Supplementary Material document for 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

August 10, 2023

S1 Description of the supporting computer code

The computer code used to reproduce the empirical analyses in the main text is in the directory `RBM_supporting_computer_code`, which can be downloaded from `http://www.ikosmidis.com/files/RBM_supporting_computer_code.zip`. That directory has six sub-directories. All references in the descriptions below are to figures, tables, and sections in main text.

- The `ratio` directory contains two R scripts for reproducing the results in Example 3.1 and Example 4.1 about the ratio of two means. The file `ratio.R` provides functions to compute the ratio estimator and conduct the simulation study, whereas `ratio_summaries.R` provides the code needed to produce Table 2 and Table 3.

- `mev` contains an R package and an R script to reproduce the simulation study in Example 5.2. The R package `PwMev` contains a C implementation of the pairwise log-likelihood function (16); the main function of the R package is `pwlik_mev`, for which documentation is provided. The file `mev_sim.R` contains the code used to run the simulations and produce Figure 1.

- The `glms` directory contains five R scripts for the probit regression simulation studies in Example 5.3 and Example 5.5. The file `probit_ms_functions.R` implements the bias-reducing penalized likelihood and provides other support functions. The code to run the simulation experiments in Example 5.3 and Example 5.5 is provided in the scripts `probit_bias_simulation.R` and `probit_ms_simulation.R`, respectively. Figure 2 and Figure 3 result from the scripts `probit_bias_summaries.R` and `probit_ms_summaries.R`, respectively.

- The `autologistic` directory provides the Gambia malaria survey data in `Gambia.csv`, which are as provided in the `geoR` R package. The `auto_symmetric_module.jl` Julia script implements M-, RBM-, and RBMp-estimation for general autologistic regression models. The `gambia_subsets_bootstrap.jl` and `gambia_simulation_subset2.jl` scripts can be used to reproduce all numerical results in Example 5.4.

*ioannis.kosmidis@warwick.ac.uk

The script `hd_simulation.jl` conducts the simulations described in Example 6.1. The R script `hd_simulation_results.R` collects the simulation results to produce Table 5.

- The `AR(1)` directory contains the R script `AR_simulation_and_summaries.R` and the R package `OLStsBR`. The former uses the latter to reproduce the simulation study in Example 7.1 and produce Table 6, Table S1, and Table S2.
- The `negbin` directory contains the R scripts that reproduce the simulation study in Section S8. The file `negbin-functions.R` implements the bias-reducing penalized likelihood and provides other support functions. The code to run the simulation experiments is in `negbin-simulation.R`, and `negbin-summaries.R` produces Figure S1. The other four R scripts with filenames starting with `arxiv` are used to define the simulation settings in Table S3.

The `images/` sub-directories appearing in some of the above directories are for saving intermediate results, while the scripts are running. All scripts require their directory to be set as the working directory in R or Julia, before running them.

S2 Assumptions

The assumptions we employ for the theoretical development in this work are listed below.

A1 Consistency: The M-estimator $\hat{\boldsymbol{\theta}}$ satisfies $\hat{\boldsymbol{\theta}} \xrightarrow{p} \bar{\boldsymbol{\theta}}$, where $\bar{\boldsymbol{\theta}}$ is such that $\mathrm{E}_G(\boldsymbol{\psi}^i) = \mathbf{0}_p$ for all $i \in \{1, \ldots, k\}$, with $\boldsymbol{\psi}^i = \boldsymbol{\psi}^i(\bar{\boldsymbol{\theta}})$ and $\mathrm{E}_G(\cdot)$ denoting expectation with respect to the unknown joint distribution function G. In particular, we assume that $\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}} = O_p(n^{-1/2})$, where n is a measure of information about $\boldsymbol{\theta}$.

A2 Local smoothness: The derivatives of $\psi^i_r(\boldsymbol{\theta})$ $(r = 1, \ldots, p)$ exist up to the 4th order in a neighbourhood $\mathcal{N}$ of $\bar{\boldsymbol{\theta}}$. In particular,

$$l_{R_a}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{\partial^{a-1} \psi^i_{r_1}(\boldsymbol{\theta})}{\partial \theta^{r_2} \cdots \partial \theta^{r_a}} \,,$$

exist for $\boldsymbol{\theta} \in \mathcal{N}$ and any set $R_a = \{r_1, \ldots, r_a\}$, with $r_j \in \{1, \ldots, p\}$ and $a \in \{1, \ldots, 5\}$, under the convention that $l_r(\boldsymbol{\theta}) = \sum_{i=1}^k \psi^i_r(\boldsymbol{\theta})$ and that the components of $\boldsymbol{\theta}$ are identified by superscripts.

A3 Asymptotic orders of centred estimating function derivatives:

$$H_{R_a} = l_{R_a} - \mu_{R_a} = O_p(n^{1/2}) \,,$$

where $\mu_{R_a} = \mathrm{E}_G(l_{R_a})$, and $l_{R_a} = l_{R_a}(\bar{\boldsymbol{\theta}})$ exist for $a \in \{1, \ldots, 5\}$. Unless otherwise stated, whenever the argument $\boldsymbol{\theta}$ is omitted from quantities that depend on it, as is the case in the right-hand side of (5), those quantities are understood as being evaluated at $\bar{\boldsymbol{\theta}}$.

A4 Asymptotic orders of joint central moments of estimating functions and their derivatives:

$$\nu_{R_{a_1}, S_{a_2}, \ldots, T_{a_b}} = \begin{cases} O(n^{(b-1)/2}) \,, & \text{if } b \text{ is odd} \\ O(n^{b/2}) \,, & \text{if } b \text{ is even} \end{cases} \,,$$

where $\nu_{R_{a_1}, S_{a_2}, \ldots, T_{a_b}} = \mathrm{E}_G(H_{R_{a_1}} H_{S_{a_2}} \cdots H_{T_{a_b}})$ are joint central moments of estimating functions and their derivatives, with $R_{a_1}, S_{a_2}, \ldots, T_{a_b}$ being subsets of $a_1, a_2, \ldots, a_b > 0$ integers.

A5 The matrix with elements μ_{rs} $(r, s = 1, \ldots, p)$ is invertible.

Below we provide an analysis of those assumptions.

