Mean squared error of ridge estimators in logistic regression

Blagus, R.; Goeman, J.J.

Received: 28 December 2018 | Revised: 26 November 2019 | Accepted: 6 December 2019

DOI: 10.1111/stan.12201

ORIGINAL ARTICLE

WILEY

Mean squared error of ridge estimators in logistic regression

Rok Blagus[1] | **Jelle J. Goeman**[2]

[1]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

[2]Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
Rok Blagus, Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, 1000 Ljubljana, Slovenia.
Email: rok.blagus@mf.uni-lj.si

Present Address
Rok Blagus, Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, 1000 Ljubljana, Slovenia

Funding information
Slovenian Research Agency, Grant/Award Number: N1-0035 and P3-0154

It is well known that the maximum likelihood estimator (MLE) is inadmissible when estimating the multidimensional Gaussian location parameter. We show that the verdict is much more subtle for the binary location parameter. We consider this problem in a regression framework by considering a ridge logistic regression (RR) with three alternative ways of shrinking the estimates of the event probabilities. While it is shown that all three variants reduce the mean squared error (MSE) of the MLE, there is at the same time, for every amount of shrinkage, a true value of the location parameter for which we are overshrinking, thus implying the minimaxity of the MLE in this family of estimators. Little shrinkage also always reduces the MSE of individual predictions for all three RR estimators; however, only the naive estimator that shrinks toward 1/2 retains this property for any generalized MSE (GMSE). In contrast, for the two RR estimators that shrink toward the common mean probability, there is always a GMSE for which even a minute amount of shrinkage increases the error. These theoretical results are illustrated on a numerical example. The estimators are also applied to a real data set, and practical implications of our results are discussed.

KEYWORDS

admissibility, generalized squared loss, James–Stein estimator, Jeffreys invariant prior, multidimensional location parameter

1 | INTRODUCTION

When estimating a multidimensional location parameter in a multivariate Gaussian distribution, it is well known that the maximum likelihood estimator (MLE) is not admissible with respect to mean squared error (MSE) loss (Brown, 1966, 1968; Brown & Fox, 1974a, 1974b; James & Stein, 1961; Stein, 1956, 1959). An estimator is said to be admissible when no other estimator that has a risk that is never larger and is sometimes smaller exists, where risk is defined as the expected value of some loss function that measures the distance between the estimate and the true parameter value. It was proved that, for a scalar and two-dimensional location parameter, the best invariant estimator is admissible, whereas it is inadmissible for a three- (or more) dimensional location parameter. Informally, it was shown that simultaneously estimating more location parameters will reduce the overall risk of the estimator, but at the expense of increasing the risk for some location parameters (and obviously decreasing it for others). In short, shrinkage estimators achieve a better MSE. One way of shrinkage is by ridge regression, which can be applied more generally in least squares problems. Theobald (1974) showed, for the ridge estimator in ordinary least squares (OLS) regression, that there always exists some value of the penalty parameter for which shrinkage will outperform the OLS estimator in terms of any positive semidefinite weighted sum of coefficient mean square and product errors; by the linearity of the link function, this then also implies that shrinkage improves the MSE of the predictions.

Shrinkage estimators are also popular in logistic regression especially with collinear and/or high-dimensional predictors (Sun & Wang, 2012; Zhou et al., 2010), where they are often used for estimating the event probability (Walter et al., 2011). Applying the idea of the ridge estimator in a linear regression model (Schaefer, 1986) obtained a straightforward ridge logistic estimator, which depended on the MLE. However, it is this dependence on the MLE that can cause issues, since the estimator of Schaefer (1986) does not exist if some of the unrestricted MLEs are infinite. To overcome this, le Cessie and van Houwelingen (1992) followed the restricted maximum likelihood approach of Duffy and Santner (1989). However, although it was shown via simulation studies that this approach makes sense (Steyerberg et al., 2001), a similar motivation as given by Theobald (1974) for the OLS regression is, so far, lacking for the estimation of the event probability.

Schaefer et al. (1984) showed, for a large sample size, a sufficient degree of collinearity and some values of the penalty parameter, the inadmissibility with respect to the MSE loss of the MLE of the logistic regression parameter. In fact, we show in the Appendix that, using asymptotic results for the ridge estimator of the logistic regression parameter (see, e.g., le Cessie & van Houwelingen, 1992), a similar result to that in the work of Theobald (1974) for the OLS regression is easily established for the ridge estimators considered by Schaefer et al. (1984) and le Cessie and van Houwelingen (1992). By the nonlinearity of the link function, this however, unlike in the OLS example, does not directly imply the improvement of the ridge estimator at the level of the estimated probabilities.

In this paper, we investigate the shrinkage properties of logistic ridge regression at the level of estimated probabilities, when using some fixed amount of shrinkage in a nonasymptotic setting. When do such shrinkage estimators outperform the MLE? We concentrate on the case of a single categorical predictor with K categories (or, equivalently, the model with multiple categorical predictors and full interactions), since for this model, we can derive explicit expressions for the estimated probabilities of the MLE and one-step approximations for the shrinkage estimators. We study the MSE and the generalized MSE (GMSE) for three variants of logistic ridge regression: The variants considered shrink estimated probabilities (a) toward 1/2, as advocated by

Elgmati et al. (2015); (b) toward the estimated probability of the reference category, as suggested by Greenland and Mansournia (2015); or toward a common mean probability, as argued for by le Cessie and van Houwelingen (1992). Special cases of Scenario (a) include Firth's estimator (FPE; Firth, 1993) and the Agresti–Coull estimator (ACE; Agresti & Coull, 1998).

We show that a similar result to that in the work of Theobald (1974) holds for our model also at the level of the estimated probabilities: For the overall MSE, a little shrinkage is always better than no shrinkage. This result holds for all three estimators. The maximal amount of shrinkage that is still beneficial, however, depends on the true values of the parameters. Any fixed amount of shrinkage may therefore result in overshrinkage. Therefore, although the MLE is never optimal for any value of the parameter, we show that it is still minimax optimal, with respect to squared loss, among all shrinkage estimators. This result also holds for all three shrinkage variants.

Next, we study how uniform the improvement of the MSE by shrinkage is by studying the GMSE. We find that, for individual estimated probabilities per category, the same result holds: A little shrinkage is always better than no shrinkage. This again holds for all three estimators. However, for general linear combinations of these probabilities, the three ridge methods start to diverge: Whereas the method that shrinks to 1/2 has at least some improvement for all possible linear combinations, the other two methods will always (aside from some special cases) have increased MSE compared to MLE for at least one linear combination.

We first derive explicit expressions for the three ridge estimators. Their properties will be analyzed by second-order moment matrices of the error. After the theoretical results, we have a numerical experiment. The value of the results in data analysis will be illustrated with a real data set.

2 | MODEL AND ASSUMPTIONS

Let $Y = \{0, 1\}$ and $X = \{1, \ldots, K\}$. We observe n independent realizations of pairs (Y_i, X_i), denoted as (y_i, x_i), where $Y_i \sim \text{Ber}(\pi_i)$ and $\pi_i = P(Y_i = 1|X_i)$, with Ber$(\cdot)$ denoting a Bernoulli distribution. We assume the logistic regression model

$$\log \frac{\pi_i}{1-\pi_i} = \beta_1 z_{i1} + \cdots + \beta_K z_{iK}, \quad i = 1, \ldots, n, \tag{1}$$

where z_{ik} denotes the dummy variable associated with the kth category, $k = 1, \ldots, K$; that is, $z_{ik} = 1$ if $x_i = k$ and $z_{ik} = 0$ otherwise. Observe that

$$\mathbf{z}_1 + \cdots + \mathbf{z}_K = \mathbf{1}; \tag{2}$$

$$\mathbf{z}_j^T \mathbf{z}_k = 0, \text{ for any } j \neq k, \tag{3}$$

where $\mathbf{z}_k = (z_{1k}, \ldots, z_{nk})^T$ and $\mathbf{1}$ is the identity vector of order n.

The likelihood is then simply

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i},$$

and we denote its natural logarithm by $l(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$. We assume we have only one categorical predictor, so our data come from a $K \times 2$ contingency table.

		y		
		0	1	
x	1	a_{10}	a_{11}	$a_{1\bullet}$
	...	...	...	...
	k	a_{k0}	a_{k1}	$a_{k\bullet}$
	...	...	...	...
	K	a_{K0}	a_{K1}	$a_{K\bullet}$
		$a_{\bullet 0}$	$a_{\bullet 1}$	n

If we would have several categorical predictors and full interactions, we would get the same type of data. For this type of data, using the parameterization given in (1), the event probability for the kth category is given by

$$\pi_k = \frac{1}{1 + \exp(-\beta_k)}, \quad k = 1, \ldots, K.$$

Two further assumptions will be made throughout.

Assumption 1. $0 < a_{k\bullet} < \infty$ for $k = 1, \ldots, K$.

Assumption 2. $0 < \pi_k < 1$ for $k = 1, \ldots, K$.

For brevity, the proofs of our theoretical results are given in the Supplementary Material. In general, we will only be interested in the estimates of the probabilities and not in those of the logistic regression parameters. Under Assumption 1, the MLE of the probabilities exists and is given by $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_K)^T$, where

$$\hat{\pi}_k = \frac{a_{k1}}{a_{k\bullet}}, k = 1, \ldots, K.$$

3 | THREE RIDGE ESTIMATORS

We now define three alternative ways of doing ridge regression in model (1). These are not all possibilities, but they are the ones we have found in the literature as actually being used. In general, ridge regression is defined as maximizing a log-likelihood minus a penalty, $P(\lambda)$, given by a sum of squares of some or all of the parameters, where the amount of penalty is determined by the penalty parameter, $\lambda > 0$. The estimator is not invariant to the parameterization chosen. We will derive the three estimators from different natural parameterizations.

If we choose the parameterization in (1) and apply the ridge penalty on this, that is,

$$P(\lambda) = \frac{\lambda}{2}\sum_{k=1}^{K}\beta_k^2 = \frac{\lambda}{2}\sum_{k=1}^{K}\log^2\left[\frac{\pi_k}{1-\pi_k}\right],$$

we obtain, using a one-step solution initializing at 1/2, the following estimators (Estimator 1): $\hat{\boldsymbol{\pi}}_1(\lambda) = (\hat{\pi}_{1,1}(\lambda), \ldots, \hat{\pi}_{1,K}(\lambda))^T$, where

$$\hat{\pi}_{1,k}(\lambda) = \frac{a_{k1} + 2\lambda}{a_{k\bullet} + 4\lambda}, \quad k = 1, \ldots, K. \tag{4}$$

We say that the shrinkage target of this estimator is 1/2 since the estimators go to 1/2 as $\lambda \to \infty$. Note that this approximate one-step estimator is more accurate when the sample size is large and/or when the targeted parameter vector is close to the starting point (see Supplementary Material for the upper bound on the error [and its propagation] made by using the one-step approximation). Ridge regression is closely related to Bayesian analysis. Consider penalizing the likelihood by the generalized Jeffreys invariant prior (Elgmati et al., 2015), where the penalized log-likelihood is given by

$$l^{P_F} = l + 2\lambda \log |I(\boldsymbol{\beta})|,$$

where $|I(\boldsymbol{\beta})|$ is the determinant of the observed information. In our setting, it can be easily seen that

$$\log |I(\boldsymbol{\beta})| = \sum_{k=1}^{K} \log \left(\pi_k - \pi_k^2\right);$$

hence, it is straightforward to show that the event probability estimates obtained when maximizing l^{P_F} are the same as in Equation (4). Note that the same solution would be obtained by adding 2λ to each cell of the contingency table and then maximizing the (unpenalized) likelihood function. Setting $\lambda = 1/4$ is the Jeffreys invariant prior used by Firth (1993; FPE). Furthermore, for $K = 1$, $\hat{\pi}_{1,1}(1)$ is the ACE (Agresti & Coull, 1998), whereas $\hat{\pi}_{1,1}(\sqrt{a_{1\bullet}}/4)$ is the unique minimax estimator with respect to $\mathcal{P} = B(\sqrt{a_{1\bullet}}/2, \sqrt{a_{1\bullet}}/2)$, where $\mathcal{P}$ is a prior distribution of π_1 and $B(\cdot)$ is a beta distribution (Bayesian minimax estimator [BE]).

Rather than shrinking toward 1/2, it may be sensible to shrink toward a common mean probability. One way to do this is to consider the alternative parameterization with a reference category, that is,

$$\log \frac{\pi_i}{1-\pi_i} = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_{K-1} z_{iK-1}, \quad i = 1, \ldots, n.$$

It follows from (2) and (3) that

$$\gamma_0 = \beta_K, \quad \gamma_k = \beta_k - \beta_K, \quad k = 1, \ldots, K-1.$$

Applying a ridge penalty to all parameters except the intercept, that is,

$$P(\lambda) = \frac{\lambda}{2}\sum_{k=1}^{K-1}\gamma_k^2 = \frac{\lambda}{2}\sum_{k=1}^{K-1}\log^2\left(\frac{\pi_k}{1-\pi_k}\frac{1-\pi_K}{\pi_K}\right) = \frac{\lambda}{2}\sum_{k=1}^{K-1}(\beta_k - \beta_K)^2,$$

results, again using a one-step solution initializing at 1/2, in the following estimators (Estimator 2): $\hat{\boldsymbol{\pi}}_2(\lambda) = (\hat{\pi}_{2,1}(\lambda), \ldots, \hat{\pi}_{2,K}(\lambda))^T$, where

$$\hat{\pi}_{2,k}(\lambda) = \frac{a_{k1} + 4\lambda\hat{\pi}_{2,K}(\lambda)}{a_{k\bullet} + 4\lambda}, \quad k = 1, \ldots, K-1,$$

$$\hat{\pi}_{2,K}(\lambda) = \frac{a_{K1} + 4\lambda\sum_{j=1}^{K-1}\frac{a_{j1}}{a_{j\bullet}+4\lambda}}{a_{K\bullet} + 4\lambda\sum_{j=1}^{K-1}\frac{a_{j\bullet}}{a_{j\bullet}+4\lambda}}.$$

As $\lambda \to \infty$, we converge to a common mean, so we can say that this is the shrinkage target. However, the shrinkage is not invariant to the choice of the reference category. Shrinkage is

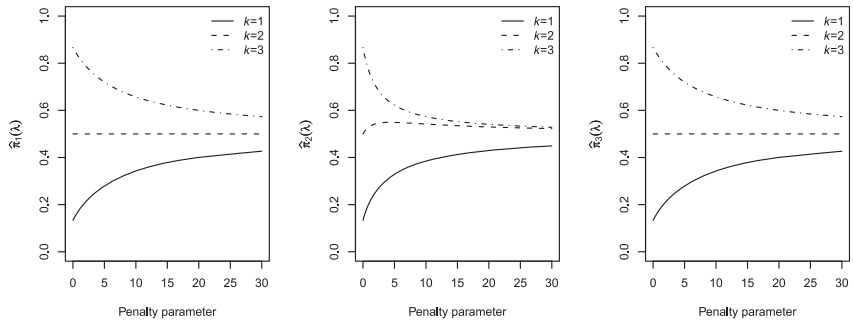

FIGURE 1 The event probability estimates for each group ($k = 1, 2, 3$) obtained by different ridge estimators as a function of the penalty parameter

uneven, with the reference category experiencing weaker shrinkage toward the common mean than the other categories. Moreover, shrinkage is not necessarily monotone: Categories may move away from the common mean for small λ, only to return to it when λ is large, as we will see in the example below.

