. (2019). What price semiparametric Cox regression? *Lifetime data analysis* **25** 406–438.

Konishi, K. and Walløe, L. (2015). Substantial decline in energy storage and stomach fullness in Antarctic minke whales during the 1990s. *Journal of Cetacean Research and Management* **15** 77–92.

Konishi, K., Tamura, T., Zenitani, R., Bando, T., Kato, H. and Walløe, L. (2008). Decline in energy storage in the Antarctic minke whale (Balaenoptera bonaerensis) in the Southern Ocean. *Polar Biology* **31** 1509–1520.

Laws, R. M. (1977). Seals and whales of the Southern Ocean. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **279** 81–96.

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *Annals of Statistics* **7** 381–394.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus With Applications in Statistics and Econometrics*. Wiley, New York.

McKinlay, J., de la Mare, W. and Welsh, A. (2017). A re-examination of minke whale body condition as reflected in the data. *IWC/SC/67A/EM02* 1–108.

Mori, M. and Butterworth, D. S. (2006). A first step towards modelling the krill–predator dynamics of the Antarctic ecosystem. *CCAMLR Science* **13** 217–277.

Müller, S., Scealy, J. L. and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science* **28** 135–167.

Government of Japan (1987). The Program for Research on the Southern Hemisphere Minke Whale and for Preliminary Research on the Marine Ecosystem in the Antarctic. *IWC/SC/39/04 [Available from the IWC Secretariat]* 1–57.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis* **109** 109–129.

Pinheiro, J. and Bates, D. (2000). *Mixed-effects Models in S and S-PLUS*. Springer.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111** 600–620.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92** 351–370.

Ver Hoef, J. M. and Boveng, P. L. (2015). Iterating on a single model is a viable alternative to multimodel inference. *The Journal of Wildlife Management* **79** 719–729.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis* **23** 541–556.

Verbeke, G., Spiessens, B. and Lesaffre, E. (2001). Conditional linear mixed models. *American Statistician* **55** 25–34.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics* **39** 174–200.

APPENDIX A: CANDIDATE MODELS FOR THE APPLICATION

The following candidate models were investigated for our whale ecology application (Section 7). The wide model M_0 is given in the main text.

M_1: simplified main effects, with quadratic year term and bigger random effect structure than the wide:

$$\begin{aligned}\texttt{fatweight} \sim\ & \texttt{year} + \texttt{year}^2 + \texttt{bodylength} + \texttt{sex} + \texttt{diatom} + \texttt{date} + \\ & \texttt{date}^2 + \texttt{bodylength} * \texttt{date} + \texttt{bodylength} * \texttt{date}^2 + \\ & (1 + \texttt{date} + \texttt{date}^2 + \texttt{bodylength} + \\ & \texttt{bodylength} * \texttt{date} + \texttt{bodylength} * \texttt{date}^2 \,|\, \texttt{year}).\end{aligned}$$

M_2: simplified main effects, with linear year term and simplified random effect structure:

$$\texttt{fatweight} \sim \texttt{year} + \texttt{bodylength} + \texttt{sex} + \texttt{date} + (1 + \texttt{date} \,|\, \texttt{year}).$$

M_3: simplified main effects, with linear year term and even simpler random effect structure:

$$\texttt{fatweight} \sim \texttt{year} + \texttt{bodylength} + \texttt{sex} + \texttt{date} + (1 \,|\, \texttt{year}).$$

M_4: only linear year term and random effect on intercept:

$$\texttt{fatweight} \sim \texttt{year} + (1 \,|\, \texttt{year}).$$

M_5: same as M_0, but without any (fixed) year term. Note that this model was necessary to specify only for the second focus parameter, $\mu_2 = P(Y \geq 1500 \,|\, x_0, z_0)$.

$$\begin{aligned}\texttt{fatweight} \sim\ & \texttt{bodylength} + \texttt{sex} + \texttt{diatom} + \texttt{date} + \texttt{date}^2 + \\ & \texttt{age} + \texttt{sex} * \texttt{diatom} + \texttt{diatom} * \texttt{date} + \texttt{diatom} * \texttt{date}^2 + \\ & \texttt{bodylength} * \texttt{sex} + \texttt{bodylength} * \texttt{date} + \\ & \texttt{bodylength} * \texttt{date}^2 + \texttt{sex} * \texttt{date} + \texttt{sex} * \texttt{date}^2 + \\ & \texttt{bodylength} * \texttt{sex} * \texttt{date} + \texttt{bodylength} * \texttt{sex} * \texttt{date}^2 + \\ & \texttt{age} * \texttt{sex} + \texttt{age} * \texttt{date} + \texttt{age} * \texttt{date}^2 + \\ & \texttt{age} * \texttt{sex} * \texttt{date} + \texttt{age} * \texttt{sex} * \texttt{date}^2 + \texttt{region} + \\ & \texttt{sex} * \texttt{region} + \texttt{diatom} * \texttt{region} + \texttt{region} * \texttt{date}^2 + \\ & \texttt{region} * \texttt{date} + (1 + \texttt{date} + \texttt{date}^2 \,|\, \texttt{year}).\end{aligned}$$

APPENDIX B: FORMULAE FOR THE $J_{M,N}$, $K_{M,N}$ AND $C_{M,N}$ MATRICES

Let the wide model and the necessary design matrices be defined as in (3.2) and similarly for the candidate model in (4.2). The wide model has the true parameter vector $\theta_{\text{true}} = (\beta_{\text{true}}, \sigma_{\text{true}}, \text{vech}(D_{\text{true}}))$ of dimension $p + 1 + k(k+1)/2$. The candidate model aims for the least false parameters $\theta_{M,0} = (\beta_{M,0}, \sigma_{M,0}, \text{vech}(D_{M,0}))$ of dimension $p_M + 1 + k_M(k_M + 1)/2$.