Assumption A1 is a working assumption that we make about the unbiasedness of the estimating functions and the consistency of the M-estimators. Consistency can sometimes be shown to hold under weak assumptions about G and the asymptotic unbiasedness of the estimating functions; see, for example, van der Vaart (1998, Section 5.2) and Huber and Ronchetti (2009, Section 6.2) for theorems on the consistency of M-estimators. We assume that there is an index n, which is typically, but not necessarily, the number of observations, that measures the rate the information about the parameter $\boldsymbol{\theta}$ accumulates, and that the difference $\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$ is $O_p(n^{-1/2})$.

Assumption A2 allows taking a sufficient number of derivatives of the estimating functions at the unknown parameter value $\bar{\boldsymbol{\theta}}$ when constructing the stochastic Taylor expansions required for the derivation of the empirical bias-reducing adjustments to the estimating functions in Section 3. Such an assumption covers many well-used estimating functions, like the ones arising in quasi-likelihood estimation (Wedderburn, 1974), estimation using generalized estimating equations (Liang and Zeger, 1986), and ML and maximum composite likelihood estimation (Lindsay, 1988; Varin et al., 2011) for a wide range of models. The local smoothness assumption may not directly cover, though, settings where the estimating function or one of its first few derivatives are non-differentiable at particular points in the parameter space. Examples of this kind are the estimating functions for quantile regression and robust regression with Huber loss; see Koenker (2005) and Huber and Ronchetti (2009) for textbook-length expositions of topics in quantile and robust regression, respectively. Nevertheless, as shown in Section 3.3, RBM-estimation ends up requiring only the first two derivatives of the estimating functions, hence its scope of application may be much wider than what is prescribed by assumptions we used to develop it. This is the topic of future work.

Assumptions A3 and A4 ensure the existence of the expectations, under the underlying process G, of products of estimating functions and their derivatives, and that $\sqrt{n}$-asymptotic arguments are valid. In the special case of ML estimation, when the model is adequate, assumptions A3 and A4 can be derived directly from Assumption A1 and A2 using the exlog relations; see, for example, Pace and Salvan (1997, Section 9.2 and Table 9.1).

Assumption A5 is a technical assumption to ensure that the expectation of the Jacobian of the estimating function is invertible, when inverting the stochastic Taylor expansion of $\mathbf{0}_p = \sum_{i=1}^k \boldsymbol{\psi}^i(\tilde{\boldsymbol{\theta}}) + \boldsymbol{A}(\tilde{\boldsymbol{\theta}})$ about $\bar{\boldsymbol{\theta}}$, and is typically assumed for estimation using ML and estimating equations (see, for example, Boos and Stefanski, 2013, Section 7.7).

S3 Stochastic Taylor expansion for $\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$

Using assumptions A1-A3 and index notation, with the indices taking values in the set $\{1, \ldots, p\}$, a calculation similar to that in McCullagh (2018, Section 7.3) can be used to show that the expansion of $\mathbf{0}_p = \sum_{i=1}^k \boldsymbol{\psi}^i(\tilde{\boldsymbol{\theta}}) + \boldsymbol{A}(\tilde{\boldsymbol{\theta}})$ about $\bar{\boldsymbol{\theta}}$ results in a stochastic Taylor expansion for $\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$ of the form

$$
\begin{aligned}
\tilde{\theta}^r - \bar{\theta}^r = {} & H^r + H^a H^r_a + \frac{1}{2} H^a H^b \mu^r_{ab} + A^r + \\
& + H^a H^b_a H^r_b + \frac{1}{2} H^a H^b H^r_c \mu^c_{ab} + \frac{1}{2} H^a H^b H^c_b \mu^r_{ac} + \\
& + \frac{1}{2} H^a H^b H^c_a \mu^r_{cb} + \frac{1}{4} H^a H^b H^c \mu^d_{bc} \mu^r_{ad} + \frac{1}{4} H^a H^b H^c \mu^d_{ab} \mu^r_{dc} + \\
& + \frac{1}{2} H^a H^b H^r_{ab} + \frac{1}{6} H^a H^b H^c \mu^r_{abc} + \\
& + A^a H^r_a + \frac{1}{2} A^a H^b \mu^r_{ba} + \frac{1}{2} A^a H^b \mu^r_{ab} + A^r_a H^a + O_p(n^{-2}) \,,
\end{aligned}
\tag{S1}
$$

where $H^r_{r_1\ldots r_a} = -\mu^{rs} H_{sr_1\cdots r_a}$, $\mu^r_{r_1\cdots r_a} = -\mu^{rs}\mu_{sr_1\cdots r_a}$, and $A^r_{r_1\cdots r_a} = -\mu^{rs} A_{sr_1\cdots r_a}$, with μ^{rs} denoting the matrix inverse of μ_{rs} (assumption A5) and $A_{r_1\cdots r_a} = \partial^{a-1} A_{r_1}(\boldsymbol{\theta})/\partial\theta^{r_2}\cdots\partial\theta^{r_a}$.

S4 Quasi Newton-Raphson for RBM estimation

Apart from special cases, like the estimation of the ratio of two means in Example 3.1 and of the parameter of an AR(1) process in Example 7.1, and as is the case for general M-estimation, the solution of the adjusted estimating equations (4) is, typically, not available in closed form. General procedures for systems of nonlinear equations can be used to solve them.

