Empirical bias-reducing adjustments to estimating functions

Ioannis Kosmidis, Nicola Lunardon

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 62–89
https://doi.org/10.1093/jrsssb/qkad083
Advance access publication 16 September 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Empirical bias-reducing adjustments to estimating functions

Ioannis Kosmidis[1] and **Nicola Lunardon**[2]

[1]Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK
[2]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Via Torino 150, Venezia Mestre 30170, Italy

Address for correspondence: Ioannis Kosmidis, Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK. Email: ioannis.kosmidis@warwick.ac.uk

Abstract

We develop a novel, general framework for reduced-bias *M*-estimation from asymptotically unbiased estimating functions. The framework relies on an empirical approximation of the bias by a function of derivatives of estimating function contributions. Reduced-bias *M*-estimation operates either implicitly, solving empirically adjusted estimating equations, or explicitly, subtracting the estimated bias from the original *M*-estimates, and applies to partially or fully specified models with likelihoods or surrogate objectives. Automatic differentiation can abstract away the algebra required to implement reduced-bias *M*-estimation. As a result, the bias-reduction methods, we introduce have broader applicability, straightforward implementation, and less algebraic or computational effort than other established bias-reduction methods that require resampling or expectations of products of log-likelihood derivatives. If *M*-estimation is by maximising an objective, then there always exists a bias-reducing penalised objective. That penalised objective relates to information criteria for model selection and can be enhanced with plug-in penalties to deliver reduced-bias *M*-estimates with extra properties, like finiteness for categorical data models. Inferential procedures and model selection procedures for *M*-estimators apply unaltered with the reduced-bias *M*-estimates. We demonstrate and assess the properties of reduced-bias *M*-estimation in well-used, prominent modelling settings of varying complexity.

Keywords: asymptotic bias, autologistic regression, composite likelihood, infinite estimates, model selection, penalised likelihood

1 Introduction

1.1 Bias reduction methods

Reduction of estimation bias in statistical modelling is a task that has attracted immense research activity since the early days of the statistical literature. This ongoing activity resulted in an abundance of general bias-reduction methods with varying levels of applicability. As is noted in Kosmidis (2014), the majority of methods start from an estimator $\hat{\boldsymbol{\theta}}$ of an unknown parameter $\boldsymbol{\theta}$ and attempt to produce an estimator $\tilde{\boldsymbol{\theta}}$, which approximates the solution of the equation

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = \boldsymbol{B}_G(\bar{\boldsymbol{\theta}}), \tag{1}$$

with respect to $\tilde{\boldsymbol{\theta}}$. In the above, G is the typically unknown distribution function of the process that generated the data, $\boldsymbol{B}_G(\boldsymbol{\theta}) = \mathrm{E}_G(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is the bias function, and $\bar{\boldsymbol{\theta}}$ is the value that $\hat{\boldsymbol{\theta}}$ is assumed to converge to in probability as information about $\boldsymbol{\theta}$ increases, typically with the volume of the data. In general, the solution of (1) requires approximation because both the value of $\bar{\boldsymbol{\theta}}$ and G are unknown or the expectation with respect to G does not have a closed form.

Received: February 10, 2023. **Accepted:** July 29, 2023

Table 1. Classification of bias reduction methods of general applicability (Method)

				Requirements		
Method	**Model**	$\boldsymbol{B}_G(\bar{\boldsymbol{\theta}})$	**Type**	$\mathrm{E}(\cdot)$	$\partial\cdot$	$\hat{\boldsymbol{\theta}}$
Asymptotic bias correction	full	analytical	explicit	yes	yes	yes
Adjusted score functions	full	analytical	implicit	yes	yes	no
Bootstrap	partial	simulation	explicit	no	no	yes
Jackknife	partial	simulation	explicit	no	no	yes
Iterated bootstrap / Indirect inference	full	simulation	implicit	no	no	yes
Explicit RB*M*-estimation	partial	analytical	explicit	no	yes	yes
Implicit RB*M*-estimation	partial	analytical	implicit	no	yes	no

Note. The classification is based on the level of model specification (Model), the way the method approximates the bias ($\boldsymbol{B}_G(\bar{\boldsymbol{\theta}})$), the type (Type) of the method according to the classification in Kosmidis (2014), and on the method's requirements in terms of computation of expectations ($\mathrm{E}(\cdot)$), differentiation ($\partial\cdot$), and access to the original estimator ($\hat{\boldsymbol{\theta}}$). Key references include: Efron (1975, Section 10) and Cordeiro and McCullagh (1991) for asymptotic bias correction; Firth (1993) and Kosmidis and Firth (2009) for adjusted score functions; Efron and Tibshirani (1993) and Hall and Martin (1988) for bootstrap; Quenouille (1956) and Efron (1982) for jackknife; Gourieroux et al. (1993) for indirect inference, Kuk (1995) and Guerrier et al. (2019) for iterated bootstrap, and MacKinnon and Smith (1998) for refinements of the indirect inference principle; and this manuscript for RB*M*-estimation.

Table 1 classifies prominent bias-reduction methods according to various criteria relating to their applicability and operation. Given the size of the literature on bias-reduction methods, we only cite key works that shaped or greatly impacted the area. Bias-reduction methods like vanilla asymptotic bias correction (Efron, 1975), the adjusted scores functions approach in Firth (1993), indirect inference in Gourieroux et al. (1993), and iterated bootstrap in Kuk (1995) and Guerrier et al. (2019), and the refinements of the indirect inference principle in MacKinnon and Smith (1998) assume that the model can be fully and correctly specified, in the sense that G results from the assumed model for specific parameter values. That assumption allows to either have access to log-likelihood derivatives and expectations of products of those or to simulate from the model. In contrast, bias reduction methods, like jackknife (Efron, 1982; Quenouille, 1956) and bootstrap (Efron and Tibshirani, 1993; Hall and Martin, 1988) can also apply to at least partially specified models; see, also, Newey and Smith (2004) for an asymptotic bias-correction method for partially specified models with independent and identically distributed random variables that applies also to over-identified problems. Such bias-reduction methods can improve estimation in involved modelling settings, where typically surrogate inference functions are used in an attempt to either limit the number of hard-to-justify modelling assumptions or because the full likelihood function is impractical or cumbersome to compute; see, for example, Wedderburn (1974) for quasi-likelihood methods, Liang and Zeger (1986) for generalised estimating equations, and Lindsay (1988) and Varin et al. (2011) for composite likelihood methods.

Kosmidis (2014) classifies bias-reduction methods according to whether they operate in an explicit or implicit manner when approximating the solution of (1). Explicit methods estimate $\boldsymbol{B}_G(\bar{\boldsymbol{\theta}})$ and subtract that estimate from $\hat{\boldsymbol{\theta}}$. Implicit methods replace $\boldsymbol{B}_G(\bar{\boldsymbol{\theta}})$ with $\hat{\boldsymbol{B}}_G(\hat{\boldsymbol{\theta}})$ for an estimator $\hat{\boldsymbol{B}}_G$ of $\boldsymbol{B}_G$ and solve the resulting implicit equation.

Bias-reduction methods can also be classified according to whether the necessary approximation of the bias term in (1) is performed analytically or through simulation. The vanilla implementations of asymptotic bias correction and adjusted score functions approximate $\boldsymbol{B}_G(\boldsymbol{\theta})$ with a function $\boldsymbol{b}(\boldsymbol{\theta})$ such that $\boldsymbol{B}_G(\boldsymbol{\theta}) = \boldsymbol{b}(\boldsymbol{\theta}) + O(n^{-3/2})$, where n is a measure of how the information about $\boldsymbol{\theta}$ accumulates. On the other hand, jackknife, bootstrap, iterated bootstrap, and indirect inference, generally, approximate the bias by simulating samples from the assumed model or an estimator of G, like the empirical distribution function. As a result, and depending on how demanding the computation of $\hat{\boldsymbol{\theta}}$ is, simulation-based methods are typically more computationally intensive than analytical methods. Also, implicit, simulation-based methods require special care and ad-hoc considerations when approximating the solution of (1), because the simulation-based estimator of $\boldsymbol{B}_G(\boldsymbol{\theta})$ is not always differentiable with respect to $\boldsymbol{\theta}$.

The requirement of differentiation of the estimating or objective functions for some of the bias-reduction methods in Table 1 has also resulted in considerable analytical effort over the years (see, for example, Kosmidis and Firth, 2009 for multivariate generalised nonlinear models, and Grün et al., 2012 for Beta regression models). Nevertheless, differentiation is nowadays a task requiring increasingly less analytical effort because of the availability of comprehensive automatic differentiation routines (Griewank and Walther, 2008) in popular computing environments. Such routines can be found, for example, in the `ForwardDiff` Julia package (Revels et al., 2016), and the C++ package `CppAD` (Bell, 2021) that enabled the development of a range of template modelling software, like the `TMB` R package (Kristensen et al., 2016; R Core Team, 2021). Even if the differentiation effort can be partly mitigated, the vanilla versions of asymptotic bias correction in Efron (1975) and the bias-reducing adjusted score functions in Firth (1993) require the computation of expectations of products of log-likelihood derivatives under the model. Those expectations are intractable or expensive to compute for models with intractable or cumbersome likelihoods and can be hard to derive even for relatively simple models; see, for example, Grün et al. (2012, Section 2.3) for how involved those expectations are for Beta regression models.

Finally, except for the adjusted scores approach in Firth (1993), all the bias-reduction methods reviewed in Table 1 require the original estimator $\hat{\boldsymbol{\theta}}$ and they cannot operate without it. For this reason, they directly inherit any of the instabilities that $\hat{\boldsymbol{\theta}}$ may have. For example, in multinomial logistic regression, there is always a positive probability of data separation (Albert and Anderson, 1984) that results in infinite maximum likelihood estimates. Then, asymptotic bias correction, bootstrap, iterated bootstrap, and jackknife cannot be applied. The direct dependence on $\hat{\boldsymbol{\theta}}$ may be more consequential for naive implementations of the latter three methods because they are simulation-based; even if data separation did not occur for the original sample, there is always a positive probability that it occurs for at least one of the subsamples or simulated samples. There is no easy way of knowing this before carrying out the simulation, and boundary estimates, if they can be identified reliably, can only be handled in an ad-hoc way.

1.2 Reduced-bias *M*-estimation

The current work develops a novel approach to the reduction of the asymptotic bias of *M*-estimators from approximately unbiased estimating functions. We call the new estimators reduced-bias *M*-estimators, or RB*M*-estimators in short. As noted in the last two rows of Table 1, RB*M*-estimation applies to models that are at least partially specified and does not require the computation of any expectations. The analytical approximation to the bias function required by RB*M*-estimation relies only on derivatives of the contributions to the estimating functions. RB*M*-estimators with bias of lower asymptotic order than the original *M*-estimators result either in an explicit or implicit manner. Explicit RB*M*-estimation proceeds by directly subtracting the value of the analytical approximation to the bias function at the *M*-estimates from the *M*-estimates. Implicit RB*M*-estimation, on the other hand, does not require the original *M*-estimator, and instead relies on an additive, bounded-in-probability adjustment to the estimating functions. The key difference to the established implicit (Firth, 1993; Gourieroux et al., 1993; Kuk, 1995) and explicit bias-reduction methods (bootstrap, jackknife, asymptotic bias correction) is that the empirical adjustments introduced here depend only on the first two derivatives of the contributions to the estimating functions. Hence, they require neither the computation of cumbersome expectations nor the, potentially computation-intensive and error-prone calculation of *M*-estimates from simulated samples.

The RB*M*-estimators have the same asymptotic distribution, and hence the same asymptotic efficiency properties, as the original *M*-estimators. The definition of asymptotically valid inference and model selection procedures follows directly from the ones developed for the original *M*-estimators (e.g. generalised Wald tests and information criteria).

Implicit and explicit RB*M*-estimation, apart from being more general than the methods listed in Table 1, are also easier to implement for arbitrary models through automatic differentiation. The only key required input for a computer implementation is an implementation of the contributions to either the estimating or objective function. The `MEstimation` Julia package (Kosmidis and Lunardon, 2022) is a proof-of-concept of such a general-purpose implementation.

If the estimating functions are the components of the gradient of an objective function, as is the case in estimation based on likelihoods or composite likelihoods, then, we show that implicit RB*M*-estimation can always be achieved by the maximisation of an appropriately penalised version of the objective. This is in contrast to the method in Firth (1993) which does not always have a penalised likelihood interpretation; see, for example, Kosmidis and Firth (2009, Theorem 1). Moreover, it is shown that the bias-reducing penalised objective closely relates to information criteria for model selection based on the Kullback–Leibler divergence. The penalties that are used for bias reduction and model selection differ only by a known scalar constant. These observations, establish, for the first time, a strong link between model selection and reduction of bias in estimation. We also show how plug-in penalties to the penalised objectives can be used to enrich RB*M*-estimation with extra, desirable properties, such as estimates that are always finite in categorical response models, without sacrificing reduction of bias.

Section 2 introduces notation, the general modelling setting we consider, and the assumptions underpinning the theoretical developments. In Section 3, we derive the leading term of the bias of the *M*-estimator, give the bias-reducing adjustments to estimating functions and their empirical versions, and describe the explicit and implicit versions of RB*M*-estimation. Section 4 shows that the asymptotic distribution of the RB*M*-estimator is the same as that of the original *M*-estimator and introduces Wald-type and generalised score approximate pivots that can be used for inference. Section 5 shows that implicit RB*M*-estimation can always be achieved using empirical bias-reducing penalties to objective functions, draws links with model selection, and discusses the enrichment of the penalised objectives with plug-in penalties that endow RB*M*-estimators with extra properties. Section 6 explores the effectiveness of RB*M*-estimation in partially specified, high-dimensional regression settings, and Section 7 discusses RB*M*-estimation in models with independent random variables to models for temporally or spatially correlated observations. We demonstrate and assess the breath and properties of the RB*M*-estimation framework with examples from well-used, important modelling settings of increasing complexity, including: estimation the ratio of two means with minimal distributional assumptions (Examples 3.1 and 4.1), estimation and model selection in generalised linear models (GLMs; Examples 5.1, 5.3, and 5.5), composite likelihood methods for the estimation of Gaussian max-stable processes (Example 5.2), estimation of autoregressive models of order one (Example 7.1), and estimation and uncertainty quantification for autologistic regression models for the analysis of spatially clustered data (Examples 5.4 and 6.1). See, also online supplementary material, Section S8 for an empirical assessment of RB*M*-estimators in negative binomial regression with many covariates. Section 8 concludes with discussion on the developments and possible extensions of this work.

2 Modelling setting and assumptions

2.1 Estimating functions

Suppose we observe realisations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ of a sequence of random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$ with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ic_i})^\top \in \mathcal{Y} \subset \Re^{c_i}$, possibly with a sequence of covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq_i})^\top \in \mathcal{X} \subset \Re^{q_i}$, and $c_i \geq 1$ and $q_i \geq 1$. We assume that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$ are independent conditionally on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_k^\top)^\top$, and X be the set of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, and denote by $G \equiv G(\boldsymbol{Y} \mid X)$ the typically unknown, underlying joint distribution function.