An alternative, more monotone way of shrinkage toward a common mean was proposed by Goeman et al. (2014). It does not involve a reference category and is symmetric in the categories. It can be constructed using the following overparameterized model:

$$\log \frac{\pi_i}{1-\pi_i} = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_K z_{iK}, \quad i = 1, \ldots, n,$$

with the ridge penalty again applying on all parameters except the intercept. Equivalently, the same estimator may be derived from the previous parameterization with an empty $(K+1)$th category. The resulting estimators are as follows (Estimator 3): $\hat{\boldsymbol{\pi}}_3(\lambda) = (\hat{\pi}_{3,1}(\lambda), \ldots, \hat{\pi}_{3,K}(\lambda))^T$, where

$$\hat{\pi}_{3,k}(\lambda) = \frac{a_{k1} + 4\lambda\bar{\pi}(\lambda)}{a_{k\bullet} + 4\lambda}, \quad k = 1, \ldots, K,$$
$$\bar{\pi}(\lambda) = \frac{\sum_{j=1}^{K} \frac{a_{j1}}{a_{j\bullet}+4\lambda}}{\sum_{j=1}^{K} \frac{a_{j\bullet}}{a_{j\bullet}+4\lambda}}.$$

Observe that, for $K = 1$, $\hat{\pi}_{2,1}(\lambda) = \hat{\pi}_{3,1}(\lambda) = \hat{\pi}_1$. The shrinkage introduced by the ridge estimators is illustrated by using the following example, where $K = 3$, $a_{k\bullet} = 30$, $k = 1, 2, 3$, and $a_{11} = 4$, $a_{21} = 15$, and $a_{31} = 26$ (see Figure 1). Observe that, since the common mean is 1/2, $\hat{\boldsymbol{\pi}}_1(\lambda)$ and $\hat{\boldsymbol{\pi}}_3(\lambda)$ perform similarly, where the second group does not experience any shrinkage (since its MLE is equal to 1/2) and the two other groups experience an equal amount of shrinkage. In contrast, when using $\hat{\boldsymbol{\pi}}_2(\lambda)$, the shrinkage is not symmetric, with the reference group ($k = 3$) experiencing less shrinkage; note that now the event probability of the second group is also affected by the penalty parameter.

We conclude this section by remarking that, assuming that $\lim_{a_{k\bullet}\to\infty} = \frac{a_{k1}}{a_{k\bullet}}$ exists and is equal to some finite constant c_k, $k = 1, \ldots, K$, then in the limit when $a_{k\bullet} \to \infty$, we have, for any $0 \leq \lambda < \infty$,

$$\lim_{a_{k\bullet}\to\infty} \hat{\pi}_k = \lim_{a_{k\bullet}\to\infty} \hat{\pi}_{1,k}(\lambda) = \lim_{a_{k\bullet}\to\infty} \hat{\pi}_{2,k}(\lambda) = \lim_{a_{k\bullet}\to\infty} \hat{\pi}_{3,k}(\lambda) = c_k, \quad k = 1, \ldots, K.$$

4 | OPTIMALITY OF RIDGE ESTIMATORS

To analyze the MSE of the three estimators, we will do this in terms of the second-order moment matrices. The second-order moment matrix of an estimator of a vector parameter $\boldsymbol{\theta}$, for example, $\hat{\boldsymbol{\theta}}$, is defined in general as

$$\mathbf{M} = E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T.$$

The matrix is useful because it relates to differences in the MSE and the GMSE. We have

$$\text{MSE} = \text{trace}(\mathbf{M}),$$

which we can use to establish differences in the MSE by the difference in the trace of $\mathbf{M}$.

First, we prove that, for every value of the true parameters, the MLE can be improved by a suitable shrinkage estimator.

Theorem 1. *Assume that* $0 < \pi_k < 1$, $k = 1, \ldots, K$, *and* $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, *and* $K \geq 2$ *if* $j \neq 1$. *Then, there exists some* $\Lambda > 0$ *such that, for all* $0 < \lambda < \Lambda$, *we have*

$$\text{MSE}\left(\hat{\boldsymbol{\pi}}_j(0)\right) > \text{MSE}\left(\hat{\boldsymbol{\pi}}_j(\lambda)\right), \quad j = 1, 2, 3.$$

Remark 1. Let $K \geq 1$, and assume that $a_{k\bullet} = a$, $k = 1, \ldots, K$. Let $\pi_k \neq 1/2$ for some $k = 1, \ldots, K$. Then, $\text{MSE}(\hat{\boldsymbol{\pi}}_1(\lambda))$ is minimized at

$$\lambda = \frac{\sum_{k=1}^{K}\left(\pi_k - \pi_k^2\right)}{\sum_{k=1}^{K}(1 - 2\pi_k)^2}.$$

This implies that, for all three estimators, there exists a $\Lambda(\boldsymbol{\pi})$ so that, for $0 < \lambda < \Lambda(\boldsymbol{\pi})$, the shrinkage estimator outperforms the MLE. For particular cases, for example, $\pi_k = 1/2$ for all k for Estimator 1 and $\pi_1 = \cdots = \pi_K$ for Estimators 2 and 3, we even have $\Lambda(\boldsymbol{\pi}) = \infty$.

We investigate $\Lambda(\boldsymbol{\pi})$ for three interesting special cases of Estimator 1, namely, FPE, ACE, and BE. In general, we have the following result.

Lemma 1. *Assume that* $0 < \pi_k < 1$, $k = 1, \ldots, K$, *and* $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, *are fixed. Then, for some* $\gamma \neq \lambda$,

$$\text{MSE}(\hat{\boldsymbol{\pi}}_1(\gamma)) \geq \text{MSE}(\hat{\boldsymbol{\pi}}_1(\lambda)),$$

if and only if

$$\sum_{k=1}^{K} \frac{a_{k\bullet}\left(\pi_k - \pi_k^2\right)\left((a_{k\bullet} + 4\lambda)^2 - (a_{k\bullet} + 4\gamma)^2\right) + 4(1 - 2\pi_k)^2\left(\gamma^2(a_{k\bullet} + 4\lambda)^2 - \lambda^2(a_{k\bullet} + 4\gamma)^2\right)}{(a_{k\bullet} + 4\gamma)^2(a_{k\bullet} + 4\lambda)^2} \geq 0.$$

Applying this to the FPE, we get the following corollary.

Corollary 1. *Assume that* $0 < \pi_k < 1$, $k = 1, \ldots, K$, *and* $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, *are fixed. Then,*

$$\text{MSE}(\hat{\boldsymbol{\pi}}1(0)) \geq \text{MSE}(\hat{\boldsymbol{\pi}}_1(1/4)),$$

if and only if

$$\sum_{k=1}^{K} \frac{4\pi_k(1 - \pi_k)(3a_{k\bullet} + 1) - a_{k\bullet}}{a_{k\bullet}(a_{k\bullet} + 1)^2} \geq 0 \tag{5}$$

holds.

Remark 2.

(a) Equation (5) holds if $0.092 \approx \frac{1}{2}(1 - \frac{\sqrt{6}}{3}) \leq \pi_k \leq \frac{1}{2}(1 + \frac{\sqrt{6}}{3}) \approx 0.908$ holds for $k = 1, \ldots, K$.
(b) Assume that $a_{k\bullet} = a$, $k = 1, \ldots, K$. Then, (5) becomes

$$\frac{1}{K}\sum_{k=1}^{K}\left(\pi_k - \pi_k^2\right) \geq \frac{a}{4(3a+1)}.$$

For $a = 1$, the right-hand side (RHS) of the above expression becomes $1/16$, whereas in the limit when $a \to \infty$, the RHS is equal to $1/12$; note also that the RHS of the above expression is an increasing function of a.

The corollary establishes conditions under which the particular value of the penalty parameter used by Firth, namely, $\lambda = 1/4$, will outperform the MLE. That is, when, for each category, the true event probabilities lie within a bound that is symmetric around 1/2, then $\hat{\boldsymbol{\pi}}_1(1/4)$ will outperform the MLE in terms of the MSE. This suggests that when the events are rare (or common), the FPE will not outperform, in terms of the MSE, the MLE. Moreover, in studies with an equal number of subjects in each category, the FPE outperforms the MLE only when the average of the variances over the categories exceeds some bound, which depends on the sample size (the bound increases with sample size). With a small sample size, the variances have to be larger in order to see the improvement over the MLE in terms of the MSE.

The following result establishes conditions under which $\lambda = 1/4$ is optimal in terms of the MSE.

Proposition 1. *Assume that* $0 < \pi_k < 1$, $k = 1, \ldots, K$, *and* $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, *are fixed. Then,*

$$\operatorname{argmin}_{\lambda}\mathrm{MSE}(\hat{\boldsymbol{\pi}}_1(\lambda)) = 1/4,$$

if and only if

$$\sum_{k=1}^{K}\frac{a_{k\bullet}}{(a_{k\bullet}+1)^3}\left(\frac{1}{8} - \left(\pi_k - \pi_k^2\right)\right) = 0. \tag{6}$$

Remark 3.

(a) Let $K > (=)1$, then (6) holds if and only if

$$\pi_k = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right) \approx 0.146 \text{ or } \pi_k = \frac{1}{2}\left(1 + \frac{1}{\sqrt{2}}\right) \approx 0.854, \quad k = 1, \ldots, K.$$

(b) Assuming that $a_{k\bullet} = a$, $k = 1, \ldots, K$, the condition becomes

$$\frac{1}{K}\sum_{k=1}^{K}\left(\pi_k - \pi_k^2\right) = \frac{1}{8}.$$

Applying Lemma 1 to the ACE, we get the following corollary.

Corollary 2. *Let* $K = 1$. *Assume that* $0 < \pi_1 < 1$ *and that* $0 < a_{1\bullet} < \infty$ *is fixed. Then,*

$$\mathrm{MSE}(\hat{\pi}_{1,1}(0)) \geq \mathrm{MSE}(\hat{\pi}_{1,1}(1)),$$

if and only if

$$\frac{1}{2}\left(1-\sqrt{\frac{a_{1\bullet}+2}{3a_{1\bullet}+2}}\right) \leq \pi_1 \leq \frac{1}{2}\left(1+\sqrt{\frac{a_{1\bullet}+2}{3a_{1\bullet}+2}}\right)$$

holds.

Remark 4.

(a) Observe that

$$\lim_{a_{1\bullet}\to\infty} \frac{a_{1\bullet}+2}{3a_{1\bullet}+2} = \frac{1}{3},$$

thence, in the limit when $a_{1\bullet} \to \infty$, the above condition becomes

$$0.211 \approx \frac{1}{2}\left(1-\sqrt{\frac{1}{3}}\right) \leq \pi_1 \leq \frac{1}{2}\left(1-\sqrt{\frac{1}{3}}\right) \approx 0.789,$$

whereas for $a_{1\bullet} = 1$, the condition is

$$0.113 \approx \frac{1}{2}\left(1-\sqrt{\frac{3}{5}}\right) \leq \pi_1 \leq \frac{1}{2}\left(1-\sqrt{\frac{3}{5}}\right) \approx 0.887.$$

(For $1 < a_{1\bullet} < \infty$, the lower and upper bounds are somewhere in between the two extreme cases.)

(b) If $K = 1$, $\mathrm{MSE}(\hat{\pi}_{1,1}(1)) < \mathrm{MSE}(\hat{\pi}_{1,1}(\lambda))$, for all $\lambda > 0$, $\lambda \neq 1$, holds if and only if

$$\pi_1 = \frac{1}{2}\left(1-\frac{\sqrt{5}}{5}\right) \approx 0.276 \text{ or } \pi_1 = \frac{1}{2}\left(1+\frac{\sqrt{5}}{5}\right) \approx 0.724.$$

Corollary 2 established conditions under which the ACE outperforms the MLE and when the particular value of the penalty parameter used in the ACE is optimal in terms of the MSE. In terms of optimality, the ACE is optimal only for two particular values of the true event probability, whereas it can outperform the MLE only for some values of the true event probability and is hence suboptimal in the scenario where events are rare (or common).

Applying Lemma 1 to the BE, we get the following corollary.

Corollary 3. *Let* $K = 1$. *Assume that* $0 < \pi_1 < 1$ *and that* $0 < a_{1\bullet} < \infty$ *is fixed. Then,*

$$\mathrm{MSE}\left(\hat{\pi}_{1,1}(0)\right) \geq \mathrm{MSE}\left(\hat{\pi}_{1,1}\left(\sqrt{a_{1\bullet}}/4\right)\right),$$

if and only if

$$\frac{1}{2}\left(1-\sqrt{\frac{2\sqrt{a_{1\bullet}}+1}{2\sqrt{a_{1\bullet}}+1+a_{1\bullet}}}\right) \leq \pi_1 \leq \frac{1}{2}\left(1+\sqrt{\frac{2\sqrt{a_{1\bullet}}+1}{2\sqrt{a_{1\bullet}}+1+a_{1\bullet}}}\right)$$

holds.

Remark 5. For $a_{1\bullet} = 1$, the condition becomes

$$0.067 \approx \frac{1}{2}\left(1-\frac{\sqrt{3}}{2}\right) \leq \pi_1 \leq \frac{1}{2}\left(1+\frac{\sqrt{3}}{2}\right) \approx 0.933,$$

whereas in the limit when $a_{1\bullet} \to \infty$, the condition becomes $\pi_1 = 1/2$.

By Corollary 3, the width of the bound on the values of π_1 for which the BE outperforms the MLE reduces with a larger sample size. This is not surprising since the prior distribution of π_1 assumed by the BE depends on the sample size, where, with a larger sample size, the prior is stronger (through the fact that the variance of the prior distribution is reduced when increasing the sample size), implying more shrinkage toward 1/2 and, hence, the possibility of overshrinking when π_1 is not equal to 1/2. With a smaller sample size, the prior is weaker, and hence, the amount of shrinkage toward 1/2 is smaller, thus increasing the width of the interval for which the BE does not overshrink.

5 | OPTIMALITY OF THE MLE

It may seem from Theorem 1 that the MLE is inadmissible since it is optimal for no value of the true parameter vector, $\boldsymbol{\pi}$. However, this thought is too simple. As we have seen in the examples of the FPE, ACE, and BE, any shrinkage estimator may also overshrink, resulting in an MSE that is larger than that of the MLE. We make this explicit in the following lemma.