A general iterative procedure of this kind results from a modification of the Newton-Raphson iteration that in the uth iteration updates the current estimate $\boldsymbol{\theta}^{(u)}$ to a new value $\boldsymbol{\theta}^{(u+1)}$ as

$$\boldsymbol{\theta}^{(u+1)} := \boldsymbol{\theta}^{(u)} + a_u \left\{ \boldsymbol{j}\left(\boldsymbol{\theta}^{(u)}\right)\right\}^{-1} \left\{ \sum_{i=1}^{k} \boldsymbol{\psi}^i\left(\boldsymbol{\theta}^{(u)}\right) + \boldsymbol{A}\left(\boldsymbol{\theta}^{(u)}\right)\right\}, \tag{S2}$$

where a_u is a deterministic sequence of positive constants that can be used to implement various schemes to further control the step size, like step-halving. Iteration (S2) defines a quasi Newton-Raphson procedure with the correct fixed point. The iteration is a relaxation of full Newton-Raphson iteration, which would have $a_u = 1$ and the matrix of derivatives of $\sum_{i=1}^k \boldsymbol{\psi}^i(\boldsymbol{\theta}) + \boldsymbol{A}(\boldsymbol{\theta})$ in the place of $\boldsymbol{j}(\boldsymbol{\theta})$. The M-estimates from the solution of $\sum_{i=1}^k \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \boldsymbol{0}_p$ are obvious starting values for the quasi Newton-Raphson procedure, and candidate stopping criteria include $|\boldsymbol{\theta}^{(u+1)} - \boldsymbol{\theta}^{(u)}|/a_u < \epsilon$ and $|| \sum_{i=1}^k \boldsymbol{\psi}^i\left(\boldsymbol{\theta}^{(u)}\right) + \boldsymbol{A}\left(\boldsymbol{\theta}^{(u)}\right) ||_1 < \epsilon$, for some $\epsilon > 0$, where $||\cdot||_1$ is the L1 norm.

Typically, quasi Newton-Raphson will have first-order convergence to the solution of the adjusted estimating equations, compared to the second-order convergence that full Newton-Raphson has. The advantage of using quasi Newton-Raphson instead of full Newton-Raphson is that all quantities required to implement (S2) are readily available once an implementation of the empirical bias-reducing adjustments is done.

The explicit RBM-estimator $\boldsymbol{\theta}^\dagger$ in (9) results as a by-product of the quasi Newton-Raphson procedure (S2). A single step of (S2) with $a_1 = 1$, starting at the M-estimator $\hat{\boldsymbol{\theta}}$ results in $\boldsymbol{\theta}^\dagger$. In fact, the quasi Newton-Raphson iteration (S2) reveals that implicit RBM-estimation can be understood as iterative bias correction, exactly as is the case for reduced-bias estimation in fully-specified models (see, for example, Kosmidis and Firth, 2010).

The fact that the empirical bias-reducing adjustment in (7) depends only on derivatives of estimating functions, enables general implementations by deriving the derivatives $\partial\psi^i_r(\boldsymbol{\theta})/\partial\theta^s$ and $\partial^2\psi^i_r(\boldsymbol{\theta})/\partial\theta^s\partial\theta^t$ $(r, s, t = 1, \ldots, p)$ either analytically or by using automatic differentiation techniques (Griewank and Walther, 2008). Those derivatives can be combined together to produce $\boldsymbol{u}_r(\boldsymbol{\theta})$, $\boldsymbol{j}(\boldsymbol{\theta})$, $\boldsymbol{e}(\boldsymbol{\theta})$ and $\boldsymbol{d}_r(\boldsymbol{\theta})$, and, then, matrix multiplication and a numerical routine for matrix inversion can be used for an easy, general implementation of (8).

For implementations using automatic differentiation, in particular, the only required input from the user is an appropriate implementation of the contributions $\boldsymbol{\psi}^i(\boldsymbol{\theta})$ to the estimating functions. The automatic differentiation routines will, then, produce implementations of the required first and second derivatives of the contributions. The `MEstimation` Julia package (`https://github.com/ikosmidis/MEstimation.jl`) provides a proof-of-concept of such an implementation.

S5 Bias-reducing penalized log-likelihoods for generalized linear models

For notational simplicity, the dependence of the various quantities on $\boldsymbol{\beta}$ and/or ϕ is suppressed.

In contrast to the bias-reduction methods proposed for generalized linear models in Kosmidis and Firth (2009), the empirical bias-reducing adjustment to the score function always corresponds to a penalty to the sum of the log-likelihood contributions (14) about $\boldsymbol{\beta}$ and ϕ. According to Section 5.1, if ϕ is unknown, the only ingredients required in the penalty are the observed information matrix about $\boldsymbol{\beta}$ and ϕ, $\boldsymbol{j}(\boldsymbol{\beta}, \phi)$, and the sum of the outer products of the gradient of (14) across observations, $\boldsymbol{e}(\boldsymbol{\beta}, \phi)$. The bias-reducing penalized log-likelihood is, then

$$\sum_{i=1}^{n} \log f_i(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) - \frac{1}{2}\text{trace}\left[\{\boldsymbol{j}(\boldsymbol{\beta}, \phi)\}^{-1} \boldsymbol{e}(\boldsymbol{\beta}, \phi)\right].$$

The closed-form expressions for $\boldsymbol{j}(\boldsymbol{\beta}, \phi)$ and $\boldsymbol{e}(\boldsymbol{\beta}, \phi)$ are

$$\boldsymbol{j} = \begin{bmatrix} \boldsymbol{j}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{j}_{\boldsymbol{\beta}\phi} \\ \boldsymbol{j}_{\boldsymbol{\beta}\phi}^\top & \boldsymbol{j}_{\phi\phi} \end{bmatrix} \quad \text{and} \quad \boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{e}_{\boldsymbol{\beta}\phi} \\ \boldsymbol{e}_{\boldsymbol{\beta}\phi}^\top & \boldsymbol{e}_{\phi\phi} \end{bmatrix},$$

where

$$\boldsymbol{j}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{1}{\phi}\boldsymbol{X}^\top \boldsymbol{Q}\boldsymbol{X}, \quad \boldsymbol{j}_{\phi\phi} = \frac{1}{\phi^3}\boldsymbol{1}_n^\top(\boldsymbol{R} - \boldsymbol{A}')\boldsymbol{1}_n + \frac{1}{2\phi^4}\boldsymbol{1}_n^\top \boldsymbol{A}''\boldsymbol{1}_n, \quad \boldsymbol{j}_{\boldsymbol{\beta}\phi} = \frac{1}{\phi^2}\boldsymbol{X}^\top \tilde{\boldsymbol{W}}\boldsymbol{1}_n,$$

$$\boldsymbol{e}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{1}{\phi^2}\boldsymbol{X}^\top \tilde{\boldsymbol{W}}^2\boldsymbol{X}, \quad \boldsymbol{e}_{\phi\phi} = \frac{1}{4\phi^4}\boldsymbol{1}_n^\top(\boldsymbol{R} - \boldsymbol{A}')^2\boldsymbol{1}_n, \quad \boldsymbol{e}_{\boldsymbol{\beta}\phi} = \frac{1}{2\phi^3}\boldsymbol{X}^\top \tilde{\boldsymbol{W}}(\boldsymbol{R} - \boldsymbol{A}')\boldsymbol{1}_n,$$