We will make use of the vector functions vec and vech. Both take matrices as their input and output vectors; vec stacks the columns of the input matrix, and vech stacks the lower triangular part of the matrix. Thus, vech(D_{true}) is the vector of unique elements defining D_{true}. We have the following relations between the vec() and vech() representations (for instance from Demidenko (2013)):

$$\text{vec}(D) = W_k \text{vech}(D) \tag{B.1}$$

where W_k is the $k^2 \times k(k+1)/2$ duplication matrix.

In order to compute the FIC scores, we need formulae for the following matrices,

$$\begin{aligned}
J_{M,n} &= -n^{-1}\sum_{i=1}^{n} \mathrm{E}_{\mathrm{wide}} \frac{\partial^2 \ell_{M,i}(\theta_{M,0})}{\partial\theta_M \partial\theta_M^{\mathrm{t}}}, \\
K_{M,n} &= n^{-1}\sum_{i=1}^{n} \mathrm{Var}_{\mathrm{wide}}\, u_{M,i}(y,\theta_{M,0}), \\
C_{M,n} &= n^{-1}\sum_{i=1}^{n} \mathrm{Cov}_{\mathrm{wide}}\left\{u_i(y,\theta_{\mathrm{true}}), u_{M,i}(y,\theta_{M,0})\right\},
\end{aligned}$$

where $u_{M,i}(y,\theta_M) = \partial\ell_{M,i}/\partial\theta_M$ is the score function of the candidate model and $\ell_{M,i}$ is the log-likelihood for the ith group of the candidate model. The expectation and variance are taken with respect to the wide linear mixed effect model.

B.1. Score function and Hessian matrix. First, we need expressions for the score vector and Hessian matrix of a general LME model like in (3.2) with parameter vector $\theta = (\beta, \sigma, \mathrm{vech}(D))$. We will find these with respect to the 'full' $(p + 1 + k^2) \times 1$ parameter vector $\theta^* = (\beta, \sigma, \mathrm{vec}(D))$. Then we will convert these vectors and matrices to get the quantities with respect to θ using duplication matrices.

The score function for group or subject i is

$$\begin{aligned}
u_i^* = u_i(y_i, \theta^*) &= \begin{pmatrix} \partial\ell_i/\partial\beta \\ \partial\ell_i/\partial\sigma \\ \{\partial\ell_i/\partial\mathrm{vec}(D)\}^{\mathrm{t}} \end{pmatrix} \\
&= \begin{pmatrix} \sigma^{-2}X_i^{\mathrm{t}}V_i^{-1}e_i \\ -m_i\sigma^{-1} + \sigma^{-3}e_i^{\mathrm{t}}V_i^{-1}e_i \\ -\frac{1}{2}\mathrm{vec}(Z_i^{\mathrm{t}}V_i^{-1}Z_i) + \frac{1}{2}\sigma^{-2}\mathrm{vec}(Z_i^{\mathrm{t}}V_i^{-1}e_ie_i^{\mathrm{t}}V_i^{-1}Z_i) \end{pmatrix},
\end{aligned}$$

where $e_i = y_i - X_i\beta$. Note that $\mathrm{vec}(Z_i^{\mathrm{t}}V_i^{-1}e_ie_i^{\mathrm{t}}V_i^{-1}Z_i) = Z_i^{\mathrm{t}}V_i^{-1}e_i \otimes Z_i^{\mathrm{t}}V_i^{-1}e_i$. Note also that we take the transpose of $\partial\ell_i/\partial\mathrm{vec}(D)$. This is because we follow the convention from Demidenko (2013) and Magnus and Neudecker (1988) where differentiation of a scalar by a column vector gives a row vector.