Lemma 2. *Assume that* $0 < \pi_k < 1$, $k = 1, \ldots, K$, *and* $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, *are fixed. Then, for every* $\lambda > 0$, *there exists some* $\boldsymbol{\pi}$ *such that*

$$\mathrm{MSE}\left(\hat{\boldsymbol{\pi}}_j(0)\right) < \mathrm{MSE}\left(\hat{\boldsymbol{\pi}}_j(\lambda)\right), \quad j = 1, 2, 3.$$

The lemma says that however little we shrink, there is always a value of the true parameter vector, $\boldsymbol{\pi}$, for which we are overshrinking. A direct consequence of this is that the MLE, which is optimal for no fixed $\boldsymbol{\pi}$, has the best minimax risk, with respect to the MSE, among all fixed shrinkage methods, which is formally given in the next theorem.

Theorem 2. *Assume that* $0 < \pi_k < 1$, $k = 1, \ldots, K$, *and* $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, *are fixed. Then, the MLE is minimax, with respect to the* MSE, *among the family* $\hat{\boldsymbol{\pi}}_j(\lambda)$ *for* $j = 1, 2, 3$.

Remark 6. The claim is only among fixed λ estimators. Adaptive estimators with data-dependent λ might uniformly improve the MLE, but it should be remembered that estimating λ also induces variability.

6 | MOMENT MATRICES AND THE GMSE

Define the GMSE as

$$\mathrm{GMSE}_j(\lambda, \mathbf{B}) = E\left(\hat{\boldsymbol{\pi}}_j(\lambda) - \boldsymbol{\pi}\right)^T \mathbf{B}\left(\hat{\boldsymbol{\pi}}_j(\lambda) - \boldsymbol{\pi}\right), \quad j = 1, 2, 3,$$

for some nonzero positive semidefinite matrix $\mathbf{B}$. (For brevity, we suppress the notation and use $\mathrm{GMSE}_j(\lambda) = \mathrm{GMSE}_j(\lambda, \mathbf{B})$ throughout.) It is a useful quantity because it allows a weighted assessment of the MSE of the parameters. The MSE in the previous section can be seen as the average MSE. The GMSE allows us to look at individual parameters or weighted combinations. For example, we can look at the MSE of π_k (only) by taking $\mathbf{B}$ with $B_{k,k} = 1$ and $B_{i,j} = 0$ elsewhere.

Importantly, we can also look at $\mathbf{M}$ to compare the GMSE. We use the following result (Theobald, 1974).

Lemma 3. *Let there be two estimators of a vector parameter $\boldsymbol{\theta}$, for example, $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ with the respective second-order moment matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ and with the error measure given by*

$$\mathrm{GMSE}_j(\mathbf{B}) = E\left(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}\right)^T \mathbf{B}\left(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}\right), \quad j = 1, 2.$$

Then, we have

$$\mathrm{GMSE}_1(\mathbf{B}) > (\geq)\mathrm{GMSE}_2(\mathbf{B}),$$

for all positive semidefinite $\mathbf{B}$ if and only if $\mathbf{M}_1 - \mathbf{M}_2$ is positive (semi)definite.

We will now analyze the three estimators in terms of their GMSE. Interestingly, we will start to see divergent behavior.

Theorem 3. *Assume that $0 < \pi_k < 1$, $k = 1, \ldots, K$, and $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, are fixed. Then, there exists some $\lambda > 0$ such that $\mathbf{M}_1(0) - \mathbf{M}_1(\lambda)$ is positive definite.*

For the FPE, the following result applies.

Proposition 2. *Assume that $0 < \pi_k < 1$, $k = 1, \ldots, K$, and $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, are fixed. Then, we have*

$$\mathrm{GMSE}(\hat{\boldsymbol{\pi}}_1(0)) \geq \mathrm{GMSE}(\hat{\boldsymbol{\pi}}_1(1/4)),$$

if and only if

$$1 - \frac{1}{8}\sum_{k=1}^{K} \frac{a_{k\bullet}(1 - 2\pi_k)^2}{\pi_k(1 - \pi_k)(a_{k\bullet} + 1/2)} \geq 0$$

holds.

Remark 7. Assume that $a_{k\bullet} = a$, $k = 1, \ldots, K$. The condition then becomes

$$\sum_{k=1}^{K} \frac{(1 - 2\pi_k)^2}{\pi_k - \pi_k^2} \leq \frac{4(2a + 1)}{a}.$$

For $a = 1$, the RHS of the above expression becomes 12, whereas in the limit when $a \to \infty$, the RHS is equal to 8; note also that the RHS of the above expression is a decreasing function of a.

In comparison with the results concerning the MSE, Proposition 2 implies that even more emphasis is given to the categories with rare (or common) events (e.g., through the added term $(1 - 2\pi_k)^2$). Hence, when the events are rare (or common), the value of the penalty parameter used by Firth is less likely to outperform the MLE in terms of the GMSE as in terms of the MSE, suggesting tighter bounds on π_k for which the FPE outperforms the MLE. Similarly, as is the case for the MSE, the bound is tighter with a larger sample size, and in this case, it is affected also by the number of categories, where more categories imply a tighter bound.

Given the unpleasant form of $\hat{\boldsymbol{\pi}}_2(\lambda)$ and $\hat{\boldsymbol{\pi}}_3(\lambda)$ for a general K, first assume that $K = 2$, where

$$\hat{\pi}_{2,k}(\lambda) = \frac{a_{j\bullet}a_{k1} + 4\lambda a_{\bullet 1}}{a_{k\bullet}a_{j\bullet} + 4\lambda n}, \quad k = 1, 2 \text{ and } j \neq k,$$

$$\hat{\pi}_{3,k}(\lambda) = \frac{a_{j\bullet}a_{k1} + 2\lambda a_{\bullet 1}}{a_{k\bullet}a_{j\bullet} + 2\lambda n}, \quad k = 1, 2 \text{ and } j \neq k.$$

Observe that

$$\hat{\pi}_{2,k}(\lambda) = \hat{\pi}_{3,k}(2\lambda), \quad k = 1, \ldots, K. \tag{7}$$

(Note however that this relation does not hold for $K > 2$.) Then, we have the following result.

Theorem 4. *Let $K = 2$, and let $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$ denote the difference between the second-order moment matrices of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_j(\lambda)$, $j = 2, 3$, respectively. Assume that $0 < \pi_1, \pi_2 < 1$ and that $0 < a_{1\bullet}, a_{2\bullet} < \infty$ are fixed. Then, we have the following exhaustive and mutually exclusive situations.*

(i) Let $\pi_1 = \pi_2$. Then, for $\lambda > 0$, $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$, $j = 2, 3$, is positive semidefinite.
(ii) Let $\pi = \pi_1 = 1 - \pi_2$, $\pi_1 \neq \pi_2$.
  (a) If $n\pi(1-\pi) - a_{1\bullet}a_{2\bullet}(2\pi - 1)^2 \geq 0$, then, for $\lambda > 0$, $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$, $j = 2, 3$, is positive semidefinite.
  (b) If $n\pi(1-\pi) - a_{1\bullet}a_{2\bullet}(2\pi - 1)^2 < 0$, there exists some $\Lambda > 0$ such that $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$, $j = 2, 3$, is positive semidefinite whenever $0 < \lambda < \Lambda$; $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$, $j = 2, 3$, is negative semidefinite whenever $\lambda > \Lambda$; and $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$, $j = 2, 3$, is a null matrix for $\lambda = \Lambda$.
(iii) Let $\pi_1 \neq \pi_2$ and $\pi_1 \neq 1 - \pi_2$. Then, for $\lambda > 0$, $\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)$, $j = 2, 3$, is indefinite.

Corollary 4. *Assume the conditions and notation of Theorem 4, Case (ii),(b). Then,*

$$\Lambda_2 = \frac{a_{1\bullet}a_{2\bullet}(\pi - \pi^2)}{2(a_{1\bullet}a_{2\bullet}(2\pi - 1)^2 - n(\pi - \pi^2))} \text{and} \Lambda_3 = \frac{a_{1\bullet}a_{2\bullet}(\pi - \pi^2)}{a_{1\bullet}a_{2\bullet}(2\pi - 1)^2 - n(\pi - \pi^2)},$$

for $\hat{\boldsymbol{\pi}}_2(\lambda)$ and $\hat{\boldsymbol{\pi}}_3(\lambda)$, respectively.

For $K > 2$, we have the following result for $\hat{\boldsymbol{\pi}}_2(\lambda)$.

Theorem 5. *Let $K > 2$, and let $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ denote the difference between the second-order moment matrices of $\hat{\boldsymbol{\pi}}_2(0)$ and $\hat{\boldsymbol{\pi}}_2(\lambda)$, respectively. Assume that $0 < \pi_k < 1$, $k = 1, \ldots, K$, and that $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, are fixed. Then, we have the following exhaustive and mutually exclusive situations.*

(i) Let $\pi_k = \pi$, $k = 1, \ldots, K$. Then, for $\lambda > 0$, $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is positive semidefinite.
(ii) Let $\pi_1(1 - \pi_1) = \cdots = \pi_K(1 - \pi_K)$, and there is some k such that $\pi_k \neq \pi_K$. Then, there exists some $\Lambda > 0$ such that, for $0 < \lambda < \Lambda$, $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is positive semidefinite.
(iii) Let $\pi_k \neq \pi_K$ and $\pi_k \neq 1 - \pi_K$ for some k. Then, for $\lambda > 0$, $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is indefinite.

For $\hat{\boldsymbol{\pi}}_3(\lambda)$, the following result applies.

Theorem 6. *Let $K > 2$, and let $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ denote the difference between the second-order moment matrices of $\hat{\boldsymbol{\pi}}_3(0)$ and $\hat{\boldsymbol{\pi}}_3(\lambda)$, respectively. Assume that $0 < \pi_k < 1$, $k = 1, \ldots, K$, and that $0 < a_{k\bullet} < \infty$, $k = 1, \ldots, K$, are fixed. Then, we have the following exhaustive and mutually exclusive situations.*

(i) Let $\pi_k = \pi$, $k = 1, \ldots, K$. Then, for $\lambda > 0$, $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ is positive semidefinite.
(ii) Let $\pi_1(1 - \pi_1) = \cdots = \pi_K(1 - \pi_K)$, and there is some pair $k \neq j$ such that $\pi_k \neq \pi_j$. Then, there exists some $\Lambda > 0$ such that, for $0 < \lambda < \Lambda$, $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ is positive semidefinite.
(iii) Let $\pi_k \neq \pi_l$ and $\pi_k \neq 1 - \pi_l$ for some pair $k \neq l$. Then, for $\lambda > 0$, $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ is indefinite.

For the variants of ridge regression that shrink toward a common mean, we now know that there are error measures for which the MLE is always better, whereas this is not observed for the naive estimator. Note that the $\mathbf{B}$ that defines the GMSE measure can depend on $\boldsymbol{\pi}$. However, not all GMSE measures are interesting. We are mostly interested in the individual $\hat{\pi}$ estimates for the categories. The following theorem shows that it is not in the MSE of these individual estimators where the problem lies.

Theorem 7. *Let* $\mathbf{B}$*, nonzero, be a diagonal positive semidefinite matrix. Assume that* $0 < \pi_k < 1$*,* $k = 1, \ldots, K$*, and that* $0 < a_{k\bullet} < \infty$*,* $k = 1, \ldots, K$*, are fixed. Then, for some* $\lambda > 0$*, we have*

$$\mathrm{GMSE}_j(0) > \mathrm{GMSE}_j(\lambda), \quad j = 1, 2, 3.$$

Remark 8. Setting one and only one diagonal element of $\mathbf{B}$ to 1 gives the MSE for a particular category. The result implies that, for some $\lambda > 0$, the MSE of each category is reduced. Obviously, the value of λ at which the MSE of each particular category is minimized will vary for different categories (unless when $a_{k\bullet} = a$ and $\pi_k = \pi$ hold for $k = 1, \ldots, K$).

7 | A NUMERICAL EXAMPLE

To illustrate the theoretical results, we calculated the relative performance (RP) of the derived estimators in comparison with the MLE, defined as

$$\mathrm{RP} = \frac{\mathrm{GMSE}\left(\hat{\boldsymbol{\pi}}_j(0)\right) - \mathrm{GMSE}\left(\hat{\boldsymbol{\pi}}_j(\lambda)\right)}{\mathrm{GMSE}\left(\hat{\boldsymbol{\pi}}_j(0)\right)} \cdot 100\%, j = 1, 2, 3,$$

for different values of the penalty parameter ($\lambda \in [0, 2]$) for an example where $K = 3$ and $a_{1\bullet} = 10$, $a_{2\bullet} = 20$, and $a_{3\bullet} = 30$ for four different scenarios: (a) $\pi_1 = \pi_2 = \pi_3 = 0.2$; (b) $\pi_1 = \pi_3 = 0.2$, $\pi_2 = 0.8$; (c) $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.9$; and (d) $\pi_1 = 0.2$, $\pi_2 = 0.3$, $\pi_3 = 0.4$. Matrix $\mathbf{B}$ in the calculation of the GMSE was set to $\mathbf{B} = \mathrm{diag}(1,1,1)$ (equivalent to comparing the overall MSEs), $\mathbf{B} = \mathrm{diag}(1,0,0)$ (equivalent to comparing the MSEs of the first category), $\mathbf{B} = \mathrm{diag}(0,1,0)$ (equivalent to comparing the MSEs of the second category), and $\mathbf{B} = \mathrm{diag}(0,0,1)$ (equivalent to comparing the MSEs of the third category). The third category was used as a reference for $\hat{\boldsymbol{\pi}}_2(\lambda)$.

As suggested by the theoretical results, the overall risk of the MLE and the risks for each individual category based on squared loss are reduced by the three shrinkage estimators for some $\lambda > 0$ (see Figure 2). The value of the penalty parameter where the largest improvement over the MLE was observed depended strongly on the scenario; however, in none of the scenarios was the value of 1/4 (Jeffreys invariant prior) optimal for either the overall or the individual risks. Moreover, the largest relative improvement was obtained at different values of the penalty parameter when considering the overall and each individual risk based on squared loss. Therefore, it was possible (depending on the scenario) that the value of the penalty parameter that yielded the largest overall gain could have led to diminished results for some categories (see Figure 2).

The following example is used to illustrate the practical implication of Theorems 3, 5, 6, and 7. Let $K = 3$, $a_{1\bullet} = 10$ and $a_{3\bullet} = a_{2\bullet} = 50$ and $\pi_1 = 0.01$ and $\pi_2 = \pi_3 = 0.5$. Additionally to matrices $\mathbf{B}$ evaluating the overall MSE and the MSE of each category as in the previous examples, let $\mathbf{B}$ also be a 3×3 matrix of ones. This definition of $\mathbf{B}$ is potentially interesting since it gives a comparison of the sum of all elements of the second-order moment matrices. The RP for this example for the three estimators is shown in Figure 3.