where $\boldsymbol{1}_n$ is a vector of n ones. The $n \times n$ diagonal matrices $\boldsymbol{R}$, $\boldsymbol{A}'$, $\boldsymbol{A}''$ have ith diagonal element $r_i = -2m_i(y_i\theta_i - \kappa(\theta_i) - c_1(y_i))$ (deviance residual), $a_i' = m_i a'(-m_1/\phi)$, $a_i'' = m_i^2 a''(-m_1/\phi)$, respectively, where $a'(u) = da(u)/du$, $a''(u) = d^2a(u)/du^2$. The $n \times n$ diagonal matrix $\boldsymbol{Q}$ and $\tilde{\boldsymbol{W}}$ have ith diagonal element $q_i = b_i d_i - b_i'(y_i - \mu_i)$ and $\tilde{w}_i = b_i(y_i - \mu_i)$, respectively, where $b_i = m_i d_i / v_i$, and $b_i' = m_i(d_i'/v_i - d_i^2 v_i'/v_i^2)$. In the latter expression, $d_i' = d^2\mu_i/d\eta_i^2$, $v_i' = dv_i/d\mu_i$,

S6 Expressions for the bias-reducing penalty for the pairwise likelihood of Padoan et al. (2010)

The joint density of the $Y_i(\boldsymbol{s}_l)$ and $Y_i(\boldsymbol{s}_m)$ $(l, m = 1, \ldots, L; l \neq m)$ is

$$f(y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m) \mid \boldsymbol{\theta}) = \exp\left\{-\frac{\Phi(w_{lm})}{y_i(\boldsymbol{s}_l)} - \frac{\Phi(v_{lm})}{y_i(\boldsymbol{s}_m)}\right\} \left[\frac{v_{lm}\phi(w_{lm})}{a_{lm}^2 y_i^2(\boldsymbol{s}_l) y_i(\boldsymbol{s}_m)} + \frac{w_{lm}\phi(v_{lm})}{a_{lm}^2 y_i^2(\boldsymbol{s}_m) y_i(\boldsymbol{s}_l)}\right. \tag{S3}$$
$$\left. + \left\{\frac{\Phi(w_{lm})}{y_i^2(\boldsymbol{s}_l)} + \frac{\phi(w_{lm})}{a_{lm} y_i^2(\boldsymbol{s}_l)} - \frac{\phi(v_{lm})}{a_{lm} y_i(\boldsymbol{s}_l) y_i(\boldsymbol{s}_m)}\right\}\left\{\frac{\Phi(v_{lm})}{y_i^2(\boldsymbol{s}_m)} + \frac{\phi(v_{lm})}{a_{lm} y_i^2(\boldsymbol{s}_m)} - \frac{\phi(w_{lm})}{a_{lm} y_i(\boldsymbol{s}_l) y_i(\boldsymbol{s}_m)}\right\}\right],$$

with $\Phi(\cdot)$ and $\phi(\cdot)$ the distribution and density function of the standard normal distribution, respectively. In the above expression, $a_{lm} = a_{lm}(\boldsymbol{\theta}) = \{(\boldsymbol{s}_l - \boldsymbol{s}_m)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})(\boldsymbol{s}_l - \boldsymbol{s}_m)\}^{1/2}$, $w_{lm} = w_{lm}(\boldsymbol{\theta}) = a_{lm}(\boldsymbol{\theta})/2 + \log\{y_i(\boldsymbol{s}_l)/y_i(\boldsymbol{s}_m)\}/a_{lm}(\boldsymbol{\theta})$, and $v_{lm} = v_{lm}(\boldsymbol{\theta}) = a_{lm}(\boldsymbol{\theta}) - w_{lm}(\boldsymbol{\theta})$.

We provide expressions for

$$l_t(\boldsymbol{\theta}; y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m)) = \frac{\partial \log f(y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m) | \boldsymbol{\theta})}{\partial \theta_t},$$

and

$$l_{tu}(\boldsymbol{\theta}; y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m)) = \frac{\partial^2 \log f(y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m) | \boldsymbol{\theta})}{\partial \theta_t \partial \theta_u},$$

for $t, u \in \{1, 2, 3\}$. These quantities are needed to form the entries of the matrices $\boldsymbol{e}(\boldsymbol{\theta})$ and $\boldsymbol{j}(\boldsymbol{\theta})$ that are required when constructing the bias-reducing penalty to the pairwise log-likelihood. Specifically, the (t, u)th elements of $\boldsymbol{j}(\boldsymbol{\theta})$ and $\boldsymbol{e}(\boldsymbol{\theta})$ are, respectively,

$$j_{tu}(\boldsymbol{\theta}) = -\sum_{i=1}^{k} \sum_{l > m} l_{tu}(\boldsymbol{\theta}; y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m)),$$

$$e_{tu}(\boldsymbol{\theta}) = -\sum_{i=1}^{k}\left[\sum_{l>m} l_t(\boldsymbol{\theta}; y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m))\right]\left[\sum_{l'>m'} l_u(\boldsymbol{\theta}; y_i(\boldsymbol{s}_{l'}), y_i(\boldsymbol{s}_{m'}))\right]^\top .$$

