Bias approximations for likelihood-based estimators

R. C. Weng[1], D. S. Coad[2]

[1]Department of Statistics, National Chengchi University, Taipei, Taiwan

[2]School of Mathematical Sciences, Queen Mary, University of London, U.K.

Abstract

Bias approximation has played an important rôle in statistical inference, and numerous bias calculation techniques have been proposed under different contexts. We provide a unified approach to approximating the bias of the maximum likelihood estimator and the l_2 penalized likelihood estimator for both linear and nonlinear models, where the design variables are allowed to be random and the sample size can be a stopping time. The proposed method is based on the Woodroofe-Stein identity and is justified by very weak approximations. The accuracy of the derived bias formulas is assessed by simulation for several examples. The bias of the ridge estimator in high-dimensional settings is also discussed.

Key words: Bias calculation; l_2 penalized likelihood; Maximum likelihood estimation; Stopping time; Very weak approximation; Woodroofe-Stein identity.

1 Introduction

The study of bias has a long history and is essential for establishing statistical properties of an estimator. It is known that maximum likelihood estimators are biased when the sample size is small or moderate. To the best of our knowledge, Bartlett (1953) was the first to give an expression for the bias to order n^{-1} of the maximum likelihood estimator in the one-parameter case. The bias in multiparameter cases of independent observations was given in Cox and Snell (1968). In subsequent work, Schaefer (1983) considered the bias correction for logistic regression, Cordeiro and McCullagh (1991) obtained the bias correction in generalized linear models, and Firth (1993) proposed a bias reduction method by modifying the score function; see also Anderson and Richardson (1979), Shenton and Bowman (1977), McLachlan (1980), among others.

Most of the above work assumes that the observations are independent. Many have extended the results to dependent cases. Cordeiro and Klein (1994) showed that the independence is not required for the results in Cox and Snell (1968), and derived the bias formula for autoregressive moving average (ARMA) models. Bao and Ullah (2007) provide a general framework to obtain the properties of a large class of estimators in linear and nonlinear time series models and they are valid for both normal and nonnormal samples of observations, and where the regressors are stochastic. The derivations rely on the assumptions in Rilstone, Srivastava, and Ullah (1996), along with the consistency of the estimators; see also Rilstone and Ullah (2005). As an application of their results, Bao and Ullah (2007) develop the approximate bias and mean square error for some time series models, such as the first-order AR and MA models, and the absolute autoregressive model. Yang (2015) proposes a hybrid approach that combines the stochastic expansion of Bao and Ullah (2007) and the bootstrap, and applies the approach to the spatial autoregressive model. A general result is derived by Bao (2018) for the approximate bias of the quasi maximum likelihood estimators in ARMA models when exogenous regressors may be included.

Maximum likelihood estimators can also be severely biased in adaptively designed models, where the design variables may depend on the previous responses. Adaptive designs have been heavily used in applications such as clinical trials. Whitehead (1986) and Todd, Whitehead, and Facey (1996) derived bias-adjusted estimators following sequential tests. In related work, Coad and Woodroofe (1998) derived the bias approximation for adaptive linear models by differentiating the fundamental identity of sequential analysis; similar techniques have been applied to a one-parameter exponential family by Woodroofe (1990) and Coad (1994).

Another widely recognized biased estimator is the penalized likelihood estimator, such as the ridge estimator (Hoerl and Kennard, 1970). These estimators allow some bias for a reduction in variance. They are also called regularized estimators and widely used in the machine learning area. Interest in estimating the bias of penalized likelihood estimators has recently arisen in statistical inference. For example, for high-dimensional linear models with fixed design matrices, Shao and Deng (2012) studied the estimation of the ridge estimator and Bühlmann (2013) proposed a bias correction term for the ridge estimator when constructing p-values; Zhao and Shojaie (2016) extended the ridge test in Bühlmann (2013) to a scenario with random design matrices; Javanmard and Montanari (2014) proposed a computational procedure to construct a de-biased estimator and form confidence intervals for lasso regression, among others.

An important feature that distinguishes the aforementioned Coad and Woodroofe (1998) from other bias approximation techniques is its use of very weak approximations. The idea of such approximations originated from Stein (1985) and Woodroofe (1989) in the study of coverage probabilities for confidence sets. Specifically, the widely used adaptive designs in the clinical trial area often involve some form of stopping time, but the asymptotic expansions for the distributions of randomly stopped sums can be quite complicated; see, for example, Woodroofe (1986). Therefore, in situations where it is difficult to determine $\mathcal{C}$ for which $P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{C}) = \alpha$, they considered $P_{\xi}(\boldsymbol{\theta} \in \mathcal{C}) = \alpha$ and argued that if the latter holds for a large class of priors ξ, then the former may hold provided that the coverage probability depends on $\boldsymbol{\theta}$ smoothly. Here, $P_{\boldsymbol{\theta}}$ denotes the probability distribution given parameter $\boldsymbol{\theta} \in \Omega$, the parameter space, and $P_{\xi}(\boldsymbol{\theta} \in \mathcal{C}) = \int_{\Omega} P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{C})\xi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Let $E_{\boldsymbol{\theta}}$ and E_{ξ} denote the expectations with respect to $P_{\boldsymbol{\theta}}$ and P_{ξ}. Coad and Woodroofe (1998) adopted the concept of very weak justification for bias approximations by considering $E_{\xi}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ in place of $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, where $\hat{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}$. However, the approach based on the fundamental identity of sequential analysis is not easily applicable to nonlinear models. The formal formulation of very weak approximations is in Section 2.2.

This paper aims to provide a unified approach to approximating the bias of the maximum likelihood estimator and the l_2 penalized likelihood estimator for both linear and nonlinear models, with either fixed or random design variables and the sample size is a stopping time. The approximations here are very weak ones, as in Coad and Woodroofe (1998), but our evaluation of $E_\xi(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is based on the Woodroofe-Stein identity and a Taylor series expansion. The bias formulas that we derive are quite general.

The organization of the paper is as follows. In Section 2, the Woodroofe-Stein identity and very weak approximations are reviewed. The general bias formulas are presented in Section 3. The specialization to a variety of models is in Section 4, including a normal model with a stopping time, the AR(1) model with t distributed errors, the dilution model, the l_2 penalized logistic model and ridge regression. In Section 5, the bias of the ridge estimator in the large p small n setting is discussed. Numerical studies are provided in Section 6. Advantages of the approach are highlighted in Section 7, together with an indication of some further possible applications and discussion of how it may be extended to obtain the covariance matrix of the estimators. Proofs of the main results are given in the Appendix.

2 Review

2.1 Woodroofe-Stein identity

The Woodroofe-Stein identity is closely related to the famous Stein's lemma (Stein, 1981). Stein's lemma is concerned with the expectation with respect to a normal distribution, which is well known for its applications to the James-Stein estimator (James and Stein, 1961). By considering the expectation with respect to a probability density of the form in (1) below, Woodroofe (1989) developed a variant of Stein's identity and applied it to set corrected confidence sets following sequential experiments. Weng and Lin (2011) called this identity the Woodroofe-Stein identity and used it to obtain a Bayesian online ranking algorithm, which is comparable to the state-of-the-art algorithm TrueSkill$^{\text{TM}}$ developed by Microsoft Research. As the identity involves complex notation, here we only sketch results necessary for bias calculation. For a complete account of the identity, we refer readers to Woodroofe and Coad (1997, Proposition 2).

In what follows, the density and distribution function of a p-variate standard normal variate are denoted by $\phi_p(\boldsymbol{z})$ and $\Phi_p(\boldsymbol{z})$, respectively. We omit the subscript for the case

$p = 1$. Throughout, let $\nabla h(\cdot)$ denote the gradient of a function h with respect to its argument, and $\nabla^2 h(\cdot)$ denote the Hessian matrix of h. Suppose that $\boldsymbol{Z}$ is a p-dimensional random vector whose density takes the form

$$p(\boldsymbol{z}) = C\phi_p(\boldsymbol{z})f(\boldsymbol{z}), \tag{1}$$

where C is the normalizing constant and f is continuously differentiable. Then an application of the Woodroofe-Stein identity gives

$$\begin{aligned} E(\boldsymbol{Z}) &= E\left\{\frac{\nabla f(\boldsymbol{Z})}{f(\boldsymbol{Z})}\right\}, & (2)\\ E(\boldsymbol{Z}\boldsymbol{Z}^T) &= I_p + E\left\{\frac{\nabla^2 f(\boldsymbol{Z})}{f(\boldsymbol{Z})}\right\}, & (3)\end{aligned}$$

provided that the components of $\nabla f(\boldsymbol{z})$ are continuously differentiable and that the expectations on both sides exist. Here, I_p denotes the $p \times p$ identity matrix. Note that, although (1) involves the normal density, the above results are not restricted to the normal distribution. This is because any density $p(\boldsymbol{z})$ can be written as $p(\boldsymbol{z}) = \phi_p(\boldsymbol{z}) \cdot \{p(\boldsymbol{z})/\phi_p(\boldsymbol{z})\}$, and the above results hold if $p(\boldsymbol{z})/\phi_p(\boldsymbol{z})$ is continuously differentiable.

The above result has a close connection with Bayesian inference. Suppose that $y_k \sim p_{\boldsymbol{\theta}}(\cdot; \boldsymbol{x}_k)$ for $k = 1, 2, \ldots$, where $\boldsymbol{x}_k \in R^q$ is a vector of adaptive design variables and $\boldsymbol{\theta}$ is a parameter with $\boldsymbol{\theta} \in \Omega$, an open subset of R^p. Assume that the log-likelihood function $\ell_n(\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator satisfying $\nabla \ell_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Assume further that the Hessian matrix $-\nabla^2 \ell_n(\hat{\boldsymbol{\theta}}_n)$ is positive definite. Now define a $p \times p$ matrix B_n and an approximate pivot $\boldsymbol{Z}_n$ as

$$B_n^T B_n = -\nabla^2 \ell_n(\hat{\boldsymbol{\theta}}_n), \tag{4}$$

$$\boldsymbol{Z}_n = B_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n). \tag{5}$$

Consider a Bayesian model in which $\boldsymbol{\theta}$ has a prior density ξ. So the posterior density of $\boldsymbol{Z}_n$ given the data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ is

$$p_\xi^n(\boldsymbol{z}_n) \propto \phi_p(\boldsymbol{z}_n)f(\boldsymbol{z}_n), \tag{6}$$

where $f(\boldsymbol{z}_n) = \xi(\boldsymbol{\theta}(\boldsymbol{z}_n))\exp\{r_n(\boldsymbol{\theta}(\boldsymbol{z}_n))\}$ and

$$r_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) - \ell_n(\hat{\boldsymbol{\theta}}_n) + \frac{1}{2}\|\boldsymbol{z}_n\|^2. \tag{7}$$

Thus, the posterior density is of the form in (1). Expectation in this model is denoted by E_ξ and conditional expectation given $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ is denoted by E^n_ξ. Hence, (2) becomes an expression for the posterior mean:

$$E^n_\xi(\boldsymbol{Z}_n) = E^n_\xi\left\{\frac{\nabla f(\boldsymbol{Z}_n)}{f(\boldsymbol{Z}_n)}\right\}. \tag{8}$$

Note that, to apply the Woodroofe-Stein identity, there are different ways to define $\boldsymbol{Z}_n$. For example, Coad and Woodroofe (1996) considered the signed root transformation.