The Hessian matrix for group or subject i, first with respect to $\theta^* = (\beta, \sigma, \mathrm{vec}(D))$, is

$$I_i^* = I_i(y_i, \theta^*) = \begin{bmatrix} I_{11}^{p\times p} & I_{12}^{p\times 1} & I_{13}^{p\times k^2} \\ I_{12}^{\mathrm{t}} & I_{22}^{1\times 1} & I_{23}^{1\times k^2} \\ I_{13}^{\mathrm{t}} & I_{23}^{\mathrm{t}} & I_{33}^{k^2\times k^2} \end{bmatrix},$$

where

$$\begin{aligned}
I_{11} &= \partial^2\ell_i/\partial\beta^2 = -\sigma^{-2}X_i^{\mathrm{t}}V_i^{-1}X_i\\
I_{12} &= \partial^2\ell_i/(\partial\beta\partial\sigma) = -2\sigma^{-3}X_i^{\mathrm{t}}V_i^{-1}e_i\\
I_{13} &= \partial^2\ell_i/(\partial\beta\partial\mathrm{vec}(D)) = -\sigma^{-2}(e_i^{\mathrm{t}}V_i^{-1}Z_i\otimes X_i^{\mathrm{t}}V_i^{-1}Z_i)\\
I_{22} &= \partial^2\ell_i/\partial\sigma^2 = m\sigma^{-2} - 3\sigma^{-4}e_i^{\mathrm{t}}V_i^{-1}e_i\\
I_{23} &= \partial^2\ell_i/(\partial\sigma\partial\mathrm{vec}(D)) = -\sigma^{-3}(e_i^{\mathrm{t}}V_i^{-1}Z_i\otimes e_i^{\mathrm{t}}V_i^{-1}Z_i)\\
I_{33} &= \partial^2\ell_i/(\partial\mathrm{vec}(D)\partial\mathrm{vec}(D)^{\mathrm{t}}) = \tfrac{1}{2}\{R_i\otimes R_i - \sigma^{-2}(Z_i^{\mathrm{t}}V_i^{-1}e_ie_i^{\mathrm{t}}V_i^{-1}Z_i\otimes R_i\\
&\quad +R_i\otimes Z_i^{\mathrm{t}}V_i^{-1}e_ie_i^{\mathrm{t}}V_i^{-1}Z_i)\},
\end{aligned}$$

with

$$e_i = y_i - X_i\beta, \qquad V_i = I + Z_iDZ_i^{\mathrm{t}}, \qquad R_i = Z_i^{\mathrm{t}}V_i^{-1}Z_i.$$

We have differentiated with respect to $\theta^* = (\beta,\sigma,\mathrm{vec}(D))$, but we actually need the differentiation with respect to $\theta = (\beta,\sigma,\mathrm{vech}(D))$. Using the relation in (B.1) and the chain rule for differentiation we get

$$\frac{\partial\ell_i}{\partial\mathrm{vech}(D)} = \frac{\partial\ell_i}{\partial\mathrm{vec}(D)}\frac{\partial\mathrm{vec}(D)}{\partial\mathrm{vech}(D)} = \frac{\partial\ell_i}{\partial\mathrm{vec}(D)}W_k.$$

Finally, to obtain a column vector we have

$$\begin{aligned}
\{\partial\ell_i/\partial\mathrm{vech}(D)\}^{\mathrm{t}} &= W_k^{\mathrm{t}}\{\partial\ell_i/\partial\mathrm{vec}(D)\}^{\mathrm{t}}\\
&= -\tfrac{1}{2}W_k^{\mathrm{t}}\mathrm{vec}(Z_i^{\mathrm{t}}V_i^{-1}Z_i) + \tfrac{1}{2}\sigma^{-2}W_k^{\mathrm{t}}\mathrm{vec}(Z_i^{\mathrm{t}}V_i^{-1}e_ie_i^{\mathrm{t}}V_i^{-1}Z_i).
\end{aligned}$$

The multiplication with the duplication matrix ensures that the elements in $\partial\ell_i/\partial\mathrm{vech}(D)$ belonging to the off-diagonal elements of D are multiplied by 2 compared with the corresponding elements in $\partial\ell_i/\partial\mathrm{vec}(D)$. We end up with the following score function

$$u_i(y_i,\theta) = \begin{pmatrix}\partial\ell_i/\partial\beta\\ \partial\ell_i/\partial\sigma\\ W_k^{\mathrm{t}}\{\partial\ell_i/\partial\mathrm{vec}(D)\}^{\mathrm{t}}\end{pmatrix}.$$

Similarly, for the Hessian matrix, we get

$$\frac{\partial^2\ell_i}{\partial\mathrm{vech}(D)\,\partial\mathrm{vech}(D)^{\mathrm{t}}} = W_k^{\mathrm{t}}\frac{\partial^2\ell_i}{\partial\mathrm{vec}(D)\,\partial\mathrm{vec}(D)^{\mathrm{t}}}W_k,$$

and

$$I_i(y_i,\theta) = \begin{bmatrix} I_{11} & I_{12} & I_{13}W_k\\ I_{12}^{\mathrm{t}} & I_{22} & I_{23}W_k\\ W_k^{\mathrm{t}}I_{13}^{\mathrm{t}} & W_k^{\mathrm{t}}I_{23}^{\mathrm{t}} & W_k^{\mathrm{t}}I_{33}W_k\end{bmatrix}.$$

B.2. Finding $\boldsymbol{J_{M,n}}$. The matrix $J_{M,n}$ is minus the expected value of the Hessian matrix of the candidate model, evaluated at the least false parameters $\theta_{M,0} = (\beta_{M,0}, \sigma_{M,0}, \mathrm{vech}(D_{M,0}))$, with the expectation taken with respect to the wide model:

$$J_{M,n} = -n^{-1}\sum_{i=1}^{n} \mathrm{E}_{\mathrm{wide}} \frac{\partial^2 \ell_{M,i}(\theta_{M,0})}{\partial\theta_M \partial\theta_M^{\mathrm{t}}}.$$