The logarithm of expression (S3) can be expressed as

$$\log f(y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m)|\boldsymbol{\theta}) = A_{lm}(\boldsymbol{\theta}) + B_{lm}(\boldsymbol{\theta}) + \log\{C_{lm}(\boldsymbol{\theta})D_{lm}(\boldsymbol{\theta}) + E_{lm}(\boldsymbol{\theta})\}, \qquad \text{(S4)}$$

where

$$\begin{aligned}
A_{lm}(\boldsymbol{\theta}) &= -\frac{\Phi(w_{lm}(\boldsymbol{\theta}))}{y_i(\boldsymbol{s}_l)}, \\
B_{lm}(\boldsymbol{\theta}) &= -\frac{\Phi(v_{lm}(\boldsymbol{\theta}))}{y_i(\boldsymbol{s}_m)}, \\
C_{lm}(\boldsymbol{\theta}) &= \frac{\Phi\{w_{lm}(\boldsymbol{\theta})\}}{y_i^2(\boldsymbol{s}_l)} + \frac{\phi\{w_{lm}(\boldsymbol{\theta})\}}{a_{lm}(\boldsymbol{\theta})y_i^2(\boldsymbol{s}_l)} - \frac{\phi\{v_{lm}(\boldsymbol{\theta})\}}{a_{lm}(\boldsymbol{\theta})y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)}, \\
D_{lm}(\boldsymbol{\theta}) &= \frac{\Phi\{v_{lm}(\boldsymbol{\theta})\}}{y_i^2(\boldsymbol{s}_m)} + \frac{\phi\{v_{lm}(\boldsymbol{\theta})\}}{a_{lm}(\boldsymbol{\theta})y_i^2(\boldsymbol{s}_m)} - \frac{\phi\{w_{lm}(\boldsymbol{\theta})\}}{a_{lm}(\boldsymbol{\theta})y_i(\boldsymbol{s}_m)y_i(\boldsymbol{s}_l)}, \\
E_{lm}(\boldsymbol{\theta}) &= \frac{v_{lm}(\boldsymbol{\theta})\phi\{w_{lm}(\boldsymbol{\theta})\}}{a_{lm}^2(\boldsymbol{\theta})y_i^2(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} + \frac{w_{lm}(\boldsymbol{\theta})\phi\{v_{lm}(\boldsymbol{\theta})\}}{a_{lm}^2(\boldsymbol{\theta})y_i(\boldsymbol{s}_l)y_i^2(\boldsymbol{s}_m)}.
\end{aligned}$$

In what follows, the dependence of the above quantities on $\boldsymbol{\theta}$, l, and m is omitted.

The first-order partial derivative of (S4) with respect to the component t of $\boldsymbol{\theta}$ is

$$l_t(\boldsymbol{\theta}; y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m)) = A_t + B_t + (CD+E)^{-1}(C_tD + CD_t + E_t),$$

where

$$\begin{aligned}
A_t &= \frac{\partial}{\partial\theta^t}A = -\frac{\phi(w)w_t}{y_i(\boldsymbol{s}_l)}, \\
B_t &= \frac{\partial}{\partial\theta^t}B = -\frac{\phi(v)v_t}{y_i(\boldsymbol{s}_m)}, \\
C_t &= \frac{\partial}{\partial\theta^t}C = \frac{\phi(w)w_t}{y_i^2(\boldsymbol{s}_l)} - \frac{w\phi(w)(w_ta - wa_t)}{a^2y_i^2(\boldsymbol{s}_l)} - \frac{v\phi(v)(v_ta - va_t)}{a^2y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)}, \\
D_t &= \frac{\partial}{\partial\theta^t}D = \frac{\phi(v)v_t}{y_i^2(\boldsymbol{s}_m)} - \frac{v\phi(v)(v_ta - va_t)}{a^2y_i^2(\boldsymbol{s}_m)} - \frac{w\phi(w)(w_ta - wa_t)}{a^2y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)}, \\
E_t &= \frac{\partial}{\partial\theta^t}E = \frac{\phi(w)\{(v_t - vww_t)a - 2va_t\}}{a^3y_i^2(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} + \frac{\phi(v)\{(w_t - wvv_t)a - 2wa_t\}}{a^3y_i(\boldsymbol{s}_l)y_i^2(\boldsymbol{s}_m)},
\end{aligned}$$

and

$$\begin{aligned}
a_t &= \frac{\partial}{\partial\theta^t}a = -\frac{1}{2a}\left\{(\boldsymbol{s}_l - \boldsymbol{s}_m)^\top\bar{\boldsymbol{\Sigma}}_t(\boldsymbol{s}_l - \boldsymbol{s}_m)\right\}, \\
\bar{\boldsymbol{\Sigma}}_t &= \frac{\partial}{\partial\theta^t}\boldsymbol{\Sigma}^{-1} = -\boldsymbol{\Sigma}^{-1}\left(\frac{\partial}{\partial\theta^t}\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}, \\
w_t &= \frac{\partial}{\partial\theta^t}w = \frac{a_t}{2} - \frac{a_t}{a^2}\log\{y_i(\boldsymbol{s}_l)/y_i(\boldsymbol{s}_m)\}, \\
v_t &= \frac{\partial}{\partial\theta^t}v = a_t - w_t.
\end{aligned}$$

The second-order partial derivative of (S4) with respect to the tth and uth component of θ is

$$\begin{aligned}
l_{tu}(\boldsymbol{\theta}; y_i(\boldsymbol{s}_l), y_i(\boldsymbol{s}_m)) &= A_{tu} + B_{tu} - (CD+E)^{-2}(C_uD + CD_u + E_u)(C_tD + CD_t + E_t) + \\
&\quad + (CD+E)^{-1}(C_{tu}D + C_tD_u + C_uD_t + CD_{tu} + E_{tu}),
\end{aligned}$$

where

$$
\begin{aligned}
A_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}A = -\frac{-w\phi(w)w_t w_u + \phi(w)w_{tu}}{y_i(\boldsymbol{s}_l)}\,,\\
B_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}B = -\frac{-v\phi(v)v_t v_u + \phi(v)v_{tu}}{a^2 y_i(\boldsymbol{s}_m)}\,,\\
C_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}C = \frac{-w\phi(w)w_u w_t + \phi(w)w_{tu}}{y_i^2(\boldsymbol{s}_l)} - \frac{w_u\phi(w)(w_t a - w a_t) - w^2\phi(w)w_u(w_t a - w a_t)}{a^2 y_i^2(\boldsymbol{s}_l)} +\\
&\quad -\frac{w\phi(w)(w_{tu}a + w_t a_u - w_u a_{tu} - w a_{tu})}{a^2 y_i^2(\boldsymbol{s}_l)} + \frac{2a_u w\phi(w)(w_t a - w a_t)}{a^3 y_i^2(\boldsymbol{s}_l)} +\\
&\quad -\frac{v_u\phi(v)(v_t a - v a_t) - v^2\phi(v)v_u(v_t a - v a_t)}{a^2 y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} +\\
&\quad -\frac{v\phi(v)(v_{tu}a + v_t a_u - v_u a_{tu} - v a_{tu})}{a^2 y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} + \frac{2a_u v\phi(v)(v_t a - v a_t)}{a^3 y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)}\,,
\end{aligned}
$$