Consistent with Theorems 5 (Case (iii)) and 6 (Case (iii)), when $\mathbf{B}$ is a matrix of ones, even for a minute amount of shrinkage, the estimators $\hat{\boldsymbol{\pi}}_2(\lambda)$ and $\hat{\boldsymbol{\pi}}_3(\lambda)$ increase the error of the MLE, which does not occur for $\hat{\boldsymbol{\pi}}_1(\lambda)$, where, for some $\lambda > 0$, the GMSE of the MLE is decreased (see Figure 3). Similarly as in the previous examples and consistent with Theorems 3 and 7, we see some improvement over the MLE for some $\lambda > 0$ when using other definitions of matrix $\mathbf{B}$. (Observe that, in this example, $\mathbf{M}_1(0) - \mathbf{M}_1(\lambda)$ is diagonal; hence, for $\hat{\boldsymbol{\pi}}_1(\lambda)$, the results when $\mathbf{B}$ is the matrix of ones and the identity matrix are the same, which holds also for the results for the

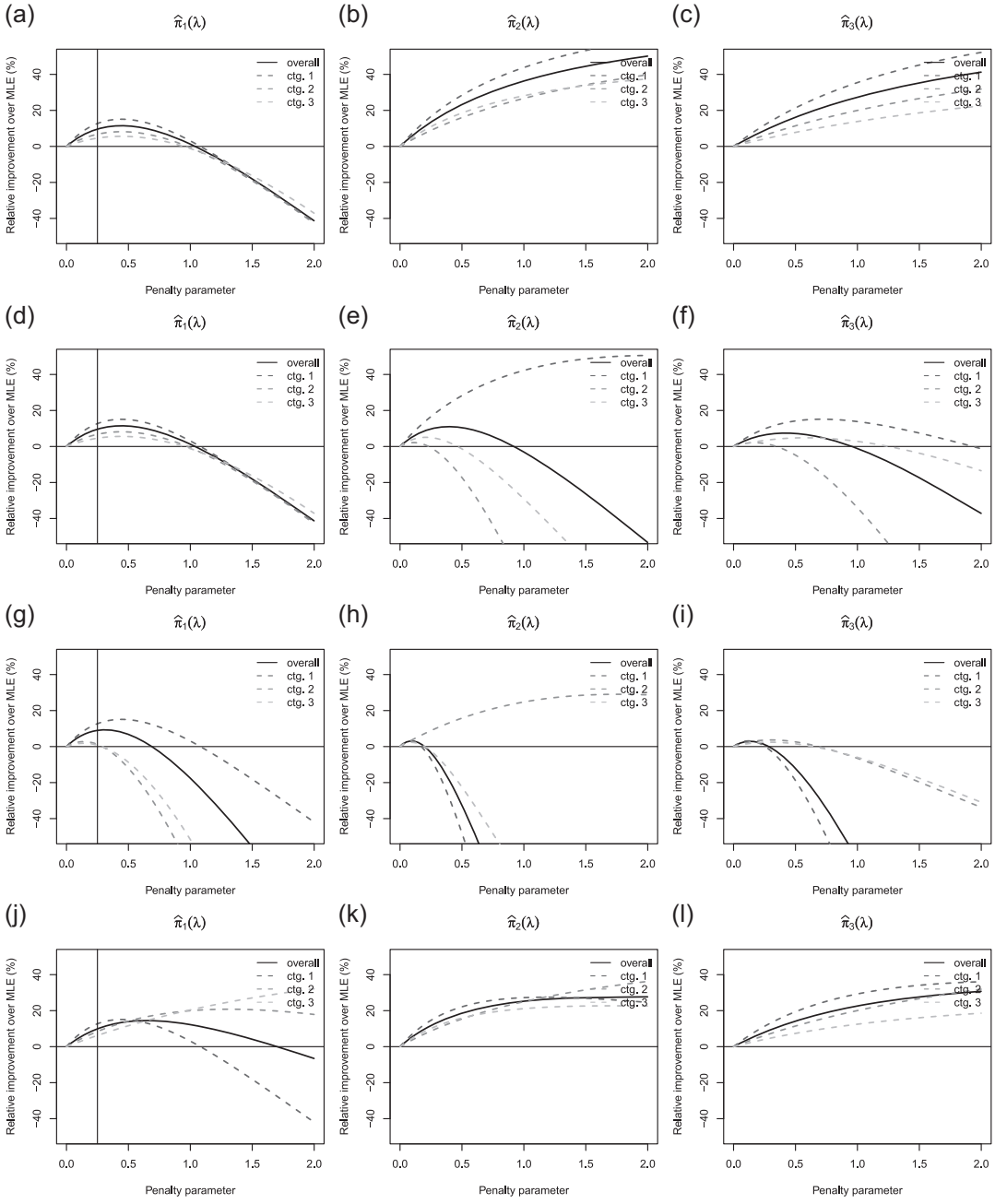

FIGURE 2 Relative improvement over the maximum likelihood estimator (MLE) (%) as a function of the penalty parameter (λ). Rows correspond to different scenarios: first row, scenario with equal event probabilities for each category ($\boldsymbol{\pi} = (0.2, 0.2, 0.2)^T$); second row, scenario with equal variances ($\boldsymbol{\pi} = (0.2, 0.8, 0.2)^T$); third row, scenario with different event probabilities and variances ($\boldsymbol{\pi} = (0.2, 0.9, 0.9)^T$); and fourth row, scenario with different event probabilities and variances ($\boldsymbol{\pi} = (0.2, 0.3, 0.4)^T$). Columns correspond to different ridge-type estimators ($\hat{\boldsymbol{\pi}}_1(\lambda)$, $\hat{\boldsymbol{\pi}}_2(\lambda)$, and $\hat{\boldsymbol{\pi}}_3(\lambda)$ for the first, second, and third columns, respectively). Solid lines refer to the overall mean squared error (MSE), whereas dashed lines are MSEs for separate categories. The vertical line in panels (A), (D), (G), and (J) is the value of the penalty parameter used when applying the Jeffreys invariant prior

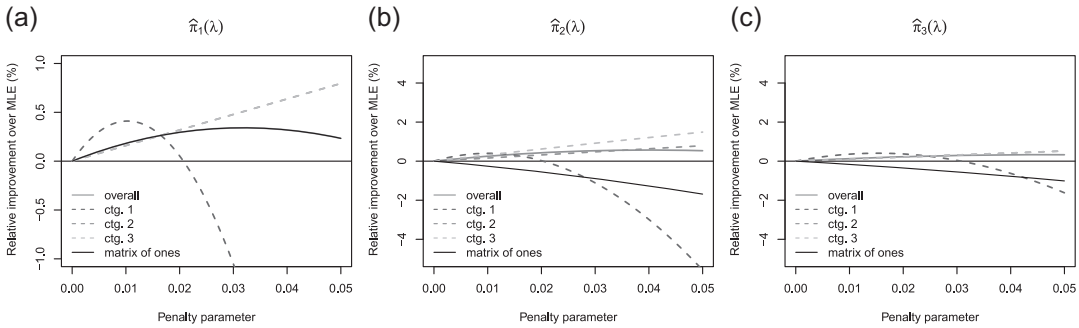

FIGURE 3 Relative improvement over the maximum likelihood estimator (MLE) (%) as a function of the penalty parameter (λ). Columns correspond to different ridge-type estimators ($\hat{\boldsymbol{\pi}}_1(\lambda)$, $\hat{\boldsymbol{\pi}}_2(\lambda)$, and $\hat{\boldsymbol{\pi}}_3(\lambda)$ for the first, second, and third columns, respectively). The two solid lines, namely, black and gray, refer to the generalized mean squared error when using the matrix of ones and overall mean squared error (MSE), respectively, whereas dashed lines are MSEs for separate categories

MSE of the second and third categories. In the examples considered previously, we observe some improvement over the MLE for all estimators also when **B** is the matrix of ones [data not shown].)

8 | AN APPLICATION

As an example, we consider a data set available within the R (R Core Team, 2015) **logistf** package (Heinze et al., 2014). The data consist of 130 sexually active college women that suffered from urinary tract infection and 109 controls, together with the covariate information on the use of contraceptives (yes and no). We are interested in predicting the event probability for women using condoms and lubricated condoms (Group 1), using either condoms or lubricated condoms (Group 2), or using neither (Group 3). The data were randomly split into two subsets (training set and validation set) of similar size, $n_T = 119$ and $n_V = 120$ so that $n_T + n_V = n$. The training set was used to estimate the event probability for each group by using different ridge estimators over a grid of penalty parameters (ranging from 0 to 2), and the validation set was then used to calculate the prediction error, defined as

$$\mathrm{PE}_j(\lambda) = \frac{1}{n_V}\sum_{i=1}^{n_V}\left(y_i^V - \hat{\pi}_{j,i}(\lambda)\right)^2, \quad \mathrm{CE}_j(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\left(\pi_k^V - \hat{\pi}_{j,k}(\lambda)\right)^2, \quad j = 1, 2, 3,$$

where y_i^V is the event indicator in the validation set and π_k^V is the proportion of events in group k in the validation set. Figure 4 shows the RP (%) over the MLE averaged over 100 random splits as a function of the penalty parameter.

In Figure 4, we can see that there is a value of the penalty parameter for which the performance of the ridge estimators is better than the MLE. For a large penalty parameter, however, the ridge estimators performed worse than the MLE. Note also the asymmetry between the gain observed with a small penalty parameter and the loss for the large values of the penalty parameter, with the loss greatly exceeding the gains.

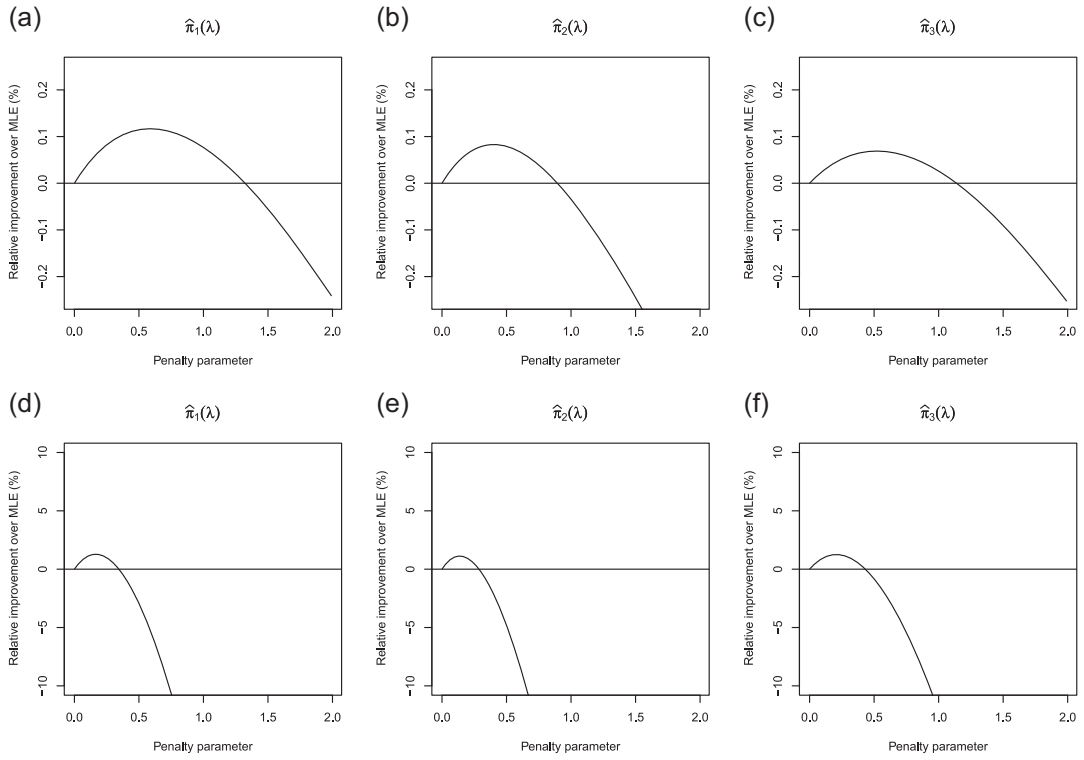

FIGURE 4 Relative improvement over the maximum likelihood estimator (MLE) (%) as a function of the penalty parameter (λ). Rows correspond to different measures: first row, prediction error; second row, Classification error (CE). Columns correspond to different ridge-type estimators ($\hat{\boldsymbol{\pi}}_1(\lambda)$, $\hat{\boldsymbol{\pi}}_2(\lambda)$, and $\hat{\boldsymbol{\pi}}_3(\lambda)$ for the first, second, and third columns, respectively)

9 | CONCLUSIONS

While the verdict on the MSE of the MLE for the Gaussian location problem is clear, the MLE is inadmissible for three or more dimensions; things are much more subtle when estimating the binary location parameter. We considered estimating the binary location parameter in a regression framework by considering logistic ridge regression in a $K \times 2$ table. Since there is no uniquely preferred way of shrinking in this problem, we looked at three variants. Estimator 1 shrunk the event probabilities toward 1/2, whereas Estimators 2 and 3 shrunk the event probability estimates toward the common mean probability. Estimators 2 and 3 differed in a way how the model was parameterized. Estimator 2 used a parameterization with the reference category; hence, the results depend on the choice of the reference category. To avoid this, Estimator 3 used an overparameterized model in order to obtain a more monotone way of shrinkage. We focused on comparing each of the shrinkage estimators with the MLE but did not compare between them as we believe that the choice of the shrinkage target should be driven by content rather than by a statistical argument.

We proved, under some mild assumptions on the marginals and the true event probabilities, that, in terms of the MSE, little shrinkage is always better than no shrinkage. However, we demonstrated that it is easy to overshrink. The MLE therefore, although it is never the optimal choice, is minimax optimal. It is preferable if all true event probabilities are close to 0 or 1.

Three well-known estimators, namely, FPE, ACE, and BE, were shown to be special cases of ridge regression. Our results give new insights on these methods. It was shown that particular values of the penalty parameter used by the FPE, ACE, and BE are optimal only in exceptional cases. The conditions under which they outperform the MLE were also derived; when the true event probabilities are close to 0 or 1, that is, when the events are rare or common, the MLE will perform better in terms of the MSE.

While the three types of ridge regression behave similarly when analysing the MSE, we see divergent results for the GMSE. Surprisingly, the most naive estimator that shrinks to 1/2 retains its property. For the other two, there is always a GMSE for which even a minute amount of shrinkage increases the error. However, it is not the individual probabilities that are problematic, as there is always little shrinkage that reduces the MSE of the individual probabilities. It is also shown that in terms of the GMSE, similarly as it was shown for the MSE, the PE, ACE, and BE do not outperform the MLE when the events are rare or common, where the bound on the true event probabilities where they do outperform the MLE is, as could be expected, tighter when considering the GMSE than the MSE.

To derive the ridge-type estimators, we used a one-step solution of the Newton–Raphson algorithm; hence, the results depend on the accuracy of this approximation. It seems natural to assume that even better results in terms of the (G)MSE can be obtained when considering the fully iterated solutions (through the fact that the variance of the fully iterated estimator will be smaller). Our numerical investigation suggests that the one-step solution is very similar to the fully iterated solution, where, rarely, more than three iterations are required to reach convergence. In general, our approximate shrinkage estimators are more accurate with a larger sample size and when the target is close to 1/2; therefore, the results of the iterated estimators may be different when the target is near 0 or 1. Our approach for obtaining the closed-form estimates fails in a setting with continuous predictor(s). Given the fact that, in this case, there are no closed-form estimates of the event probabilities even for the MLE, which was required for the approach taken here, it seems unlikely that they could be obtained for the penalized estimators in a more general setting.