2.2 Very weak approximation

Let $o_p(1)$ denote convergence in $P_{\boldsymbol{\theta}}$-probability and $o(1)$ denote convergence to zero of a sequence of real numbers. Let $\boldsymbol{Z}_n$ be as in (5) and h be a real-valued function defined on R^p. Suppose that $w(\boldsymbol{\theta})$ satisfies

$$\int_\Omega \left[E_{\boldsymbol{\theta}}\{h(\boldsymbol{Z}_n)\} - w(\boldsymbol{\theta})\right] \xi(\boldsymbol{\theta}) d\boldsymbol{\theta} = o(n^{-1}) \tag{9}$$

for a class of prior densities ξ. Woodroofe (1989) calls approximations of the form (9) "very weak" and writes

$$E_{\boldsymbol{\theta}}\{h(\boldsymbol{Z}_n)\} = w(\boldsymbol{\theta}) + o(n^{-1}), \text{ very weakly, or } E_{\boldsymbol{\theta}}\{h(\boldsymbol{Z}_n)\} \simeq w(\boldsymbol{\theta}).$$

The term "very weak" comes from the fact that, if (9) holds for a class of priors, then it can be regarded as a form of weak convergence. For bias approximation, the very weak justification seeks to find $\boldsymbol{b}(\boldsymbol{\theta})$ for which

$$E_\xi(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \int_\Omega E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\xi(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_\Omega \boldsymbol{b}(\boldsymbol{\theta})\xi(\boldsymbol{\theta})d\boldsymbol{\theta} + o(n^{-1})$$

for a wide class of priors. The above line is written as $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \simeq \boldsymbol{b}(\boldsymbol{\theta})$. The condition that we impose on ξ is that it is continuously differentiable with a compact support. By letting ξ be highly concentrated around $\boldsymbol{\theta}$, and assuming that both $\boldsymbol{b}(\boldsymbol{\theta})$ and the bias depend on $\boldsymbol{\theta}$ smoothly, it may be possible to deduce that $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \boldsymbol{b}(\boldsymbol{\theta}) + o(n^{-1})$ for fixed $\boldsymbol{\theta}$ under some regularity conditions. From this, we see that our bias formula does not depend on a specific prior.

Indeed, the standard approximation that gives $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n)$ is a stronger result than the "very weak" approximation here. Although the very weak approximation may yield the standard one provided that both the bias and $\boldsymbol{b}(\boldsymbol{\theta})$ depend on $\boldsymbol{\theta}$ smoothly, in situations where the bias or $\boldsymbol{b}(\boldsymbol{\theta})$ is not smooth, the very weak approximation would not work. For example, for the AR(1) model $y_k = \theta y_{k-1} + e_k$, it is known that the bias is not smooth at $\theta = 1$; therefore, the very weak approximation would fail.

3 Bias calculation

3.1 Preliminaries

To evaluate $E_\xi(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, the idea of our approach is to use $\boldsymbol{Z}_n$ in (5) and write

$$E_\xi(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) = E_\xi(B_n^{-1}\boldsymbol{Z}_n) = E_\xi\{B_n^{-1}E_\xi^n(\boldsymbol{Z}_n)\}, \tag{10}$$

and then express the posterior mean $E_\xi^n(\boldsymbol{Z}_n)$ by means of the Woodroofe-Stein identity. Specifically, by (6)-(8), the posterior mean can be written as

$$E_\xi^n(\boldsymbol{Z}_n) = (B_n^T)^{-1}E_\xi^n\left\{\frac{\nabla\xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})} + \nabla r_n(\boldsymbol{\theta})\right\}, \tag{11}$$

where the gradient of $r_n(\boldsymbol{\theta})$ can be derived from (4), (5) and (7) as

$$\nabla r_n(\boldsymbol{\theta}) = \nabla\ell_n(\boldsymbol{\theta}) - \nabla^2\ell_n(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n). \tag{12}$$

To proceed further, some notation is needed. Let $\ell^{(3)}_{n,i_1i_2i_3}(\boldsymbol{\theta})$ denote the third partial derivative of $\ell_n(\boldsymbol{\theta})$ with respect to $\theta_{i_1}, \theta_{i_2}, \theta_{i_3}$, and $\ell^{(3)}_{n,k\cdot\cdot}(\boldsymbol{\theta})$ denote the $p \times p$ matrix whose (i, j) element is $\ell^{(3)}_{n,kij}(\boldsymbol{\theta})$. Define $p \times p$ matrices

$$W_{n,k} = (B_n^T)^{-1}\ell^{(3)}_{n,k\cdot\cdot}(\hat{\boldsymbol{\theta}}_n)B_n^{-1}, \; k = 1, ..., p. \tag{13}$$

Next, let

$$M_n = n\left(B_n^T B_n\right)^{-1}, \tag{14}$$

which is n times minus the inverse of the Hessian matrix. When the design variable $\boldsymbol{x}$ is random, define

$$M(\boldsymbol{\theta}) = \lim_n M_n \text{ in } P_{\boldsymbol{\theta}}\text{-probability}, \tag{15}$$

$$V_k(\boldsymbol{\theta}) = \lim_n \operatorname{tr}(W_{n,k}) \text{ for } k = 1, ..., p \quad \text{and} \quad V(\boldsymbol{\theta}) = (V_1(\boldsymbol{\theta}), \cdots, V_p(\boldsymbol{\theta}))^T. \tag{16}$$

When $\boldsymbol{x}$ is fixed, let

$$\bar{M}_n(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[n\{-\nabla^2 \ell_n(\boldsymbol{\theta})\}^{-1}], \tag{17}$$

and, for $k = 1, ..., p$,

$$\bar{V}_{n,k}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\operatorname{tr}[\{-\nabla^2 \ell_n(\boldsymbol{\theta})\}^{-1} \ell^{(3)}_{n,k\cdot\cdot}(\boldsymbol{\theta})]) \text{ and } \bar{V}_n(\boldsymbol{\theta}) = (\bar{V}_{n,1}(\boldsymbol{\theta}), \cdots, \bar{V}_{n,p}(\boldsymbol{\theta}))^T. \tag{18}$$

Note that $M(\boldsymbol{\theta})$ and $V(\boldsymbol{\theta})$ are functions of only $\boldsymbol{\theta}$, but $\bar{M}_n(\boldsymbol{\theta})$ and $\bar{V}_n(\boldsymbol{\theta})$ also involve the design variables $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Further, for linear models, $\ell^{(3)}_{n,k\cdot\cdot}$ is zero, and hence $V(\boldsymbol{\theta})$ and $\bar{V}_n(\boldsymbol{\theta})$ vanish.

3.2 Bias for maximum likelihood estimation

We will present the bias formulas for $\hat{\boldsymbol{\theta}}_n$ for random and fixed $\boldsymbol{x}$. To begin, multiplying both sides of (10) by n, together with (11) and the definition of M_n in (14), we obtain

$$nE_\xi(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) = E_\xi \left[M_n E^n_\xi \left\{ \frac{\nabla \xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})} + \nabla r_n(\boldsymbol{\theta}) \right\} \right]. \tag{19}$$

The following lemma is needed. The proof is straightforward and omitted.

Lemma 1 *Let $\boldsymbol{y}$ be a p-dimensional random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Let A be a $p \times p$ matrix. Then $E(\boldsymbol{y}^T A \boldsymbol{y}) = \operatorname{tr}(A\Sigma) + \boldsymbol{\mu}^T A \boldsymbol{\mu}$.*

From (12) and the fact that $\nabla \ell_n(\hat{\boldsymbol{\theta}}_n) = 0$, the kth component of $\nabla r_n(\boldsymbol{\theta})$ can be written as

$$\frac{\partial r_n(\boldsymbol{\theta})}{\partial \theta_k} = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \theta_k} - \frac{\partial \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \theta_k} - \sum_{i=1}^p \frac{\partial^2 \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \theta_k \partial \theta_i} (\theta_i - \hat{\theta}_{ni}).$$

Then a Taylor series expansion gives

$$\frac{\partial r_n(\boldsymbol{\theta})}{\partial \theta_k} = \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^T \ell^{(3)}_{n,k\cdot\cdot}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + \text{Rem} = \frac{1}{2} \boldsymbol{Z}_n^T W_{n,k} \boldsymbol{Z}_n + o_p(1), \tag{20}$$

where $\boldsymbol{Z}_n$ and $W_{n,k}$ are as in (5) and (13), and $|\text{Rem}| \le C \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\|^3 \left|\partial^4 \ell_n(\boldsymbol{\eta}) / \partial\theta_i \partial\theta_j \partial\theta_k \partial\theta_l\right|$ for some $\boldsymbol{\eta}$ lying on the line segment joining $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_n$; therefore, $|\text{Rem}| = o_p(1)$, provided that $\sqrt{n}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\| = O_p(1)$ and $\partial^4 \ell_n(\boldsymbol{\theta}) / \partial\theta_i \partial\theta_j \partial\theta_k \partial\theta_l = O_p(n)$. Corollary 1 below follows from Lemma 1 and (20).

Corollary 1 $E_\xi^n\left(\boldsymbol{Z}_n^T W_{n,k}\boldsymbol{Z}_n\right) = \mathrm{tr}(W_{n,k}) + o_p(1)$.