$$
\begin{aligned}
D_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}D = \frac{-v\phi(v)v_u v_t + \phi(v)v_{tu}}{y_i^2(\boldsymbol{s}_m)} - \frac{v_u\phi(v)(v_t a - v a_t) - v^2\phi(v)v_u(v_t a - v a_t)}{a^2 y_i^2(\boldsymbol{s}_m)} +\\
&\quad -\frac{v\phi(v)(v_{tu}a + v_t a_u - v_u a_{tu} - v a_{tu})}{a^2 y_i^2(\boldsymbol{s}_m)} + \frac{2a_u v\phi(v)(v_t a - v a_t)}{a^3 y_i^2(\boldsymbol{s}_m)} +\\
&\quad -\frac{w_u\phi(w)(w_t a - w a_t) - w^2\phi(w)w_u(w_t a - w a_t)}{a^2 y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} +\\
&\quad -\frac{w\phi(w)(w_{tu}a + w_t a_u - w_u a_{tu} - w a_{tu})}{a^2 y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} + \frac{2a_u w\phi(w)(w_t a - w a_t)}{a^3 y_i(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)}\,,\\
E_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}E = \frac{v_{tu}\phi(w) - v_t w\phi(w)w_u - v_u w\phi(w)w_t - vw\phi(w)w_t w_u - vw\phi(w)w_{tu}}{a^4 y_i^2(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} +\\
&\quad -2a_u\frac{v_t\phi(w) - vw\phi(w)w_t}{a^3 y_i^2(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} - 2\frac{v_u\phi(w)a_t - vw\phi(w)w_u a_t + v\phi(w)a_{tu}}{a^6 y_i^2(\boldsymbol{s}_l)y_i(\boldsymbol{s}_m)} +\\
&\quad \frac{w_{tu}\phi(v) - w_t v\phi(v)v_u - w_u v\phi(v)v_t - wv\phi(v)v_t v_u - wv\phi(v)v_{tu}}{a^4 y_i(\boldsymbol{s}_l)y_i^2(\boldsymbol{s}_m)} +\\
&\quad -2a_u\frac{w_t\phi(v) - wv\phi(v)v_t}{a^3 y_i(\boldsymbol{s}_l)y_i^2(\boldsymbol{s}_m)} - 2\frac{w_u\phi(v)a_t - wv\phi(v)v_u a_t + w\phi(v)a_{tu}}{a^6 y_i(\boldsymbol{s}_l)y_i^2(\boldsymbol{s}_m)}\,,
\end{aligned}
$$

and

$$
\begin{aligned}
a_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}a = -\frac{a_u}{2a^2}\left\{(\boldsymbol{s}_l - \boldsymbol{s}_m)^\top\bar{\boldsymbol{\Sigma}}_t(\boldsymbol{s}_l - \boldsymbol{s}_m)\right\} + \frac{1}{2a}\left\{(\boldsymbol{s}_l - \boldsymbol{s}_m)^\top\bar{\boldsymbol{\Sigma}}_{tu}(\boldsymbol{s}_l - \boldsymbol{s}_m)\right\}\,,\\
\bar{\boldsymbol{\Sigma}}_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}\boldsymbol{\Sigma}^{-1} = -\left(\frac{\partial}{\partial\theta^t}\boldsymbol{\Sigma}^{-1}\right)\left(\frac{\partial}{\partial\theta^u}\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\left(\frac{\partial}{\partial\theta^t}\boldsymbol{\Sigma}\right)\left(\frac{\partial}{\partial\theta^u}\boldsymbol{\Sigma}^{-1}\right)\,,\\
w_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}w = \frac{a_{tu}}{2} - \frac{a_{tu}a - 2a_t a_u}{a^4}\log\{y_i(\boldsymbol{s}_l)/y_i(\boldsymbol{s}_m)\}\,,\\
v_{tu} &= \frac{\partial^2}{\partial\theta^t\partial\theta^u}v = a_{tu} - w_{tu}\,.
\end{aligned}
$$

S7 Additional simulation results for Example 7.1

Tables S1 and S2 provide simulation results like those in Table 6 of the main text, for $\theta \in \{0.2, 0.9\}$. Note that in Table S2 the figures are conditional on the ordinary least square estimate, $\hat{\theta}$, begin less than 1 in absolute value, i.e., the estimated autoregressive process is stationary.

Table S1: $\theta = 0.2$. Figures are reported in 2 decimal places, and the figures 0.00 and -0.00 are for estimated biases less than 0.0024 and -0.0024, respectively. The simulation error for the estimates of the bias is between 2.40×10^{-4} and 7.11×10^{-4}.

	α	Errors	T					Slope	
			50	100	200	400	800	Est	Exp
$\hat{\theta}$			-0.81	-0.39	-0.20	-0.08	-0.06	0.98	-1
$\tilde{\theta}$	1/3		-0.55	-0.20	-0.09	-0.01	-0.03	-1.27	$-3/2$
	1/2		-0.47	-0.16	-0.07	0.00	-0.02	-1.64	$-3/2$
$\theta^{\dagger}$	1/3		-0.53	-0.19	-0.08	-0.01	-0.03	-1.25	$-3/2$
	1/2		-0.42	-0.13	-0.05	0.01	-0.02	-1.25	$-3/2$
$\theta^{(J)}$			-0.11	-0.07	-0.04	0.01	-0.00	-1.64	< -1
$\theta^{(M)}$			-2.18	-1.12	-0.59	-0.27	-0.16	-0.96	< -1
$\theta^{(S)}$	1/3		3.97	3.65	3.20	2.84	2.20	-0.21	< -1
	$\log(T/2)/\log(T)$		0.48	0.31	0.17	0.11	0.03	-0.95	< -1
θ^{*}		Normal	0.05	0.03	-0.04	0.01	-0.00	-1.64	-2
		Student-t	0.78	0.32	0.14	0.09	0.02	-1.24	
		Laplace	-4.56	-2.81	-1.79	-1.19	-0.80	-0.63	

Table S2: $\theta = 0.9$. Figures are reported in 2 decimal places, and the figure 0.00 is for estimated bias of -0.0024. The simulation error for the estimates of the bias is between 1.0×10^{-4} and 7.41×10^{-3}.