One of the typical aims in statistical modelling is to estimate at least a sub-vector of an unknown p-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \Re^p$ using data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. This is most commonly achieved through an *M*-estimator $\hat{\boldsymbol{\theta}}$ (van der Vaart, 1998, Chapter 5) that results from the solution of a system of p estimating equations

$$\sum_{i=1}^{k} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = 0_p, \tag{2}$$

with respect to $\boldsymbol{\theta}$. In equation (2), 0_p is a p-vector of zeros, $\boldsymbol{\psi}^i(\boldsymbol{\theta}) = (\psi_1^i(\boldsymbol{\theta}), \ldots, \psi_p^i(\boldsymbol{\theta}))^\top$, $\boldsymbol{\psi}^i(\boldsymbol{\theta}) = \boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{Y}_i, \boldsymbol{x}_i)$, and $\psi_r^i(\boldsymbol{\theta}) = \psi_r(\boldsymbol{\theta}, \boldsymbol{Y}_i, \boldsymbol{x}_i)$, $r \in \{1, \ldots, p\}$. Examples of estimation methods that fall within the above framework are the estimation via quasi-likelihood methods (Wedderburn, 1974) and generalised estimating equations (Liang and Zeger, 1986). Stefanski

and Boos (2002) provide an accessible overview of estimating functions and demonstrate their generality and the handling of challenging estimation problems with less assumptions than likelihood-based estimation.

One way to derive estimating functions is through a, typically stronger, modelling assumption that $\boldsymbol{Y}_i$ has a distribution function with joint mixed density $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ or that the joint mixed densities of sub-vectors of $\boldsymbol{Y}_i$ can be composed into a composite function $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$. The estimator $\hat{\boldsymbol{\theta}}$ can then be taken to be the maximiser of the objective function

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} \log f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}). \tag{3}$$

The word 'mixed' is used here to allow some of the components of $\boldsymbol{y}_i$ to be continuous and some discrete. If the objective function (3) is used, then the estimating functions in (2) have $\boldsymbol{\psi}^i(\boldsymbol{\theta}) = \nabla \log f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$, assuming that the gradient exists in Θ. Prominent examples of estimation methods that involve an objective function of the form (3) are maximum likelihood and maximum composite likelihood (see, for example, Lindsay, 1988; Varin et al., 2011).

2.2 Assumptions

The sufficient conditions we employ for the theoretical developments in this work are standard and cover a wealth of M-estimation problems. The assumptions are listed in detail in online supplementary material, Section S2 and include: the unbiasedness of the estimating function and convergence of the M-estimator to the root of the expectation of the estimating function with respect to G (see online supplementary material, Assumption A1), local smoothness of the estimating functions and existence of a sufficient number of derivatives of those around that root (see online supplementary material, Assumption A2), existence of moments of products of derivatives of estimating functions under the unknown data generating process along with conditions on their growth rate in terms of the amount of information about the parameter (see online supplementary material, Assumptions A3–A4). Many of those conditions can be derived from others for particular models and estimation methods. An annotated analysis of the links among them, and their wide scope and applicability are provided in online supplementary material, Section S2. Unless otherwise stated, in what follows, we assume that these assumptions hold.

3 Adjusted estimating equations for bias reduction

3.1 Asymptotic bias

It can be shown that the bias of the M-estimator $\hat{\boldsymbol{\theta}}$ is $\mathrm{E}_G(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = O(n^{-1})$, where $\bar{\boldsymbol{\theta}}$ is such that $\mathrm{E}_G(\boldsymbol{\psi}^i(\bar{\boldsymbol{\theta}})) = 0_p$ for all $i \in \{1, \ldots, k\}$. This provides some reassurance that, as the information about the parameter $\boldsymbol{\theta}$ grows, estimation bias vanishes. Nevertheless, the finite sample bias of $\hat{\boldsymbol{\theta}}$ is typically not zero. It is also possible to write down the bias in the more specific form $\mathrm{E}_G(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = \boldsymbol{b}(\bar{\boldsymbol{\theta}}) + O(n^{-3/2})$, where $\boldsymbol{b}(\bar{\boldsymbol{\theta}})$ depends on joint moments, with respect to G, of estimating functions and derivatives of those. Then, because $\boldsymbol{b}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \boldsymbol{b}(\bar{\boldsymbol{\theta}})$ when $\hat{\boldsymbol{\theta}} \xrightarrow{p} \bar{\boldsymbol{\theta}}$, we can define a reduced-bias estimator $\hat{\boldsymbol{\theta}} - \boldsymbol{b}(\hat{\boldsymbol{\theta}})$.

The estimator $\hat{\boldsymbol{\theta}} - \boldsymbol{b}(\hat{\boldsymbol{\theta}})$ has been shown to, indeed, have better bias properties when estimation is by maximum likelihood (ML) and the model is correctly specified (see Efron, 1975, Section 10). When the model is correctly specified, the unknown joint distribution function $G(\boldsymbol{Y} \mid X)$ is assumed to be a particular member of the family of distributions specified fully by $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ when forming (3). It is, then, possible to evaluate $\boldsymbol{b}(\boldsymbol{\theta})$ and, hence, compute $\hat{\boldsymbol{\theta}} - \boldsymbol{b}(\hat{\boldsymbol{\theta}})$ in light of data, because the expectations involved in the joint null moments are with respect to the modelling assumption. This is the basis of more refined bias reduction methods, like the adjusted score function approach that has been derived in Firth (1993) and explored further in Kosmidis and Firth (2009).

In the more general setting of Section 2, where the model can be only partially specified, naive evaluation of $\boldsymbol{b}(\boldsymbol{\theta})$ using $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ not only does not lead to the reduction of bias, in general, but it can also inflate the bias; see, for example, Lunardon and Scharfstein (2017, Section 2.1), and Example 7.1 for the estimation of autoregressive processes. Even if the researcher is comfortable

to assume that $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ is correctly specified, the applicability of standard bias-reduction methods is hampered whenever $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ is impossible or impractical to compute in closed form. Estimation and inference, in those cases, is typically based on pseudolikelihood or estimating functions. An example of this kind is within the framework of max-stable processes for which the evaluation of the joint density becomes quickly infeasible when the number of site locations increases (Davison and Gholamrezaee, 2012); see, also, Example 5.2.

3.2 Family of bias-reducing adjustments to estimating functions

Consider the estimator $\tilde{\boldsymbol{\theta}}$ that results from the solution of the adjusted estimating equations

$$\sum_{i=1}^{k} \boldsymbol{\psi}^i(\boldsymbol{\theta}) + \boldsymbol{A}(\boldsymbol{\theta}) = 0_p, \tag{4}$$

where both $\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta}, \boldsymbol{Y}, X)$ and its derivatives with respect to $\boldsymbol{\theta}$ are $O_p(1)$ as n grows. A calculation similar to that in McCullagh (2018, Section 7.3) leads to a stochastic Taylor expansion for $\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$ (see online supplementary material, expression (S1) in Section S3), whose expectation with respect to G gives that the bias of $\tilde{\boldsymbol{\theta}}$ is

$$\mathrm{E}_G(\tilde{\theta}^r - \bar{\theta}^r) = -\mu^{ra}\mathrm{E}_G(A_a) + \frac{1}{2}\mu^{ra}\mu^{bc}\left(2v_{ab,c} - \mu^{de}v_{c,e}\mu_{abd}\right) + O(n^{-3/2}). \tag{5}$$

In the above expression, we employ index notation. The quantities $\mu_{R_a}(\boldsymbol{\theta}) = \mathrm{E}_G(l_{R_a}(\boldsymbol{\theta}))$ with $l_{R_a}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \partial^{a-1}\psi^i_{r_1}(\boldsymbol{\theta})/\partial\theta^{r_2}\cdots\partial\theta^{r_a}$ are expectations of derivatives of the estimating function, $(R_a = \{r_1, \ldots, r_a\}; r_j \in \{1, \ldots, p\})$, $\mu^{ra}(\boldsymbol{\theta})$ is the inverse of $\mu_{ar}(\boldsymbol{\theta})$, and $v_{c,e}(\boldsymbol{\theta}) = \mathrm{E}_G(l_c(\boldsymbol{\theta})l_e(\boldsymbol{\theta}))$ and $v_{ab,c}(\boldsymbol{\theta}) = \mathrm{E}_G(l_{ab}(\boldsymbol{\theta})l_c(\boldsymbol{\theta}))$. Unless otherwise stated, the dependence on the parameter is omitted whenever quantities are evaluated at $\bar{\boldsymbol{\theta}}$, as is the case in the right-hand side of (5). For example, $\mu^{ra} = \mu^{ra}(\bar{\boldsymbol{\theta}})$. Expansion (5) implies that an estimator $\tilde{\boldsymbol{\theta}}$ with bias $\mathrm{E}_G(\tilde{\theta}^r - \bar{\theta}^r) = O(n^{-3/2})$ is obtained by the solution of (4) with any adjustment $\boldsymbol{A}(\theta)$ satisfying

$$\mathrm{E}_G(A_r(\theta)) = \frac{1}{2}\mu^{ab}(\theta)\left\{2v_{ra,b}(\theta) - \mu^{cd}(\theta)v_{b,d}(\theta)\mu_{rac}(\theta)\right\} + O(n^{-1/2}). \tag{6}$$

Then $\tilde{\boldsymbol{\theta}}$ has, asymptotically, smaller bias than $\hat{\boldsymbol{\theta}}$. In other words, expression (6) defines a family of bias-reducing adjustments to estimating functions. Expansion (5), and, hence, all results below remain valid if the estimating functions are asymptotically unbiased with $O(n^{-1/2})$ expectation.

An obvious candidate for $\boldsymbol{A}(\boldsymbol{\theta})$ has rth component in the first term on the right-hand side of (6). If $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ fully specifies G and the estimation method is ML, then the Bartlett relation $\mu_{cd}(\boldsymbol{\theta}) + v_{c,d}(\boldsymbol{\theta}) = 0$ holds (see, for example, Pace and Salvan, 1997, Section 9.2), and the adjustment satisfying (6) becomes $\mu^{ab}(\theta)\{2v_{ra,b}(\theta) + \mu_{rab}(\theta)\}/2$. The latter expression is the bias-reducing adjustment for ML derived in Firth (1993) for fully specified models.

For the more general setting of Section 2, however, the underlying distribution G is typically only partially specified through $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ or relations of moments of the random component. Hence, the expectations involved in the right-hand side of (6) cannot be computed.

3.3 Empirical bias-reducing adjustments to estimating functions

The weak law of large numbers and a straightforward calculation can be used to produce an easy-to-implement family of empirical adjustments that delivers bias reduction of M-estimators in the general framework of Section 2. In particular, the empirical adjustment of the form

$$A_r(\boldsymbol{\theta}) = \frac{1}{2}l^{ab}(\boldsymbol{\theta})\left\{2l_{ra|b}(\boldsymbol{\theta}) - l^{cd}(\boldsymbol{\theta})l_{b|d}(\boldsymbol{\theta})l_{rac}(\boldsymbol{\theta})\right\}, \tag{7}$$

satisfies equation (6). In the above expression, $l_{r|s}(\theta) = \sum_{i=1}^{k} \psi_r^i(\theta)\psi_s^i(\theta)$ and $l_{rs|t}(\theta) = \sum_{i=1}^{k} \{\partial\psi_r^i(\boldsymbol{\theta})/\partial\theta^s\}\psi_t^i(\boldsymbol{\theta})$ are estimators of $v_{r,s}(\theta) = \mathrm{E}_G(\sum_{i=1}^{k} \psi_r^i(\theta) \sum_{j=1}^{k} \psi_s^j(\theta))$ and $v_{rs,t}(\theta) = \mathrm{E}_G(\sum_{i=1}^{k} (\partial\psi_r^i(\theta)/\partial\theta^s) \sum_{j=1}^{k} \psi_t^j(\theta))$, respectively, $l^{st}(\boldsymbol{\theta})$ is the matrix inverse of $l_{ts}(\boldsymbol{\theta})$ and an estimator of $\mu^{rs}(\boldsymbol{\theta})$, and $l_{stu}(\boldsymbol{\theta})$ is an estimator of $\mu_{stu}(\boldsymbol{\theta})$. The matrix form of expression (7) sets the rth element of the vector of empirical bias-reducing adjustments to

$$A_r(\boldsymbol{\theta}) = -\mathrm{trace}\left\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{d}_r(\boldsymbol{\theta})\right\} - \frac{1}{2}\mathrm{trace}\left[\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\left\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\right\}^{\top}\boldsymbol{u}_r(\boldsymbol{\theta})\right], \tag{8}$$

where $\boldsymbol{u}_r(\boldsymbol{\theta}) = \sum_{i=1}^{k} \nabla\nabla^{\top}\psi_r^i(\boldsymbol{\theta})$, and $\boldsymbol{j}(\boldsymbol{\theta})$ is the matrix with sth row $-\sum_{i=1}^{k} \nabla\psi_s^i(\boldsymbol{\theta})$. The matrix $\boldsymbol{j}(\boldsymbol{\theta})$, assumed to be invertible but is not necessarily symmetric, coincides with the negative of the hessian matrix $\nabla\nabla^{\top}l(\boldsymbol{\theta})$ when estimation is through the maximisation of an objective function like (3) and is the observed information in maximum likelihood estimation. The matrices $\boldsymbol{e}(\boldsymbol{\theta})$ and $\boldsymbol{d}_r(\boldsymbol{\theta})$ correspond to the quantities $l_{r|s}(\boldsymbol{\theta})$ and $l_{rs|t}(\boldsymbol{\theta})$ in (7), respectively.

3.4 Explicit and implicit reduced-bias *M*-estimation

According to the classification of bias-reduction methods in Kosmidis (2014), the solution of the adjusted estimating equations $\sum_{i=1}^{k} \boldsymbol{\psi}^i(\boldsymbol{\theta}) + \boldsymbol{A}(\boldsymbol{\theta}) = 0_p$, with $\boldsymbol{A}(\boldsymbol{\theta})$ as in (8), results in an implicit bias-reduction method. Hence, the resulting estimator $\tilde{\boldsymbol{\theta}}$ is called the implicit RB*M*-estimator. From expression (5), another estimator with $o(n^{-1})$ bias is

$$\boldsymbol{\theta}^{\dagger} = \hat{\boldsymbol{\theta}} + \boldsymbol{j}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{A}(\hat{\boldsymbol{\theta}}). \tag{9}$$

The estimator (9) defines an explicit bias-reduction method for partially specified models and is called the explicit RB*M*-estimator. Both explicit and implicit RB*M*-estimation are attractive compared to other bias-reduction methods whose applicability is limited to either cases where $\sum_{i=1}^{k} \boldsymbol{\psi}^i(\boldsymbol{\theta})$ is the gradient of the log-likelihood function of a correctly specified model (see, e.g. Cordeiro and McCullagh, 1991; Firth, 1993), or samples from a correctly specified model can be simulated (see, e.g. Gourieroux et al., 1993; Guerrier et al., 2019; Kuk, 1995). Also, this generalisation comes with less implementation requirements because both explicit and implicit RB*M*-estimation require only the estimating functions and the first two derivatives of those. The RB*M*-estimators can directly be computed and used in settings that involve realisations of k independent random vectors with dependent components, like the generalised estimating equations in Liang and Zeger (1986) and composite likelihood methods (Varin et al., 2011).