We shall need $\mathrm{E}_w\, y_i = X_i\beta_{\mathrm{true}}$, $\mathrm{Var}_w\, y_i = \sigma^2_{\mathrm{true}}(I + Z_i D_{\mathrm{true}} Z_i^{\mathrm{t}}) = \sigma^2_{\mathrm{true}} V_i$, where we here and several places below use 'w' as shorthand for the subscript 'wide'. Let also

$$e_{M,i} = y_i - X_{M,i}\beta_{M,0},$$

along with $\mu_{e,i} = \mathrm{E}_w\, e_{M,i} = X_i\beta_{\mathrm{true}} - X_{M,i}\beta_{M,0}$ and $V_{M,0,i} = I + Z_{M,i} D_{M,0} Z_{M,i}^{\mathrm{t}}$. We will make use of the following general formulae. (i) The expectation of a quadratic form: Let A be a matrix and x a random vector with expectation μ and covariance matrix Σ. Then $\mathrm{E}\, x^{\mathrm{t}} A x = \mathrm{Tr}(A\Sigma) + \mu^{\mathrm{t}} A\mu$. (ii) With X a random matrix, $\mathrm{E}\,(A \otimes X) = A \otimes \mathrm{E}\, X$ and $\mathrm{E}\,(X \otimes A) = \mathrm{E}\, X \otimes A$; cf. Magnus and Neudecker (1979, Theorem 4.3). (iii) $\mathrm{E}\,\mathrm{vec}(X) = \mathrm{vec}(\mathrm{E}\, X)$.

We get

$$\begin{aligned} J_{M,n} &= -n^{-1}\textstyle\sum_{i=1}^n \begin{bmatrix} \mathrm{E}_w(I_{11,i}) & \mathrm{E}_w(I_{12,i}) & \mathrm{E}_w(I_{13,i})W_{k_M} \\ \mathrm{E}_w(I_{12,i})^{\mathrm{t}} & \mathrm{E}_w(I_{22,i}) & \mathrm{E}_w(I_{23,i})W_{k_M} \\ W_{k_M}^{\mathrm{t}}\mathrm{E}_w(I_{13,i})^{\mathrm{t}} & W_{k_M}^{\mathrm{t}}\mathrm{E}_w(I_{23,i})^{\mathrm{t}} & W_{k_M}^{\mathrm{t}}\mathrm{E}_w(I_{33,i})W_{k_M} \end{bmatrix} \\ &= n^{-1}\textstyle\sum_{i=1}^n \begin{bmatrix} J_{11,i} & J_{12,i} & J_{13,i}W_{k_M} \\ J_{12,i}^{\mathrm{t}} & J_{22,i} & J_{23,i}W_{k_M} \\ W_{k_M}^{\mathrm{t}} J_{13,i}^{\mathrm{t}} & W_{k_M}^{\mathrm{t}} J_{23,i}^{\mathrm{t}} & W_{k_M}^{\mathrm{t}} J_{33,i} W_{k_M} \end{bmatrix}, \end{aligned}$$

with

$$\begin{aligned} J_{11,i} &= \sigma_{M,0}^{-2} X_{M,i}^{\mathrm{t}} V_{M,i}^{-1} X_{M,i}, \\ J_{12,i} &= 2\sigma_{M,0}^{-3} X_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1}\mu_{e,i}, \\ J_{13,i} &= \sigma_{M,0}^{-2}(\mu_{e,i}^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i} \otimes X_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i}), \\ J_{22,i} &= -m_i\sigma_{M,0}^{-2} + 3\sigma_{M,0}^{-4}\{\sigma^2_{\mathrm{true}}\mathrm{Tr}(V_{M,0,i}^{-1}V_i) + \mu_{e,i}^{\mathrm{t}} V_{M,0,i}^{-1}\mu_{e,i}\}, \\ J_{23,i} &= \sigma_{M,0}^{-3}[\sigma^2_{\mathrm{true}}\mathrm{vec}(Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} V_i V_{M,0,i}^{-1} Z_{M,i}) \\ &\quad +\mu_{e,i}^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i} \otimes \mu_{e,i}^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i}], \\ J_{33,i} &= \tfrac{1}{2}[\sigma_{M,0}^{-2}(Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1}\{\sigma^2_{\mathrm{true}}V_i + \mu_{e,i}\mu_{e,i}^{\mathrm{t}}\}V_{M,0,i}Z_{M,i} \otimes R_{M,i} \\ &\quad +R_{M,i} \otimes Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1}(\sigma^2_{\mathrm{true}}V_i + \mu_{e,i}\mu_{e,i}^{\mathrm{t}})V_{M,0,i}^{-1}Z_{M,i}) - R_{M,i} \otimes R_{M,i}]. \end{aligned}$$

Note that $\sum_{i=1}^n J_{12,i} = 0$.