	α	Errors	T					Slope	
			50	100	200	400	800	Est	Exp
$\hat{\theta}$			-3.22	-1.70	-0.87	-0.44	-0.22	-0.97	-1
$\tilde{\theta}$	1/3		-2.94	-1.44	-0.68	-0.31	-0.13	-1.21	$-3/2$
	1/2		-2.66	-1.21	-0.52	-0.19	-0.07	-1.32	$-3/2$
$\theta^{\dagger}$	1/3		-2.91	-1.42	-0.67	-0.31	-0.13	-1.12	$-3/2$
	1/2		-2.54	-1.13	-0.47	-0.17	-0.06	-1.35	$-3/2$
$\theta^{(J)}$			-0.68	-0.19	-0.06	-0.01	0.00	-2.11	< -1
$\theta^{(M)}$			-2.14	-1.22	-0.70	-0.39	-0.21	-0.83	< -1
$\theta^{(S)}$	1/3		15.14	14.71	13.56	12.26	9.79	-0.15	< -1
	$\log(T/2)/\log(T)$		1.93	1.24	0.70	0.39	0.20	-0.82	< -1
θ^{*}		Normal	0.60	0.56	0.12	0.03	0.01	-1.60	-2
		Student-t	-0.18	1.07	0.95	0.43	0.18	-0.13	
		Laplace	-18.34	-12.59	-8.49	-5.83	-3.82	-0.56	

Whenever the estimated process is non stationary with $\hat{\theta} > 1$, it is not possible to generate stationary bootstrap series for the computation of the estimator $\theta^{(M)}$. The estimated non-

Table S3: Simulation settings for the negative binomial model in Section S8. The settings are as in Guerrier et al. (2020, Section 6) (2010.13867v2, §6) and Zhang et al. (2022, Section J.4 in Supplementary Materials) (2204.07907v1, §J.4).

	2010.13867v2, §6	2204.07907v1, §J.4	2204.07907v1, §J.4	2204.07907v1, §J.4
n	100	200	400	800
p	20	41	51	61
$\boldsymbol{x}_i$	$x_{i1} = 1$ $x_{i2} \sim N(0, 1)$ $x_{i3} = I(i > 50)$ $x_{ij} \sim N(0, 4/25)$ $(j = 4, \ldots, 20)$	$x_{i1} = 1$ $x_{ij} \sim N(0, 1/5)$ $(j = 2, \ldots, 41)$	$x_{i1} = 1$ $x_{ij} \sim N(0, 4/25)$ $(j = 2, \ldots, 51)$	$x_{i1} = 1$ $x_{ij} \sim N(0, 2/15)$ $(j = 2, \ldots, 61)$
β_1	1.5	2	2	2
β_2	2.5	1	1	1
β_3	-2.5	-1	-1	-1
β_4	0	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_p	0	0	0	0
κ	0.6	0.7	0.7	0.7

stationary processes are 497 (out of $250 \times 2^{16}/50$) for $T = 50$, 1 (out of $250 \times 2^{16}/100$) for $T = 100$, and 0 for other sample sizes.

S8 Negative binomial regression

The performance of the explicit and implicit RBM estimators is assessed here in the context of negative binomial regression with many covariates using the experiments in Guerrier et al. (2020, Section 6) and Zhang et al. (2022, Section J.4 in Supplementary Materials). Both Guerrier et al. (2020) and Zhang et al. (2022) are unpublished preprints at the time of writing the current Supplementary Material document, and, hence, they are subject to change or become unavailable. For this reason, in what follows, we fully describe the simulation settings we consider from those preprints.

Suppose that $y_1, \ldots, y_n$ are realizations of $Y_1, \ldots, Y_n$, which are assumed to be conditionally independent given covariates $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$. Assume that $Y_i \mid \boldsymbol{x}_i$ is distributed according to a negative binomial distribution with probability mass function

$$f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, \kappa) = \frac{\Gamma(y_i + \kappa^{-1})}{\Gamma(y_i + 1)\Gamma(\kappa^{-1})} \left(\frac{\kappa^{-1}}{\kappa^{-1} + \mu_i} \right)^{\kappa^{-1}} \left(\frac{\mu_i}{\kappa^{-1} + \mu_i} \right)^{y_i} ,$$

where $\mu_i = \mathrm{E}(Y_i | \boldsymbol{x}_i; \boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. The variance of $Y_i \mid \boldsymbol{x}_i$ is $\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \mu_i + \kappa \mu_i^2$, and κ is an overdispersion parameter. When $\kappa \to 0$, the distribution of $Y_i \mid \boldsymbol{x}_i$ converges to the Poisson distribution. So, negative binomial regression can be viewed as a fully-parametric extension to Poisson regression with log link that accounts for over-dispersion.

The simulation settings we consider are as shown in Table S3, and are exactly as in Guerrier et al. (2020, Section 6) and Zhang et al. (2022, Section J.4 in Supplementary Materials). All covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ have entries that are generated independently from each other