The choice between using $\boldsymbol{\theta}^{\dagger}$ and $\tilde{\boldsymbol{\theta}}$ is application-dependent. For example, $\tilde{\boldsymbol{\theta}}$ in Example 3.1 below has certain robustness properties that $\boldsymbol{\theta}^{\dagger}$ lacks. Furthermore, as is shown in Section 5, in *M*-estimation problems that involve an objective function $\tilde{\boldsymbol{\theta}}$ can be viewed as the maximiser of a bias-reducing penalised objective that has strong connections to model selection procedures. On the other hand, for general models for which $\boldsymbol{\theta}^{\dagger}$ and $\tilde{\boldsymbol{\theta}}$ are not available in closed form, $\boldsymbol{\theta}^{\dagger}$ is typically faster to compute because it results from a single step of a quasi-Newton–Raphson iteration with stationary point $\tilde{\boldsymbol{\theta}}$ (see online supplementary material, Section S4). The `MEstimation` Julia package implements explicit and implicit RB*M*-estimation for arbitrary models using automatic differentiation.

Example 3.1 (Ratio of two means). Consider a setting where independent random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed, and suppose that interest is in the ratio of the mean of Y_i to the mean of X_i, that is $\theta = \mu_Y/\mu_X$, with $\mu_Y = \mathrm{E}_G(Y_i)$ and $\mu_X = \mathrm{E}_G(X_i) \neq 0$ $(i = 1, \ldots, n)$.

Assuming that sampling is from an infinite population, one way of estimating θ without any further assumptions about the joint distribution of

(X_i, Y_i) is to set up an unbiased estimating equation of the form (2), with $\psi^i(\theta) = Y_i - \theta X_i$. Then, the M-estimator is

$$\hat{\theta} = \arg\operatorname*{solve}_{\theta \in \Re} \left\{ \sum_{i=1}^{n} \psi^i(\theta) = 0 \right\} = \frac{s_Y}{s_X}, \tag{10}$$

where $s_X = \sum_{i=1}^n X_i$ and $s_Y = \sum_{i=1}^n Y_i$. The estimator $\hat{\theta}$ is generally biased and efforts have been made in reducing its bias; see, for example, Durbin (1959), for an early work in that direction using the jackknife.

For the estimating function in (10), $j(\theta) = s_X$, $u(\theta) = 0$, $e(\theta) = s_{YY} + \theta^2 s_{XX} - 2\theta s_{XY}$, and $d(\theta) = -s_{XY} + \theta s_{XX}$, where $s_{XX} = \sum_{i=1}^n X_i^2$, $s_{YY} = \sum_{i=1}^n Y_i^2$ and $s_{XY} = \sum_{i=1}^n X_i Y_i$. So, the empirical bias-reducing adjustment in (7) is $s_{XY}/s_X - \theta s_{XX}/s_X$. The solution of the adjusted estimating equation $\sum_{i=1}^n \psi^i(\theta) + A(\theta) = 0$ results in the implicit RBM-estimator $\tilde{\theta} = (s_Y + s_{XY}/s_X)/(s_X + s_{XX}/s_X)$. A direct substitution of $j(\theta)$ and $A(\theta)$ in (9) gives that the explicit RBM-estimator has the form $\theta^\dagger = \hat{\theta}(1 - s_{XX}/s_X^2) + s_{XY}/s_X^2$.

Implicit RBM-estimation has the side-effect of producing an estimator that is more robust to small values of s_X than the standard M-estimator and the explicit RBM-estimator are. In particular, as s_X becomes smaller in absolute value, $\hat{\theta}$ and $\theta^\dagger$ diverge, while $\tilde{\theta}$ converges to s_{XY}/s_{XX}, which is the slope of the regression line through the origin of y on x. As a result, when μ_X is small in absolute value, $\tilde{\theta}$ has not only smaller bias, as granted by the developments in the current paper, but also smaller variance than $\hat{\theta}$, and, hence, smaller mean squared error.

To illustrate the performance of the implicit and explicit RBM-estimators of the ratio θ we assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent random vectors from a bivariate distribution constructed through a Gaussian copula with correlation 0.5, to have an exponential marginal with rate 1/2 for X_i and a normal marginal for Y_i with mean 10 and variance 1. Hence, $\bar{\theta} = 5$. For each $n \in \{10, 20, 40, 80, 160, 320\}$, we simulate $N_n = 250 \times 2^{16}/n$ samples. Calibrating the simulation size in this way guarantees a fixed simulation error for the simulation-based estimate of the bias of the M-estimator. For each sample, we estimate θ using the M-estimator $\hat{\theta}$, the implicit and explicit RBM-estimators $\tilde{\theta}$ and $\theta^\dagger$, respectively, the jackknife estimator $\theta^{(J)} = n\hat{\theta} - (n-1)\sum_{i=1}^n \hat{\theta}_{(-i)}/n$, where $\hat{\theta}_{(-i)} = \sum_{j \neq i} Y_i / \sum_{j \neq i} X_i$, and the 'better' nonparametric bootstrap ratio estimator $\theta^{(B)}$ in Efron and Tibshirani (1993, Section 10.4) using 500 resamples. Note here that both vanilla asymptotic bias correction (Efron, 1975) and the bias-reducing adjusted score function in Firth (1993) require expectation of products of log-likelihood derivatives, which require specifying the bivariate distribution for (X_i, Y_i). Despite the simplicity of the current setting, even if one could confidently specify that distribution, the required expectations involve non-trivial analytic calculations and may not even be available in closed form.

Table 2 shows the simulation-based estimates of the bias $\mathrm{E}_G(\theta_n - \bar{\theta})$, mean squared error $\mathrm{E}_G\{(\theta_n - \bar{\theta})^2\}$, mean absolute deviation $\mathrm{E}_G(|\theta_n - \bar{\theta}|)$, and probability of underestimation $P_G(\theta_n < \bar{\theta})$ for θ_n being $\hat{\theta}$, $\theta^{(J)}$, $\theta^{(B)}$, $\theta^\dagger$ and $\tilde{\theta}$. As with the jackknife and bootstrap, explicit and implicit RBM-estimation result in a marked reduction of the bias. The estimated slopes of the regression lines of the logarithm of the absolute value of the estimated biases for $\hat{\theta}$, $\theta^{(J)}$, $\theta^{(B)}$, $\theta^\dagger$, and $\tilde{\theta}$ on the values of $\log n$ are −1.039, −1.696, −1.413, −1.856, and −2.118, respectively, which are in agreement to the theoretical slopes of −1, −3/2, −3/2, −3/2, and −3/2.

Reduction of the bias in this setting leads also in marked reduction in mean squared error and mean absolute deviation, with $\tilde{\theta}$, $\theta^\dagger$, $\theta^{(J)}$ and $\theta^{(B)}$

Table 2. Simulation-based estimates of the bias (B), mean squared error (MSE), mean absolute error (MAE), and probability of underestimation (PU) for the estimation of a ratio of means in the setting of Example 3.1

	θ_n	n						
		5	10	20	40	80	160	320
B	$\hat{\theta}$	1.19	0.53	0.25	0.12	0.06	0.03	0.01
	$\theta^{(J)}$	−0.38	−0.07	−0.01	0.00	0.00	0.00	0.00
	$\theta^{(B)}$	−0.19	0.01	0.01	0.00	0.00	0.00	0.00
	$\theta^{\dagger}$	0.62	0.17	0.05	0.01	0.00	0.00	0.00
	$\tilde{\theta}$	0.40	0.10	0.03	0.01	0.00	0.00	0.00
MSE	$\hat{\theta}$	12.99	3.76	1.47	0.65	0.30	0.15	0.07
	$\theta^{(J)}$	12.69	3.04	1.29	0.61	0.29	0.15	0.07
	$\theta^{(B)}$	59.22	3.00	1.29	0.61	0.29	0.15	0.07
	$\theta^{\dagger}$	10.09	3.11	1.31	0.61	0.29	0.15	0.07
	$\tilde{\theta}$	9.37	3.04	1.30	0.61	0.29	0.15	0.07
MAE	$\hat{\theta}$	2.23	1.37	0.91	0.62	0.43	0.30	0.22
	$\theta^{(J)}$	2.29	1.31	0.88	0.61	0.43	0.30	0.21
	$\theta^{(B)}$	2.20	1.29	0.88	0.61	0.43	0.30	0.21
	$\theta^{\dagger}$	2.03	1.29	0.88	0.61	0.43	0.30	0.21
	$\tilde{\theta}$	2.00	1.29	0.88	0.61	0.43	0.30	0.21
PU	$\hat{\theta}$	0.44	0.46	0.47	0.48	0.48	0.49	0.49
	$\theta^{(J)}$	0.64	0.60	0.56	0.54	0.53	0.52	0.51
	$\theta^{(B)}$	0.62	0.58	0.56	0.54	0.53	0.52	0.51
	$\theta^{\dagger}$	0.53	0.54	0.54	0.54	0.53	0.52	0.51
	$\tilde{\theta}$	0.56	0.56	0.55	0.54	0.53	0.52	0.51

Note. The estimator θ_n is either the M-estimator $\hat{\theta}$, the Jackknife estimator $\theta^{(J)}$, the 'better' nonparametric bootstrap estimators $\theta^{(B)}$ in Efron and Tibshirani (1993, Section 10.4) using 500 resamples, the explicit RBM-estimator $\theta^{\dagger}$, or the implicit RBM-estimator $\tilde{\theta}$. Figures are reported in 2 decimal places, and a figure of 0.00 indicates an estimated bias between −0.005 and 0.005. The simulation error for the bias estimates is in (0.001, 0.002).

performing similarly for $n \geq 10$, and markedly better than $\hat{\theta}$. The large mean squared errors of $\theta^{(J)}$ and $\theta^{(B)}$ for $n = 5$ are due to extreme ratio estimates appearing with positive probability under the resampling distribution. The RBM-estimators also have a probability of underestimation closer to 0.5, hence they are closer to being median unbiased than $\hat{\theta}$, $\theta^{(J)}$ and $\theta^{(B)}$ are.

4 Asymptotic distribution of reduced-bias *M*-estimators

The stochastic Taylor expansion (S1) in online supplementary material implies that $\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$ and the RBM-estimation counterparts $\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^{\dagger} - \bar{\boldsymbol{\theta}}$ have exactly the same $O_p(n^{-1/2})$ term in their expansions because $\boldsymbol{A}(\boldsymbol{\theta}) = O_p(1)$. Hence, the argument in Stefanski and Boos (2002, Section 2) applied to that first term gives that the implicit RBM-estimator $\tilde{\boldsymbol{\theta}}$ is such that

$$\boldsymbol{Q}(\bar{\boldsymbol{\theta}})^{1/2}(\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \xrightarrow{d} \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{I}_p), \tag{11}$$

and the same holds for the M-estimator and the explicit RBM-estimator $\theta^{\dagger}$. In (11), $\boldsymbol{I}_p$ is the $p \times p$ identity matrix, $\boldsymbol{Q}(\boldsymbol{\theta}) = \boldsymbol{V}(\boldsymbol{\theta})^{-1}$, where $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta})^{-1}\boldsymbol{M}(\boldsymbol{\theta})\{\boldsymbol{B}(\boldsymbol{\theta})^{-1}\}^{\top}$, with $\boldsymbol{M}(\boldsymbol{\theta}) = \mathrm{E}_G[\sum_{i=1}^{k}\sum_{j=1}^{k}\psi^i(\boldsymbol{\theta})\{\psi^j(\boldsymbol{\theta})\}^{\top}]$, and $\boldsymbol{B}(\boldsymbol{\theta})$ is a $p \times p$ matrix with rth row $-\sum_{i=1}^{k}\mathrm{E}_G\{\nabla\psi_r^i(\boldsymbol{\theta})\}$.

An implication of (11) is that $\hat{\boldsymbol{V}}(\boldsymbol{\theta}) = \boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\}^{\top}$ evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ (or $\boldsymbol{\theta} = \boldsymbol{\theta}^{\dagger}$) is a consistent estimator of the variance-covariance matrix of $\tilde{\boldsymbol{\theta}}$ (or $\theta^{\dagger}$). This is exactly as in the case where

Table 3. Simulation-based estimates of the variances of the *M*-estimator $\hat{\theta}$ and the RB*M*-estimator $\tilde{\theta}$ ($\text{var}_G(\hat{\theta})$ and $\text{var}_G(\tilde{\theta})$, respectively) and of the mean of $\hat{V}(\hat{\theta})$ and $\hat{V}(\tilde{\theta})$ ($\text{E}_G\{\hat{V}(\hat{\theta})\}$ and $\text{E}_G\{\hat{V}(\tilde{\theta})\}$, respectively), from the simulation study of Example 3.1

	n					
	10	20	40	80	160	320
$\text{var}_G(\hat{\theta})$	3.49	1.41	0.64	0.30	0.15	0.07
$\text{E}_G\{\hat{V}(\hat{\theta})\}$	2.55	1.20	0.59	0.29	0.14	0.07
$\text{var}_G(\tilde{\theta})$	3.09	1.31	0.61	0.30	0.15	0.07
$\text{E}_G\{\hat{V}(\tilde{\theta})\}$	2.16	1.10	0.56	0.28	0.14	0.07

Note. See, also Table 2.

$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})$ is used as an estimator of the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ in the framework of *M*-estimation (see, for example, Stefanski and Boos, 2002, Section 2). The expression for $\hat{\boldsymbol{V}}(\boldsymbol{\theta})$ appears unaltered in the second term of the right-hand-side of expression (8) for the empirical bias-reducing adjustment. As a result, the value of $\hat{\boldsymbol{V}}(\tilde{\boldsymbol{\theta}})$ or $\hat{\boldsymbol{V}}(\theta^{\dagger})$ and that of estimated standard errors for the estimators are available at the final quasi-Newton–Raphson update for computing the RB*M*-estimates, as detailed in online supplementary material, Section S4.

In addition, if the model is correctly specified, and $l(\boldsymbol{\theta})$ in (3) is the log-likelihood, then the second Bartlett identity gives $\boldsymbol{M}(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta})$, which implies that $\boldsymbol{Q}(\boldsymbol{\theta})$ is the expected information matrix. As a result, the RB*M*-estimator is asymptotically efficient, exactly as the ML estimator and the reduced-bias estimator in Firth (1993) are.

Example 4.1 (Ratio of two means (continued)). Table 3 shows the estimates of the actual variances of $\hat{\theta}$ and $\tilde{\theta}$ for $n \in \{10, 20, 40, 80, 160, 320\}$ and the estimate of the mean of $\hat{V}(\hat{\theta})$ and $\hat{V}(\tilde{\theta})$, respectively. As expected by the arguments above, the large sample approximations to the variance of the estimator converge to the actual variances as the sample size increases.