B.3. Finding $K_{M,n}$. The matrix $K_{M,n}$ is the variance of the score function of the candidate model, evaluated at the least false parameters $\theta_{M,0} = (\beta_{M,0}, \sigma_{M,0}, \mathrm{vech}(D_{M,0})$, with the variance taken with respect to the wide model:

$$\begin{aligned} K_{M,n} &= n^{-1} \sum_{i=1}^{n} \mathrm{Var}_{\mathrm{wide}}\, u_{M,i}(y, \theta_{M,0}) \\ &= n^{-1} \sum_{i=1}^{n} \begin{bmatrix} K_{11,i} & K_{12,i} & K_{13,i} W_{k_M} \\ K_{12,i}^{\mathrm{t}} & K_{22,i} & K_{23,i} W_{k_M} \\ W_{k_M}^{\mathrm{t}} K_{13,i}^{\mathrm{t}} & W_{k_M}^{\mathrm{t}} K_{23,i}^{\mathrm{t}} & W_{k_M}^{\mathrm{t}} K_{33,i} W_{k_M} \end{bmatrix}. \end{aligned}$$

We will then make use of the following well-known facts, found in e.g. Magnus and Neudecker (1979): $\mathrm{vec}(ABC) = (C^{\mathrm{t}} \otimes A)\mathrm{vec}(B)$, and $\mathrm{vec}(xy^{\mathrm{t}}) = y \otimes x$. Note also that $e_{M,i} \sim \mathrm{N}_{m_i}(\mu_{e,i}, \sigma_{\mathrm{true}}^2 V_i)$, where we will need the following formula for the variance of a quadratic form: If A is a matrix and x a random vector with expectation μ and covariance matrix Σ, then $\mathrm{Var}\, x^{\mathrm{t}} A x = 2\,\mathrm{Tr}(A\Sigma A\Sigma) + 4\mu^{\mathrm{t}} A \Sigma A \mu$. In the first step we get

$$\begin{aligned} K_{11,i} &= \sigma_{M,0}^{-4} \sigma_{\mathrm{true}}^2 X_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} V_i V_{M,0,i}^{-1} X_{M,i}, \\ K_{12,i} &= \sigma_{M,0}^{-5} X_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} \mathrm{Cov}(e_{M,i}, e_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} e_{M,i}), \\ K_{13,i} &= \tfrac{1}{2} \sigma_{M,0}^{-4} X_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} \mathrm{Cov}(e_{M,i}, e_{M,i} \otimes e_{M,i}) (Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} \otimes Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1})^{\mathrm{t}}, \\ K_{22,i} &= \sigma_{M,0}^{-6} [2\sigma_{\mathrm{true}}^4 \mathrm{Tr}(V_{M,0,i}^{-1} V_i V_{M,0,i}^{-1} V_i) + 4\sigma_{\mathrm{true}}^2 \mu_{e,i}^{\mathrm{t}} V_{M,0,i}^{-1} V_i V_{M,0,i}^{-1} \mu_{e,i}], \\ K_{23,i} &= \tfrac{1}{2} \sigma_{M,0}^{-5} \mathrm{Cov}(e_{M,i}^{\mathrm{t}} V_i^{-1} e_{M,i}, e_{M,i} \otimes e_{M,i}) (Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} \otimes Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1})^{\mathrm{t}}, \\ K_{33,i} &= \tfrac{1}{4} \sigma_{M,0}^{-4} (Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} \otimes Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1}) \mathrm{Var}(e_{M,i} \otimes e_{M,i}) \\ &\quad (Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} \otimes Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1})^{\mathrm{t}}. \end{aligned}$$

It is straightforward to show that

$$\mathrm{Cov}\,(e_{M,i}, e_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} e_{M,i}) = 2\sigma_{\mathrm{true}}^2 V_i V_{M,0,i}^{-1} \mu_{e,i}.$$

The next covariance term needs more efforts to work out,

$$\mathrm{Cov}\,(e_{M,i}, e_{M,i} \otimes e_{M,i}) = \sigma_{\mathrm{true}}^2 (V_i \otimes \mu_{e,i}^{\mathrm{t}} + \mu_{e,i}^{\mathrm{t}} \otimes V_i).$$

One way to show this is to write $e_{M,i} = \mu_{e,i} + Sz$, where $z \sim \mathrm{N}_{m_i}(0, I)$ and S is chosen such that $SS^{\mathrm{t}} = \sigma_{\mathrm{true}}^2 V_i$. Then may then write

$$\begin{aligned} \mathrm{Cov}\,(e_{M,i}, e_{M,i} \otimes e_{M,i}) = {} & S\,\mathrm{Cov}(z, \mathrm{vec}(\mu_{e,i} z^{\mathrm{t}} S^{\mathrm{t}})) + S\,\mathrm{Cov}(z, \mathrm{vec}(S z \mu_{e,i}^{\mathrm{t}})) \\ & + S\,\mathrm{Cov}(z, \mathrm{vec}(S z z^{\mathrm{t}} S^{\mathrm{t}})). \end{aligned}$$