Thus far, nothing has been said about specifying the penalty parameter λ. In our calculations, we neglected the fact that the penalty parameter can, in practice, be estimated from the data, for example, by selecting the penalty parameter that gives the smallest cross-validated likelihood error and is therefore itself a random variable. It seems reasonable to assume that this will lead to less bias but more variance. In the Appendix, we show, for the five examples considered in Section 7, how the optimal penalty parameter λ can be determined by using the estimated second-order matrices or leave-one-out cross-validated deviance with numerical optimization methods. It is shown that while the first approach provides, on average, values of the penalty parameter λ that are closer to being optimal in terms of the theoretically best possible improvement over the MLE, extra variability introduced by estimating λ, combined with the fact that, as observed also in our numerical illustration and implied by our theoretical results, it is easy to overshrink for a particular data set, results in either a much smaller improvement over the MLE than it would be theoretically feasible or an even worse performance in terms of some GMSE(s) when compared with the MLE. In the examples considered here, using the leave-one-out cross-validated deviance performed poorly, generally increasing the GMSE(s) in comparison with the MLE. While an analytical investigation of this problem seems too difficult, a large simulation study in a more general setting and using also other options for a data-adaptive choice of the penalty parameter would be in order.

From our numerical example, as well as an application to a real data set, we could have observed a large asymmetry between the gain and the loss for small and large values of the penalty

parameter, respectively, with the loss greatly exceeding the gain. Hence, it seems that, in the absence of a reliable way of determining the penalty parameter, it makes, on average, more sense not to use shrinkage, which was observed sometimes also in our small-scale investigation of estimating the amount of shrinkage from the data (see Appendix). On any given data set, however, anything can happen; hence, it remains to be seen how reliable the procedures are, which are, in practice, used for estimating the penalty parameter, and/or if reliable ways of estimating it can be obtained in the future. For the examples considered here, it seems that, at least for the two approaches for estimating the penalty parameter from the data we considered, the additionally introduced variability due to estimating the penalty parameter is so large that it is not worthwhile to use shrinkage at all.

ACKNOWLEDGMENTS

We thank the Associate Editor and two anonymous reviewers for their thoughtful comments that improved the presentation of this paper. Rok Blagus acknowledges the financial support from the Slovenian Research Agency ("Predicting rare events more accurately," N1-0035; "Methodology for data analysis in medical sciences," P3-0154).

ORCID

Rok Blagus https://orcid.org/0000-0001-7200-894X

REFERENCES

Agresti, A., & Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *The Annals of Mathematical Statistics*, *37*(5), 1087–1136. https://doi.org/10.1214/aoms/1177699259

Brown, L. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *The Annals of Mathematical Statistics*, *39*(1), 29–48. https://doi.org/10.1214/aoms/1177698503

Brown, L. D., & Fox, M. (1974a). Admissibility in statistical problems involving a location or scale parameter. *The Annals of Statistics*, *2*(4), 807–814. https://doi.org/10.1214/aos/1176342768

Brown, L. D., & Fox, M. (1974b). Admissibility of procedures in two-dimensional location parameter problems. *The Annals of Statistics*, *2*(2), 248–266. https://doi.org/10.1214/aos/1176342661

Duffy, D. E., & Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics - Theory and Methods*, *18*(3), 959–980. https://doi.org/10.1080/03610928908829944

Elgmati, E., Fiaccone, R. L., Henderson, R., & Matthews, J. N. S. (2015). Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. *Lifetime Data Analysis*, *21*(4), 542–560. https://doi.org/10.1007/s10985-015-9321-4

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Goeman, J., Meijer, R., Chaturvedi, N., & Lueder, M. (2014). Penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-45. http://CRAN.R-project.org/package=penalized

Greenland, S., & Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, *34*(23), 3133–3143. https://doi.org/10.1002/sim.6537

Heinze, G., Ploner, M., Dunkler, D., & Southworth, H. (2014). logistf: Firth's bias reduced logistic regression. R package version 1.22. http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/fllogistf/

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379. https://projecteuclid.org/euclid.bsmsp/1200512173

le Cessie, S., & van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*(1), 191–201.

R Core Team (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Schaefer, R. L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, *25*(1-2), 75–91. https://doi.org/10.1080/00949658608810925

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics - Theory and Methods*, *13*, 99–113. https://doi.org/10.1080/03610928408828664

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pp. 197–206. https://projecteuclid.org/euclid.bsmsp/1200501656

Stein, C. (1959). The admissibility of Pitman's estimator of a single location parameter. *The Annals of Mathematical Statistics*, *30*(4), 970–979. https://doi.org/10.1214/aoms/1177706080

Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2001). Application of shrinkage techniques in logistic regression analysis: A case study. *Statistica Neerlandica*, *55*(1), 76–88. https://doi.org/10.1111/1467-9574.00157

Sun, H., & Wang, S. (2012). Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics*, *28*(10), 1368–1375. https://doi.org/10.1093/bioinformatics/bts145

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society Series B (Methodological)*, *36*(1), 103–106. http://www.jstor.org/stable/2984775

Walter, R. B., Othus, M., Borthakur, G., Ravandi, F., Cortes, J. E., Pierce, S. A., … Estey, E. H. (2011). Prediction of early death after induction therapy for newly diagnosed acute myeloid leukemia with pretreatment risk scores: A novel paradigm for treatment assignment. *Journal of Clinical Oncology*, *29*(33), 4417–4424. https://doi.org/10.1200/JCO.2011.35.7525

Zhou, H., Sehl, M. E., Sinsheimer, J. S., & Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, *26*(19), 2375–2382. https://doi.org/10.1093/bioinformatics/btq448

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Blagus R, Goeman JJ. Mean squared error of ridge estimators in logistic regression. *Statistica Neerlandica*. 2020;74:159–191. https://doi.org/10.1111/stan.12201

APPENDIX A

PROOFS OF THE THEORETICAL RESULTS

In this section, we give the proofs. Where necessary, lemmas are stated and proved. It will be helpful to write $\hat{\boldsymbol{\pi}}_2(\lambda)$ and $\hat{\boldsymbol{\pi}}_3(\lambda)$ in matrix notation. Define K vectors $\mathbf{a} = (a_{11}, \dots, a_{K1})^T$ and

$\mathbf{1} = (1, \ldots, 1)^T$ as $K \times K$ diagonal matrices $\mathbf{A} = \text{diag}(a_{1\bullet}, \ldots, a_{K\bullet})$ and $\mathbf{I} = \text{diag}(1, \ldots, 1)$ and a $K \times K$ singular arrowhead matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & -\mathbf{1} \\ -\mathbf{1} & K-1 \end{bmatrix}.$$

Then,

$$\hat{\boldsymbol{\pi}}_2(\lambda) = (\mathbf{A} + 4\lambda\mathbf{P})^{-1}\mathbf{a},$$
$$\hat{\boldsymbol{\pi}}_3(\lambda) = \left(\mathbf{A} + 4\lambda\left[\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right]\right)^{-1}\mathbf{a}.$$

A trivial result from real analysis will be used in the proof of Theorem 1.

Lemma 4. *Let $\phi(x)$ be some continuous function defined on $x \in [0, \infty)$. Let $\frac{d\phi(x)}{dx}$ be the derivative of $\phi(x)$ according to x, and assume that $\frac{d\phi(x)}{dx}$ is continuous. If $\frac{d\phi(x)}{dx}|_{x=0} < 0$ holds, then $\phi(0) > \phi(\epsilon)$, for some $\epsilon > 0$.*

Proof. The proof is trivial and follows from the continuity of $\phi(x)$ and its derivative. □

Proof of Theorem 1. The second-order moment matrix of $\hat{\boldsymbol{\pi}}_1(\lambda)$ is

$$\mathbf{M}_1(\lambda) = \mathbf{D}(\lambda) + \mathbf{b}(\lambda)\mathbf{b}(\lambda)^T, \tag{A1}$$

where

$$\mathbf{D}(\lambda) = \text{diag}\left(\frac{a_{1\bullet}\pi_1(1-\pi_1)}{(a_{1\bullet}+4\lambda)^2}, \ldots, \frac{a_{K\bullet}\pi_K(1-\pi_K)}{(a_{K\bullet}+4\lambda)^2}\right),$$
$$\mathbf{b}(\lambda) = \left(\frac{2\lambda(1-2\pi_1)}{a_{1\bullet}+4\lambda}, \ldots, \frac{2\lambda(1-2\pi_K)}{a_{K\bullet}+4\lambda}\right)^T.$$

Hence, it can be shown that

$$\frac{d}{d\lambda}\text{trace}\left(\mathbf{M}_1(0) - \mathbf{M}_1(\lambda)\right) = -\frac{d}{d\lambda}\text{trace}\left(\mathbf{M}_1(\lambda)\right) = 8\sum_{k=1}^{K}\frac{a_{k\bullet}\left(\pi_k - \pi_k^2 - \lambda(1-2\pi_k)^2\right)}{(a_{k\bullet}+4\lambda)^3}.$$

For $\lambda = 0$, it is obvious that $\frac{d}{d\lambda}\text{trace}\left(\mathbf{M}_1(0) - \mathbf{M}_1(\lambda)\right)|_{\lambda=0} > 0$; thus, by applying Lemma 4, we prove the theorem for $\hat{\boldsymbol{\pi}}_1(\lambda)$.

The second-order moment matrix of $\hat{\boldsymbol{\pi}}_2(\lambda)$ is

$$\mathbf{M}_2(\lambda) = \mathbf{U}(\lambda)^{-1}(\mathbf{A}\mathbf{W} + 16\lambda^2\mathbf{P}\boldsymbol{\pi}\boldsymbol{\pi}^T\mathbf{P})\mathbf{U}(\lambda)^{-1}, \tag{A2}$$

where

$$\mathbf{U}(\lambda) = \mathbf{A} + 4\lambda\mathbf{P}, \quad \mathbf{W} = \text{diag}(\pi_1(1-\pi_1), \ldots, \pi_K(1-\pi_K)), \quad \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T.$$

It can be shown that

$$\frac{d}{d\lambda}\text{trace}\left(\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)\right)|_{\lambda=0} = 8\left(\sum_{k=1}^{K-1}\frac{\pi_k - \pi_k^2}{a_{k\bullet}^2} + (K-1)\frac{\pi_K - \pi_K^2}{a_{K\bullet}^2}\right) \geq 0,$$

with equality applying if and only if $K = 1$; hence, by applying Lemma 4, we prove the theorem for $\hat{\boldsymbol{\pi}}_2(\lambda)$ when $K > 1$.

The second-order moment matrix of $\hat{\boldsymbol{\pi}}_3(\lambda)$ is

$$\mathbf{M}_3(\lambda) = \mathbf{U}_1(\lambda)^{-1}\left(\mathbf{AW} + 16\lambda^2\left[\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right]\boldsymbol{\pi}\boldsymbol{\pi}^T\left[\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right]\right)\mathbf{U}_1(\lambda)^{-1}, \tag{A3}$$

where

$$\mathbf{U}_1(\lambda) = \mathbf{A} + 4\lambda\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right).$$

It can be shown that

$$\frac{d}{d\lambda}\text{trace}\left(\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)\right)|_{\lambda=0} = 8\left(1 - \frac{1}{K}\right)\sum_{k=1}^{K}\frac{\pi_k - \pi_k^2}{a_{k\bullet}^2} \geq 0,$$

with equality applying if and only if $K = 1$, thus proving, after applying Lemma 4, the theorem for $\hat{\boldsymbol{\pi}}_3(\lambda)$ when $K > 1$. □

Proof of Lemma 1. The proof follows from the calculation of

$$\text{MSE}(\hat{\boldsymbol{\pi}}_1(\gamma)) - \text{MSE}(\hat{\boldsymbol{\pi}}_1(\lambda)) = \text{trace}\left(\mathbf{M}_1(\gamma) - \mathbf{M}_1(\lambda)\right)$$

and by simple algebra. □

Proof of Corollary 1. Set $\gamma = 0$ and $\lambda = 1/4$, and apply Lemma 1. □

Proof of Proposition 1. Calculate

$$\frac{d}{d\lambda}\text{trace}\left(\mathbf{M}_1(\lambda)\right) = 8\sum_{k=1}^{K}\frac{a_{k\bullet}}{(a_{k\bullet} + 4\lambda)^3}\left(\lambda(2\pi_k - 1)^2 - \left(\pi_k - \pi_k^2\right)\right).$$

Set $\frac{d\text{MSE}(\hat{\boldsymbol{\pi}}_1(\lambda))}{d\lambda} = 0$, and let $\lambda = 1/4$, to complete the proof. □

Proof of Corollary 2. Set $K = 1$, $\gamma = 0$, and $\lambda = 1$, and apply Lemma 1. The proof is then completed by simple algebra. □

Proof of Corollary 3. Set $K = 1$, $\gamma = 0$, and $\lambda = \sqrt{a_{1\bullet}}/4$, and apply Lemma 1. The proof is completed by simple algebra. □

Proof of Lemma 2. For $\hat{\boldsymbol{\pi}}_1(\lambda)$, calculate

$$\text{MSE}(\hat{\boldsymbol{\pi}}_1(0)) - \text{MSE}(\hat{\boldsymbol{\pi}}_1(\lambda)) = 4\lambda\sum_{k=1}^{K}\frac{2\left(\pi_k - \pi_k^2\right)(a_{k\bullet} + 2\lambda) - \lambda a_{k\bullet}(1 - 2\pi_k)^2}{a_{k\bullet}(a_{k\bullet} + 4\lambda)^2}.$$

Express the numerator for each k as

$$\left(\pi_k - \pi_k^2\right)(4\lambda(a_{k\bullet} + 1) + 2a_{k\bullet}) - \lambda a_{k\bullet}.$$

Since, for $\lambda > 0$, $\lambda a_{k\bullet} > 0$, it follows that, for each $\lambda > 0$, there exists some $\pi_k - \pi_k^2 > 0$ such that the numerator for each k is negative, thus proving the claim for $\hat{\boldsymbol{\pi}}_1(\lambda)$.

For $\hat{\boldsymbol{\pi}}_3(\lambda)$, calculate

$$\text{MSE}(\hat{\boldsymbol{\pi}}_3(0)) - \text{MSE}(\hat{\boldsymbol{\pi}}_3(\lambda)) = 8\lambda\sum_{k=1}^{K}\frac{\left(\pi_k - \pi_k^2\right)(1 - w_k) - 2\lambda g_k}{(a_{k\bullet} + 4\lambda)^2}, \tag{A4}$$

where

$$g_k = \sum_{j=1}^{K} \frac{w_j^2\left(\pi_j - \pi_j^2\right)}{a_{j\bullet}} + \left[\pi_k - \sum_{j=1}^{K} w_j \pi_j\right]^2, \quad k = 1, \ldots, K,$$

$$w_k = \frac{a_{k\bullet}}{a_{k\bullet} + 4\lambda} \Big/ \sum_{j=1}^{K} \frac{a_{j\bullet}}{a_{j\bullet} + 4\lambda}, \quad k = 1, \ldots, K.$$

Note that $0 < w_k < 1$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} w_k = 1$. Set $\pi_1 = \epsilon$ and $\pi_k = 1 - \epsilon$, $k = 2, \ldots, K$, and consider the limiting case where $\epsilon \to 0$. Then, for every $k = 1, \ldots, K$, we have $\pi_k - \pi_k^2 \to 0$, but $\lim_{\epsilon \to 0} g_k > 0$ since not all π_k are equal. It follows that the numerator of (A4) has a strictly negative limit for every k. By continuity, there is an $\epsilon > 0$ and, therefore, a $\boldsymbol{\pi}$ for which (A4) is negative, proving the claim for $\hat{\boldsymbol{\pi}}_3(\lambda)$.