Theorem 1 *Let M, V, $\bar{M}_n$ and $\bar{V}_n$ be as in* (15)-(18). *Let*

$$M_{ij}^{\#} = \frac{\partial M_{ij}(\boldsymbol{\theta})}{\partial \theta_j} \quad \text{and} \quad \bar{M}_{n,ij}^{\#} = \frac{\partial \bar{M}_{n,ij}(\boldsymbol{\theta})}{\partial \theta_j},\ i,j = 1,...,p, \tag{21}$$

and $\mathbf{1}$ *be the unit p-vector. Suppose that ξ is continuously differentiable with a compact support in Ω. Then, for the case of random design variables, we have*

$$nE_\xi(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \int_\Omega \left\{ M^{\#}(\boldsymbol{\theta})\mathbf{1} - \frac{1}{2}M(\boldsymbol{\theta})V(\boldsymbol{\theta})\right\}\xi(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1), \tag{22}$$

which is written as

$$E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \simeq \frac{1}{n}\left\{ M^{\#}(\boldsymbol{\theta})\mathbf{1} - \frac{1}{2}M(\boldsymbol{\theta})V(\boldsymbol{\theta})\right\}; \tag{23}$$

and for the case of fixed design variables, we have

$$nE_\xi(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \int_\Omega \left\{ \bar{M}_n^{\#}(\boldsymbol{\theta})\mathbf{1} - \frac{1}{2}\bar{M}_n(\boldsymbol{\theta})\bar{V}_n(\boldsymbol{\theta})\right\}\xi(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1),$$

which is written as

$$E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \simeq \frac{1}{n}\left\{ \bar{M}_n^{\#}(\boldsymbol{\theta})\mathbf{1} - \frac{1}{2}\bar{M}_n(\boldsymbol{\theta})\bar{V}_n(\boldsymbol{\theta})\right\}. \tag{24}$$

The proof is in the Appendix. Cox and Snell (1968) obtained the bias for the maximum likelihood estimator for independent observations with fixed design variables. We will show that, for the fixed design scenario, our (24) agrees with their bias to order n^{-1}. To begin, we introduce some of their notation. Let the y_k be independent observations, but not necessarily identically distributed. Define

$$\begin{aligned} I_{rs} &= \sum_{k=1}^n E_{\boldsymbol{\theta}}\left\{ -\frac{\partial^2 \log p_{\boldsymbol{\theta}}(y_k)}{\partial\theta_r\partial\theta_s}\right\}, \\ K_{rst} &= \sum_{k=1}^n E_{\boldsymbol{\theta}}\left\{ \frac{\partial^3 \log p_{\boldsymbol{\theta}}(y_k)}{\partial\theta_r\partial\theta_s\partial\theta_t}\right\}, \\ J_{r,st} &= \sum_{k=1}^n E_{\boldsymbol{\theta}}\left\{ \frac{\partial \log p_{\boldsymbol{\theta}}(y_k)}{\partial\theta_r}\frac{\partial^2 \log p_{\boldsymbol{\theta}}(y_k)}{\partial\theta_s\partial\theta_t}\right\}. \end{aligned} \tag{25}$$

Note that I, J, K refer to totals over the sample and are of order n. Cox and Snell (1968) gave the bias as

$$E_{\boldsymbol{\theta}}(\hat{\theta}_{n,s} - \theta_s) = \frac{1}{2} \sum_{r,t,u} I^{rs} I^{tu} (K_{rtu} + 2J_{t,ru}) + o(n^{-1}), \tag{26}$$

where the superscripts denote matrix inversion, that is, I^{rs} is the (r, s) element of I^{-1}. From (24), the bias for individual θ_s is

$$E_{\boldsymbol{\theta}}(\hat{\theta}_{n,s} - \theta_s) \simeq \frac{1}{n} \left\{ \sum_t \bar{M}^{\#}_{n,st}(\boldsymbol{\theta}) - \frac{1}{2} \sum_r \bar{M}_{n,sr}(\boldsymbol{\theta}) \bar{V}_{n,r}(\boldsymbol{\theta}) \right\}. \tag{27}$$

Theorem 2 *The leading terms in* (26) *and* (27) *agree to order* n^{-1}.

The proof is in the Appendix. Since the bias formula for fixed design variables (24) resembles that for the random case (23), in Section 3.3 we will only present results for the case of random design variables.

3.3 Bias for l_2 penalized likelihood estimation

The technique described above can be applied to l_2 penalized likelihood estimators. To fix ideas, consider a regression model

$$y_k \sim p_{\boldsymbol{\theta}}(\cdot; \boldsymbol{x}_k), \tag{28}$$

where $p_{\boldsymbol{\theta}}$ is a known probability distribution, the y_k are independent responses, the $\boldsymbol{x}_k$ are the p-dimensional random covariates, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau)^T \in R^{p+1}$ is the unknown parameter with $\boldsymbol{\beta} \in R^p$ associated with the covariates and τ a nuisance parameter. Here, we assume that p is fixed and $p << n$. Define the penalized log-likelihood and its maximizer as

$$\ell_n^\lambda(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) - \frac{\lambda}{2\tau} \sum_{j=1}^p \beta_j^2 \quad \text{and} \quad \hat{\boldsymbol{\theta}}_n^\lambda = (\hat{\boldsymbol{\beta}}_n^\lambda, \hat{\tau}_n^\lambda)^T = \text{argmax}_{\boldsymbol{\theta}} \ell_n^\lambda(\boldsymbol{\theta}). \tag{29}$$

For example, if y_k in (28) has a normal distribution with mean $\boldsymbol{x}_k^T \boldsymbol{\beta}$ and variance τ, then $\hat{\boldsymbol{\beta}}_n^\lambda$ is the ridge estimator; and, if $y_k \sim \text{Bernoulli}(\boldsymbol{x}_k^T \boldsymbol{\beta})$, then, by setting $\tau = 1$ in (29), we have the l_2 penalized likelihood estimator for logistic regression.

Now consider a Bayesian model with prior

$$\xi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \exp\left(-\frac{\lambda}{2\tau}\sum_{j=1}^{p}\beta_j^2\right), \tag{30}$$

where $\lambda > 0$ and π is continuously differentiable and vanishes off of a compact support. Then the posterior density of $\boldsymbol{\theta}$ is

$$p_\xi^n(\boldsymbol{\theta}) \propto \xi(\boldsymbol{\theta})e^{\ell_n(\boldsymbol{\theta})} \propto \pi(\boldsymbol{\theta})e^{\ell_n^\lambda(\boldsymbol{\theta})}.$$

Next, define B_n^λ and $\boldsymbol{Z}_n^\lambda$ in a similar way to (4) and (5), but with ℓ_n and $\hat{\boldsymbol{\theta}}_n$ replaced by ℓ_n^λ and $\hat{\boldsymbol{\theta}}_n^\lambda$, that is,

$$(B_n^\lambda)^T B_n^\lambda = -\nabla^2\ell_n^\lambda(\hat{\boldsymbol{\theta}}_n^\lambda) \quad \text{and} \quad \boldsymbol{Z}_n^\lambda = B_n^\lambda(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n^\lambda).$$

Here, B_n^λ is $(p+1)\times(p+1)$ and $\boldsymbol{Z}_n^\lambda$ is a $(p+1)$-dimensional vector. Similarly, define an analogue of r_n in (7) as $r_n^\lambda(\boldsymbol{\theta}) = \ell_n^\lambda(\boldsymbol{\theta}) - \ell_n^\lambda(\hat{\boldsymbol{\theta}}_n^\lambda) - \|\boldsymbol{z}_n^\lambda\|^2/2$. Define also an analogue of M_n as M_n^λ and an analogue of $W_{n,k}$ as $W_{n,k}^\lambda$, that is,

$$M_n^\lambda = n\{-\nabla^2\ell_n^\lambda(\hat{\boldsymbol{\theta}}_n^\lambda)\}^{-1} \quad \text{and} \quad W_{n,k}^\lambda = \{(B_n^\lambda)^T\}^{-1}\ell_{n,k\cdot\cdot}^{\lambda(3)}(\hat{\boldsymbol{\theta}}_n^\lambda)(B_n^\lambda)^{-1}. \tag{31}$$

From (29), it can be seen that ℓ_n^λ and ℓ_n differ only by a penalty term not depending on n. So we have

$$\lim_n M_n^\lambda = \lim_n M_n = M(\boldsymbol{\theta}) \quad \text{and} \quad \lim_n \operatorname{tr}(W_{n,k}^\lambda) = \lim_n \operatorname{tr}(W_{n,k}) = V_k(\boldsymbol{\theta}), \tag{32}$$

where $M(\boldsymbol{\theta})$ and $V_k(\boldsymbol{\theta})$ are defined in (15) and (16). Then write the posterior density of $\boldsymbol{Z}_n^\lambda$ as

$$p_\xi^n(\boldsymbol{z}_n^\lambda) \propto \phi_{p+1}(\boldsymbol{z}_n^\lambda)f(\boldsymbol{z}_n^\lambda),$$

where $f(\boldsymbol{z}_n^\lambda) = \pi(\boldsymbol{\theta})e^{r_n^\lambda(\boldsymbol{\theta})}$. Therefore,

$$\frac{\nabla f_n(\boldsymbol{Z}_n^\lambda)}{f_n(\boldsymbol{Z}_n^\lambda)} = \{(B_n^\lambda)^T\}^{-1}\left\{\frac{\nabla\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} + \nabla r_n^\lambda(\boldsymbol{\theta})\right\}$$

and we have the following analogue of (19):

$$nE_\xi(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n^\lambda) = E_\xi\left[M_n^\lambda E_\xi^n\left\{\frac{\nabla\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} + \nabla r_n^\lambda(\boldsymbol{\theta})\right\}\right]. \tag{33}$$

Theorem 3 *Suppose that π is continuously differentiable with a compact support. Then*

$$nE_{\xi}(\hat{\boldsymbol{\theta}}_n^{\lambda} - \boldsymbol{\theta}) = \int_{\Omega} \left\{ M^{\#}(\boldsymbol{\theta})\mathbf{1} - \frac{1}{2}M(\boldsymbol{\theta})V(\boldsymbol{\theta}) - M(\boldsymbol{\theta})\eta_{\lambda}(\boldsymbol{\theta}) \right\} \xi(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1), \tag{34}$$

where η_λ is a $(p+1)$-dimensional vector whose jth component is

$$\eta_{\lambda,j}(\boldsymbol{\theta}) = \begin{cases} \lambda\beta_j/\tau & \text{if } j = 1, ..., p, \\ -\lambda\sum_{k=1}^{p}\beta_k^2/(2\tau^2) & \text{if } j = p+1. \end{cases} \tag{35}$$

The proof is in the Appendix. Observe from (34) that the bias induced from the use of the l_2 penalty appears only in η_λ, and that, if $\lambda = 0$, then $\eta_{\lambda,j} = 0$ for all j and (34) reduces to (22).

3.4 Bias for experiments with stopping times

The proposed method can be easily applied to models in which the sample size is a stopping time. It is known that the use of stopping times does not affect the form of the likelihood function; see, for example, Berger and Wolpert (1984). So the maximum likelihood estimator can be obtained as if the experiment is a fixed sample one. However, its sampling distribution may be affected in a complicated way.

Theorem 4 *Suppose that the sample size is a stopping time depending on a, and denote it as $t = t_a$. Further suppose that $a/t \to \rho(\boldsymbol{\theta})$ in $P_{\boldsymbol{\theta}}$-probability. Then the bias of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_t$ is*

$$E_{\xi}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}) = \frac{1}{a}\int_{\Omega} \left[\{\rho(\boldsymbol{\theta})M(\boldsymbol{\theta})\}^{\sharp}\mathbf{1} - \frac{1}{2}\rho(\boldsymbol{\theta})M(\boldsymbol{\theta})V(\boldsymbol{\theta}) \right] \xi(\boldsymbol{\theta})d\boldsymbol{\theta} + o\left(\frac{1}{a}\right). \tag{36}$$

The proof is in the Appendix. Note that, if the stopping times are not present, then $t = n$. Further, by taking $a = n$, we have $\rho(\boldsymbol{\theta}) = 1$ and (36) reduces to (22).