The last term disappears because $S\,\mathrm{Cov}(z,\mathrm{vec}(zz^{\mathrm{t}}))(S\otimes S)^{\mathrm{t}}$ as well as $\mathrm{Cov}(z,\mathrm{vec}(zz^{\mathrm{t}}))$ are equal to zero. We are left with $\mathrm{Cov}\,(e_{M,i},e_{M,i}\otimes e_{M,i})=S(S^{\mathrm{t}}\otimes\mu_{e,i}^{\mathrm{t}})+S(\mu_{e,i}^{\mathrm{t}}\otimes S^{\mathrm{t}})$, and using $S(S^{\mathrm{t}}\otimes\mu_{e,i}^{\mathrm{t}})=(S\otimes 1)(S^{\mathrm{t}}\otimes\mu_{e,i}^{\mathrm{t}})=SS^{\mathrm{t}}\otimes\mu_{e,i}^{\mathrm{t}}=\sigma_{\mathrm{true}}^2V_i\otimes\mu_{e,i}^{\mathrm{t}}$ we find the result stated above. Noticing that

$$\begin{aligned}e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i}&=\mathrm{vec}(e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i})\\&=(e_{M,i}^{\mathrm{t}}\otimes e_{M,i}^{\mathrm{t}})\mathrm{vec}(V_{M,0,i}^{-1})=\mathrm{vec}(V_{M,0,i}^{-1})^{\mathrm{t}}(e_{M,i}\otimes e_{M,i}),\end{aligned}$$

one can find the third covariance term,

$$\mathrm{Cov}\,(e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i},e_{M,i}\otimes e_{M,i})=\mathrm{vec}(V_{M,0,i}^{-1})^{\mathrm{t}}\,\mathrm{Var}(e_{M,i}\otimes e_{M,i}).$$

Theorem 4.3 in Magnus and Neudecker (1979) leads to

$$\mathrm{Var}\,e_{M,i}\otimes e_{M,i}=\sigma_{\mathrm{true}}^2(I_{m_i}+K_{m_i})(\sigma_{\mathrm{true}}^2V_i\otimes V_i+V_i\otimes\mu_{e,i}\mu_{e,i}^{\mathrm{t}}+\mu_{e,i}\mu_{e,i}^{\mathrm{t}}\otimes V_i)$$

where K_{m_i} is a commutation matrix. Let

$$Q=2\sigma_{\mathrm{true}}^2(\sigma_{\mathrm{true}}^2V_i\otimes V_i+V_i\otimes\mu_{e,i}\mu_{e,i}^{\mathrm{t}}+\mu_{e,i}\mu_{e,i}^{\mathrm{t}}\otimes V_i),$$

and note that using properties of commutation matrices K_{m_i} will disappear in the expressions (partly due to the duplication matrices). Finally we have

$$\begin{aligned}K_{11,i}&=\sigma_{M,0}^{-4}\sigma_{\mathrm{true}}^2X_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}V_iV_{M,0,i}^{-1}X_{M,i},\\K_{12,i}&=2\sigma_{M,0}^{-5}\sigma_{\mathrm{true}}^2X_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}V_iV_{M,0,i}^{-1}\mu_{e,i},\\K_{13,i}&=\tfrac{1}{2}\sigma_{M,0}^{-4}\sigma_{\mathrm{true}}^2X_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}(V_i\otimes\mu_{e,i}^{\mathrm{t}}+\mu_{e,i}^{\mathrm{t}}\otimes V_i)(Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})^{\mathrm{t}},\\K_{22,i}&=\sigma_{M,0}^{-6}\{2\sigma_{\mathrm{true}}^4\,\mathrm{Tr}(V_{M,0,i}^{-1}V_iV_{M,0,i}^{-1}V_i)+4\sigma_{\mathrm{true}}^2\mu_{e,i}^{\mathrm{t}}V_{M,0,i}^{-1}V_iV_{M,0,i}^{-1}\mu_{e,i}\},\\K_{23,i}&=\tfrac{1}{2}\sigma_{M,0}^{-5}\mathrm{vec}(V_{M,0,i}^{-1})^{\mathrm{t}}Q(Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})^{\mathrm{t}},\\K_{33,i}&=\tfrac{1}{4}\sigma_{M,0}^{-4}(Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})Q(Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})^{\mathrm{t}}.\end{aligned}$$

B.4. Finding $\boldsymbol{C_{M,N}}$. The matrix $C_{M,n}$ is the covariance between the score function of the wide model and the score function of the candidate model, evaluated at the true parameter vector $\theta_{\mathrm{true}}=(\beta_{\mathrm{true}},\sigma_{\mathrm{true}},\mathrm{vech}(D_{\mathrm{true}}))$ and at the least false parameters $\theta_{M,0}=(\beta_{M,0},\sigma_{M,0},\mathrm{vech}(D_{M,0}))$ respectively, with the covariance taken with respect to the wide model:

$$C_{M,n}=n^{-1}\mathrm{Cov}\Big\{\sum_{i=1}^n u_i(y_i,\theta_{\mathrm{true}}),\sum_{i=1}^n u_{M,i}(y_i,\theta_{M,0})\Big\}.$$