Another implication of (11) is that asymptotically valid inferential procedures, like hypothesis tests and confidence regions for the model parameters, can be constructed based on Wald-type and generalised score approximate pivots of the form

$$\begin{aligned} W_{(e)}(\boldsymbol{\theta}) &= (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top}\{\hat{\boldsymbol{V}}(\tilde{\boldsymbol{\theta}})\}^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}), \\ W_{(s)}(\boldsymbol{\theta}) &= \left\{\sum_{i=1}^{k} \boldsymbol{\psi}^{i}(\boldsymbol{\theta}) + \boldsymbol{A}(\boldsymbol{\theta})\right\}^{\top} \{\boldsymbol{e}(\tilde{\boldsymbol{\theta}})\}^{-1} \left\{\sum_{i=1}^{k} \boldsymbol{\psi}^{i}(\boldsymbol{\theta}) + \boldsymbol{A}(\boldsymbol{\theta})\right\}, \end{aligned} \tag{12}$$

respectively, which, asymptotically, have a χ^2_p distribution. The same holds when $\boldsymbol{\theta}^{\dagger}$ is used in place of $\tilde{\boldsymbol{\theta}}$. These pivots are direct extensions of the Wald-type and generalised score pivots, respectively, that are typically used in *M*-estimation; see Boos (1992) for a discussion about generalised score tests in *M*-estimation.

5 Bias-reducing penalties to objective functions

5.1 Penalised objectives

When *M*-estimation is through the maximisation of an objective function of the form (3), implicit RB*M*-estimation is equivalent to the maximisation of the penalised objective function

$$l(\boldsymbol{\theta}) - \frac{1}{2}\text{trace}\left\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\right\}, \tag{13}$$

assuming that the maximum exists. This equivalence can be proved by noting that $\boldsymbol{j}(\boldsymbol{\theta})$ is a symmetric matrix, and differentiating the penalty in (13) to get adjustment (8).

Example 5.1 (Generalised linear models). Suppose that the distribution of $Y_i \mid \boldsymbol{x}_i$ is fully specified to be an exponential family with mean $\mu_i = h(\eta_i)$ for a known inverse link function $h(\cdot)$, with $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, and variance $\phi v(\mu_i)/m_i$. Then, the ith log-likelihood contribution is

$$\log f_i(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \frac{m_i}{\phi}\{y_i\theta_i - \kappa(\theta_i) - c_1(y_i)\} - \frac{1}{2}a\left(-\frac{m_i}{\phi}\right), \qquad (14)$$

for sufficiently smooth functions $\kappa(\cdot)$, $c_1(\cdot)$, and $a(\cdot)$, and known prior weights $m_1, \ldots, m_n$ with $m_i > 0$, where $\mu_i = d\kappa(\theta_i)/\mathrm{d}\theta_i$ and $v(\mu_i) = d^2\kappa(\theta_i)/d\theta_i^2$. Online supplementary material, Section S5 gives the form of $\boldsymbol{j}(\boldsymbol{\beta}, \phi)$ and $\boldsymbol{e}(\boldsymbol{\beta}, \phi)$ for GLMs with unknown dispersion parameter ϕ. If the dispersion parameter ϕ is known, as is, for example, for the binomial and Poisson distributions, then the bias-reducing penalty in (13) involves only the $(\boldsymbol{\beta}, \boldsymbol{\beta})$ blocks $\boldsymbol{j}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta})$ and $\boldsymbol{e}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta})$ of $\boldsymbol{j}(\boldsymbol{\beta}, \phi)$ and $\boldsymbol{e}(\boldsymbol{\beta}, \phi)$, respectively. Simple algebra shows that the RBM-estimator of $\boldsymbol{\beta}$ results as the maximiser of the bias-reducing penalised log-likelihood

$$\sum_{i=1}^{n} m_i\left\{y_i\theta_i - \kappa(\theta_i) - \frac{1}{2}s_i\frac{d_i}{v_i}(y_i - \mu_i)\right\}, \qquad (15)$$

where $d_i = \mathrm{d}h(\eta_i)/\mathrm{d}\eta_i$, and $v_i = v(\mu_i)$. The quantity s_i is the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{W}}$, where $\boldsymbol{X}$ is the $n \times p$ model matrix with rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and the diagonal matrices $\tilde{\boldsymbol{W}}$ and $\boldsymbol{Q}$ are as defined in online supplementary material, Section S5.

Expression (15) is in contrast to the method in Firth (1993), for which a penalised log-likelihood does not always exist; see, Kosmidis and Firth (2009, Theorem 1) for the necessary and sufficient condition for the existence of a bias-reducing penalised likelihood for GLMs.

5.2 Estimation

The implicit RBM-estimates can be computed using general numerical optimisation procedures for the maximisation of the penalised objective function (13) that operate by numerically approximating gradients, like those provided by the `optim` function in R or the `Optim` Julia package (Mogensen and Riseth, 2018). In such cases, RBM-estimation requires only routines for matrix multiplication and inversion, and the contributions to the objective function and their first two derivatives for the implementation of $\boldsymbol{j}(\boldsymbol{\theta})$ and $\boldsymbol{e}(\boldsymbol{\theta})$, which can be obtained using automatic differentiation. The `MEstimation` Julia package provides a proof-of-concept of an implementation relying on automatic differentiation.

The explicit RBM-estimator $\boldsymbol{\theta}^\dagger = \hat{\boldsymbol{\theta}} + \boldsymbol{j}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{A}(\hat{\boldsymbol{\theta}})$ in (9) can be obtained by computing $\boldsymbol{j}(\hat{\boldsymbol{\theta}})$, and numerically differentiating the penalty term $-\mathrm{trace}\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\}/2$ in (13) at the M-estimate $\hat{\boldsymbol{\theta}}$ to get the value of $\boldsymbol{A}(\hat{\boldsymbol{\theta}})$.

Example 5.2 (Gaussian max-stable processes). Vanilla likelihood and Bayesian approaches for spatial extreme processes face challenges because the direct generalisation of the classical multivariate extreme value distributions to the spatial case is a max-stable process for which the evaluation of the likelihood becomes increasingly more intractable as the number of site locations increases (Davison and Gholamrezaee, 2012). Several works have proposed the use of computationally appealing surrogates to the likelihood, like composite likelihoods, which are formed by specifying marginal or conditional densities for subsets of site locations (see Davison and Gholamrezaee, 2012; Genton et al., 2011; Huser and Davison, 2013; Padoan et al., 2010, among others). Nevertheless, standard bias-reduction

methods, like the one in Firth (1993) and Kuk (1995) in Table 1, are either infeasible or computationally expensive because the calculation of the bias function involves either integrals with respect to the true underlying joint density or requires repeated sampling and refitting.

An example of a surrogate to the likelihood is the pairwise likelihood introduced in Padoan et al. (2010) under a block maxima approach for modelling extremes. Suppose that $y_1(s), \ldots, y_k(s)$ with $s \in \{s_1, \ldots, s_L\}, s_j \in \Re^2$, are k independent observations at each of L site locations. The pairwise log-likelihood formed from the collection of $L(L-1)/2$ distinct pairs of locations is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} \sum_{l>m} \log f(y_i(s_l), y_i(s_m) \mid \boldsymbol{\theta}), \tag{16}$$

where $f(y_i(s_l), y_i(s_m) \mid \boldsymbol{\theta})$ is the joint density of $Y_i(s_l)$ and $Y_i(s_m)$ $(l, m = 1, \ldots, L; l \neq m)$, given in expression (S3) of online supplementary material. The spatial dependence between $Y_i(s_l)$ and $Y_i(s_m)$ is characterised by the 2×2 matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ with diagonal elements σ_1^2 and σ_2^2, and σ_{12}^2 in the off-diagonals. The maximiser of (16) with respect to $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \sigma_{12}^2)^\top$, is the maximum pairwise likelihood estimator $\hat{\boldsymbol{\theta}}$, and the RB*M*-estimator $\tilde{\boldsymbol{\theta}}$ maximises (13). Expressions for $\boldsymbol{j}(\boldsymbol{\theta})$ and $\boldsymbol{e}(\boldsymbol{\theta})$ are given in online supplementary material, Section S6.

Simulations are run by generating independent observations $y_1(s), \ldots, y_k(s)$ from a Gaussian max-stable process observed at $L = 50$ site locations. The locations are generated uniformly on a $[0, 40] \times [0, 40]$ region. We consider sample sizes $k \in \{10, 20, 40, 80, 160\}$ with corresponding number of simulations equal to 4000 k, and true parameter settings $\bar{\boldsymbol{\theta}} = (2000, 3000, 1500)^\top$ and $\bar{\boldsymbol{\theta}} = (20, 30, 15)^\top$, imposing strong and weak spatial dependence, respectively. These parameter values correspond to $\boldsymbol{\Sigma}_4$ and $\boldsymbol{\Sigma}_5$ in Table 1 of Padoan et al. (2010). In Figure 1, we show the simulation-based estimates of the logarithms of the absolute biases as functions of $\log n$, where $n = k$. These curves have roughly slopes -1, and between $-3/2$ and -2, respectively, as expected by the asymptotic theory in Section 3, demonstrating the reduction of the bias that $\tilde{\boldsymbol{\theta}}$ delivers. Furthermore, $\tilde{\boldsymbol{\theta}}$ appears to have smaller finite-sample mean squared error, and hence smaller variance than the maximum pairwise likelihood estimator $\hat{\boldsymbol{\theta}}$. The simulation results provide evidence for the superiority of the RB*M*-estimator.

We are not including simulation- or resampling-based methods for bias reduction (see Table 1) in the simulation experiments here because they require the repeated calculation of maximum pairwise likelihood estimates, and hence are expensive computationally. Furthermore, the bias reduction method in Firth (1993) does not apply here because its application requires expectations of products of derivatives of the full likelihood function for the Gaussian max-stable process, which poses more severe tractability issues than ML.

5.3 Plug-in penalties and finite estimates in binomial-response GLMs

The family of bias-reducing adjustments to the estimating functions defined by expression (6) allows for some creativity in the construction of adjustments. For example, when *M*-estimation is through an objective function, then maximisation of (13) after adding a plug-in penalty $P(\theta)$ with $\nabla P(\theta) = \boldsymbol{O}_p(\boldsymbol{n}^{-1/2})$ still results in estimators with bias of order $O(n^{-3/2})$. In this way, we can construct RB*M*-estimators with extra properties (RB*M*p-estimators).

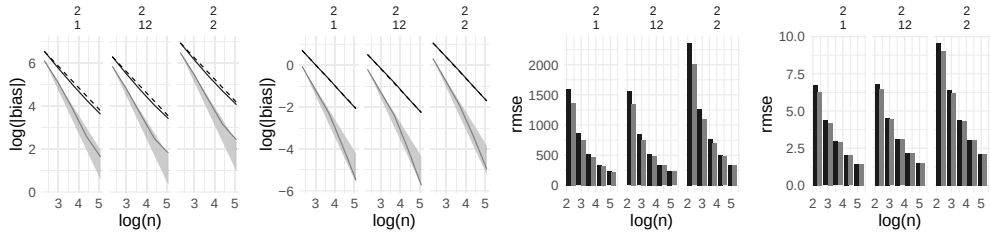

Figure 1. The two left-most panels refer to Σ_4 and Σ_5, respectively, and show simulation-based estimates of the logarithm of the absolute bias of $\hat{\boldsymbol{\theta}}$ (black) and $\tilde{\boldsymbol{\theta}}$ (grey) from the experiments in Example 5.2. The dashed line has slope −1 corresponding to the theoretical rate of the bias of $\hat{\boldsymbol{\theta}}$, and the shadowed region is defined by lines with slopes −3/2 and −2, corresponding to the theoretical rate of the bias of $\tilde{\boldsymbol{\theta}}$. The two right-most panels show simulation-based estimates of the root mean squared error of $\hat{\boldsymbol{\theta}}$ (black) and $\tilde{\boldsymbol{\theta}}$ (grey) at Σ_4 and Σ_5.

Firth (1993) showed that maximising the logistic regression likelihood after penalising it by the Jeffreys' invariant prior results in reduced-bias estimates. As proved in Kosmidis and Firth (2021), such penalisation has the useful side-effect that the reduced-bias estimates are always finite, even in cases where the ML estimates are infinite. The finiteness property of the reduced-bias estimator is rather attractive for applied work because it circumvents all the numerical and inferential issues that ML encounters with separated datasets (see, Mansournia et al., 2018, for a recent review of the practical issues associated with infinite estimates in logistic regression). In Kosmidis and Firth (2021), it is also shown that the finiteness of the maximum penalised likelihood estimates extends more generally to penalties that are any positive power of the Jeffreys' prior penalty, and to binomial-response GLMs with links other than the logistic. The resulting estimators, though, do not necessarily have better bias properties than the ML estimator, and that bias can be considerable for large effects; for example, Cordeiro and McCullagh (1991, Section 8.1) show that, in logistic regressions, the first term in the bias expansion of the maximum likelihood estimator grows roughly linearly with the effect size.

To this end, we can appropriately select a plug-in penalty $P(\theta)$ to (13) and construct finite RB*M*p-estimators for a wide range of binomial-response generalised linear models, including logistic, probit, complementary log-log, and cauchit regression. Consider a binomial-response GLM with ith likelihood contribution $f_i(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta})$ as in (14) and binomial totals $m_1, \ldots, m_n$. Consider also that estimation is by maximisation of the penalised log-likelihood

$$\sum_{i=1}^{n} \log f_i(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) - \frac{1}{2}\mathrm{trace}\left\{\boldsymbol{j}_{\beta\beta}(\boldsymbol{\beta})^{-1}\boldsymbol{e}_{\beta\beta}(\boldsymbol{\beta})\right\} + \frac{1}{N}\log|\boldsymbol{X}^\top \boldsymbol{W}(\beta)\boldsymbol{X}|, \tag{17}$$

for some $N > 0$, where $\boldsymbol{X}$ is the $n \times p$ matrix with rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, assumed to have full rank, and $\boldsymbol{W}(\beta)$ is a diagonal matrix with ith diagonal element the working weight $m_i\omega(\eta_i)$ $(i = 1, \ldots, n)$ with $\omega(\eta) = h'(\eta)^2/[h(\eta)\{1 - h(\eta)\}]$ and $h'(\eta) = dh(\eta)/\mathrm{d}\eta$. The first two terms in (17) form the bias-reducing penalised log-likelihood for GLMs with known dispersion derived in Example 5.1. What is important to note here is that, directly by its definition, $\boldsymbol{e}_{\beta\beta}(\boldsymbol{\beta})$ is symmetric and positive semi-definite. Furthermore, as shown in Pratt (1981), the log-likelihood for binomial generalised linear models is concave for many of the commonly used link functions, including logit, probit, complementary log-log, and cauchit links. Hence, for those link functions $\boldsymbol{j}_{\beta\beta}(\boldsymbol{\beta})$, apart from being symmetric, is also positive semi-definite. As a result $\mathrm{trace}\{\boldsymbol{j}_{\beta\beta}(\boldsymbol{\beta})^{-1}\boldsymbol{e}_{\beta\beta}(\boldsymbol{\beta})\} \geq 0$, and the bias-reducing penalised log-likelihood (13) is bounded above by zero. For the aforementioned link functions, it also holds that $\omega(\eta)$ converges to zero as η diverges to either $-\infty$ or $+\infty$. According to Kosmidis and Firth (2021, Theorem 1 and Section 3.1) the plug-in penalty $\log|\boldsymbol{X}^\top \boldsymbol{W}(\beta)\boldsymbol{X}|/N$ diverges to $-\infty$ as any element of β diverges. Hence, the maximiser of (17) has all of its components finite for any $N > 0$. Choosing, for example, $N = \sqrt{\sum_{i=1}^{n} m_i}$ or $N = \sum_{i=1}^{n} m_i$ also guarantees that the maximiser of (17), apart from finite components, also has bias that is free from the first-order

term. Examples 5.3 and 5.4 below demonstrate the performance of implicit RB*M*-estimators with plug-in penalties in fully specified probit regression, and partially specified autologistic regression.