3.5 Usefulness of bias formulas

Theorems 1, 3 and 4 provide formulas for the bias to order n^{-1} for the maximum likelihood estimator for both fixed and random design variables, the l_2 penalized likelihood estimator and the maximum likelihood estimator when the sample size is a stopping time. The basic formula (23) involves the matrices M and V, where V vanishes for linear models. One way to interpret this formula is that there are two contributions to the bias, one from the nonlinearity of the estimator and one from the design. A natural question is whether we can design the experiment so that the latter is minimized.

In the next section, a variety of examples are analysed in order to demonstrate the scope of application of the proposed methodology. Of course, to apply the formulas, it is necessary to calculate the matrices M and V. Although analytical expressions may be possible only in specific cases, an alternative approach is to approximate the matrices by simulation. This is the approach taken in Section 5 in the context of ridge regression with $p > n$.

4 Examples

4.1 Normal model

Suppose that $y_1, ..., y_n$ are independent normal random variables with unknown mean μ and unknown variance τ. Let $\boldsymbol{\theta} = (\mu, \tau)^T$. Then the log-likelihood of $\boldsymbol{\theta}$ is $\ell_n(\boldsymbol{\theta}) = -(n/2)\log(2\pi\tau) - \sum_{k=1}^{n}(y_k - \mu)^2/(2\tau)$. Straightforward calculation shows that $M(\boldsymbol{\theta}) = \mathrm{diag}(\tau, 2\tau^2)$, a 2×2 diagonal matrix, and $V(\boldsymbol{\theta}) = (0, 5/\tau)^T$. So the approximate biases of $\hat{\mu}_n$ and $\hat{\tau}_n$ are

$$nE_{\boldsymbol{\theta}}\left\{\begin{pmatrix}\hat{\mu}_n - \mu \\ \hat{\tau}_n - \tau\end{pmatrix}\right\} \simeq M^{\#}(\boldsymbol{\theta})\mathbf{1} - \frac{1}{2}M(\boldsymbol{\theta})V(\boldsymbol{\theta}) = \begin{pmatrix} 0 \\ -\tau \end{pmatrix},$$

which is the exact result for normal models.

Next, suppose that $\tau = 1$ and consider the stopping time

$$t = t_a = \inf\{n \geq 1 : \Big|\sum_{k=1}^{n} y_k\Big| \geq a\}, \tag{37}$$

where $a > 0$. Therefore, we have $\theta = \mu$, $M(\theta) = \lim_t M_t = \lim_t t\{-\ell_t''(\hat{\theta}_t)\}^{-1} = 1$ and $a/t \to \rho(\theta) = |\theta|$ in P_θ-probability. Then simple calculation shows that the approximate bias of $\hat{\theta}_t$ is

$$E_\theta(\hat{\theta}_t - \theta) \simeq \frac{1}{a}\text{sign}(\theta) \equiv b(\theta). \tag{38}$$

4.2 Autoregressive model

Consider the first-order autoregressive model

$$y_k = \theta y_{k-1} + e_k,$$

where $|\theta| < 1$. We will derive the bias of the maximum likelihood estimator $\hat{\theta}_n$ when e_k is assumed to follow $t(\nu)$, Student's t distribution with degrees of freedom ν. The first four moments of e_k are $E(e_k) = E(e_k^3) = 0$, $E(e_k^2) = \nu/(\nu-2)$ and $E(e_k^4) = 3(\nu-2)/(\nu-4)$, provided that $\nu > 4$. The log-likelihood function of θ and its first and second derivatives are

$$\begin{aligned}
\ell_n(\theta) &= C - \frac{v+1}{2}\sum_{k=1}^n \log\left\{1 + \frac{(y_k - \theta y_{k-1})^2}{v}\right\}, \\
\ell_n'(\theta) &= \frac{v+1}{v}\sum_{k=1}^n \left\{\frac{y_{k-1}(y_k - \theta y_{k-1})}{1 + \frac{(y_k - \theta y_{k-1})^2}{v}}\right\}, \\
\ell_n''(\theta) &= \frac{v+1}{v}\sum_{k=1}^n y_{k-1}^2 \left[\frac{-1}{1 + \frac{(y_k - \theta y_{k-1})^2}{v}} + \frac{\frac{2}{v}(y_k - \theta y_{k-1})^2}{\left\{1 + \frac{(y_k - \theta y_{k-1})^2}{v}\right\}^2}\right],
\end{aligned}$$

where C is a constant. Straightforward calculation shows that

$$M(\theta) = \lim_n n\left\{-\ell_n''(\hat{\theta}_n)\right\}^{-1} = (1-\theta^2)w(\nu),$$

where

$$w(\nu) = \frac{(\nu-2)(\nu+3)}{\nu(\nu+1)}. \tag{39}$$

It is not difficult to see that $w(\nu) \to 1$ as $\nu \to \infty$. Some of the values of $w(\nu)$ are given in Table 1. Since the third derivative of ℓ_n involves y_k^3, whose expected value is zero, its effect

Table 1: Values of $w(\nu)$.

ν	5	6	7	8	9	10	20	30
$w(\nu)$	0.800	0.857	0.893	0.917	0.933	0.945	0.986	0.994

on the bias is negligible. So the approximate bias of $\hat{\theta}_n$ is

$$E_\theta(\hat{\theta}_n - \theta) \simeq \frac{1}{n} M^\sharp(\theta) = \frac{1}{n} M'(\theta) = -\frac{2\theta}{n} w(\nu) \equiv b(\theta). \tag{40}$$

Note that $b(\theta)$ approaches $-2\theta/n$ as $\nu \to \infty$, which is the bias under the assumption of normality.

4.3 Dilution model

This model has been discussed by Abdelbasit and Plackett (1983), and Coad (2014). Let $x > 0$ be the dilution level, $y = 0, 1$ be the response to the dilution, and $\theta > 0$ be the density. The probability model has the form

$$p_\theta(y; x) = (e^{-\theta x})^y (1 - e^{-\theta x})^{1-y}.$$

So the log-likelihood function is

$$\ell_n(\theta) = -\theta \sum_{k=1}^n x_k y_k + \sum_{k=1}^n (1 - y_k) \log(1 - e^{-\theta x_k}).$$

The $(k+1)$st design point is $x_{k+1} = 1.59/\hat{\theta}_k$. Straightforward calculation gives

$$M(\theta) = \frac{\theta^2(e^{1.59} - 1)}{1.59^2} \quad \text{and} \quad V(\theta) = \frac{1.59}{\theta} \left(\frac{e^{1.59} + 1}{e^{1.59} - 1} \right).$$

Therefore, the approximate bias of $\hat{\theta}_n$ is

$$E_\theta(\hat{\theta}_n - \theta) \simeq \frac{1}{n} \left\{ \frac{2\theta(e^{1.59} - 1)}{1.59^2} - \frac{1}{2} \frac{\theta(e^{1.59} + 1)}{1.59} \right\} \equiv b(\theta). \tag{41}$$

4.4 Generalized linear models

Consider a generalized linear model in which the kth response has probability distribution

$$p_{\boldsymbol{\theta}}(y_k; \boldsymbol{x}_k) = \exp[\{y_k \eta_k - a(\eta_k) + b(y_k)\}/\varphi],$$

where $\boldsymbol{x}_k \in R^p$ is a vector of adaptive design variables, $\boldsymbol{\theta} \in R^p$ is the unknown parameter, $\eta_k = \eta(\boldsymbol{x}_k^T\boldsymbol{\theta})$ and φ is known. So the log-likelihood function of $\boldsymbol{\theta}$ is

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{\varphi}\sum_{k=1}^{n}\left\{y_k\eta_k - a(\eta_k) + b(y_k)\right\}, \tag{42}$$

and Theorems 1 and 3 can be applied.

As an illustrative example, we consider a two-point design for the logistic model studied in Abdelbasit and Plackett (1983). Let y be a binary response variable with

$$p(y=1|x) = e^{\beta(x-\mu)}/\{1+e^{\beta(x-\mu)}\}. \tag{43}$$

Abdelbasit and Plackett (1983) obtained a two-point D-optimal design when the sample size is $n/2$ for each design point. They showed that the two-point designs symmetric about μ are such that x_1^* and x_2^* correspond to probabilities of response $p^* = 0.824$ and $q^* = 1 - p^*$. Further, given the current estimates $\hat{\beta}_k$ and $\hat{\mu}_k$, their sequential procedure suggests taking the next two design points as $x_{k+1} = \hat{\mu}_k - (1/\hat{\beta}_k)\log(p^*/q^*)$ and $x_{k+2} = \hat{\mu}_k + (1/\hat{\beta}_k)\log(p^*/q^*)$.

Let $\boldsymbol{\theta} = (\mu, \beta)^T$. The log-likelihood function of $\boldsymbol{\theta}$ for model (43) is

$$\ell_n(\boldsymbol{\theta}) = \sum_{k=1}^{n}[y_k(x_k-\mu)\beta - \log\{1+e^{\beta(x_k-\mu)}\}],$$

which is of the form (42) with $\eta_k = (x_k - \mu)\beta$ and $a(\eta_k) = \log(1+e^{\eta_k})$. Straightforward calculation shows that

$$M(\boldsymbol{\theta}) = \operatorname{diag}\left(\frac{1}{p^*q^*\beta^2}, \frac{\beta^2}{p^*q^*\{\log(p^*/q^*)\}^2}\right),$$

a 2×2 diagonal matrix, and $V(\boldsymbol{\theta}) = (0, 2/\beta)^T$. Hence, $M_{11}^\sharp(\boldsymbol{\theta}) = M_{12}^\sharp(\boldsymbol{\theta}) = M_{21}^\sharp(\boldsymbol{\theta}) = 0$, $M_{22}^\sharp(\boldsymbol{\theta}) = 2\beta/[p^*q^*\{\log(p^*/q^*)\}^2]$. Then, by (23), we have

$$\begin{aligned} nE_{\boldsymbol{\theta}}(\hat{\mu}_n - \mu) &\simeq 0, \\ nE_{\boldsymbol{\theta}}(\hat{\beta}_n - \beta) &\simeq \frac{\beta}{p^*q^*\{\log(p^*/q^*)\}^2}. \end{aligned}$$