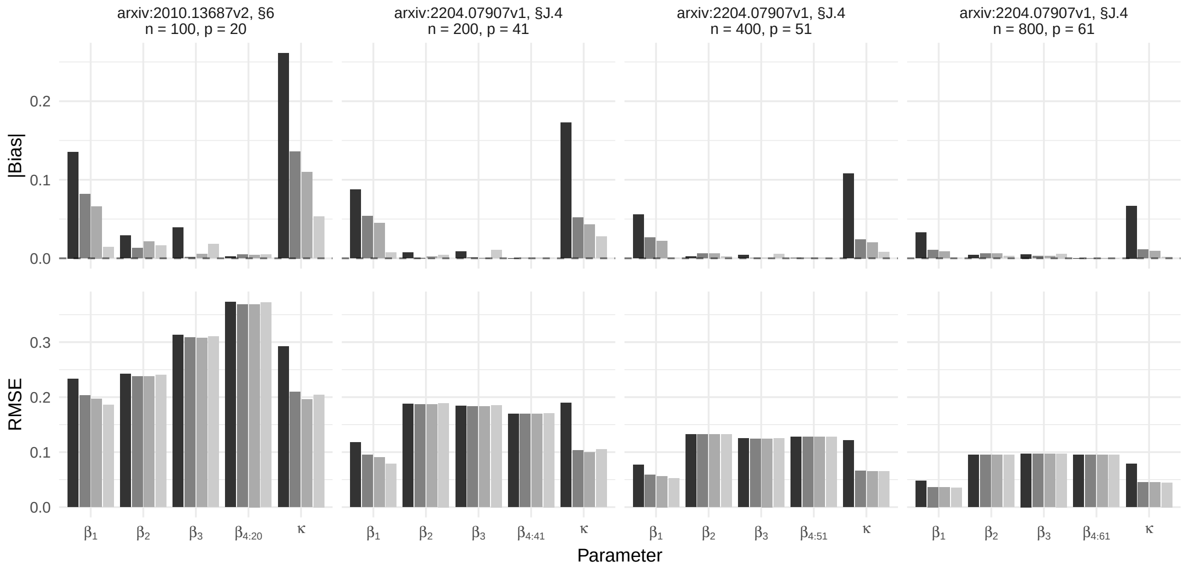

Figure S1: Absolute bias (|Bias|) and empirical root mean squared error (RMSE) of various estimators of $\boldsymbol{\beta}$ and κ for the simulation settings in Table S3 of Section S8. Results are shown, from darker to lighter grey, for the ML estimator, the implicit RBM-estimator, the explicit RBM-estimator, and the adjusted score functions estimator of Firth (1993).

according to Table S3. The covariate vectors are generated once and held fixed across the 1 000 simulations of the response vector $(y_1, \dots, y_n)^\top$ at each of the four sets of values for $\boldsymbol{\beta}$ and κ.

For each sample, we estimate $\boldsymbol{\beta}$ and κ using ML, as implemented in the `glm.nb` function from the `MASS` R package (Venables and Ripley, 2002), the adjusted score functions approach in Firth (1993), as implemented in the `brnb` function from the `brglm2` R package (Kosmidis, 2023), and explicit and implicit RBM-estimation.

To our knowledge there is no formal way to date that can detect whether the ML estimate of the negative binomial regression has elements on the boundary of the parameter space, which includes $\kappa = 0$ and/or $|\beta_j| = \infty$ for at least one $j \in \{1, \dots, p\}$. For this reason, boundary estimates where declared in an ad-hoc way, by checking if either the estimate of κ is smaller than 10^{-3} or the estimated standard error for β_j was larger than 100 ($j = 1, \dots, p$). In our simulation studies, this only happened for the setting 2010.13867v2, §6 in Table S3. There were $3, 7, 5, 4$ samples out of a 1000, where at least one of the components of ML, the adjusted score functions approach in Firth (1993), and the explicit and implicit RBM-estimation, respectively, were declared as being on the boundary.

Figure S1 shows estimates of the absolute bias and root mean squared error of the four estimators. Similarly to Example 5.3 of the main text, the summaries are computed after removing the samples where estimates have been declared as being on the boundary. As is apparent, the adjusted scores approach of Firth (1993), and explicit and implicit RBM-estimation result in estimators with substantially smaller bias and mean squared error than the ML estimator. In all cases, the reduction in bias is substantial in the estimation of β_1 and κ. The adjusted scores approach of Firth (1993), which relies on expectations of products of log-likelihood derivatives with respect to the correct model, is able to almost completely remove finite sample bias in all four simulation settings of Table S3. As in Example 5.3 of the main text, the differences between the various reduced-bias estimators in terms of bias and root mean squared error diminish fast as the sample size increases.

The arguments in Section 5.3 can be extended to develop a composite plug-in penalty that returns implicit RBM-estimators, which apart from components away from the boundary, also have bias that is free from the first-order term. The penalty can consist of a term that diverges to $-\infty$ as $\kappa \to 0$, and, as the results in Joshi et al. (2022) on Poisson regression suggest, a scaled-version of Jeffreys' invariant prior.

References

Boos, D. D. and L. A. Stefanski (2013). *Essential Statistical Inference.* New York, NY: Springer.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Griewank, A. and A. Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation* (2nd ed.). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Guerrier, S., M. Karemera, S. Orso, M.-P. Victoria-Feser, and Y. Zhang (2020). A general approach for simulation-based bias correction in high dimensional settings. *arXiv e-prints*, arXiv:2010.13687.

Huber, P. J. and E. Ronchetti (2009). *Robust Statistics* (2nd ed.). Hoboken, NJ: Wiley.

Joshi, A., A. Geroldinger, L. Jiricka, P. Senchaudhuri, C. Corcoran, and G. Heinze (2022). Solutions to problems of nonexistence of parameter estimates and sparse data bias in poisson regression. *Statistical Methods in Medical Research 31*(2), 253–266.

Koenker, R. (2005). *Quantile Regression.* Cambridge, NY: Cambridge University Press.

Kosmidis, I. (2023). *brglm2: Bias Reduction in Generalized Linear Models.* R package version 0.9.

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika 96*(4), 793–804.

Kosmidis, I. and D. Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics 4*, 1097–1112.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lindsay, B. G. (1988). Composite likelihood methods. In N. U. Prabhu (Ed.), *Statistical Inference from Stochastic Processes*, Volume 80 of *Contemporary Mathematics.* American Mathematical Society.

McCullagh, P. (2018). *Tensor methods in statistics* (2nd ed.). Mineola, NY: Dover Publications.

Pace, L. and A. Salvan (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective.* London, UK: World Scientific.

Padoan, S. A., M. Ribatet, and S. A. Sisson (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association 105*(489), 263–277.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge, UK: Cambridge University Press.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika 61*(3), 439–447.

Zhang, Y., Y. Ma, S. Orso, M. Karemera, M.-P. Victoria-Feser, and S. Guerrier (2022). A flexible bias correction method based on inconsistent estimators. *arXiv e-prints*, arXiv:2204.07907.