For $\hat{\boldsymbol{\pi}}_2(\lambda)$, calculate

$$\mathrm{MSE}(\hat{\boldsymbol{\pi}}_2(0)) - \mathrm{MSE}(\hat{\boldsymbol{\pi}}_2(\lambda)) = \frac{8\lambda}{c^2}\left((\pi_K - \pi_K)^2 b_K + \sum_{k=1}^{K-1}\left(\pi_k - \pi_k^2\right) b_k - 2\lambda g\right),$$

where

$$c = a_{K\bullet} + 4\lambda \sum_{k=1}^{K-1} \frac{a_{k\bullet}}{a_{k\bullet} + 4\lambda},$$

$$b_K = \sum_{k=1}^{K-1} \frac{a_{k\bullet}}{a_{k\bullet} + 4\lambda} + 2\lambda\left[\frac{1}{a_{K\bullet}}\left(\sum_{j=1}^{K-1} \frac{a_{j\bullet}}{a_{j\bullet} + 4\lambda}\right)^2 - a_{K\bullet} \sum_{k=1}^{K-1} \frac{1}{(a_{k\bullet} + 4\lambda)^2}\right],$$

$$b_k = \frac{1}{(a_{k\bullet} + 4\lambda)^2}\left(\frac{c^2(a_{k\bullet} + 2\lambda)}{a_{k\bullet}} - 2\lambda a_{k\bullet}\left[16\lambda^2 \sum_{j=1}^{K-1} \frac{1}{(a_{j\bullet} + 4\lambda)^2} + \frac{2c}{a_{k\bullet} + 4\lambda}\right]\right), \quad k = 1, \ldots, K-1,$$

$$g = \sum_{k=1}^{K-1} \frac{\left(a_{K\bullet}\pi_K + 4\lambda \sum_{j=1}^{K-1} \frac{a_{j\bullet}\pi_j}{a_{j\bullet} + 4\lambda} - c\pi_k\right)^2}{(a_{k\bullet} + 4\lambda)^2} + \left(\sum_{j=1}^{K-1} \frac{a_{j\bullet}\pi_j}{a_{j\bullet} + 4\lambda} - \pi_K \sum_{j=1}^{K-1} \frac{a_{j\bullet}}{a_{j\bullet} + 4\lambda}\right)^2.$$

Set $\pi_K = 1 - \epsilon$ and $\pi_k = \epsilon$, $k = 1, \ldots, K-1$, and consider the limiting case where $\epsilon \to 0$. Then, for every $k = 1, \ldots, K$, we have $\pi_k - \pi_k^2 \to 0$, but

$$\lim_{\epsilon \to 0} g = \lim_{\epsilon \to 0} (1 - 2\epsilon)^2 \left(a_{K\bullet}^2 \sum_{k=1}^{K-1} \frac{1}{(a_{k\bullet} + 4\lambda)^2} + \left(\sum_{k=1}^{K-1} \frac{a_{k\bullet}}{a_{k\bullet} + 4\lambda}\right)^2\right) > 0.$$

Use the same argument as for $\hat{\boldsymbol{\pi}}_3(\lambda)$ to complete the proof. □

Proof of Theorem 2. Let $R_j(\lambda, \boldsymbol{\pi}) = \mathrm{MSE}(\hat{\boldsymbol{\pi}}_j(\lambda))$ be the risk depending on parameters $\boldsymbol{\pi}$ and λ, $j = 1, 2, 3$. By Lemma 2, for every λ, there is a $\boldsymbol{\pi}_0(\lambda)$ such that

$$R_j(\lambda, \boldsymbol{\pi}_0(\lambda)) > R_j(0, \boldsymbol{\pi}_0(\lambda)), \quad j = 1, 2, 3.$$

Now, consider the maximal relative risk

$$\max_{\boldsymbol{\pi}} \frac{R_j(\lambda, \boldsymbol{\pi})}{R_j(0, \boldsymbol{\pi})}, \quad j = 1, 2, 3,$$

which is always 1 for $\lambda = 0$. For $\lambda \neq 0$, we have

$$\max_{\boldsymbol{\pi}} \frac{R_j(\lambda, \boldsymbol{\pi})}{R_j(0, \boldsymbol{\pi})} \geq \frac{R_j(\lambda, \boldsymbol{\pi}_0(\lambda))}{R_j(0, \boldsymbol{\pi}_0(\lambda))} > 1, \quad j = 1, 2, 3.$$

Therefore,

$$\min_{\lambda} \max_{\boldsymbol{\pi}} \frac{R_j(\lambda, \boldsymbol{\pi})}{R_j(0, \boldsymbol{\pi})} = 1, \quad j = 1, 2, 3,$$

and this minimax risk is attained at $\lambda = 0$. □

Proof of Lemma 3. See the work of Theobald (1974, p. 104). □

Proof of Theorem 3. Write

$$\mathbf{M}_1(0) - \mathbf{M}_1(\lambda) = \mathbf{D}(0) - \mathbf{D}(\lambda) - \mathbf{b}(\lambda)\mathbf{b}(\lambda)^T,$$

$\lambda > 0$. Now, since, for $\lambda > 0$,

$$\{\mathbf{D}(0) - \mathbf{D}(\lambda)\}_{kk} = \frac{8\lambda\pi_k(1 - \pi_k)(a_{k\bullet} + 2\lambda)}{a_{k\bullet}(a_{k\bullet} + 4\lambda)^2} > 0, \quad k = 1, \ldots, K,$$

thence $\mathbf{D}(0) - \mathbf{D}(\lambda)$ is positive definite and

$$\begin{aligned}\mathbf{M}_1(0) - \mathbf{M}_1(\lambda) =& (\mathbf{D}(0) - \mathbf{D}(\lambda))^{1/2}(\mathbf{I} - (\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1/2}\mathbf{b}(\lambda)\mathbf{b}(\lambda)^T(\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1/2}) \\ & (\mathbf{D}(0) - \mathbf{D}(\lambda))^{1/2}.\end{aligned}$$

Hence, $\mathbf{M}_1(0) - \mathbf{M}_1(\lambda)$ is positive definite if and only if

$$\mathbf{I} - (\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1/2}\mathbf{b}(\lambda)\mathbf{b}(\lambda)^T(\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1/2}$$

is positive definite. Then, since the eigenvalues of

$$(\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1/2}\mathbf{b}(\lambda)\mathbf{b}(\lambda)^T(\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1/2}$$

are zero (with multiplicity $K - 1$) and

$$\mathbf{b}(\lambda)^T(\mathbf{D}(0) - \mathbf{D}(\lambda))^{-1}\mathbf{b}(\lambda),$$

it follows that if

$$1 - \frac{\lambda}{2}\sum_{k=1}^{K} \frac{a_{k\bullet}(1 - 2\pi_k)^2}{\pi_k(1 - \pi_k)(a_{k\bullet} + 2\lambda)} > 0 \tag{A5}$$

holds, then $\mathbf{M}_1(0) - \mathbf{M}_1(\lambda)$ is positive definite. The left-hand side (LHS) of (A5) for $\lambda = 0$ is positive; hence, it follows from the continuity of the LHS of (A5) that there exists $\lambda > 0$ for which (A5) also holds. □

Proof of Proposition 2. Using the result from the proof of Theorem 3 and setting $\lambda = 1/4$, $\mathbf{M}_1(0) - \mathbf{M}_1(1/4)$ is nonnegative definite if and only if

$$1 - \frac{1}{8}\sum_{k=1}^{K} \frac{a_{k\bullet}(1 - 2\pi_k)^2}{\pi_k(1 - \pi_k)(a_{k\bullet} + 1/2)} \geq 0$$

holds, thus proving the claim. □

Proof of Theorem 4. By (7), it is sufficient to prove the results only for $\hat{\boldsymbol{\pi}}_2(\lambda)$. After some algebra,

$$\mathbf{M}_2(0) - \mathbf{M}_2(\lambda) = \frac{4\lambda}{(a_{k\bullet}a_{j\bullet} + 4\lambda n)^2}\mathbf{G}(\lambda),$$

where $\mathbf{G} = \mathbf{G}(\lambda)$ is a 2×2 symmetric matrix with entries

$$\begin{aligned}
\{\mathbf{G}\}_{11} &= \frac{2a_{2\bullet}}{a_{1\bullet}}\left(a_{1\bullet}a_{2\bullet}\pi_1(1-\pi_1) + 2\lambda c_1\right),\\
\{\mathbf{G}\}_{22} &= \frac{2a_{1\bullet}}{a_{2\bullet}}\left(a_{1\bullet}a_{2\bullet}\pi_2(1-\pi_2) + 2\lambda c_2\right),\\
\{\mathbf{G}\}_{12} &= -\left(a_{1\bullet}a_{2\bullet}c_3 + 4\lambda c_4\right),
\end{aligned}$$

where

$$\begin{aligned}
c_1 &= \pi_1(1-\pi_1)(n + a_{1\bullet}) - a_{1\bullet}(\pi_2(1-\pi_2) + a_{2\bullet}(\pi_1 - \pi_2)^2),\\
c_2 &= \pi_2(1-\pi_2)(n + a_{2\bullet}) - a_{2\bullet}(\pi_1(1-\pi_1) + a_{1\bullet}(\pi_1 - \pi_2)^2),\\
c_3 &= \pi_1(1-\pi_1) + \pi_2(1-\pi_2) < 2,\\
c_4 &= a_{1\bullet}\pi_1(1-\pi_1) + a_{2\bullet}\pi_2(1-\pi_2) - a_{1\bullet}a_{2\bullet}(\pi_1 - \pi_2)^2.
\end{aligned}$$

Calculate

$$\begin{aligned}
\det(\mathbf{G}) = 16\lambda^2\left(c_1c_2 - c_4^2\right) &+ 8\lambda a_{1\bullet}a_{2\bullet}(\pi_1(1-\pi_1)c_2 + \pi_2(1-\pi_2)c_1 - c_3c_4)\\
&+ a_{1\bullet}^2a_{2\bullet}^2\left(4\pi_1(1-\pi_1)\pi_2(1-\pi_2) - c_3^2\right).
\end{aligned}$$

Case (i): Assume that $\pi_1 = \pi_2 = \pi$. Then, we have

$$c_1 = c_2 = c_4 = n\pi(1-\pi), \quad \{\mathbf{G}\}_{11} > 0, \det(\mathbf{G}) = 0;$$

hence, for $\lambda > 0$, by the principal minors argument, $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is positive semidefinite. *Case (ii)*: Now, assume that $\pi_1 = 1 - \pi_2$ and $\pi_1 \neq \pi_2$, which implies that $\pi_1, \pi_2 \neq 1/2$. Hence, $\pi_1(1-\pi_1) = \pi_2(1-\pi_2) = \pi(1-\pi)$. Then,

$$c = c_1 = c_2 = c_4 = n\pi_1\pi_2 - a_{1\bullet}a_{2\bullet}(\pi_1 - \pi_2)^2, \det(\mathbf{m}) = 0,$$

and the sign of $\mathbf{G}_{11}$ depends on the sign of c. Then, using the principal minors argument, if

$$n\pi(1-\pi) - a_{1\bullet}a_{2\bullet}(\pi_1 - \pi_2)^2 \geq 0,$$

then, for $\lambda > 0$, $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is positive semidefinite, whereas for

$$n\pi(1-\pi) - a_{1\bullet}a_{2\bullet}(\pi_1 - \pi_2)^2 < 0,$$

$\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is positive semidefinite for some $0 < \lambda < \Lambda$, where $\Lambda > 0$ depends on the data and the true event probabilities. Then, it is obvious that, for $\lambda > \Lambda$, $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is negative semidefinite, and the proof of Case (ii),(b) is complete by remarking that, in this case, $\mathbf{G}_{11} = 0$ implies $\mathbf{G}_{22} = 0$. *Case (iii)*: Now, consider the case where $\pi_1 \neq \pi_2$ and $\pi_1 \neq 1 - \pi_2$. Then, after some algebra,

$$\begin{aligned}
c_1c_2 - c_4^2 &= -n^2(\pi_1(1-\pi_1) - \pi_2(1-\pi_2))^2 < 0,\\
\pi_1(1-\pi_1)c_2 + \pi_2(1-\pi_2)c_1 - c_3c_4 &= -n(\pi_1(1-\pi_1) - \pi_2(1-\pi_2))^2 < 0,\\
4\pi_1(1-\pi_1)\pi_2(1-\pi_2) - c_3^2 &= -(\pi_1(1-\pi_1) - \pi_2(1-\pi_2))^2 < 0.
\end{aligned}$$

This implies that, for $\lambda > 0$, $\det(\mathbf{G}) < 0$, which implies that $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is indefinite. □

Proof of Corollary 4. We prove the claim for $\hat{\boldsymbol{\pi}}_3(\lambda)$. The proof of the claim for $\hat{\boldsymbol{\pi}}_2(\lambda)$ then follows easily by (7). It can be shown that, under the conditions of Theorem 4, Case (ii),(b),

$$\mathbf{M}_3(0) - \mathbf{M}_3(\lambda) = 4\lambda \mathbf{U}_1(\lambda)^{-1}\mathbf{G}(\lambda)\mathbf{U}_1(\lambda)^{-1},$$

where $\mathbf{G} = \mathbf{G}(\lambda)$ is a 2×2 symmetric matrix with entries

$$\{\mathbf{G}\}_{k,k} = \pi - \pi^2 + \lambda\left(-(2\pi-1)^2 + \sum_{j=1}^{2}\frac{\pi-\pi^2}{a_{j\bullet}}\right), \quad k = 1,2, \quad l \neq k,$$

$$\{\mathbf{G}\}_{k,l} = -(\pi - \pi^2) + \lambda\left((2\pi-1)^2 - \sum_{j=1}^{2}\frac{\pi-\pi^2}{a_{j\bullet}}\right), \quad k,l = 1,2, \quad l \neq k.$$

Observe that $\{\mathbf{G}\}_{k,k} = -\{\mathbf{G}\}_{k,l} = a$ so that

$$\mathbf{G} = 2a\mathbf{I} - a\mathbf{1}\mathbf{1}^T.$$

One eigenvalue of $\mathbf{G}$ is $2a$, whereas one eigenvalue is

$$2a - 2a = 0.$$

Now, $2a = 0$ holds if and only if

$$\pi - \pi^2 + \lambda\left(\sum_{j=1}^{2}\frac{\pi-\pi^2}{a_{j\bullet}} - (2\pi-1)^2\right) = 0$$

holds. The solution is then

$$\Lambda = \frac{\pi - \pi^2}{(2\pi-1)^2 - \sum_{j=1}^{2}\frac{\pi-\pi^2}{a_{j\bullet}}}.$$

The proof is completed by simple algebra. □

The following lemma, which can be proved by using the results from Theorem 4, Case (iii), will be used for proving the result for $\hat{\boldsymbol{\pi}}_2(\lambda)$ when $K > 2$.