Example 5.3 (Probit regression). The performance of the explicit and implicit RB*M* estimators is assessed here in a fully specified, Bernoulli-response GLM with $\mu_i = \Phi(\beta_1 + \sum_{t=2}^{5} \beta_t x_{it})$ $(i = 1, \ldots, n)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard Normal distribution. The covariate values $x_{i2}, x_{i3}, x_{i4}, x_{i5}$ $(i = 1, \ldots, n)$ are generated independently and independent of each other from a standard normal distribution, Bernoulli distributions with probabilities 1/4 and 3/4, and an exponential distribution with rate 1, respectively. Note that the necessary and sufficient condition of Kosmidis and Firth (2009, Theorem 1) is not satisfied for this model, and hence, there is no bias-reducing penalised log-likelihood that corresponds to the adjusted score functions of Firth (1993). Instead, the bias-reducing penalised log-likelihood (15) and its version with a plug-in penalty in (17) are well-defined.

For $n \in \{100, 150, 200, 250, 300\}$, we simulate n covariate values, as detailed in the previous paragraph, and conditional on those we simulate 1,000 samples of n response values from Bernoulli distributions with probabilities prescribed by a probit regression model with $\boldsymbol{\beta} = (-2, 2, 2, -0.5, -0.5)^\top$. For each sample, we estimate $\boldsymbol{\beta}$ using ML, explicit and implicit RB*M*-estimation, implicit RB*M*-estimation with plug-in penalty with $N = n$, and the adjusted score functions approach in Firth (1993), which relies on expectations of products of log-likelihood derivatives with respect to the correct model.

Maximum likelihood estimates are computed using the `glm` function in R (R Core Team, 2021), and the Firth (1993) adjusted scores estimates are computed using the `brglm_fit` method from the `brglm2` R package (Kosmidis, 2023). The RB*M*-estimates with and without plug-in penalty result from the numerical maximisation of (17) and (15), respectively; see the R scripts for probit regression supplied in online supplementary material.

Infinite ML estimates were observed for 15, and 4 samples when $n = 100$ and $n = 150$, respectively. The detection of infinite estimates was done using the linear programming algorithms in Konis (2007), as implemented in the `detectseparation` R package (Kosmidis et al., 2022). In those cases, maximising the likelihood and the bias-reducing penalised likelihood (15) result in estimates on the boundary of the parameter space and also explicit RB*M*-estimates cannot be computed. As expected by the arguments in Section 5.3, the maximisation of the bias-reducing penalised likelihood with the plug-in penalty in (17) always results in finite estimates, as did the adjusted score approach of Firth (1993).

Figure 2 shows estimates of the absolute bias and root mean squared error of the five estimators. The summaries are conditional on the ML estimates not being on the boundary of the parameter space because otherwise the bias and the root mean squared error are not formally defined for ML, explicit RB*M* estimation, and implicit RB*M*-estimation with no plug-in penalty. As is apparent, the adjusted scores approach of Firth (1993), and explicit and implicit RB*M*-estimation, with or without plug-in penalty, result in estimators with substantially smaller conditional bias and mean squared error than the ML estimator. The finite-sample bias and mean squared error of the RBM-estimators tend to be slightly larger than that of the adjusted scores approach of Firth (1993). The differences diminish fast as the sample size increases.

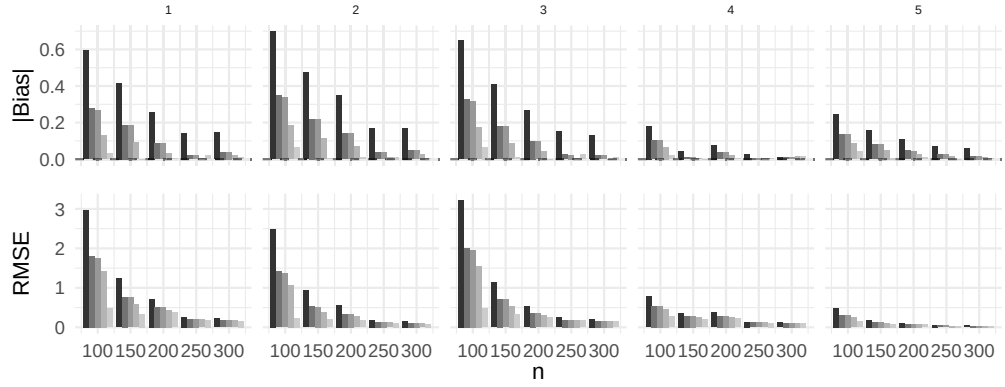

Figure 2. Absolute bias (|Bias|) and empirical root mean squared error (RMSE) of various estimators of $\beta_1, \ldots, \beta_5$ for $n \in \{100, 150, 200, 250, 300\}$ from the simulation setting in Example 5.3. Results are shown, from darker to lighter grey, for the ML estimator, the implicit RB*M*-estimator, the implicit RB*M*-estimator with plug-in penalty, the explicit RB*M*-estimator, and the adjusted score functions estimator of Firth (1993).

Example 5.4 (Autologistic regression models). The arguments in Section 5.3 are used here to deliver finite RB*M*p-estimators for autologistic regression models, which are typically estimated by composite likelihoods. Autologistic regression (J. Besag, 1974; J. E. Besag, 1972) is an important model for analysing binary responses with spatial or network correlation, used extensively in a range of disciplines, including ecology, anthropology, and computer vision; see Wolters (2017) for references. Suppose that $y(s) \in \{-1, 1\}$ is an observation at location $s \in \mathcal{S} = \{s_{11}, \ldots, s_{1c_1}, \ldots, s_{k1}, \ldots, s_{kc_k}\}$, which is assumed to be a realisation of a random variable $Y(s)$ with conditional probability mass function

$$f(y \mid \{y(u) : u \in G(s)\}, \boldsymbol{x}(s), \boldsymbol{\theta}) = \frac{e^{y\zeta(s)}}{e^{-\zeta(s)} + e^{\zeta(s)}} \quad \text{with} \quad \zeta(s) = \boldsymbol{x}(s)^\top \boldsymbol{\beta} + \lambda \sum_{u \in G(s)} y(u), \tag{18}$$

where $\boldsymbol{x}(s)$ is a p-vector of covariate observed at location s, $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \lambda)^\top$, and λ is the association parameter. The mapping $G(s)$ is assumed to be known and returns the set of all locations that are neighbours of s, excluding s. We allow for $k \in \{1, 2, \ldots\}$ disconnected clusters of observations, each with its own neighbourhood structure, by assuming $G(s_{ij}) \cap G(s_{i'j'}) = \emptyset$, for $i \neq i'$, $j \in \{1, \ldots, c_i\}$ and $j' \in \{1, \ldots, c_{i'}\}$. Negative values of λ promote discordant responses in the neighbourhoods, while positive values promote concordant ones. Wolters (2017) shows that the $\{-1, 1\}$ coding of the binary responses in combination with the parameterisation in (18) has notable advantages over alternative proposals for autologistic regression models (for example, the traditional specification with $\{0, 1\}$-coding in J. Besag, 1974, and the, more recent, centred autologistic model with $\{0, 1\}$-coding in Caragea and Kaiser, 2009), both in terms of estimation and interpretation of the model parameters. What is worth noting here is that the regression and association effects in the corresponding $\{0, 1\}$-coding model are $2\boldsymbol{\beta}$ and 4λ, respectively.

The joint probability mass function $f(y(s_{11}), \ldots, y(s_{kc_k}) \mid \boldsymbol{x}(s_{11}), \ldots, \boldsymbol{x}(s_{kc_k}), \boldsymbol{\theta})$ with full conditionals of the form (18) involves a typically intractable normalising constant. Nevertheless, sampling from it is possible using

Gibbs sampling or perfect sampling approaches; see, Hughes et al. (2011) and Wolters (2017), and references therein, and the `Autologistic` Julia package (Wolters, 2022), which implements a range of sampling techniques. So, estimation for autologistic regression models can be performed using Monte Carlo approximations to the likelihood function. A computationally more attractive approach is to maximise the pseudolikelihood

$$\exp\{l(\boldsymbol{\theta})\} = \prod_{i=1}^{k}\prod_{j=1}^{c_i} f(y(\boldsymbol{s}_{ij}) \mid \{y(u) : u \in G(\boldsymbol{s}_{ij})\}, \boldsymbol{x}(\boldsymbol{s}_{ij}), \boldsymbol{\theta}), \tag{19}$$

with respect to $\boldsymbol{\theta}$, which, in the terminology of Varin et al. (2011), is a composite conditional likelihood. J. Besag (1975) establishes the consistency of the M-estimators from the maximisation of $l(\boldsymbol{\theta})$. A straightforward calculation gives that

$$\nabla \log f(y(\boldsymbol{s}_{ij}) \mid \{y(u) : u \in G(\boldsymbol{s}_{ij})\}, \boldsymbol{x}(\boldsymbol{s}_{ij}), \boldsymbol{\theta}) = \{z(\boldsymbol{s}_{ij}) - \pi(\boldsymbol{s}_{ij})\}\tilde{\boldsymbol{x}}(\boldsymbol{s}_{ij}),$$

where $z(\boldsymbol{s}) = \{1 + y(\boldsymbol{s})\}/2$, $\pi(\boldsymbol{s}) = 1/\{1 + e^{-2\zeta(\boldsymbol{s})}\}$, and $\tilde{\boldsymbol{x}}(\boldsymbol{s}) = (2\boldsymbol{x}(\boldsymbol{s})^\top, 2\sum_{u\in G(\boldsymbol{s})} y(u))^\top$. So, the contribution to the estimating functions from the ith cluster of observations is $\boldsymbol{\psi}^i(\boldsymbol{\theta}) = \sum_{j=1}^{c_i}\{z(\boldsymbol{s}_{ij}) - \pi(\boldsymbol{s}_{ij})\}\tilde{\boldsymbol{x}}(\boldsymbol{s}_{ij})$, which has expectation zero under the joint probability mass function. The bias-reducing penalised pseudolikelihood is $l^{(\mathrm{RBM})}(\boldsymbol{\theta}) = l(\theta) - \mathrm{trace}\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\}/2$ in (13), with $\boldsymbol{e}(\boldsymbol{\theta}) = \sum_{i=1}^{k}\boldsymbol{\psi}^i(\boldsymbol{\theta})\boldsymbol{\psi}^i(\boldsymbol{\theta})^\top$ and $\boldsymbol{j}(\boldsymbol{\theta}) = \tilde{\boldsymbol{X}}^\top\boldsymbol{W}(\theta)\tilde{\boldsymbol{X}}$, where the matrix $\tilde{\boldsymbol{X}}$ has (i, j)th element $\tilde{\boldsymbol{x}}(\boldsymbol{s}_{ij})$ and $\boldsymbol{W}(\theta)$ has diagonal elements $\pi(\boldsymbol{s}_{11})\{1 - \pi(\boldsymbol{s}_{11})\}, \ldots, \pi(\boldsymbol{s}_{kc_k})\{1 - \pi(\boldsymbol{s}_{kc_k})\}$.

As with logistic regression, the estimates from the maximisation of (19) can be infinite with positive probability. This is particularly dangerous in the practice of these models, because parametric bootstrap is, currently, the most widely used approach for uncertainty quantification and inference about the model parameters. There is a positive probability that optimisation algorithms will fail or return a very large, in absolute value, estimate, instead of plus or minus infinity, because their convergence criteria have been satisfied. Then, naive handling of the bootstrap estimates will return large bootstrap standard errors or nonsensical bootstrap intervals, when the latter are simply formally not well-defined. To the best of authors' knowledge, this is something that has not been reported before in the autologistic regression literature. Because $l(\theta)$ is bounded above by zero and is concave (see, Wolters, 2017, Theorem 5), the same arguments as in Section 5.3 show that the RBMp-estimates from the maximisation of $l^{(\mathrm{RBMp})}(\boldsymbol{\theta}) = l^{(\mathrm{RBM})}(\boldsymbol{\theta}) + \log|\tilde{\boldsymbol{X}}^\top\boldsymbol{W}(\boldsymbol{\theta})\tilde{\boldsymbol{X}}|/\sum_{i=1}^{k} c_i$ have both improved bias properties and are always finite. Hence, they are safe to use for bootstrap uncertainty quantification and inferences.

For illustration purposes, we focus on two subsets of the Gambia malaria survey data (Thomson et al., 1999), which is provided in the `geoR` R package (Ribeiro Jr et al., 2022). The two subsets consist of villages in the central (subset 1, with 270 children in 8 villages) and south western (subset 2, with 321 children in 11 villages) region of the Gambia, respectively. The variable of interest is an indicator denoting the presence (coded as 1) or not (coded as −1) of malaria in the blood sample taken from each child. The covariate information includes: the age of the child in years (age), an indicator denoting whether (coded as 1) or not (coded as 0) the child regularly sleeps under a bed-net (netuse), an indicator denoting whether (coded as 1) or not (coded as 0) the bed-net is treated (treated), a satellite-derived

Table 4. *M*- and RB*M*p-estimates and corresponding nominally 95% normal and percentile bootstrap confidence intervals from an autologistic regression model for two subsets of the Gambia malaria survey data (Thomson et al., 1999), as detailed in Example 5.4

	Estimates		Normal		Percentile	
Subset 1 (intercept and coefficient for netuse in $\arg\max l$ **are** $-\infty$ **and** ∞**, respectively)**						
	M	**RBMp**	**M [500]**	**RBMp [499]**	**M [500]**	**RBMp [499]**
intercept	−11.43	−2.60	(−13.86, −9.58)	(−5.39, −0.02)	(−12.83, −8.78)	(−4.99, 0.50)
age	0.16	0.16	(0.02, 0.31)	(0.03, 0.30)	(0.02, 0.31)	(0.03, 0.30)
netuse	12.24	3.24	(9.81, 14.09)	(2.66, 3.77)	(10.84, 14.88)	(2.83, 3.71)
treated	–0.41	–0.39	(−1.01, 0.45)	(−0.92, 0.27)	(−1.31, 0.09)	(−1.12, 0.05)
green	–0.04	–0.03	(−0.13, 0.10)	(−0.09, 0.04)	(−0.19, 0.01)	(−0.12, 0.00)
phc	0.05	0.04	(−0.86, 0.76)	(−0.51, 0.52)	(−0.45, 1.04)	(−0.35, 0.58)
λ	0.01	0.01	(−0.05, 0.13)	(−0.03, 0.07)	(−0.12, 0.02)	(−0.06, 0.03)
Subset 2 (no infinite components in $\arg\max l$**)**						
	M	**RBMp**	**M [1]**	**RBMp [0]**	**M [1]**	**RBMp [0]**
intercept	–9.27	–9.99	(−41.20, 16.98)	(−27.23, 11.31)	(−30.22, 26.54)	(−34.65, 3.38)
age	0.16	0.16	(0.04, 0.27)	(0.03, 0.30)	(0.05, 0.28)	(0.03, 0.30)
netuse	0.16	0.11	(−0.75, 0.94)	(−0.49, 0.72)	(−0.41, 0.81)	(−0.54, 0.63)
treated	0.01	–0.05	(−0.54, 0.67)	(−0.62, 0.53)	(−0.62, 0.60)	(−0.62, 0.59)
green	0.21	0.23	(−0.43, 1.00)	(−0.29, 0.65)	(−0.67, 0.72)	(−0.10, 0.83)
phc	0.01	0.14	(−1.10, 1.50)	(−0.60, 0.72)	(−1.64, 0.47)	(−0.43, 0.95)
λ	0.02	0.02	(−0.02, 0.08)	(0.01, 0.04)	(−0.05, 0.03)	(0.01, 0.03)

Note. The numbers in [·] are the number of simulated samples with at least one infinite component in the *M*-estimate.

measure of the greenness of vegetation in the immediate vicinity of the village in arbitrary units (green), and an indicator for the presence (coded as 1) or absence (coded as 0) of a health centre in the village (phc).