It is also possible to consider the l_2 penalized likelihood estimators $\hat{\mu}_n^\lambda$ and $\hat{\beta}_n^\lambda$. By Theorem 3, the additional term in the bias is

$$M(\boldsymbol{\theta})\eta_\lambda(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{p^*q^*\beta^2} & 0 \\ 0 & \frac{\beta^2}{p^*q^*\{\log(p^*/q^*)\}^2} \end{pmatrix}\begin{pmatrix} \lambda\mu \\ \lambda\beta \end{pmatrix} = \begin{pmatrix} \frac{\lambda\mu}{p^*q^*\beta^2} \\ \frac{\lambda\beta^3}{p^*q^*\{\log(p^*/q^*)\}^2} \end{pmatrix}.$$

Therefore, the approximate biases for $\hat{\mu}_n^\lambda$ and $\hat{\beta}_n^\lambda$ are

$$\begin{aligned} E_{\boldsymbol{\theta}}(\hat{\mu}_n^\lambda - \mu) &\simeq -\frac{\lambda\mu}{np^*q^*\beta^2} \equiv b_1(\boldsymbol{\theta}), \\ E_{\boldsymbol{\theta}}(\hat{\beta}_n^\lambda - \beta) &\simeq \frac{\beta(1-\lambda\beta^2)}{np^*q^*\{\log(p^*/q^*)\}^2} \equiv b_2(\boldsymbol{\theta}). \end{aligned} \tag{44}$$

4.5 Ridge regression

Suppose that model (28) has the form

$$y_k = \boldsymbol{x}_k^T \boldsymbol{\beta} + e_k, \tag{45}$$

where e_k is normally distributed with mean 0 and unknown variance τ, and the $\boldsymbol{x}_k$ are p-dimensional random covariates with mean $\mathbf{0}$ and known covariance matrix Σ. Let $\boldsymbol{y} = (y_1, ..., y_n)^T$ and X_n be the design matrix whose kth row is $\boldsymbol{x}_k^T$, $k = 1, ..., n$. Then, by (29), we have

$$\hat{\boldsymbol{\beta}}_n^\lambda = (X_n^T X_n + \lambda I_p)^{-1} X_n^T \boldsymbol{y} \quad \text{and} \quad \hat{\tau}_n^\lambda = \frac{1}{n}\left\{\sum_{k=1}^n (y_k - \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_n^\lambda)^2 + \lambda \sum_{j=1}^p (\hat{\beta}_{n,j}^\lambda)^2\right\}. \tag{46}$$

Note that $\hat{\boldsymbol{\beta}}_n^\lambda$ is the ridge estimator. The following result is a simple consequence of Theorem 3.

Corollary 2 *The approximate biases of $\hat{\boldsymbol{\beta}}_n^\lambda$ and $\hat{\tau}_n^\lambda$ are*

$$\begin{aligned} E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_n^\lambda - \boldsymbol{\beta}) &\simeq -\frac{\lambda}{n}\Sigma^{-1}\boldsymbol{\beta}, \\ E_{\boldsymbol{\theta}}(\hat{\tau}_n^\lambda - \tau) &\simeq \frac{1}{n}\left(\lambda \sum_{j=1}^p \beta_j^2 - p\tau\right). \end{aligned} \tag{47}$$

The proof is sketched below. First, from (29), (31) and (32), we obtain

$$M_n^\lambda = \begin{pmatrix} n\hat{\tau}_n^\lambda (X_n^T X_n + \lambda I_p)^{-1} & \mathbf{0} \\ \mathbf{0}^T & 2\hat{\tau}_n^2 \end{pmatrix} \quad \text{and} \quad \lim_n M_n^\lambda = \begin{pmatrix} \tau\Sigma^{-1} & \mathbf{0} \\ \mathbf{0}^T & 2\tau^2 \end{pmatrix}, \tag{48}$$

where $\mathbf{0}$ is a $p \times 1$ zero vector. Next, from (31) and (32), we have $V_i(\boldsymbol{\theta}) = 0$ for $i = 1, ..., p$ and

$$V_{p+1}(\boldsymbol{\theta}) = \operatorname{tr}\left(\lim_n W_{n,p+1}^\lambda\right) = \operatorname{tr}\begin{pmatrix} \frac{1}{\tau} I_p & \mathbf{0} \\ \mathbf{0}^T & \frac{4}{\tau} \end{pmatrix} = \frac{p+4}{\tau}. \tag{49}$$

Then plugging these into (34) gives the desired results.

5 Ridge regression with $p > n$

For the linear regression model (45), the bias in (47) is justified for large n and small p. Specifically, the expressions in (32), (48) and (49) are in the limit of n. In this section, we will modify the procedures in Section 3 to approximate the bias of the ridge estimator $\hat{\boldsymbol{\beta}}_n^\lambda$ in (46) when $p > n$. To begin, we observe that the expectation of $\hat{\boldsymbol{\beta}}_n^\lambda$ in (46) does not depend on τ. So here we take τ as known and write $\boldsymbol{\theta} = \boldsymbol{\beta}$. Then M_n^λ in (31) becomes

$$M_n^\lambda = n\tau(X_n^T X_n + \lambda I_p)^{-1}, \tag{50}$$

and, as r_n^λ in (33) vanishes in the context of normal linear models, (33) simplifies to

$$nE_\xi(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^\lambda) = E_\xi\left[M_n^\lambda E_\xi^n\left\{\frac{\nabla\pi(\boldsymbol{\beta})}{\pi(\boldsymbol{\beta})}\right\}\right]. \tag{51}$$

Proposition 1 *Define* $\bar{M}_n = nE_{\boldsymbol{\beta}}\left(X_n^T X_n + \lambda I_p\right)^{-1}$, *which does not depend on* $\boldsymbol{\beta}$. *Then*

$$nE_\xi(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^\lambda) = \int_\Omega \lambda\bar{M}_n\boldsymbol{\beta}\xi(\boldsymbol{\beta})d\boldsymbol{\beta},$$

that is,

$$E_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_n^\lambda - \boldsymbol{\beta}) \simeq -\frac{\lambda}{n}\bar{M}_n\boldsymbol{\beta} \equiv \boldsymbol{b}(\boldsymbol{\beta}). \tag{52}$$

The proof is in the Appendix. The matrix $\bar{M}_n$ in (52) can be approximated through simulation, provided that the distribution of the covariate $\boldsymbol{x}_k$ is known.

6 Simulations

This section reports the accuracy of the bias formulas and the mean squared errors of the bias corrected estimator $\check{\theta}_n$ defined by $\check{\theta}_n = \hat{\theta}_n - b(\hat{\theta}_n)$. Whitehead (1986) suggested a bias-adjusted estimator $\tilde{\theta}_n$ obtained by solving

$$\hat{\theta}_n = \tilde{\theta}_n + b(\tilde{\theta}_n). \tag{53}$$

The estimator $\tilde{\theta}_n$ is sometimes called the indirect inference estimator; see, for instance, Phillips (2012). We will also report the performance of $\tilde{\theta}_n$ when it is easily obtainable. We found that $\tilde{\theta}_n$ may have a smaller variance than $\hat{\theta}_n$ in some cases. In what follows, let $\hat{b} = b(\hat{\theta}_n)$.

6.1 Normal model with stopping times

Consider a normal model with the stopping time $t = t_a$ defined in (37). In the simulations, we take $a = 50$. So the bias formula (38) gives

$$b(\theta) = \frac{1}{a}\text{sign}(\theta) = (0.02) \cdot \text{sign}(\theta).$$

Table 2 presents the bias of $\hat{\theta}_t$ for various θ values based on 10,000 replicates. The column b stands for the bias $b(\theta)$. The next two columns report the averaged $\hat{\theta}_t$ and $\check{\theta}_t$, and the associated variances are in parentheses. The results show that $\check{\theta}_t$ has a smaller bias and about the same variance.

6.2 Autoregressive model

We consider the first-order autoregressive model in Section 4.2. From (53) and the bias formula (40), we obtain the bias-adjusted estimator

$$\tilde{\theta}_n = \frac{\hat{\theta}_n}{1 - \frac{2w(\nu)}{n}},$$

where $w(\nu)$ is in (39). Table 3 presents the bias of $\hat{\theta}_n$ for various θ values when $n = 50$, with $\nu = 5$ and 10, based on 10,000 replicates. The column b stands for the bias $b(\theta)$. The next three columns show the averaged $\hat{\theta}_n$, $\check{\theta}_n$ and $\tilde{\theta}_n$, with the associated variances in parentheses. The results show that both bias-corrected estimators, $\check{\theta}_n$ and $\tilde{\theta}_n$, have smaller biases but slightly larger variances. As the squared bias is much smaller than the variance, both corrected estimators have larger mean squared errors.

Table 2: Normal model with stopping times.

θ	b	$\hat{\theta}_t$	$\check{\theta}_t$
1	0.02	1.019 (0.020)	0.999 (0.020)
4	0.02	4.016 (0.076)	3.996 (0.076)
-1	-0.02	-1.019 (0.020)	-0.999 (0.020)
-4	-0.02	-4.021 (0.077)	-4.001 (0.077)

Table 3: AR(1) model with $e_k \sim t(\nu)$. Left: $\nu = 5$; Right: $\nu = 10$.

θ	b	$\hat{\theta}_n$	$\check{\theta}_n$	$\tilde{\theta}_n$	θ	b	$\hat{\theta}_n$	$\check{\theta}_n$	$\tilde{\theta}_n$
0.5	-0.016	0.473	0.488	0.489	0.5	-0.019	0.471	0.489	0.489
		(0.014)	(0.014)	(0.014)			(0.015)	(0.016)	(0.016)
-0.5	0.016	-0.473	-0.488	-0.489	-0.5	0.019	-0.469	-0.487	-0.487
		(0.013)	(0.014)	(0.014)			(0.014)	(0.015)	(0.015)
0.8	-0.026	0.761	0.786	0.787	0.8	-0.030	0.755	0.784	0.785
		(0.008)	(0.009)	(0.009)			(0.009)	(0.010)	(0.010)
-0.8	0.026	-0.761	-0.786	-0.786	-0.8	0.030	-0.756	-0.785	-0.786
		(0.008)	(0.009)	(0.009)			(0.009)	(0.010)	(0.010)

6.3 Dilution model

We study the bias of $\hat{\theta}_n$ for various θ values. We take the first two design points as 1 to obtain an initial estimate of θ. The subsequent points are taken sequentially by the procedure described in Section 4.3. From (41), $b(\theta) = a\theta/n$, where a is about 1.23. So, by (53), the bias-adjusted estimator $\tilde{\theta}_n$ satisfies $\hat{\theta}_n = \tilde{\theta}_n(1 + a/n)$; therefore, $\tilde{\theta}_n$ has a smaller variance, as $\mathrm{var}(\tilde{\theta}_n) = \mathrm{var}(\hat{\theta}_n)(1+a/n)^{-2} < \mathrm{var}(\hat{\theta}_n)$. Table 4 presents the results for various θ values, with $n = 25$ and 50, based on 10,000 replicates. The true θ values are in the first column. The second column b stands for the bias $b(\theta)$ given in (41). The next three columns report the averaged $\hat{\theta}_n$, $\check{\theta}_n$ and $\tilde{\theta}_n$, and the associated variances are in parentheses.