Lemma 5. *Define*

$$\phi(c_1,c_2,\lambda) = \left(\left(\pi_1 - \pi_1^2\right)c_1 - \left(\pi_2 - \pi_2^2\right)c_2\right)(c_1 - c_2) + 2\lambda(c_1-c_2)^2\left(\frac{\pi_1-\pi_1^2}{a_{1\bullet}} + \frac{\pi_2-\pi_2^2}{a_{2\bullet}} - (\pi_1-\pi_2)^2\right),$$

for some $c_1 \in \mathbb{R}$, $c_2 \in \mathbb{R}$, $\pi_1 \in \mathbb{R}$, $\pi_2 \in \mathbb{R}$, $a_{1\bullet} \in \mathbb{R} - \{0\}$, $a_{2\bullet} \in \mathbb{R} - \{0\}$ and $\lambda > 0$. Assume that $\pi_1 \neq \pi_2$ and $\pi_1 \neq 1-\pi_2$. Then, for each $\lambda > 0$, there exists some $(c_1,c_2)^T$ such that $\phi(c_1,c_2,\lambda) > 0$ and some other $(c_1,c_2)^T$ such that $\phi(c_1,c_2,\lambda) < 0$.

Proof. Let $\mathbf{G} = \mathbf{G}(\lambda)$ denote a 2×2 symmetric matrix with entries

$$\{\mathbf{G}\}_{11} = 2\left(\pi_1 - \pi_1^2\right) + 4\lambda\left(\frac{\pi_1-\pi_1^2}{a_{1\bullet}} + \frac{\pi_2-\pi_2^2}{a_{2\bullet}} - (\pi_1-\pi_2)^2\right),$$

$$\{\mathbf{G}\}_{12} = -\left(\pi_1 - \pi_1^2\right) - \left(\pi_2 - \pi_2^2\right) + 4\lambda\left(-\frac{\pi_2-\pi_2^2}{a_{2\bullet}} - \frac{\pi_1-\pi_1^2}{a_{1\bullet}} + (\pi_1-\pi_2)^2\right),$$

$$\{\mathbf{G}\}_{22} = 2\left(\pi_2 - \pi_2^2\right) + 4\lambda\left(\frac{\pi_2-\pi_2^2}{a_{2\bullet}} + \frac{\pi_1-\pi_1^2}{a_{1\bullet}} - (\pi_1-\pi_2)^2\right).$$

It can be shown that

$$\mathbf{M}_2(0) - \mathbf{M}_2(\lambda) = 4\lambda \mathbf{U}(\lambda)^{-1}\mathbf{G}(\lambda)\mathbf{U}(\lambda)^{-1}.$$

Assuming that $\pi_1 \neq \pi_2$ and $\pi_1 \neq 1 - \pi_2$, then, for all $\lambda > 0$ by Theorem 4, Case (iii), $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ and, hence, $\mathbf{G}(\lambda)$ are indefinite. This implies that, for each $\lambda > 0$, there exists some $c = (c_1, c_2)^T$ such that $c^T\mathbf{G}(\lambda)c > 0$ and some other $c = (c_1, c_2)^T$ such that $c^T\mathbf{G}(\lambda)c < 0$. Simple calculation then shows that $\phi(c_1, c_2, \lambda) = \frac{1}{2}c^T\mathbf{G}(\lambda)c$. □

Proof of Theorem 5. Calculate

$$\mathbf{M}_2(0) - \mathbf{M}_2(\lambda) = 4\lambda \mathbf{U}(\lambda)^{-1}\mathbf{G}(\lambda)\mathbf{U}(\lambda)^{-1},$$

where $\mathbf{G} = \mathbf{G}(\lambda)$ is a $K \times K$ symmetric matrix with entries

$$\{\mathbf{G}\}_{kk} = 2\left(\pi_k - \pi_k^2\right) + 4\lambda\left(\frac{\pi_k - \pi_k^2}{a_{k\bullet}} + \frac{\pi_K - \pi_K^2}{a_{K\bullet}} - (\pi_k - \pi_K)^2\right), \quad k = 1, \ldots, K-1,$$

$$\{\mathbf{G}\}_{kj} = 4\lambda\left(\frac{\pi_K - \pi_K^2}{a_{K\bullet}} - (\pi_k - \pi_K)(\pi_j - \pi_K)\right), \quad j, k = 1, \ldots, K-1, \quad j \neq k,$$

$$\{\mathbf{G}\}_{kK} = -\left(\pi_k - \pi_k^2\right) - \left(\pi_K - \pi_K^2\right) + 4\lambda\left(-(K-1)\frac{\pi_K - \pi_K^2}{a_{K\bullet}} - \frac{\pi_k - \pi_k^2}{a_{k\bullet}} + (\pi_k - \pi_K)\left(\sum_{k=1}^{K-1}\pi_k - (K-1)\pi_K\right)\right),$$

$$k = 1, \ldots, K-1,$$

$$\{\mathbf{G}\}_{KK} = 2(K-1)\left(\pi_K - \pi_K^2\right) + 4\lambda\left((K-1)^2\frac{\pi_K - \pi_K^2}{a_{K\bullet}} + \sum_{k=1}^{K-1}\frac{\pi_k - \pi_k^2}{a_{k\bullet}} - \left(\sum_{k=1}^{K-1}\pi_k - (K-1)\pi_K\right)^2\right).$$

Let $c = (c_1, \ldots, c_K)^T$. Then,

$$\begin{aligned}\frac{1}{2}c^T\mathbf{G}c = &\sum_{k=1}^{K-1}\left(\left(\pi_k - \pi_k^2\right)c_k - \left(\pi_K - \pi_K^2\right)c_K\right)(c_k - c_K) \\ &+ 2\lambda\left(\frac{\pi_K - \pi_K^2}{a_{K\bullet}}\left(\sum_{k=1}^{K-1}c_k - (K-1)c_K\right)^2 + \sum_{k=1}^{K-1}\frac{\pi_k - \pi_k^2}{a_{k\bullet}}(c_k - c_K)^2\right. \\ &\left. - \left(\sum_{k=1}^{K-1}(\pi_k - \pi_K)c_k - c_K\left(\sum_{k=1}^{K-1}\pi_k - (K-1)\pi_K\right)\right)^2\right).\end{aligned}$$

Obviously for $c_k = a$, $k = 1, \ldots, K$, for some constant a, $c^T\mathbf{G}c = 0$. *Cases (i) and (ii)*: Assume that $\pi_k - \pi_k^2 = \pi - \pi^2$, $k = 1, \ldots, K$. Then, we have

$$c^T\mathbf{G}c = 2(\pi - \pi^2)\sum_{k=1}^{K-1}(c_k - c_K)^2 + 4\lambda g,$$

for some constant

$$g = (\pi - \pi^2)\left(\frac{\left(\sum_{k=1}^{K-1}(c_k - c_K)\right)^2}{a_{K\bullet}} + \sum_{k=1}^{K-1}\frac{(c_k - c_K)^2}{a_{k\bullet}}\right) - \left(\sum_{k=1}^{K-1}(\pi_k - \pi_K)(c_k - c_K)\right)^2.$$

Now, for $\lambda = 0$, $c^T\mathbf{G}c \geq 0$, with equality applying if and only if $c_k = c_K$, $k = 1, \ldots, K-1$; hence, $\mathbf{G}$ is positive semidefinite. Thence, it follows from the continuity argument that there exists some $\Lambda > 0$, such that, for $0 < \lambda < \Lambda$, $\mathbf{G}$ is also positive semidefinite. If $\pi_k = \pi$, $k = 1, \ldots, K$, then $g > 0$ and $c^T\mathbf{G}c \geq 0$ holds for $\lambda > 0$. *Case (iii)*: Assume that $\pi_k \neq \pi_K$ and $\pi_k \neq 1 - \pi_K$ for some k. Set $c_j = c_K$ for $j \notin \{k, K\}$. Then,

$$\frac{1}{2}c^T\mathbf{G}c = \left(\left(\pi_k - \pi_k^2\right)c_k - \left(\pi_K - \pi_K^2\right)c_K\right)(c_k - c_K) + 2\lambda(c_k - c_K)^2\left(\frac{\pi_k - \pi_k^2}{a_{k\bullet}} + \frac{\pi_K - \pi_K^2}{a_{K\bullet}} - (\pi_k - \pi_K)^2\right).$$

By Lemma 5, for $\lambda > 0$, $\mathbf{G}$ is indefinite. The proof is completed after remarking that the definiteness of $\mathbf{M}_2(0) - \mathbf{M}_2(\lambda)$ is determined by the definiteness of $\mathbf{G}(\lambda)$. □

The following result will be used for proving the result for $\hat{\boldsymbol{\pi}}_3(\lambda)$ when $K > 2$.

Lemma 6. *Define*

$$\psi(c_1, c_2, \lambda) = (c_1 - c_2)(c_1 w_{11} - c_2 w_{22}) + \lambda(c_1 - c_2)^2\left(\frac{w_{11}}{a_{11}} + \frac{w_{22}}{a_{22}} - (p_1 - p_2)^2\right),$$

for some $c_1 \in \mathbb{R}$, $c_2 \in \mathbb{R}$, $w_{11} \in \mathbb{R}$, $w_{22} \in \mathbb{R}$, $p_1 \in \mathbb{R}$, $p_2 \in \mathbb{R}$, $a_{11} \in \mathbb{R} - \{0\}$, $a_{22} \in \mathbb{R} - \{0\}$, $\lambda > 0$. If $w_{11} \neq w_{22}$, then, for each $\lambda > 0$, there exists some c_1 and c_2 such that $\phi(c_1, c_2, \lambda) > 0$ and some other c_1 and c_2 such that $\phi(c_1, c_2, \lambda) < 0$.

Proof. Let $\mathbf{H} = \mathbf{H}(\lambda)$ be a symmetric 2×2 matrix with entries

$$\mathbf{H}_{12} = w_{11} + \lambda\left(\frac{w_{11}}{a_{11}} + \frac{w_{22}}{a_{22}} - (p_1 - p_2)^2\right),$$
$$\mathbf{H}_{22} = w_{22} + \lambda\left(\frac{w_{11}}{a_{11}} + \frac{w_{22}}{a_{22}} - (p_1 - p_2)^2\right),$$
$$\mathbf{H}_{12} = \mathbf{H}_{21} = -\frac{1}{2}(w_{11} + w_{22}) - \lambda\left(\frac{w_{11}}{a_{11}} + \frac{w_{22}}{a_{22}} - (p_1 - p_2)^2\right).$$

The determinant of $\mathbf{H}$ is

$$\det(\mathbf{H}) = -\frac{1}{4}(w_{11} - w_{22})^2 \leq 0,$$

with equality applying if and only if $w_{11} = w_{22}$. Hence, whenever $w_{11} \neq w_{22}$, then, for each $\lambda > 0$, there exists some c_1 and c_2 such that $c^T\mathbf{H}c > 0$ and some other c_1 and c_2 such that $c^T\mathbf{H}c < 0$. The proof is completed after remarking that

$$\psi(c_1, c_2, \lambda) = c^T\mathbf{H}c$$

holds. □

Proof of Theorem 6. Calculate

$$\mathbf{M}_3(0) - \mathbf{M}_3(\lambda) = 4\lambda\mathbf{U}_1(\lambda)^{-1}\mathbf{G}(\lambda)\mathbf{U}_1(\lambda)^{-1},$$

where $\mathbf{G} = \mathbf{G}(\lambda)$ is a $K \times K$ symmetric matrix with entries

$$\{\mathbf{G}\}_{k,k} = 2\left(\pi_k - \pi_k^2\right)\frac{K-1}{K} + 4\lambda\left(\frac{(K-1)^2}{K^2}\frac{\left(\pi_k - \pi_k^2\right)}{a_{k\bullet}} + \frac{1}{K^2}\sum_{j=1, j\neq k}^{K}\frac{\pi_j - \pi_j^2}{a_{j\bullet}} - (\pi_k - \bar{\pi})^2\right), \quad k = 1, \ldots, K,$$

$$\{\mathbf{G}\}_{k,l} = 4\lambda\left(-\frac{(K-1)}{K^2}\left(\frac{\pi_k - \pi_k^2}{a_{k\bullet}} + \frac{\pi_l - \pi_l^2}{a_{l\bullet}}\right) + \frac{1}{K^2}\sum_{j=1, j\neq k, j\neq l}^{K}\frac{\pi_j - \pi_j^2}{a_{j\bullet}} - \pi_k\pi_l - \bar{\pi}^2 + \bar{\pi}(\pi_k + \pi_l)\right) - \frac{1}{K}\left(\pi_k - \pi_k^2 + \pi_l - \pi_l^2\right), \quad k, l = 1, \ldots, K, \quad k \neq l.$$

Let $c = (c_1, \ldots, c_K)^T$, for some constants c_k, $k = 1, \ldots, K$. Then, it can be shown that

$$\frac{1}{2}c^T\mathbf{G}c = \sum_{k=1}^{K}\left((\pi_k - \pi_k^2)c_k\left(c_k - \frac{1}{K}\sum_{k=1}^{K}c_k\right)\right) + 2\lambda\left[\sum_{k=1}^{K}\frac{\pi_k - \pi_k^2}{a_{k\bullet}}\left(c_k - \frac{1}{K}\sum_{k=1}^{K}c_k\right)^2 - \left(\sum_{k=1}^{K}c_k\pi_k - \bar{\pi}\sum_{k=1}^{K}c_k\right)^2\right].$$

If we set $c_k = a$, $k = 1, \ldots, K$, for some constant a, then it can be seen that

$$c^T\mathbf{G}_{-(K+1),-(K+1)}c = 0.$$

Cases (i) and (ii): Assume that $\pi_k - \pi_k^2 = \pi - \pi^2$, $k = 1, \ldots, K$. Then,

$$c^T\mathbf{G}c = 2(\pi - \pi^2)\left(\sum_{k=1}^{K}c_k^2 - \frac{1}{K}\left(\sum_{i=1}^{K}c_k\right)^2\right) + 4\lambda g,$$

for some constant

$$g = \sum_{k=1}^{K}\frac{\pi - \pi^2}{a_{k\bullet}}\left(c_k - \frac{1}{K}\sum_{k=1}^{K}c_k\right)^2 - \left(\sum_{k=1}^{K}c_k\pi_k - \bar{\pi}\sum_{k=1}^{K}c_k\right)^2.$$