Table 4 shows the estimates from fitting an autologistic regression model by the numerical maximisation of $l(\boldsymbol{\theta})$ and $l^{(\mathrm{RBMp})}(\boldsymbol{\theta})$ on each of the subsets. For both subsets, the first element of $\boldsymbol{x}(\boldsymbol{s}_{ij})$ is set to 1, corresponding to an intercept parameter, and the remainder elements are the covariate values for child j in village i. The mapping $G(\boldsymbol{s})$ has been constructed under the assumption that any two children in the data are neighbours only if they come from the same village. All estimates from the maximisation of $l(\boldsymbol{\theta})$ for subset 2 have finite components, while, for subset 1, the estimates −11.43 and 12.24 for the intercept and the parameter for netuse, respectively, are in reality $-\infty$ and ∞ (detection of infinite estimates took place using the `detectseparation` R package for the logistic regression of $z(\boldsymbol{s})$ on $\tilde{\boldsymbol{x}}(\boldsymbol{s})$). Those apparently finite values are artefacts due to the numerical optimisation procedure stopping prematurely by meeting the optimiser's convergence criteria. In contrast, all RB*M*p-estimates are finite and can be trusted at the reported accuracy.

Table 4 also reports 95% bootstrap confidence intervals using normal approximation and 95% bootstrap percentile intervals (see Davison and Hinkley, 1997, expression (5.5) and expression (5.18), respectively). The intervals have been computed using parametric bootstrap of size 500 at the reported estimates, which is the current state of the art for autologistic regression, without employing any special convention for the handling of

any infinite *M*-estimates that may arise in the bootstrap samples. The effect of infinite estimates is dramatic in subset 1, where all bootstrap samples at the *M*-estimates result in infinite *M*-estimates for the intercept and the coefficient of netuse. The impact is that the apparent evidence that bootstrap at the *M*-estimates provides about the respective parameters being zero are grossly exaggerated, simply reflecting the optimiser's stopping criteria. Overall, the bootstrap intervals based on RB*M*p, apart from being always well-defined tend to be shorter in length, and in some cases result in different inferential conclusions than the corresponding ones based on the *M*-estimates. For example, both bootstrap intervals based on the RB*M*p estimates provide evidence of positive association in subset 2, while the ones based on *M*-estimates provide no evidence of association.

The bootstrap intervals based on the RB*M*p estimator are also found to perform better in terms of coverage than the typically wider intervals based on the *M*-estimator. We simulated 1000 samples at RB*M*p estimates from subset 2 in Table 4, and for each sample we computed the nominally 95% bootstrap percentile intervals based on the *M*- and the RB*M*p-estimators using a bootstrap of size 500. No infinite *M*-estimates were detected in the simulated samples. The estimated coverage probabilities for *M*-estimation in the order that the parameters appear in Table 4 are 0.84, 0.95, 0.92, 0.92, 0.82, 0.81, 0.82, respectively. The corresponding estimated coverage probabilities based on RB*M*p-estimation are found to be markedly closer to the nominal level with 0.90, 0.95, 0.94, 0.94, 0.89, 0.91, 0.90, respectively, with the intervals also having shorter lengths at 75%, 99%, 72%, 17%, 75%, 18%, 77%, respectively, of the length of the *M*-estimation ones. Properties of the various variants of bootstrap intervals and their suitability for autologistic regression is an interesting topic that is beyond the scope of the current work.

Note here that the adjusted score equations of Firth (1993) do not apply easily because the joint probability mass function with full conditionals of the form (18) is typically intractable. Furthermore, alternative bias reduction methods such bootstrap, indirect inference, and the jackknife are not well-defined, due to the positive probability of infinite maximum pseudolikelihood estimates. In fact, if infinite estimates go undetected, such methods can easily return nonsensical estimates.

5.4 Links to model selection using Kullback–Leibler divergence

Suppose that $l(\boldsymbol{\theta})$ is the log-likelihood function based on an assumed parametric model F. Takeuchi (1976) showed that

$$-2l(\hat{\boldsymbol{\theta}}) + 2\text{trace}\left\{\boldsymbol{j}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{e}(\hat{\boldsymbol{\theta}})\right\} \tag{20}$$

is an estimator of the expected Kullback–Leibler divergence of the underlying process G to the assumed model F, where $\hat{\boldsymbol{\theta}}$ is the ML estimator. Expression (20) is known as the Takeuchi information criterion (TIC), and, in contrast to the Akaike Information Criterion (AIC; Akaike, 1974), is robust against deviations from the assumption that the model is correct. Claeskens and Hjort (2008, Section 2.5) thoroughly discuss the relationship between TIC and AIC.

Model selection from a set of parametric models proceeds by computing $\hat{\boldsymbol{\theta}}$ for each model and selecting the model with the smallest TIC value (20), or equivalently, with the largest

$$l(\hat{\boldsymbol{\theta}}) - \text{trace}\left\{\boldsymbol{j}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{e}(\hat{\boldsymbol{\theta}})\right\}. \tag{21}$$

A direct comparison of expressions (21) and (13) reveals a previously unnoticed close connection between bias reduction in ML estimation and model selection. Specifically, both bias reduction

and TIC model selection rely on exactly the same penalty trace$\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\}$, but differ in the strength of penalisation; bias reduction is achieved by using half that penalty, while valid model selection requires stronger penalisation by using one times the penalty.

As discussed in Section 4, the explicit and implicit RB*M*-estimators $\boldsymbol{\theta}^\dagger$ and $\tilde{\boldsymbol{\theta}}$, respectively, have the same asymptotic distribution as $\hat{\boldsymbol{\theta}}$. Then, the derivation of TIC (see, for example, Claeskens and Hjort, 2008, Section 2.3) works also with the RB*M*-estimators in place of the ML estimator. As a result, TIC and, under extra assumptions, AIC at the RB*M*-estimates are asymptotically equivalent to their versions at the maximum likelihood estimates. The same holds for reduced-bias estimators of Firth (1993). In other words, TIC model selection can proceed by selecting the model with the largest value of

$$l(\tilde{\boldsymbol{\theta}}) - \text{trace}\left\{\boldsymbol{j}(\tilde{\boldsymbol{\theta}})^{-1}\boldsymbol{e}(\tilde{\boldsymbol{\theta}})\right\}, \tag{22}$$

and AIC model selection using the largest value of $l(\tilde{\boldsymbol{\theta}}) - p$, and the same holds when $\boldsymbol{\theta}^\dagger$ is used in place of $\tilde{\boldsymbol{\theta}}$. The quantity in (22) is readily available once (13) has been maximised to obtain the implicit RB*M*-estimates; the only requirement for model selection is to adjust, from 1/2 to 1, the factor of the value of trace$\{\boldsymbol{j}(\boldsymbol{\theta})^{-1}\boldsymbol{e}(\boldsymbol{\theta})\}$ after maximisation. The same holds when $\tilde{\boldsymbol{\theta}}$ is obtained using plug-in penalties, as in (17).

Varin and Vidoni (2005) developed a model selection procedure when the objective $l(\boldsymbol{\theta})$ is a composite likelihood (see, Varin et al., 2011, for a review of composite likelihood methods). The composite likelihood information criterion (CLIC) derived in Varin and Vidoni (2005) has the same functional form as TIC in (20). So, the link between model selection and bias reduction exists also when $l(\boldsymbol{\theta})$ is the logarithm of a composite likelihood. From the discussion in Section 5.2 it follows that both implicit and explicit RB*M*-estimation are readily available when there is a ready implementation of TIC for likelihood problems, and CLIC for composite likelihood problems (see, e.g. Padoan and Bevilacqua, 2015, for estimation of random fields based on composite likelihoods).

Example 5.5 (Model selection in probit regression). The performance of model selection procedures is assessed here using the probit regression model in Example 5.3. There are 16 possible nested probit regression models with an intercept β_1, depending on which of $\beta_2, \ldots, \beta_5$ are zero or non-zero. For $n \in \{75, 150, 300, 600\}$, we simulate n covariate values, as detailed in Example 5.3, and conditional on those we simulate 10,000 samples of n response values from Bernoulli distributions with probabilities prescribed by a probit regression model with $\boldsymbol{\beta} = (-0.5, 0, 0, 0.5, 0.5)^\top$. For each sample, we estimate all 16 possible models using maximum likelihood, the adjusted score functions approach in Firth (1993), and by maximising the bias-reducing penalised likelihoods with and without plug-in penalty, in (17) and (15), respectively, where we use $N = n$ for the plug-in penalty.

There were 7 separated data sets for at least one of the 16 models, detected using the `detectseparation` R package. Figure 3 shows the selection proportion among the 16 models based on AIC and TIC at the ML estimates, the implicit RB*M*-estimates, the implicit RB*M*-estimates with plug-in penalty with $N = n$, and the adjusted score functions estimates in Firth (1993). As expected by the discussion in Section 5.4, the probability of selecting the model with $\beta_2 = \beta_3 = 0$ increases with the sample size for both information criteria and for all estimation methods. There are only small discrepancies on the selection proportions between estimation methods, which tend to disappear as the sample size increases. Finally, it is worth noting that AIC model selection tends to be more confident on what the true model is than TIC, illustrating less variability in the selected proportions. In the current study, for the larger samples sizes this results in selecting the correct model more often than TIC does. However, in smaller samples sizes, AIC selects the model with $\beta_2 = \beta_3 = \beta_4 = 0$ more often.

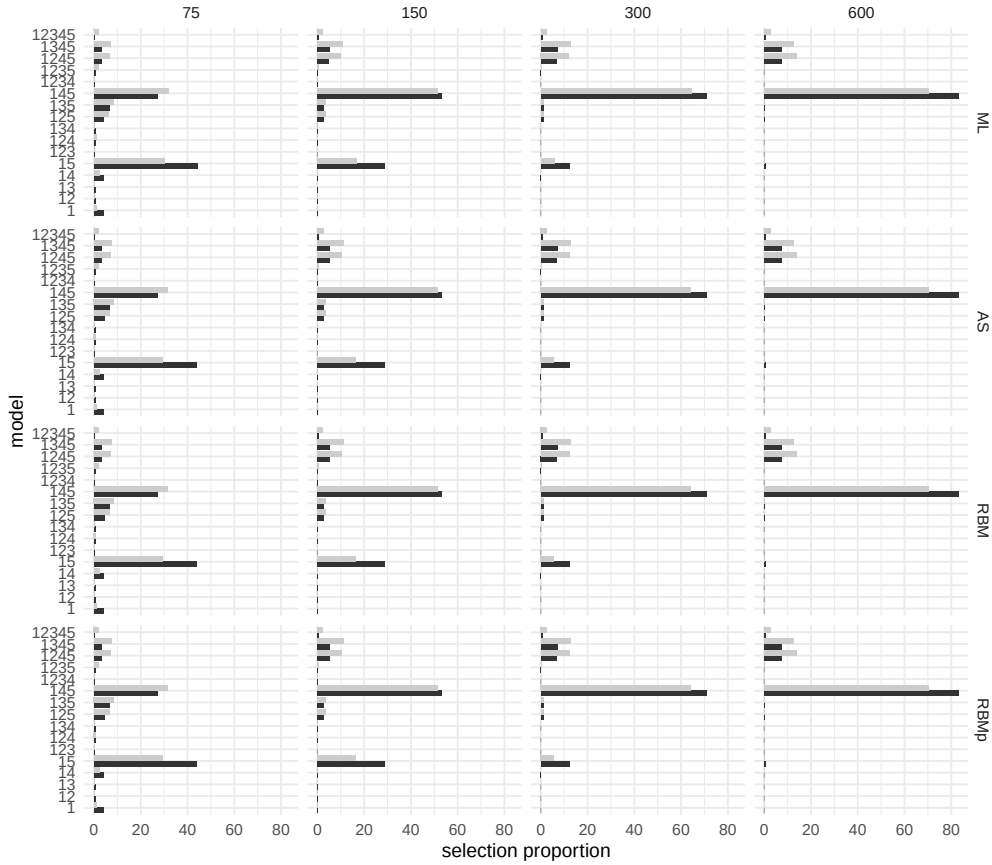

Figure 3. Model selection proportions based on AIC (black) and TIC (grey) among the 16 nested models of the probit regression in the simulation setting of Example 5.5 when estimation is through ML, maximum bias-reducing penalised likelihood without (RB*M*) and with (RB*M*p) plug-in penalty, and the adjusted scores (AS) approach of Firth (1993). The *y*-axes give the indices of the elements of β present in the estimated model.

6 High-dimensional regression settings

For the estimation of logistic regression models with $p/n \to \kappa \in (0, 1)$, experiments reported in the supplementary information of Sur and Candès (2019) and the supplementary material of Kosmidis and Firth (2021) illustrate that the bias-reduction method of Firth (1993) performs similarly to the methods of Sur and Candès (2019), which are based on approximate message passing algorithms, and markedly better than ML. Reduced-bias *M*-estimators are found here to perform well in high-dimensional, partially specified regression settings, where the bias-reduction methods of Firth (1993) and the methods of Sur and Candès (2019) do not apply directly. We note here that the sufficient conditions we used (see online supplementary material, Section S2) for developing reduced-bias *M*-estimation do not cover for $p/n \to \kappa \in (0, 1)$.