Since observations from different groups are independent, we only need to consider the covariances between the score functions from the same group, so

$$\begin{aligned} C_{M,n} &= n^{-1}\sum_{i=1}^{n} \mathrm{Cov}\,\{u_i(y_i,\theta_{\mathrm{true}}), u_{M,i}(y_i,\theta_{M,0})\} \\ &= n^{-1}\sum_{i=1}^{n} \begin{bmatrix} C_{11,i} & C_{12,i} & C_{13,i}W_{k_M} \\ C_{21,i} & C_{22,i} & C_{23,i}W_{k_M} \\ W_k^{\mathrm{t}}C_{31,i} & W_k^{\mathrm{t}}C_{32,i} & W_k^{\mathrm{t}}C_{33,i}W_{k_M} \end{bmatrix}, \end{aligned}$$

which will not be symmetric. With the required stamina and algebraic efforts we find

$$\begin{aligned} C_{11,i} &= \sigma_{M,0}^{-2}\sigma_{\mathrm{true}}^{-2}X_i^{\mathrm{t}}V_i^{-1}\mathrm{Cov}(e_i,e_{M,i})V_{M,0,i}^{-1}X_{M,i}, \\ C_{12,i} &= \sigma_{M,0}^{-3}\sigma_{\mathrm{true}}^{-2}X_i^{\mathrm{t}}V_i^{-1}\mathrm{Cov}(e_i,e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i}), \\ C_{13,i} &= \tfrac{1}{2}\sigma_{M,0}^{-2}\sigma_{\mathrm{true}}^{-2}X_i^{\mathrm{t}}V_i^{-1}\mathrm{Cov}(e_i,e_{M,i}\otimes e_{M,i})(Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})^{\mathrm{t}}, \\ C_{22,i} &= \sigma_{M,0}^{-3}\sigma_{\mathrm{true}}^{-3}\mathrm{Cov}(e_i^{\mathrm{t}}V_i^{-1}e_i, e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i}), \\ C_{23,i} &= \tfrac{1}{2}\sigma_{M,0}^{-2}\sigma_{\mathrm{true}}^{-3}\mathrm{Cov}(e_i^{\mathrm{t}}V_i^{-1}e_i, e_{M,i}\otimes e_{M,i}), (Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})^{\mathrm{t}} \\ C_{33,i} &= \tfrac{1}{4}\sigma_{M,0}^{-2}\sigma_{\mathrm{true}}^{-2}(Z_i^{\mathrm{t}}V_i^{-1}\otimes Z_i^{\mathrm{t}}V_i^{-1})\mathrm{Cov}(e_i\otimes e_i, e_{M,i}\otimes e_{M,i}) \\ &\quad (Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}\otimes Z_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1})^{\mathrm{t}}, \\ C_{21,i} &= \sigma_{M,0}^{-2}\sigma_{\mathrm{true}}^{-3}\mathrm{Cov}(e_i^{\mathrm{t}}V_i^{-1}e_i, e_{M,i})V_{M,0,i}^{-1}X_{M,i}, \\ C_{31,i} &= \tfrac{1}{2}\sigma_{M,0}^{-2}\sigma_{\mathrm{true}}^{-2}(Z_i^{\mathrm{t}}V_i^{-1}\otimes Z_i^{\mathrm{t}}V_i^{-1})\mathrm{Cov}(e_i\otimes e_i, e_{M,i})V_{M,0,i}^{-1}X_{M,i}, \\ C_{32,i} &= \tfrac{1}{2}\sigma_{M,0}^{-3}\sigma_{\mathrm{true}}^{-2}(Z_i^{\mathrm{t}}V_i^{-1}\otimes Z_i^{\mathrm{t}}V_i^{-1})\mathrm{Cov}(e_i\otimes e_i, e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i}). \end{aligned}$$

Here we need to find a little list of expressions for different covariances, involving the quantities $e_i \sim \mathrm{N}_{m_i}(0,\sigma_{\mathrm{true}}^2V_i)$, $e_{M,i} \sim \mathrm{N}_{m_i}(\mu_e,\sigma_{\mathrm{true}}^2V_i)$, with $\mu_e = X_i\beta_{\mathrm{true}} - X_{M,i}\beta_{M,0}$. Note that $e_{M,i} = e_i + \mu_e$.