Observe that, for $\lambda = 0$, $c^T\mathbf{G}c \geq 0$, with equality applying if and only if $c_k = a$, $k = 1, \ldots, K$, for some constant a. Then, it follows from the continuity argument that there exists some $\Lambda > 0$ such that, for $0 < \lambda < \Lambda$, $\mathbf{G}$ and thence also $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ are positive semidefinite. Assuming further that $\pi_k = \pi$, $k = 1, \ldots, K$, then $g > 0$; hence, for $\lambda > 0$, $\mathbf{G}$ and thence also $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ are positive semidefinite. *Case (iii)*: Now, let $\pi_k \neq \pi_l$ and $\pi_k \neq 1 - \pi_l$ for some pair $k \neq l$. Without loss of generality, assume that $\pi_1 - \pi_1^2 < \pi_2 - \pi_2^2$. Set c_1 to some constant and $c_2 = \cdots = c_K = c_0$ for some constant c_0. Then,

$$\frac{1}{2}c^T\mathbf{G}c = \frac{c_1 - c_0}{K}\left(c_1(K-1)\left(\pi_1 - \pi_1^2\right) - c_0\sum_{j=2}^{K}\left(\pi_j - \pi_j^2\right)\right) + 2\lambda\frac{(c_1 - c_0)^2}{K^2}\left((K-1)^2\frac{\pi_1 - \pi_1^2}{a_{1\bullet}} + \sum_{j=2}^{K}\frac{\pi_j - \pi_j^2}{a_{j\bullet}} - (\pi_1 - \bar{\pi})^2\right).$$

Now, let

$$w_{11} = \frac{K-1}{K}\left(\pi_1 - \pi_1^2\right), \quad w_{22} = \frac{1}{K}\sum_{j=2}^{K}\left(\pi_j - \pi_j^2\right), \quad p_1 = \frac{\sqrt{2}\pi_1}{K},$$

$$p_2 = \frac{\sqrt{2}\bar{\pi}}{K}, \quad a_{11} = \frac{Ka_{1\bullet}}{2(K-1)}, \quad a_{22} = \frac{Ka'_{2\bullet}}{2},$$

where $a'_{2\bullet} = \min(a_{2\bullet}, \ldots, a_{K\bullet})$ so that

$$\psi(c_1, c_0, \lambda) = \frac{c_1 - c_0}{K}\left(c_1(K-1)\left(\pi_1 - \pi_1^2\right) - c_0\sum_{j=2}^{K}\left(\pi_j - \pi_j^2\right)\right) + 2\lambda\frac{(c_1 - c_0)^2}{K^2}\left((K-1)^2\frac{\pi_1 - \pi_1^2}{a_{1\bullet}} + \sum_{j=2}^{K}\frac{\pi_j - \pi_j^2}{a'_{2\bullet}} - (\pi_1 - \bar{\pi})^2\right).$$

Since, by Case (iii),

$$(K-1)\left(\pi_1 - \pi_1^2\right) \neq \sum_{j=2}^{K}\left(\pi_j - \pi_j^2\right)$$

holds, then, by Lemma 6, for all $\lambda > 0$, there exists some c_1 and c_0 such that $\psi(c_1, c_0, \lambda) < 0$ holds. Since

$$\sum_{j=2}^{K}\frac{\pi_j - \pi_j^2}{a_{j\bullet}} < \sum_{j=2}^{K}\frac{\pi_j - \pi_j^2}{a'_{2\bullet}}$$

holds, thence, for all $\lambda > 0$, there exists some c_1 and c_0 such that $c^T\mathbf{G}c < 0$. Now, let

$$a'_{2\bullet} = \max(a_{2\bullet}, \ldots, a_{K\bullet}).$$

By Lemma 6, for all $\lambda > 0$, there exists some c_1 and c_0 such that $\psi(c_1, c_0, \lambda) > 0$ holds. Since

$$\sum_{j=2}^{K}\frac{\pi_j - \pi_j^2}{a_{j\bullet}} > \sum_{j=2}^{K}\frac{\pi_j - \pi_j^2}{a'_{2\bullet}}$$

holds, thence there exists some c_1 and c_0 such that $c^T\mathbf{G}c > 0$. Hence, $\mathbf{G}$ and thence also $\mathbf{M}_3(0) - \mathbf{M}_3(\lambda)$ are indefinite. □

Proof of Theorem 7. Since $\mathbf{B}$ is diagonal, we have

$$\text{trace}(\mathbf{B}(\mathbf{M}_j(0) - \mathbf{M}_j(\lambda))) = \sum_{k=1}^{K} b_k\{\mathbf{M}_j(0) - \mathbf{M}_j(\lambda)\}_{kk}, \quad j = 1, 2, 3.$$

By the positive semidefiniteness of $\mathbf{B}$, $b_k \geq 0$, $k = 1, \ldots, K$, with strict inequality applying to at least one k. Proceed as in the proof of Theorem 1 to complete the proof. □

APPENDIX B

A NOTE ON DETERMINING THE OPTIMAL AMOUNT OF PENALTY

Here, we show how it is possible to use numerical methods to optimize λ in terms of the estimated GMSE of the estimated probabilities for the particular data considered in this paper, by using the estimated second-order moment matrices. Let $\hat{\mathbf{M}}_j(\lambda)$, $j = 1, 2, 3$, denote the estimate of the second-order moment matrix $\mathbf{M}_j(\lambda)$, $j = 1, 2, 3$ defined in Equations (A1), (A2), and (A3),

TABLE B1 Optimized values of λ, using *oracle* (λ_O^*), estimated second-order moment matrices (λ_M^*), and leave-one-out cross-validation (λ_D^*), as well as the relative improvement of empirical generalized mean squared errors over the maximum likelihood estimator (RP_O, RP_M, and RP_D, when using λ_O^*, λ_M^*, and λ_D^*, respectively) averaged over 1,000 simulated data sets for different examples, when using different shrinkage estimators and matrices **B** (the numbers in brackets are standard deviations)

		$\hat{\boldsymbol{\pi}}_1(\lambda)$			$\hat{\boldsymbol{\pi}}_2(\lambda)$			$\hat{\boldsymbol{\pi}}_3(\lambda)$		
Example		$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$
(i)	λ_O^*	0.44	0.44	0.18	10	10	10	10	10	10
	λ_M^*	0.51 (0.43)	1.19 (2.13)	0.22 (0.23)	4.62 (3.55)	5.35 (4.1)	4.83 (4.19)	5.78 (3.56)	5.81 (3.91)	5.93 (4.21)
	λ_D^*	0.97 (0.61)	0.97 (0.61)	0.97 (0.61)	1.6 (2.08)	1.6 (2.08)	1.6 (2.08)	2.53 (2.66)	2.53 (2.66)	2.53 (2.66)
	RP_O	11	15	4	70	80	15	67	79	14
	RP_M	−10	−40	−25	40	37	6	40	37	6
	RP_D	−21	−30	−114	22	22	3	23	25	3
(ii)	λ_O^*	0.44	0.44	1.05	0.39	2.04	10	0.42	0.71	10
	λ_M^*	0.52 (0.45)	1.22 (2.14)	2.72 (3.48)	0.36 (0.25)	3.19 (3.28)	5.73 (4.38)	0.49 (0.44)	2.32 (3.48)	6.11 (4.47)
	λ_D^*	0.97 (0.64)	0.97 (0.64)	0.97 (0.64)	0.4 (0.23)	0.4 (0.23)	0.4 (0.23)	0.56 (0.2)	0.56 (0.2)	0.56 (0.2)
	RP_O	12	14	20	11	49	15	8	14	12
	RP_M	−10	−39	−12	0	8	5	−9	−27	3
	RP_D	−21	−28	5	5	17	2	−1	3	2
(iii)	λ_O^*	0.31	0.44	1.67	0.09	0.08	0.1	0.14	0.12	0.17
	λ_M^*	0.34 (0.24)	1.22 (2.11)	2.09 (1.56)	0.12 (0.11)	0.1 (0.1)	0.14 (0.16)	0.17 (0.16)	0.15 (0.15)	0.22 (0.25)
	λ_D^*	0.6 (0.29)	0.6 (0.29)	0.6 (0.29)	0.37 (0.25)	0.37 (0.25)	0.37 (0.25)	0.48 (0.24)	0.48 (0.24)	0.48 (0.24)
	RP_O	9	15	43	3	3	2	3	3	1
	RP_M	−12	−40	5	−17	−26	−15	−17	−26	−15
	RP_D	−18	−6	21	−38	−56	−23	−32	−47	−13

(Continues)

TABLE B1 (Continued)

(iv)	λ_O^*	0.65	0.44	0.32	1.92	1.17	0.54	4.01	2.38	1.27
	λ_M^*	1.03 (1.39)	1.28 (2.14)	0.61 (1.19)	2.88 (3.39)	3.51 (4.17)	2.5 (3.74)	3.91 (3.52)	3.98 (3.98)	3.22 (3.97)
	λ_D^*	2.8 (3.12)	2.8 (3.12)	2.8 (3.12)	0.83 (1)	0.83 (1)	0.83 (1)	1.53 (1.46)	1.53 (1.46)	1.53 (1.46)
	RP_O	13	13	6	27	26	1	33	35	1
	RP_M	−10	−42	−31	13	0	−5	17	7	−4
	RP_D	−36	−88	−132	12	10	0	15	13	0
(v)	λ_O^*	0.03	0.01	0.03	0.04	0.01	0	0.05	0.02	0
	λ_M^*	0.04 (0.05)	0.01 (0.05)	0.04 (0.06)	0.05 (0.06)	0.02 (0.05)	0.01 (0.03)	0.05 (0.08)	0.02 (0.08)	0.01 (0.05)
	λ_D^*	0.08 (0.26)	0.08 (0.26)	0.08 (0.26)	0.05 (0.16)	0.05 (0.16)	0.05 (0.16)	0.08 (0.23)	0.08 (0.23)	0.08 (0.23)
	RP_O	0	0	0	1	0	0	0	0	0
	RP_M	−4	−48	−6	−4	−47	−2	−4	−47	−2
	RP_D	−24	−282	−25	−14	−167	−12	−15	−171	−12

Note. Example (i) is the example with equal event probabilities for each category ($\boldsymbol{\pi} = (0.2, 0.2, 0.2)^T$), Example (ii) is the scenario with equal variances ($\boldsymbol{\pi} = (0.2, 0.8, 0.2)^T$), Example (iii) is the scenario with different event probabilities and variances ($\boldsymbol{\pi} = (0.2, 0.9, 0.9)^T$), Example (iv) is the scenario with different event probabilities and variances ($\boldsymbol{\pi} = (0.2, 0.3, 0.4)^T$), and Example (v) is the scenario with a category with rare events ($\boldsymbol{\pi} = (0.01, 0.5, 0.5)^T$). In Examples (i)–(iv), the number of subjects per group was $a_{1\bullet} = 10$, $a_{1\bullet} = 20$, and $a_{2\bullet} = 30$, whereas in Example (v), $a_{1\bullet} = 10$ and $a_{2\bullet} = a_{3\bullet} = 50$ were used. Matrices $\mathbf{B}_j$, $j = 1, 2, 3$, were set to an identity matrix, a matrix with the first diagonal element equal to 1 and 0 elsewhere, and a matrix of ones, respectively.

respectively, which is obtained by replacing the unknown π_k, $k = 1, \ldots, K$, with their respective MLEs (bootstrap could easily be used to account for potential issues with overfitting of the MLE). Observe that the only unknown parameter in $\hat{\mathbf{M}}_j(\lambda)$ is λ; hence, trace($\mathbf{B}\hat{\mathbf{M}}_j(\lambda)$) can be optimized for any positive semidefinite matrix $\mathbf{B}$ over some range of λ by using numerical methods. For the type of data considered here, this numerical optimization is also feasible when using leave-one-out cross-validated deviance, since the leave-one-out cross-validated estimates for all the considered estimators have, in this case, nice closed-form expressions. We compare these approaches with the *oracle* approach, where λ is optimized based on the (true) second-order moment matrix.

Applying this idea to the examples considered in Section 7 (simulating 1,000 data sets for each example), we obtain (using the R function **optimize** over [0, 10]), for various choices of $\mathbf{B}$, the optimal values of λ, which (averaged over the 1,000 simulated data sets) are presented in Table B1. We can see in Table B1 that the averaged optimal values of λ differ from the optimal values obtained by the *oracle* approach and that there is a lot of variability introduced by estimating λ. This results in either a much smaller empirical relative improvement over the MLE than would be theoretically feasible by using the *oracle* approach or even in a diminished performance in comparison with the MLE (see Table B1).

APPENDIX C

(ASYMPTOTIC) GMSE OF THE ESTIMATED COEFFICIENTS

Here, we consider a model with $p < n$ predictor variables, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$, $i = 1, \ldots, n$, so that

$$\log \frac{\pi_i}{1-\pi_i} = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \text{ for } i = 1, \ldots, n, \tag{C1}$$

where $\pi_i = P(Y_i = 1|\mathbf{X}_i)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of logistic regression coefficients. Assume that the $n \times p$ matrix $\boldsymbol{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ is of full rank. Let the log-likelihood function based on (C1) be $l(\boldsymbol{\beta})$, and let

$$l^{\lambda}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

be the penalized log-likelihood for some nonnegative penalty parameter $\lambda > 0$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(\lambda)$ denote the MLE and the penalized MLE of $\boldsymbol{\beta}$, respectively. The asymptotic second-order moment matrix of $\hat{\boldsymbol{\beta}}(\lambda)$ is, under some regularity conditions (see (le Cessie & van Houwelingen, 1992),

$$\mathbf{M}(\lambda) = \mathbf{v}(\lambda) + \mathbf{b}(\lambda)\mathbf{b}(\lambda)^T,$$

where $\mathbf{v}(\lambda)$ and $\mathbf{b}(\lambda)$ are the asymptotic variance and the bias of the ridge estimator given by

$$\mathbf{b}(\lambda) = -\lambda(I(\boldsymbol{\beta}) + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}$$

and

$$\mathbf{v}(\lambda) = (I(\boldsymbol{\beta}) + \lambda\mathbf{I})^{-1}I(\boldsymbol{\beta})(I(\boldsymbol{\beta}) + \lambda\mathbf{I})^{-1},$$

respectively, where $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$ and $\mathbf{I}$ is the identity matrix.

Proposition 3. *Assume that (1) $\boldsymbol{\beta}^T\boldsymbol{\beta} < \infty$ holds. Then, for some $\lambda > 0$, $\mathbf{M}(0) - \mathbf{M}(\lambda)$ is positive definite.*

Proof. Simple calculation yields

$$\mathbf{M}(0) - \mathbf{M}(\lambda) = \lambda(I(\boldsymbol{\beta}) + \lambda\mathbf{I})^{-1}\left[2\mathbf{I} + \lambda I(\boldsymbol{\beta})^{-1} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}^T\right](I(\boldsymbol{\beta}) + \lambda\mathbf{I})^{-1},$$

$\lambda > 0$. By Condition (1), we need to show that there is some $\lambda > 0$ such that $2\mathbf{I} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}^T$ is positive definite, since (1) implies that $I(\boldsymbol{\beta})$ is positive definite. Now, since p eigenvalues of $\lambda\boldsymbol{\beta}\boldsymbol{\beta}^T$ are zero and one is $\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$, the condition is $2 - \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} > 0$. By Condition (1), there exists some $\lambda > 0$ such that the condition holds, thus proving the proposition. □