Example 6.1 (High-dimensional autologistic regression). To illustrate, we consider the autologistic regression model in (18), with $k = 100$ fully connected groups, with $c = c_1 = \ldots = c_{100} = 10$, where each covariate vector $\boldsymbol{x}(s_{ij})$ has $p = 100$ entries simulated independently from a Normal distribution with mean 0 and variance $(2kc)^{-1}$. We consider five high-dimensional autologistic regression models, for $\lambda \in \{-0.2, -0.1, 0, 0.1, 0.2\}$ and where $\boldsymbol{\beta}$ has the first 20 entries having value 10, the next 20 entries having value 5, and the remaining 60 entries having value 0. For these parameter settings, the marginal probabilities within each group range from spanning (0, 1) ($\lambda = -0.2$,

Table 5. Variance and the mean squared error to variance ratio (in parenthesis) for the *M*-estimator and the RB*M*p-estimator of $\boldsymbol{\beta}$ for the distinct values in the true value for $\boldsymbol{\beta}$ in the simulation setting of Example 6.1

λ	*M*-estimator			*M*p-estimator			RB*M*p-estimator		
	0	5	10	0	5	10	0	5	10
−0.20	6.23	6.83	6.72	6.23	6.83	6.71	4.80	5.24	5.07
	(1.00)	(1.15)	(1.52)	(1.00)	(1.15)	(1.51)	(1.00)	(1.01)	(1.03)
−0.10	5.80	6.35	6.24	5.80	6.35	6.23	4.51	4.92	4.75
	(1.00)	(1.14)	(1.45)	(1.00)	(1.14)	(1.45)	(1.00)	(1.01)	(1.02)
0.00	5.37	5.84	5.78	5.37	5.84	5.78	4.36	4.83	4.87
	(1.00)	(1.12)	(1.44)	(1.00)	(1.12)	(1.44)	(1.00)	(1.01)	(1.03)
0.10	6.21	6.41	6.31	6.21	6.40	6.31	4.79	4.92	4.80
	(1.00)	(1.15)	(1.52)	(1.00)	(1.15)	(1.52)	(1.00)	(1.01)	(1.03)
0.20	19.94	21.04	24.85	19.90	21.00	24.76	13.27	13.71	15.61
	(1.00)	(1.21)	(1.74)	(1.00)	(1.21)	(1.74)	(1.00)	(1.04)	(1.16)

Note. *M*p is *M*-estimation penalised just by the plug-in penalty that ensures finiteness of the estimates.

which promotes discordant, high-variance neighbourhoods) to being similarly extreme in value ($\lambda = 0.2$, which promotes concordant, low-variance neighbourhoods).

We simulate 100 independent response samples at each $\lambda \in \{-0.2, -0.1, 0, 0.1, 0.2\}$ and for each sample we estimate $\boldsymbol{\beta}$ and λ. Table 5 shows the variance and the mean squared error to variance ratio for the *M*-estimator and the RB*M*p-estimator of $\boldsymbol{\beta}$ for the distinct values in the true value for $\boldsymbol{\beta}$. All estimates had finite components. A mean squared error to variance ratio close to one indicates that the bias has a small contribution to the mean squared error. Both *M* and RB*M*p estimators are close to being unbiased for the zero effects in $\boldsymbol{\beta}$. Nevertheless, the contribution of the bias to the mean squared error of the *M*-estimator increases rapidly as the effects sizes grow; for the entries of $\boldsymbol{\beta}$ with value 10, the mean squared error of the *M*-estimator is as much as 74% larger than its variance. Remarkably, the RB*M*p-estimator apart from smaller variance for all parameters has a mean squared error to variance ratios close to one. For reference, Table 5 also reports figures for *M*-estimation penalised just by the plug-in penalty. The figures are almost identical to those for *M*-estimation indicating that the excellent bias and variance properties the RB*M*p-estimator has in this setting are due to bias-reducing penalty $-\text{trace}\{\boldsymbol{j}(\theta)^{-1}\boldsymbol{e}(\theta)\}/2$, with the plug-penalty having almost no impact on the *M*-estimates.

7 Correlated random components

In Section 3.3, we have shown that when $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are realisations of independent random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, the implicit and explicit RB*M*-estimation will result in estimates with superior bias properties than the corresponding *M*-estimates. This is true regardless of the dependence structure imposed by the model on the components of each random vector $\boldsymbol{Y}_i$; this fact is illustrated in Example 5.2 in the context of Gaussian max-stable processes. However, it is often the case that $k = 1$, in the sense that the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_T)^\top$ is assumed to be a single realisation of a random vector $\boldsymbol{Y}$ with correlated components, like when a single time series or spatial process is observed, and the information about θ is assumed or expected to increase with T. In such cases, $\boldsymbol{j}(\theta)$ and $\boldsymbol{u}_r(\theta)$ in the empirical adjustment (8) remain valid estimators of $\mu_{rs}(\theta)$ and $\mu_{rst}(\theta)$ in (6), respectively. However, $\boldsymbol{e}(\theta)$ and $\boldsymbol{d}_r(\boldsymbol{\theta})$ tend to be rather imprecise estimators of $v_{r,s}(\theta)$ and $v_{rs,t}(\theta)$, and their use may duly result in the original estimator $\hat{\boldsymbol{\theta}}$ because $\boldsymbol{e}(\hat{\boldsymbol{\theta}}) = \boldsymbol{d}_r(\hat{\boldsymbol{\theta}}) = 0_{p\times p}$ when $k = 1$.

In fact, estimation of $v_{r,s}(\theta)$ when $k = 1$ is a topic that has attracted much research, which mainly focuses on estimating the covariance matrix of composite likelihood estimators (see, for example, Varin et al., 2011, Section 5), or of M-estimators in regression problems more generally (see, for example, Heagerty and Lumley, 2000). In such cases, weak stationarity conditions (Carlstein, 1986; Heagerty and Lumley, 2000) on the model allow the use of window subsampling to evaluate the estimating functions (or derivatives of those) over multiple, overlapping or non-overlapping subsets of the temporal or spatial domain, and aggregate the resulting contributions. Similar procedures can be used for the estimator of $v_{rs,t}(\theta)$.

Example 7.1 (Least squares for autoregressive processes). Let $\{Y_1, \ldots, Y_T\}$ be an autoregressive process of order one, in the sense that $Y_t = \theta Y_{t-1} + \epsilon_t$ with $\theta \in (-1, 1)$, where $\epsilon_1, \ldots, \epsilon_T$ are independent and identically distributed random variables with zero mean and finite variance. The ordinary least squares estimator of θ is $\hat{\theta} = S_1(\Omega)/S_2(\Omega)$, where $S_1(\omega) = \sum_{t \in \omega} Y_{t+1} Y_t$, $S_2(\omega) = \sum_{t \in \omega} Y_t^2$, $\Omega = \{1, \ldots, T-1\}$, and $Y_t = 0$ if $t \notin \Omega \cup \{T\}$. The estimator $\hat{\theta}$ results by maximising the objective $l(\theta) = -\sum_{t \in \Omega} (Y_{t+1} - \theta Y_t)^2$ or, equivalently, by finding the root of the estimating function $\psi(\theta) = 2S_1(\Omega) - 2\theta S_2(\Omega)$.

The implicit and explicit RBM-estimators can be computed using the bias-reducing penalised objective in (13). For this model, $j(\theta) = 2S_2(\Omega)$, and in place of $e(\theta)$ we use $e^{(ws)}(\theta) = \kappa_T \sum_{r=1}^{R} \{2S_1(\omega_r) - 2\theta S_2(\omega_r)\}^2$ where $\kappa_T > 0$ is a suitable constant. The quantity $e^{(ws)}(\theta)$ is a window subsampling estimator of $\mathrm{E}_G[\{2S_1(\Omega) - 2\theta S_2(\Omega)\}^2]$ (Carlstein, 1986), where $\omega_1, \ldots, \omega_R$ is a partitioning of Ω, with ω_r having indices of observations that are consecutive in time. Then, the implicit and explicit RBM-estimators are

$$\tilde{\theta} = \frac{S_{12}(\Omega) + \kappa_T S_{12}(\omega_1, \ldots, \omega_R)}{S_{22}(\Omega) + \kappa_T S_{22}(\omega_1, \ldots, \omega_R)} \text{ and}$$

$$\theta^\dagger = \hat{\theta}\left\{1 + \kappa_T \frac{S_{22}(\omega_1, \ldots, \omega_R)}{S_{22}(\Omega)}\right\} - \kappa_T \frac{S_{12}(\omega_1, \ldots, \omega_R)}{S_{22}(\Omega)},$$

respectively, where $S_{lm}(\omega_1, \ldots, \omega_R) = \sum_{r=1}^{R} S_l(\omega_r) S_m(\omega_r)$. Note that the reduced-bias estimator θ^* of Firth (1993) can only be obtained under specific assumptions about the distribution of $\epsilon_1, \ldots, \epsilon_T$. If $\epsilon_t \sim N(0, 1)$, then the root of the adjusted score function $2S_1(\Omega) - 2\theta S_2(\Omega) - 4(T-1)/\{T(1-\theta^2)\}$ has second-order bias. The latter expression is a polynomial of order 3 and, hence, it can have up to 3 distinct roots, with at least one of them being real. In contrast, $\tilde{\theta}$ and $\theta^\dagger$ always have a single value.

We conduct a simulation study to assess the finite-sample bias properties of $\hat{\theta}$, $\tilde{\theta}$, $\theta^\dagger$, and θ^*. We simulate N_T independent time series of lengths $T \in \{50, 100, 200, 400, 800\}$, with varying strength of dependence $\theta \in \{0.2, 0.5, 0.9\}$. The simulation size is set to $N_T = 250 \times 2^{16}/T$ to control the simulation error, similarly to Example 3.1. The errors ϵ_t ($t = 1, \ldots, T$) are simulated so that the joint distribution of the random vector $(Y_1, \ldots, Y_T)^\top$ is multivariate normal, multivariate Student-t with 5 degrees of freedom, and multivariate asymmetric Laplace, all with mean $\mathbf{0}_T$ and covariance matrix with (t, k) entry $\mathrm{Cov}(Y_t, Y_k) = \theta^{|t-k|}$, which is the exponential decaying function of an autoregressive process of order one. We compute θ^* for all three error structures with the bias-reducing adjustment computed at the model with normal errors to illustrate the impact of using the wrong bias adjustment, and we only keep the root of the polynomial that is returned by starting at $\hat{\theta}$. In order to investigate the impact of the subsampling scheme on $\tilde{\theta}$ and $\theta^\dagger$, we consider $m_T = T^\alpha$ with the $\alpha \in \{1/3, 1/2\}$. The choice $\alpha = 1/3$ is the one that minimises the mean

Table 6. Simulation-based estimates of the bias (×100) of the various estimators for the parameter of the AR(1) process in Example 7.1

	α	Errors	T					Slope	
			50	100	200	400	800	Est	Exp
$\hat{\theta}$			−1.95	−0.98	−0.50	−0.23	−0.13	−0.99	−1
$\tilde{\theta}$	1/3		−1.41	−0.56	−0.23	−0.08	−0.04	−1.32	−3/2
	1/2		−1.20	−0.44	−0.16	−0.04	−0.02	−1.49	−3/2
$\theta^{\dagger}$	1/3		−1.36	−0.54	−0.22	−0.08	−0.04	−1.32	−3/2
	1/2		−1.07	−0.38	−0.14	−0.03	−0.02	−1.53	−3/2
$\theta^{(J)}$			−0.22	−0.04	−0.02	0.02	−0.00	−1.468	<−1
$\theta^{(M)}$			−2.47	−1.33	−0.71	−0.34	−0.18	−0.945	<−1
$\theta^{(S)}$	1/3		10.44	9.28	8.08	7.08	5.53	−0.22	<−1
	$\log(T/2)/\log(T)$		1.33	0.79	0.43	0.24	0.11	−0.89	<−1
		Normal	0.30	0.07	0.01	0.02	−0.00	−1.60	−2
θ^{*}		Student-t	2.33	0.88	0.38	0.19	0.08	−1.18	
		Laplace	−12.55	−8.49	−5.68	−3.98	−2.75	−0.55	

Note. Figures are reported in 2 decimal places, and the figures −0.00 are for estimated biases in the interval $(-0.0025, -0.0025)$. The simulation error for the estimates of the bias is between 2.76×10^{-4} and 1.37×10^{-3}. The last two columns are the estimated (Est) and expected (Exp) slope of the regression line $\log|\text{bias}| \sim \log T$.

squared error of $e^{(ws)}(\theta)$ (Carlstein, 1986) for AR(1) models. In the comparison, we also include the bootstrap bias-corrected estimators $\theta^{(M)}$ using residual resampling, and $\theta^{(S)}$ using the stationary bootstrap (see Davison and Hinkley, 1997, Chapter 8), computed from 500 bootstrap samples, and as implemented in the `tsboot` function in the `boot` R package (Canty and Ripley, 2022). The block scheme for the stationary bootstrap is chosen to have average length T^{α}, with $\alpha \in \{1/3, \log(T/2)/\log(T)\}$. Average block length proportional to $T^{1/3}$ is optimal under the general guidelines in Kunsch (1989) and Politis and Romano (1994) for the standard error of $\hat{\theta}$. The other choice for α sets the average block length equal to $T/2$. We also include the half-sample jackknife estimator $\theta^{(J)}$ (Quenouille, 1949).

Table 6 gives the simulation-based estimates of the bias of estimators at $\theta = 0.5$. The results for $\theta \in \{0.2, 0.9\}$ are in online supplementary material. The estimated biases $\hat{\theta}$, $\tilde{\theta}$, $\theta^{\dagger}$, $\theta^{(J)}$, $\theta^{(M)}$, and $\theta^{(S)}$ in Table 6 are identical for all three error distributions we consider. The reason is that by mere inspection of their expressions, the RB*M*-estimates, like the ordinary least squares estimates and, in turn, bootstrap- and jackknife-based versions, are invariant to scaling the time series by a scalar random variable C that is independent to the distribution of $(\epsilon_1, \ldots, \epsilon_T)$, and both the multivariate Student-t and asymmetric Laplace distributions are infinite mixtures of the multivariate Normal distribution. The estimator θ^{*} does not have such invariance.

Table 6 also reports the estimated and theoretical slopes from the regression of the logarithm of the absolute bias to $\log T$. We observe a close agreement between estimated and expected slopes for $\hat{\theta}$, $\tilde{\theta}$, $\theta^{\dagger}$, and θ^{*} under Normal errors. We also note that RB*M*-estimators deliver the bias reduction expected by the theory in Section 3 for both window sizes considered. As the error distribution departs from Normal, the use of an incorrect adjustment impacts θ^{*} both in terms of finite sample bias and the asymptotic rates of the bias, which drop dramatically in absolute value from what is expected under Normal errors. The jackknife estimator $\theta^{(J)}$ is as effective

as $\tilde{\theta}$ and $\theta^{\dagger}$ in terms of reducing bias. On the other hand, use of $\theta^{(M)}$ appears to have not much effect in reducing the bias of $\hat{\theta}$ with an estimated slope close to -1, and $\theta^{(S)}$ delivers both larger finite sample bias than $\hat{\theta}$ and worse asymptotic rates for the bias.

When the ad-hoc average block length $T/2$ is used, the performance of $\theta^{(S)}$ in terms of bias improves. Hence, $\theta^{(S)}$ appears to require a different tuning of the average block length than $T^{1/3}$, which is optimal for standard error estimation.