The results show that $\hat{\theta}_n$ tends to have an upward bias, and both $\check{\theta}_n$ and $\tilde{\theta}_n$ substantially reduce the bias and variance of $\hat{\theta}_n$. We remark that the large variances for $(n, \theta) = (25, 3)$ are due to two extremely large $\hat{\theta}_n$ values that exceed 135.

Table 4: Dilution example. Left: $n = 25$; Right: $n = 50$.

θ	b	$\hat{\theta}_n$	$\check{\theta}_n$	$\tilde{\theta}_n$	θ	b	$\hat{\theta}_n$	$\check{\theta}_n$	$\tilde{\theta}_n$
1	0.049	1.053	1.001	1.003	1	0.025	1.028	1.003	1.003
		(0.081)	(0.073)	(0.074)			(0.036)	(0.034)	(0.034)
2.5	0.123	2.657	2.526	2.533	2.5	0.062	2.576	2.513	2.514
		(1.760)	(1.590)	(1.598)			(0.226)	(0.215)	(0.215)
3	0.148	3.220	3.062	3.069	3	0.074	3.082	3.006	3.008
		(5.568)	(5.033)	(5.058)			(0.321)	(0.305)	(0.305)

6.4 Generalized linear models

We consider the logistic model with a two-point sequential design discussed in Section 4.4. Preliminary simulation shows that the sequential sampling may cause computational problems with maximum likelihood methods. The computational issues of these methods for logistic models and some modifications have been studied in the literature; see, for example, Clogg, Rubin, Schenker, Schultz, and Weidman (1991). Since we have derived bias approximations for l_2 penalized likelihood estimators, here we modify the likelihood by adding the l_2 penalized term. To further ensure the convergence of the estimators, we take five initial design points at each of x_1^* and x_2^*. Then, the remaining design points are taken sequentially according to the procedure described in Section 4.4.

We take $\lambda = 1$, $n = 50$ and 80, and various choices of (μ, β). For each (μ, β) in Table 5, the columns b_1 and b_2 stand for $b_1(\boldsymbol{\theta})$ and $b_2(\boldsymbol{\theta})$ in (44), which shows that $\hat{\mu}_n^\lambda$ tends to have a downward bias, and, for $\lambda = 1$, $\hat{\beta}_n^\lambda$ tends to have an upward bias when $|\beta| < 1$ and a downward bias when $|\beta| > 1$. The remaining columns report the averaged $\hat{\boldsymbol{\theta}}_n^\lambda = (\hat{\mu}_n^\lambda, \hat{\beta}_n^\lambda)^T$ and $\check{\boldsymbol{\theta}}_n^\lambda = (\check{\mu}_n^\lambda, \check{\beta}_n^\lambda)^T$ based on 10,000 replicates, with the associated variances in parentheses. The simulation results show that $\check{\theta}_n$ has reduced the bias, but often tends to have a larger variance; the resulting mean squared error is also larger.

A biased-adjusted estimator $\tilde{\boldsymbol{\theta}}_n^\lambda = (\tilde{\mu}_n^\lambda, \tilde{\beta}_n^\lambda)^T$ can be calculated using (44). Although this reduces the bias in some cases, $b_1(\tilde{\boldsymbol{\theta}}_n^\lambda)$ is not guaranteed to be closer to $b_1(\boldsymbol{\theta})$ than $b_1(\hat{\boldsymbol{\theta}}_n^\lambda)$. Table 5 shows that a worthwhile reduction in the biases of $\hat{\mu}_n^\lambda$ and $\hat{\beta}_n^\lambda$ can be obtained by simply subtracting $\hat{b}_1$ and $\hat{b}_2$, respectively.

Table 5: Two-point design. Upper: $n = 50$; Lower: $n = 80$.

μ	β	b_1	b_2	$\hat{\mu}_n^\lambda$	$\hat{\beta}_n^\lambda$	$\check{\mu}_n^\lambda$	$\check{\beta}_n^\lambda$
1	0.5	-0.552	0.022	0.645	0.546	0.99	0.526
				(0.235)	(0.023)	(0.565)	(0.024)
2	0.8	-0.431	0.017	1.592	0.827	1.966	0.819
				(0.168)	(0.047)	(0.242)	(0.057)
3	1.2	-0.287	-0.031	2.409	1.049	2.853	1.079
				(0.335)	(0.131)	(0.239)	(0.174)
1	0.5	-0.345	0.014	0.75	0.533	1	0.52
				(0.189)	(0.013)	(0.338)	(0.013)
2	0.8	-0.269	0.01	1.745	0.829	1.988	0.823
				(0.114)	(0.030)	(0.142)	(0.033)
3	1.2	-0.18	-0.019	2.609	1.108	2.87	1.129
				(0.316)	(0.107)	(0.168)	(0.127)

6.5 Ridge regression

We consider the ridge regression model (45) with two random covariates $(x_1, x_2)^T \sim N_2(0, 0, \sigma_1^2, \sigma_2^2, \rho)$, where ρ represents the correlation coefficient between x_1 and x_2, and σ_1, σ_2 and ρ are assumed to be known. Then, by (47), the approximate biases of $\hat{\beta}_{n,i}^\lambda$ and $\hat{\tau}_n^\lambda$ are

$$\begin{aligned} E_{\boldsymbol{\theta}}(\hat{\beta}_{n,1}^\lambda - \beta_1) &\simeq \frac{\lambda}{n(1-\rho^2)} \left(\frac{\rho\beta_2}{\sigma_1\sigma_2} - \frac{\beta_1}{\sigma_1^2} \right) \equiv b_1(\boldsymbol{\theta}), \\ E_{\boldsymbol{\theta}}(\hat{\beta}_{n,2}^\lambda - \beta_2) &\simeq \frac{\lambda}{n(1-\rho^2)} \left(\frac{\rho\beta_1}{\sigma_1\sigma_2} - \frac{\beta_2}{\sigma_2^2} \right) \equiv b_2(\boldsymbol{\theta}), \end{aligned} \tag{54}$$

$$E_{\boldsymbol{\theta}}(\hat{\tau}_n^\lambda - \tau) \simeq \frac{1}{n} \left(\lambda \sum_{k=1}^{2} \beta_k^2 - 2\tau \right) \equiv b_3(\boldsymbol{\theta}). \tag{55}$$

In particular, if $\rho = 0$, then the bias of $\hat{\tau}_n^\lambda$ remains the same, but (54) becomes

$$E_{\boldsymbol{\theta}}(\hat{\beta}_{n,i}^\lambda - \beta_i) \simeq -\frac{\lambda\beta_i}{n\sigma_i^2},\ i = 1, 2.$$

In the simulations, we take $\lambda = 2$, $\tau = 1$, $(\sigma_1, \sigma_2) = (1, 2)$, $n = 15$ and 25, with various $\boldsymbol{\beta}$ values. Tables 6 and 7 report results for $\rho = 0$ and 0.5, respectively, based on 10,000

Table 6: Ridge regression. $\rho = 0$. Upper: $n = 15$; Lower: $n = 25$.

β_1	β_2	τ	b_1	b_2	b_3	$\hat{\beta}^{\lambda}_{n1}$	$\hat{\beta}^{\lambda}_{n2}$	$\hat{\tau}^{\lambda}_{n}$	$\check{\beta}^{\lambda}_{n1}$	$\check{\beta}^{\lambda}_{n2}$	$\check{\tau}^{\lambda}_{n}$
2	3	1	-0.267	-0.100	1.600	1.722	2.884	2.491	1.988	2.984	0.875
						(0.072)	(0.022)	(0.143)	(0.093)	(0.023)	(0.155)
-1	2	1	0.133	-0.067	0.533	-0.864	1.920	1.501	-0.997	1.986	0.953
						(0.062)	(0.020)	(0.124)	(0.082)	(0.021)	(0.154)
3	5	1	-0.400	-0.167	4.400	2.582	4.802	5.111	2.982	4.969	0.699
						(0.090)	(0.028)	(0.196)	(0.110)	(0.029)	(0.170)
2	3	1	-0.160	-0.060	0.960	1.837	2.936	1.926	1.997	2.996	0.961
						(0.040)	(0.011)	(0.080)	(0.047)	(0.012)	(0.087)
-1	2	1	0.080	-0.040	0.320	-0.919	1.955	1.309	-0.999	1.995	0.984
						(0.038)	(0.011)	(0.075)	(0.045)	(0.012)	(0.086)
3	5	1	-0.240	-0.100	2.640	2.750	4.891	3.538	2.990	4.991	0.892
						(0.045)	(0.012)	(0.090)	(0.052)	(0.013)	(0.089)

Table 7: Ridge regression. $\rho = 0.5$. Upper: $n = 15$; Lower: $n = 25$.

β_1	β_2	τ	b_1	b_2	b_3	$\hat{\beta}^{\lambda}_{n1}$	$\hat{\beta}^{\lambda}_{n2}$	$\hat{\tau}^{\lambda}_{n}$	$\check{\beta}^{\lambda}_{n1}$	$\check{\beta}^{\lambda}_{n2}$	$\check{\tau}^{\lambda}_{n}$
2	3	1	-0.222	-0.044	1.600	1.777	2.938	2.523	1.999	2.982	0.903
						(0.076)	(0.024)	(0.133)	(0.112)	(0.029)	(0.151)
-1	2	1	0.267	-0.133	0.533	-0.738	1.858	1.479	-1.004	1.992	0.928
						(0.079)	(0.025)	(0.130)	(0.116)	(0.030)	(0.154)
3	5	1	-0.311	-0.089	4.400	2.682	4.880	5.208	2.992	4.969	0.795
						(0.093)	(0.029)	(0.174)	(0.130)	(0.033)	(0.160)
2	3	1	-0.133	-0.027	0.960	1.862	2.967	1.939	1.995	2.994	0.975
						(0.048)	(0.014)	(0.078)	(0.060)	(0.015)	(0.088)
-1	2	1	0.160	-0.080	0.320	-0.843	1.918	1.299	-1.003	1.998	0.972
						(0.050)	(0.014)	(0.080)	(0.063)	(0.015)	(0.089)
3	5	1	-0.187	-0.053	2.640	2.810	4.934	3.572	2.997	4.987	0.928
						(0.052)	(0.015)	(0.085)	(0.065)	(0.016)	(0.088)

replicates. The first three columns are the true $\boldsymbol{\theta}$ values and the next three columns, b_i, $i = 1, 2, 3$, stand for $b_i(\boldsymbol{\theta})$ in (54) and (55). The remaining columns report the averaged values of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n^\lambda = (\hat{\beta}_{n1}^\lambda, \hat{\beta}_{n1}^\lambda, \hat{\tau}_n^\lambda)^T$ and the bias-corrected estimate $\check{\boldsymbol{\theta}}_n^\lambda = (\check{\beta}_{n1}^\lambda, \check{\beta}_{n1}^\lambda, \check{\tau}_n^\lambda)^T$, with the associated variances in parentheses. Here, the bias-corrected estimate is defined as $\check{\theta}_{ni}^\lambda = \hat{\theta}_{ni}^\lambda - b_i(\hat{\boldsymbol{\theta}}_n^0)$ with $\hat{\boldsymbol{\theta}}_n^0 = (\hat{\boldsymbol{\beta}}_n^0, n\hat{\tau}_n^0/(n-2))^T$, the unbiased estimator of $\boldsymbol{\theta}$. Note that, from (46), $(\hat{\boldsymbol{\beta}}_n^0, \hat{\tau}_n^0)^T$ is the maximum likelihood estimator. The results show that the bias approximations are pretty accurate, but the corrected estimators tend to have larger variances and mean squared errors.