1. $\mathrm{Cov}(e_i,e_{M,i}) = \mathrm{Var}\,y_i = \sigma_{\mathrm{true}}^2V_i$.
2. $\mathrm{Cov}(e_i,e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i}) = 2\mathrm{Cov}(e_i,\mu_e^{\mathrm{t}}V_{M,0,i}^{-1}e_i) = 2\sigma_{\mathrm{true}}^2V_iV_{M,0,i}^{-1}\mu_e$.
3. $\mathrm{Cov}(e_i^{\mathrm{t}}V_i^{-1}e_i, e_{M,i}^{\mathrm{t}}V_{M,0,i}^{-1}e_{M,i}) = \mathrm{Cov}(e_i^{\mathrm{t}}V_i^{-1}e_i, e_i^{\mathrm{t}}V_{M,0,i}^{-1}e_i) = 2\sigma^4$ $\mathrm{Tr}(V_{M,0,i}^{-1}V_i)$.
4. $\mathrm{Cov}(e_i,e_{M,i}\otimes e_{M,i}) = \sigma_{\mathrm{true}}^2V_i[\mu_e^{\mathrm{t}}\otimes I_{m_i} + I_{m_i}\otimes\mu_e^{\mathrm{t}}]$.
5. $\mathrm{Cov}(e_i\otimes e_i, e_{M,i}\otimes e_{M,i}) = \mathrm{Cov}(e_i\otimes e_i, e_i\otimes e_i) = \mathrm{Var}(e_i\otimes e_i) = \sigma_{\mathrm{true}}^4(I_{m_i}+K_{m_i})(V_i\otimes V_i)$ (again, K_{m_i} is a commutation matrix).
6. $\mathrm{Cov}(e_i^{\mathrm{t}}V_i^{-1}e_i, e_{M,i}\otimes e_{M,i}) = \mathrm{vec}(V_i^{-1})^{\mathrm{t}}\mathrm{Cov}(e_i\otimes e_i, e_{M,i}\otimes e_{M,i}) = \sigma_{\mathrm{true}}^4$ $\mathrm{vec}(V_i^{-1})^{\mathrm{t}}(I_{m_i}+K_{m_i})(V_i\otimes V_i) = 2\sigma_{\mathrm{true}}^4\mathrm{vec}(V_i^{-1})^{\mathrm{t}}(V_i\otimes V_i)$ (using the definition of a commutation matrix).

7. $\mathrm{Cov}(e_i^{\mathrm{t}} V_i^{-1} e_i, e_{M,i}) = \mathrm{Cov}(e_i^{\mathrm{t}} V_i^{-1} e_i, e_i) = 0$ (by a property of quadratic forms used above).
8. $\mathrm{Cov}(e_i \otimes e_i, e_{M,i}) = \mathrm{Cov}(e_i \otimes e_i, e_i) = 0$.
9. $\mathrm{Cov}(e_i \otimes e_i, e_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} e_{M,i}) = \mathrm{Cov}(e_i \otimes e_i, e_{M,i} \otimes e_{M,i})\mathrm{vec}(V_{M,0,i}^{-1}) = \sigma_{\mathrm{true}}^4 (I_{m_i} + K_{m_i})(V_i \otimes V_i)\mathrm{vec}(V_{M,0,i}^{-1}) = 2\sigma_{\mathrm{true}}^4 (V_i \otimes V_i)\mathrm{vec}(V_{M,0,i}^{-1})$ (using the definition of a commutation matrix).

Point 4 relies on the following fact, writing out $\mathrm{Cov}(e_i, e_{M,i} \otimes e_{M,i}) = \mathrm{Cov}(e_i, \mathrm{vec}(e_i e_i^{\mathrm{t}} + e_i \mu_e^{\mathrm{t}} + \mu_e e_i^{\mathrm{t}}))$ and using $\mathrm{vec}(ABC) = (C^{\mathrm{t}} \otimes A)\mathrm{vec}(B)$ and that $\mathrm{Cov}(e_i, \mathrm{vec}(e_i e_i^{\mathrm{t}})) = 0$. Point 5 is found in a similar manner. After some further simplifications (also dealing with the commutation matrices), we get

$$\begin{aligned}
C_{11,M,i} &= \sigma_{M,0}^{-2} X_i^{\mathrm{t}} V_{M,0,i}^{-1} X_{M,i}, \\
C_{12,M,i} &= 2\sigma_{M,0}^{-3} X_i^{\mathrm{t}} V_{M,0,i}^{-1} \mu_e, \\
C_{13,M,i} &= \tfrac{1}{2}\sigma_{M,0}^{-2} X_i^{\mathrm{t}} (\mu_e^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i} \otimes V_{M,0,i}^{-1} Z_{M,i} + V_{M,0,i}^{-1} Z_{M,i} \otimes \mu_e^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i}), \\
C_{22,M,i} &= 2\sigma_{M,0}^{-3} \sigma_{\mathrm{true}} \mathrm{Tr}(V_{M,0,i}^{-1} V_i), \\
C_{23,M,i} &= \sigma_{M,0}^{-2} \sigma_{\mathrm{true}} \mathrm{vec}(Z_{M,i}^{\mathrm{t}} V_{M,0,i}^{-1} V_i V_{M,0,i}^{-1} Z_{M,i})^{\mathrm{t}}, \\
C_{33,M,i} &= \tfrac{1}{2}\sigma_{M,0}^{-2} \sigma_{\mathrm{true}}^2 (Z_i^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i} \otimes Z_i^{\mathrm{t}} V_{M,0,i}^{-1} Z_{M,i}), \\
C_{21,M,i} &= 0, \\
C_{31,M,i} &= 0, \\
C_{32,M,i} &= \sigma_{M,0}^{-3} \sigma_{\mathrm{true}}^2 \mathrm{vec}(Z_i^{\mathrm{t}} V_{M,0,i}^{-1} Z_i).
\end{aligned}$$

When the candidate model is the same as the wide model, we get back the information matrix under the wide model, as expected.