8 Discussion and further work

We have developed a novel and general framework for the reduction of the asymptotic bias of general *M*-estimators from asymptotically unbiased estimating functions. The framework relies on a novel approximation of the bias function that requires only the contributions to the estimating functions and the first two derivatives of those, which we used to derive an explicit and implicit method for reduced-bias *M*-estimation of general applicability. Explicit RB*M*-estimation proceeds by subtracting the approximation of the bias function at the *M*-estimates from the *M*-estimates. Implicit RB*M*-estimates result as roots of additively adjusted estimating functions, with the empirical bias-reducing adjustment (8). Both explicit and implicit RB*M*-estimates can be computed using the quasi Newton–Raphson iteration in online supplementary material, Section S4. The RB*M*-estimators have the same asymptotic distribution, and, hence, they are asymptotically as efficient as the initial *M*-estimators. As detailed in Section 4, uncertainty quantification can be carried out using the empirical estimate $\hat{\boldsymbol{V}}(\boldsymbol{\theta})$ of the variance-covariance matrix of that asymptotic distribution. The expression for $\hat{\boldsymbol{V}}(\boldsymbol{\theta})$ is part of expression (8) for the empirical bias-reducing adjustment, and, hence, is readily available at the last iteration of the quasi-Newton–Raphson iterative procedure. Inferences can be constructed using the Wald-type and generalised score pivots in expression (12).

If *M*-estimation is by the maximisation of an objective function, then implicit RB*M*-estimation can always be achieved by the maximisation of the bias-reducing penalised objective (13), which closely relates to model selection procedures using the Kullback–Leibler divergence; the functions of the parameters and the data that are used for bias reduction and model selection differ only by a known scalar constant. In particular, we show that bias reduction in estimation is closely related to model selection using TIC if estimation is via ML, and CLIC (Varin and Vidoni, 2005) if estimation is via maximum composite likelihood. Furthermore, TIC and CLIC are still consistent information criteria when evaluated at explicit or implicit RB*M*-estimates. The same justification we provided in Section 5.4 for the use of information criteria at the reduced-bias estimates can be used to justify the use of information criteria at estimates arising from the additive adjustment of estimating functions by alternative $O_p(1)$ quantities, like the median reduced-bias estimates discussed in Kenne Pagui et al. (2017) and Kosmidis et al. (2020), and the mean reduced-bias estimators in Firth (1993).

As shown in Section 5.3, the bias-reducing penalised objectives, can be further enriched with plug-in penalties of small asymptotic orders to return RB*M*p-estimates with enhanced properties. In Section 5.3, we used this fact to construct reduced-bias estimates that are always finite in binomial regression models, like logistic, probit, complementary log-log, and cauchit regression, and in autologistic regression models for clustered spatial data in Example 5.4. Note that, to the authors' knowledge, there is no proof that the reduced-bias estimator in Firth (1993) for binomial regression models with link functions other than logit always takes finite values; see, Kosmidis and Firth (2021) for a proof of the finiteness of the reduced-bias estimator of Firth (1993) in logistic regression when the model matrix is of full rank.

As we discussed in Section 7, when there is only $k = 1$ observation with correlated components, application-dependent conditions need to be used for the appropriate definition of $\boldsymbol{e}(\boldsymbol{\theta})$ in the penalised objective function in (13) or of $\boldsymbol{e}(\boldsymbol{\theta})$ and $\boldsymbol{d}_r(\boldsymbol{\theta})$ in the empirical bias-reducing adjustment (8). In the context of time series and spatial data that do not seriously depart from the condition of stationarity, one can consider window subsampling for the definition of $\boldsymbol{e}(\boldsymbol{\theta})$ and $\boldsymbol{d}_r(\boldsymbol{\theta})$ (see, for example Carlstein, 1986; Heagerty and Lumley, 2000 for definitions and guidance on the choice of the window size).

Our empirical experience with fully specified models is that while the RB*M*-estimators are first-order unbiased, they may not deliver an as strong bias correction in small samples as other methods in Table 1 that require full specification of the correct model. Example 5.3 and online supplementary material, Section S8 provide evidence about this effect.

From the simulation-based methods, the linear-bias-correcting (LBC) estimator of MacKinnon and Smith (1998) has interesting properties that may warrant further investigation. In particular, if the initial estimator $\hat{\theta}$ has a linear bias function, always takes values away from the boundary of the parameter space, and samples from the correct model or an appropriate empirical distribution function can be taken, then the LBC estimator of MacKinnon and Smith (1998) can deliver estimators with virtually zero bias. If the assumption of linear bias function is not satisfied, then the LBC estimator is free from first-order bias, just like RB*M*-estimators are. These properties of LBC come at the expense of the need to simulation from the full model and more computation. MacKinnon and Smith (1998, Section 6) (and Section 1.1) provide a clear discussion on why estimators based on asymptotic approximation of the bias (like RB*M*-estimators) are typically more attractive than simulation-based estimators in applications.

The differences between RB*M*-estimators and other bias-reduction methods that aim to remove the first term or more from the bias function in fully or partially specified models, are beyond the first term of the bias expansion. Hence, while a detailed theoretical assessment of those differences may be interesting for particular models, it is subtle, mathematically involved and potentially inconclusive at the level of generality RB*M* estimation has been developed.

Lunardon (2018) showed that bias reduction in ML estimation using the adjustments in Firth (1993) can be particularly effective for inference about a low-dimensional parameter of interest in the presence of high-dimensional nuisance parameters, while providing, at the same time, improved estimates of the nuisance parameters. Current research investigates the performance of the RB*M*-estimator for general *M*-estimation in stratified settings, extending the optimality results in Lunardon (2018) when maximum composite likelihood or other *M*-estimators are used.

The sufficient conditions we used for RB*M*-estimation do not necessarily cover for asymptotic regimes where $p/n \to \kappa \in (0, 1)$ (see, for example, Sur and Candès, 2019). RB*M*-estimators have been found, though, to deliver remarkable corrections in both bias and variance with high-dimensional covariate specifications in partially specified autologistic regression (see Example 6.1), and in fully specified negative binomial regression (see online supplementary material, Section S8). A formal assessment of the RB*M*-estimation framework in high-dimensional models is the subject of future work.

Acknowledgments

We thank David Firth and Elvezio Ronchetti for helpful discussions and comments on an earlier version of this work. Part of this work was completed when Nicola Lunardon was at the Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca. The authors acknowledge the DEMS Data Science Lab at the University of Milano-Bicocca for supporting this work by providing computational resources. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising from this submission.

Conflicts of interest: None declared.

Data availability

The Gambia malaria survey data (Thomson et al., 1999) is provided in the geoR R package (Ribeiro Jr et al., 2022)

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

References

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Albert A., & Anderson J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1–10. https://doi.org/10.1093/biomet/71.1.1
Bell B. (2021). *CppAD: A package for C++ algorithmic differentiation (20230812)*. https://cppad.readthedocs.io
Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 192–236. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x
Besag J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(3), 179–195. https://doi.org/10.2307/2987782
Besag J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(1), 75–83. https://doi.org/10.1111/j.2517-6161.1972.tb00889.x
Boos D. D. (1992). On generalized score tests. *The American Statistician*, *46*(4), 8. https://doi.org/10.2307/2685328
Canty A., & Ripley B. D. (2022). *Boot: Bootstrap R (S-Plus) functions* (R package version 1.3-28.1).
Caragea P. C., & Kaiser M. S. (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, *14*(3), 281–300. https://doi.org/10.1198/jabes.2009.07032
Carlstein E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, *14*(3), 1171–1179. https://doi.org/10.1214/aos/1176350057
Claeskens G., & Hjort N. L. (2008). *Model selection and model averaging*. Cambridge University Press.
Cordeiro G. M., & McCullagh P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 629–643. https://doi.org/10.1111/j.2517-6161.1991.tb01852.x
Davison A. C., & Gholamrezaee M. M. (2012). Geostatistics of extremes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *468*(2138), 581–608. https://doi.org/10.1098/rspa.2011.0412
Davison A. C., & Hinkley D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.
Durbin J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, *46*(3), 477–480. https://doi.org/10.1093/biomet/46.3-4.477
Efron B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, *3*(6), 1189–1242. https://doi.org/10.1214/aos/1176343282
Efron B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics.
Efron B., & Tibshirani R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
Firth D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27
Genton M. G., Ma Y., & Sang H. (2011). On the likelihood function of gaussian max-stable processes. *Biometrika*, *98*(2), 481–488. https://doi.org/10.1093/biomet/asr020
Gourieroux C., Monfort A., & Renault E. (1993). Indirect inference. *Journal of Applied Econometrics*, *8*(S1), S85–S118. https://doi.org/10.1002/jae.3950080507
Griewank A., & Walther A. (2008). *Evaluating derivatives: Principles and techniques of algorithmic differentiation* (2nd ed.). Society for Industrial and Applied Mathematics.
Grün B., Kosmidis I., & Zeileis A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, *48*(11), 1–25. https://doi.org/10.18637/jss.v048.i11
Guerrier S., Dupuis-Lozeron E., Ma Y., & Victoria-Feser M.-P. (2019). Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association*, *114*(525), 146–157. https://doi.org/10.1080/01621459.2017.1380031
Hall P., & Martin M. A. (1988). On bootstrap resampling and iteration. *Biometrika*, *75*(4), 661–671. https://doi.org/10.1093/biomet/75.4.661
Heagerty P. J., & Lumley T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, *95*(449), 197–211. https://doi.org/10.1080/01621459.2000.10473914
Hughes J., Haran M., & Caragea P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics*, *22*(7), 857–871. https://doi.org/10.1002/env.1102
Huser R., & Davison A. C. (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika*, *100*(2), 511–518. https://doi.org/10.1093/biomet/ass089
Kenne Pagui E. C., Salvan A., & Sartori N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, *104*(4), 923–938. https://doi.org/10.1093/biomet/asx046
Konis K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models* [Ph.D. thesis]. University of Oxford.
Kosmidis I. (2014). Bias in parametric estimation: Reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(3), 185–196. https://doi.org/10.1002/wics.1296
Kosmidis I. (2023). *brglm2: Bias reduction in generalized linear models* (R package version 0.9).

Kosmidis I., & Firth D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, *96*(4), 793–804. https://doi.org/10.1093/biomet/asp055
Kosmidis I., & Firth D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, *108*(1), 71–82. https://doi.org/10.1093/biomet/asaa052
Kosmidis I., Kenne Pagui E. C., & Sartori N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, *30*(1), 43–59. https://doi.org/10.1007/s11222-019-09860-6
Kosmidis I., & Lunardon N. (2022). *MEstimation.jl: Methods for M-estimation of statistical models* (Julia package version 0.2.0).
Kosmidis I., Schumacher D., & Schwendinger F. (2022). *detectseparation: Detect and check for separation and infinite maximum likelihood estimates* (R package version 0.3).
Kristensen K., Nielsen A., Berg C. W., Skaug H., & Bell B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*(5), 1–21. https://doi.org/10.18637/jss.v070.i05
Kuk A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(2), 395–407. https://doi.org/10.1111/j.2517-6161.1995.tb02035.x
Kunsch H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241. https://doi.org/10.1214/aos/1176347265
Liang K.-Y., & Zeger S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13
Lindsay B. G. (1988). Composite likelihood methods. In N. U. Prabhu (Ed.), *Statistical inference from stochastic processes: Vol. 80. Contemporary mathematics*. American Mathematical Society.
Lunardon N. (2018). On bias reduction and incidental parameters. *Biometrika*, *105*(1), 233–238. https://doi.org/10.1093/biomet/asx079
Lunardon N., & Scharfstein D. (2017). Comment on 'Small sample GEE estimation of regression parameters for longitudinal data'. *Statistics in Medicine*, *36*(22), 3596–3600. https://doi.org/10.1002/sim.7366
MacKinnon J. G., & Smith A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, *85*(2), 205–230. https://doi.org/10.1016/S0304-4076(97)00099-7
Mansournia M. A., Geroldinger A., Greenland S., & Heinze G. (2018). Separation in logistic regression: Causes, consequences, and control. *American Journal of Epidemiology*, *187*(4), 864–870. https://doi.org/10.1093/aje/kwx299
McCullagh P. (2018). *Tensor methods in statistics* (2nd ed.). Dover Publications.
Mogensen P. K., & Riseth A. N. (2018). Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, *3*(24), 615. https://doi.org/10.21105/joss.00615
Newey W. K., & Smith R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, *72*(1), 219–255. https://doi.org/10.1111/j.1468-0262.2004.00482.x
Pace L., & Salvan A. (1997). *Principles of statistical inference from a neo-Fisherian perspective*. World Scientific.
Padoan S. A., & Bevilacqua M. (2015). Analysis of random fields using CompRandFld. *Journal of Statistical Software*, *63*(9), 1–27. https://doi.org/10.18637/jss.v063.i09
Padoan S. A., Ribatet M., & Sisson S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, *105*(489), 263–277. https://doi.org/10.1198/jasa.2009.tm08577
Politis D. N., & Romano J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870
Pratt J. W. (1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, *76*(373), 103–106. https://doi.org/10.1080/01621459.1981.10477613
Quenouille M. H. (1949). Approximate tests of correlation in time-series 3. *Mathematical Proceedings of the Cambridge Philosophical Society*, *45*(3), 483–484. https://doi.org/10.1017/S0305004100025123
Quenouille M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*(3), 353–360. https://doi.org/10.1093/biomet/43.3-4.353
R Core Team (2023). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
Revels J., Lubin M., & Papamarkou T. (2016). 'Forward-mode automatic differentiation in Julia', arXiv, arXiv:1607.07892, preprint: not peer reviewed.
Ribeiro P. J., Jr., Diggle P., Christensen O., Schlather M., Bivand R., & Ripley B. (2022). *geoR: Analysis of geostatistical data* (R package version 1.9-2).
Stefanski L. A., & Boos D. D. (2002). The calculus of M-estimation. *The American Statistician*, *56*(1), 29–38. https://doi.org/10.1198/000313002753631330
Sur P., & Candès E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, *116*(29), 14516–14525. https://doi.org/10.1073/pnas.1810420116
Takeuchi K. (1976). Distribution of informational statistics and a criterion for model fitting [in Japanese]. *Suri-Kagaku*, *153*, 12–18.

Thomson M. C., Connor S. J., D'Alessandro U., Rowlingson B., Diggle P., Cresswell M., & Greenwood B. (1999). Predicting malaria infection in gambian children from satellite data and bed net use surveys: The importance of spatial correlation in the interpretation of results. *The American Journal of Tropical Medicine and Hygiene*, *61*(1), 2–8. https://doi.org/10.4269/ajtmh.1999.61.2

van der Vaart A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Varin C., Reid N., & Firth D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5–42. https://www.jstor.org/stable/24309261

Varin C., & Vidoni P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, *92*(3), 519–528. https://doi.org/10.1093/biomet/92.3.519

Wedderburn R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447. https://doi.org/10.1093/biomet/61.3.439

Wolters M. (2022). *Autologistic.jl* (Julia package version 0.5.1).

Wolters M. A. (2017). Better autologistic regression. *Frontiers in Applied Mathematics and Statistics*, *3*, 24. https://doi.org/10.3389/fams.2017.00024