The R code for the simulation study is available at `http://www3.nccu.edu.tw/~chweng/publication.htm`.

7 Discussion

For the AR(1) model with t distributed errors, the estimator can be expressed as the solution to an estimating equation. So it is possible to obtain an approximate bias using the formula of Bao and Ullah (2007). Explicitly, using their (4), the bias for this model can be written as

$$E_\theta(\hat{\theta}_n - \theta) = n^{-2}\{M(\theta)\}^2 E\{\ell_n'(\theta)\ell_n''(\theta)\} + o(n^{-1}) \tag{56}$$

in our notation. As direct calculation of $E\{\ell_n'(\theta)\ell_n''(\theta)\}$ may be complicated for dependent observations, in the following, we will use similar techniques to those in the proof of Theorem 2 in the Appendix to show the equivalence of (56) to our (40). First, employing the third Bartlett identity (Bartlett, 1953) gives

$$E\{\ell_n'(\theta)\ell_n''(\theta)\} = \frac{d}{d\theta}E\{\ell_n''(\theta)\} - E\{\ell_n'''(\theta)\},$$

where $E\{\ell_n'''(\theta)\} = 0$ in this case. Together with the fact that $E\{n^{-1}\ell_n''(\theta)\} = M^{-1}(\theta) + o(1)$, (56) becomes

$$E_\theta(\hat{\theta}_n - \theta) = n^{-1}\{M(\theta)\}^2\{M^{-1}(\theta)\}' + o(n^{-1}) = n^{-1}M'(\theta) + o(n^{-1}),$$

where the second equality follows by direct computation or by the one-dimensional version of Lemma 2 in the Appendix. The above arguments show that our formula is simplified dramatically. One nice feature of the bias approximations is that they can be expressed in terms of just two matrices M and V.

The bias of the l_2 penalized likelihood estimator was obtained in Section 3.3. It is of interest to see whether the techniques in Rilstone, Srivastava, and Ullah (1996) can be applied to this case. Their paper studied a class of estimators $\hat{\boldsymbol{\beta}}_n \in R^p$ that can be written as the solution to a set of moment equations of the form

$$\Psi_n(\hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{k=1}^{n} q_k(\hat{\boldsymbol{\beta}}_n) = 0, \tag{57}$$

where $q_k(\boldsymbol{\beta}) = q(Z_k; \boldsymbol{\beta})$ is a known $p \times 1$ vector-valued function of the observable data Z_k and a parameter vector $\boldsymbol{\beta} \in R^p$ with a true β_0 defined such that $E\{q(Z_k; \boldsymbol{\beta})\} = 0$ only at $\boldsymbol{\beta}_0$ for all k. Now consider ridge regression in Section 4.5 with a known variance $\tau = 1$. It is easily seen that the estimator $\hat{\boldsymbol{\beta}}_n^\lambda$ satisfies $\sum_{k=1}^n q_k(\hat{\boldsymbol{\beta}}_n^\lambda) = 0$ with

$$q_k(\boldsymbol{\beta}) = \nabla \ell_{nk}(\boldsymbol{\beta}) - \frac{2\lambda}{n}\boldsymbol{\beta},$$

where ℓ_{nk} is the log-likelihood based on the kth observation $(\boldsymbol{x}_k, y_k)$. However, the condition $E\{q(Z_k; \boldsymbol{\beta})\} = 0$ does not hold in this case. So the results in Rilstone, Srivastava, and Ullah (1996) need some modifications to obtain the bias for penalized estimators. It is also possible to modify the approach of Cox and Snell (1968) to obtain the bias in the penalized case.

It may be possible to apply the approach in Section 3.3 to lasso regression, since its penalty term $\sum_{j=1}^p |\beta_j|$ satisfies the requirements of the Woodroofe-Stein identity. Although the bias approximation is likely to be complicated in this case, we hope to report the details in a separate paper.

In Section 4.5, it was assumed that the covariance matrix of the $\boldsymbol{x}_k$ is known. This may be relaxed by assuming that the covariates follow a probability distribution with unknown parameter $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)^T$. By letting $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \tau, \psi_1, \ldots, \psi_r)^T$, it can be shown that the bias calculations for $\hat{\boldsymbol{\psi}}_n^\lambda$ and $(\hat{\boldsymbol{\beta}}_n^\lambda, \hat{\tau}_n^\lambda)^T$ may be handled separately. The details are omitted.

Although the focus of this paper has been bias approximations, it is possible to use the proposed approach to approximate the covariance matrix of the estimators up to order n^{-2}.

The derivation is based on an expression for the second moment as in (3). Currently, we are able to obtain the approximate covariance matrix for normal linear models. However, the details of the approximation have not been included here, because some technical difficulties associated with other models have not yet been fully resolved.

In this paper, analytical approximations to the bias have been provided. Such an approach gives some information on both the form and the direction of the bias. This information can be very helpful when constructing bias-adjusted estimators, as shown in Section 6. An alternative approach would be to use the bootstrap to estimate the bias. Such estimates are usually harder to compute in practice.

Appendix: Proofs

Proof of Theorem 1 We prove only (23). The proof of (24) is similar and we omit it. By (19), it suffices to show that

$$E_\xi \left\{ M_n E_\xi^n \left(\frac{\nabla \xi}{\xi} \right) \right\} = - \int_\Omega M^\#(\boldsymbol{\theta}) \mathbf{1} \xi(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(1) \tag{58}$$

and

$$E_\xi \left[M_n E_\xi^n \{ \nabla r_n(\boldsymbol{\theta}) \} \right] = \frac{1}{2} \int_\Omega M(\boldsymbol{\theta}) V(\boldsymbol{\theta}) \xi(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(1). \tag{59}$$

For (58), write

$$\begin{aligned} E_\xi \left[M_n E_\xi^n \left\{ \frac{\nabla \xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})} \right\} \right] &= E_\xi \left\{ M_n \frac{\nabla \xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})} \right\} = \int_\Omega E_{\boldsymbol{\theta}}(M_n) \nabla \xi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ = \int_\Omega M(\boldsymbol{\theta}) \nabla \xi(\boldsymbol{\theta}) d\boldsymbol{\theta} &+ \int_\Omega \{ E_{\boldsymbol{\theta}}(M_n) - M(\boldsymbol{\theta}) \} \nabla \xi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \end{aligned} \tag{60}$$

where

$$\int_\Omega M(\boldsymbol{\theta}) \nabla \xi(\boldsymbol{\theta}) d\boldsymbol{\theta} = - \int_\Omega M^\#(\boldsymbol{\theta}) \mathbf{1} \xi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

follows by integration by parts. Suppose that $E_{\boldsymbol{\theta}}(M_n)$ is finite and continuous in $\boldsymbol{\theta}$. Together with the assumption that ξ is continuously differentiable with a compact support in Ω, we have that $\nabla \xi$ is bounded on Ω and that the second term on the right-hand side of (60) converges to zero.

For (59), by (20) and Corollary 1, we have

$$E_\xi\left[M_n E_\xi^n\{\nabla r_n(\boldsymbol{\theta})\}\right] = \frac{1}{2}E_\xi\left\{M_n\begin{pmatrix}\mathrm{tr}(W_{n,1})\\ \vdots \\ \mathrm{tr}(W_{n,p})\end{pmatrix}\right\} + o(1),$$

where $\mathrm{tr}(W_{n,k}) \to V_k(\boldsymbol{\theta})$ as in (16). Suppose that $E_{\boldsymbol{\theta}}\{\mathrm{tr}(W_{n,k})\}$ is finite and continuous in $\boldsymbol{\theta}$. Then the above line converges to $E_\xi\left\{M(\boldsymbol{\theta})V(\boldsymbol{\theta})\right\}/2$. □

To prove Theorem 2, we need the following formula for the derivative of the inverse of a matrix; see, for example, Dhrymes (1978).

Lemma 2 *Let A be a $p \times p$ matrix whose elements are functions of the scalar parameter x. Then*

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}.$$

Proof of Theorem 2 First, from the third Bartlett identity (Bartlett, 1953), we have

$$J_{t,ru} = -K_{tru} - \frac{\partial I_{ru}}{\partial \theta_t},$$

and hence (26) can be written as

$$\begin{aligned}E_{\boldsymbol{\theta}}(\hat{\theta}_{n,s} - \theta_s) &= -\frac{1}{2}\sum_{r,t,u} I^{rs}I^{tu}\left(K_{rtu} + 2\frac{\partial I_{ru}}{\partial \theta_t}\right) + o\left(\frac{1}{n}\right)\\ &= -\sum_{r,t,u} I^{rs}I^{tu}\frac{\partial I_{ru}}{\partial \theta_t} - \frac{1}{2}\sum_{r,t,u} I^{rs}I^{tu}K_{rtu} + o\left(\frac{1}{n}\right).\end{aligned}$$

So it suffices to show that

$$\begin{aligned}&\frac{1}{n}\left\{\sum_t \bar{M}^{\#}_{n,st}(\boldsymbol{\theta}) - \frac{1}{2}\sum_r \bar{M}_{n,sr}(\boldsymbol{\theta})\bar{V}_{n,r}(\boldsymbol{\theta})\right\}\\ &= -\sum_{r,t,u} I^{rs}I^{tu}\frac{\partial I_{ru}}{\partial \theta_t} - \frac{1}{2}\sum_{r,t,u} I^{rs}I^{tu}K_{rtu} + o\left(\frac{1}{n}\right).\end{aligned}$$

Now, from the definitions of $\bar{M}_n(\boldsymbol{\theta})$ and I in (17) and (25), we have $\bar{M}_n(\boldsymbol{\theta}) = nI^{-1}(\boldsymbol{\theta}) + o(1)$, under some regularity conditions. Then, by Lemma 2 and the definition of $\bar{M}^{\#}_n$ in (21), we have

$$\bar{M}^{\#}_{n,st} = \frac{\partial \bar{M}_{n,st}}{\partial \theta_t} = \frac{\partial (nI^{-1})_{st}}{\partial \theta_t} + o(1) = -n\sum_{r,u} I^{sr}I^{ut}\frac{\partial I_{ru}}{\partial \theta_t} + o(1).$$

Therefore,

$$\frac{1}{n}\sum_{t}\bar{M}^{\#}_{n,st} = -\sum_{t,r,u} I^{sr}I^{ut}\frac{\partial I_{ru}}{\partial\theta_t} + o\left(\frac{1}{n}\right).$$

Finally, from the definition of $\bar{V}_{n,r}(\boldsymbol{\theta})$ in (18), we have

$$\bar{V}_{n,r}(\boldsymbol{\theta}) = \operatorname{tr}\left(E_{\boldsymbol{\theta}}[\{-\nabla^2\ell_n(\boldsymbol{\theta})\}^{-1}]E_{\boldsymbol{\theta}}\{\ell^{(3)}_{n,r..}(\boldsymbol{\theta})\}\right) + o(1) = \sum_{t,u} I^{tu}K_{rtu} + o(1).$$

So

$$\frac{1}{n}\sum_{r}\bar{M}_{n,sr}\bar{V}_{n,r} = \sum_{r,t,u} I^{sr}I^{tu}K_{rtu} + o\left(\frac{1}{n}\right).$$

This completes the proof. □

Proof of Theorem 3 From (33), write

$$nE_\xi(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{\lambda}_n) = E_\xi\left[M^{\lambda}_n E^n_\xi\left\{\frac{\nabla\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} + \nabla r^{\lambda}_n(\boldsymbol{\theta})\right\}\right] = I_1 + I_2,$$

say. By (30) and similar techniques above, we obtain

$$I_1 = E_\xi\left[M^{\lambda}_n E^n_\xi\left\{\frac{\nabla\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right\}\right] \to \int_\Omega e^{-\frac{\lambda}{2\tau}\sum_{k=1}^{p}\beta_k^2}M(\boldsymbol{\theta})\nabla\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The ith component for the term on the right-hand side of the above line is

$$\int_\Omega\sum_{j=1}^{p+1}\left(m_{ij}\frac{\partial\pi}{\partial\theta_j}e^{-\frac{\lambda}{2\tau}\sum_{k=1}^{p}\beta_k^2}\right)d\boldsymbol{\theta} = \int_\Omega\xi(\boldsymbol{\theta})\sum_{j=1}^{p+1}\left\{\eta_{\lambda,j}(\boldsymbol{\theta})m_{ij}(\boldsymbol{\theta}) - m^{\#}_{ij}(\boldsymbol{\theta})\right\}d\boldsymbol{\theta},$$

where η_λ is as in (35). The treatment of I_2 is the same as (59), except that ℓ_n is now replaced by ℓ^{λ}_n. □

Proof of Theorem 4 The proof is similar to that for Theorem 1. First, replacing n in (19) with t gives

$$E_\xi(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t) = E_\xi\left[\frac{1}{t}M_t E^t_\xi\left\{\frac{\nabla\xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})} + \nabla r_t(\boldsymbol{\theta})\right\}\right] = \frac{1}{a}E_\xi\left[\frac{a}{t}M_t E^t_\xi\left\{\frac{\nabla\xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})} + \nabla r_t(\boldsymbol{\theta})\right\}\right].$$

Then the first term on the right-hand side of the above line can be written as

$$\begin{aligned}
&\frac{1}{a}E_\xi\left[\frac{a}{t}M_t E^t_\xi\left\{\frac{\nabla\xi(\boldsymbol{\theta})}{\xi(\boldsymbol{\theta})}\right\}\right] = \frac{1}{a}\int_\Omega E_{\boldsymbol{\theta}}\left(\frac{a}{t}M_t\right)\nabla\xi(\boldsymbol{\theta})d\boldsymbol{\theta} \\
=&\frac{1}{a}\int_\Omega\rho(\boldsymbol{\theta})M(\boldsymbol{\theta})\nabla\xi(\boldsymbol{\theta})d\boldsymbol{\theta} + \frac{1}{a}\int_\Omega\left\{E_{\boldsymbol{\theta}}\left(\frac{a}{t}M_t\right) - \rho(\boldsymbol{\theta})M(\boldsymbol{\theta})\right\}\nabla\xi(\boldsymbol{\theta})d\boldsymbol{\theta} \\
=&-\frac{1}{a}\int_\Omega\{\rho(\boldsymbol{\theta})M(\boldsymbol{\theta})\}^{\sharp}\,\mathbf{1}\xi(\boldsymbol{\theta})d\boldsymbol{\theta} + o\left(\frac{1}{a}\right)
\end{aligned}$$

and the second term is

$$\frac{1}{a}E_\xi\left[\frac{a}{t}M_tE_\xi^t\left\{\nabla r_t(\boldsymbol{\theta})\right\}\right]=\frac{1}{2a}\int_\Omega \rho(\boldsymbol{\theta})M(\boldsymbol{\theta})V(\boldsymbol{\theta})\xi(\boldsymbol{\theta})d\boldsymbol{\theta}+o\left(\frac{1}{a}\right).$$

Hence, the desired result follows. □

Proof of Proposition 1 First recall that $M_n^\lambda = n\tau(X_n^TX_n+\lambda I_p)^{-1}$ in (50). With the definition of $\bar{M}_n$, we now rewrite the right-hand side of (51) as

$$\begin{aligned}E_\xi\left[M_n^\lambda E_\xi^n\left\{\frac{\nabla\pi(\boldsymbol{\beta})}{\pi(\boldsymbol{\beta})}\right\}\right] &= E_\xi\left[\tau\bar{M}_n\left\{\frac{\nabla\pi(\boldsymbol{\beta})}{\pi(\boldsymbol{\beta})}\right\}\right]\\ =\int_\Omega\left\{-(\tau\bar{M}_n)^{\#}(\boldsymbol{\beta})\mathbf{1}+\tau\bar{M}_n\eta_\lambda(\boldsymbol{\beta})\right\}\xi(\boldsymbol{\beta})d\boldsymbol{\beta}, &\qquad (61)\end{aligned}$$

where the last line follows by integration by parts, with $\eta_\lambda(\boldsymbol{\beta})=\lambda\boldsymbol{\beta}/\tau$, a $p\times 1$ vector. Since $\tau\bar{M}_n$ does not involve $\boldsymbol{\beta}$, we have $(\tau\bar{M}_n)_{ij}^{\#}=0$ for all i and j. By (61), (51) becomes

$$nE_\xi(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_n^\lambda)=\int_\Omega \lambda\bar{M}_n\boldsymbol{\beta}\xi(\boldsymbol{\beta})d\boldsymbol{\beta},$$

and Proposition 1 follows. □

Acknowledgements

The first author is partially supported by the Ministry of Science and Technology of Taiwan. Part of the work was carried out while the second author was visiting the National Chengchi University during August and September 2017. The authors also wish to thank Professor M. B. Woodroofe and two referees for their comments, which have led to an improved paper.

References

Abdelbasit, K. M., & Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association, 78*, 90–98.

Anderson, J., & Richardson, S. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics, 21*, 71–78.

Bao, Y. (2018). A general result on the estimation bias of ARMA models. *Journal of Statistical Planning and Inference, 197*, 107–125.

Bao, Y., & Ullah, A. (2007). The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics, 140*, 650–669.

Bartlett, M. (1953). Approximate confidence intervals. *Biometrika, 40*, 12–19.

Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle.* Hayward, CA: Institute of Mathematical Statistics.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli, 19*, 1212–1242.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. *Journal of the American Statistical Association, 86*, 68–78.

Coad, D. S. (1994). Estimation following sequential tests involving data-dependent treatment allocation. *Statistica Sinica, 4*, 693–700.

Coad, D. S. (2014). Corrected confidence intervals based on the signed root transformation for multi-parameter sequentially designed experiments. *Journal of Statistical Planning and Inference, 147*, 173–187.

Coad, D. S., & Woodroofe, M. B. (1996). Corrected confidence intervals after sequential testing with applications to survival analysis. *Biometrika, 83*, 763–777.

Coad, D. S., & Woodroofe, M. B. (1998). Approximate bias calculations for sequentially designed experiments. *Sequential Analysis, 17*, 1–31.

Cordeiro, G. M., & Klein, R. (1994). Bias correction in ARMA models. *Statistics and Probability Letters, 19*, 169–176.

Cordeiro, G. M., & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological), 53*, 629–643.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological), 30*, 248–275.

Dhrymes, P. J. (1978). *Mathematics for econometrics.* New York: Springer.

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

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

James, W., & Stein, C. M. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley, CA: University of California Press, pp. 361–379.

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research, 15*, 2869–2909.

McLachlan, G. (1980). A note on bias correction in maximum likelihood estimation with logistic discrimination. *Technometrics, 22*, 621–627.

Phillips, P. C. B. (2012). Folklore theorems, implicit maps, and indirect inference. *Econometrica, 80*, 425–454.

Rilstone, P., Srivastava, V. K., & Ullah, A. (1996). The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics, 75*, 369–395.

Rilstone, P., & Ullah, A. (2005). Corrigendum to 'The second-order bias and mean squared error of nonlinear estimators' by P. Rilstone, V.-K. Srivastava, & A. Ullah. *Journal of Econometrics, 124*, 203–204.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine, 2*, 71–78.

Shao, J. & Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics, 40*, 812–831.

Shenton, L. R., & Bowman, K. O. (1977). *Maximum likelihood estimation in small samples.* New York, NY: Macmillan.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics, 9*, 1135–1151.

Stein, C. M. (1985). On the coverage probability of confidence sets based on a prior distribution. *Banach Center Publications, 16*, 485–514.

Todd, S., Whitehead, J., & Facey, K. M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika, 83*, 453–461.

Weng, R. C., & Lin, C. J. (2011). A Bayesian approximation method for online ranking. *Journal of Machine Learning Research, 12*, 267–300.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika, 73*, 573–581.

Woodroofe, M. (1986). Very weak expansions for sequential confidence levels. *The Annals of Statistics, 14*, 1049–1067.

Woodroofe, M. (1989). Very weak expansions for sequentially designed experiments: Linear models. *The Annals of Statistics, 17*, 1087–1102.

Woodroofe, M. (1990). On stopping times and stochastic monotonicity. *Sequential Analysis, 9*, 335–342.

Woodroofe, M., & Coad, D. S. (1997). Corrected confidence sets for sequentially designed experiments. *Statistica Sinica, 7*, 53–74.

Yang, Z. (2015). A general method for third-order bias and variance corrections on a nonlinear estimator. *Journal of Econometrics, 186*, 178–200.

Zhao, S., & Shojaie, A. (2016). A significance test for graph-constrained estimation. *Biometrics, 72*, 484–493.