Journal of Statistical Planning and Inference 220 (2022) 1–14

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Robust active learning with binary responses

Jesús López-Fidalgo [a,b,*], Douglas P. Wiens [c,1]

[a] Institute of Data Science and Artificial Intelligence, University of Navarre, Edificio Ismael Sánchez Bella, Pamplona, 31009, Spain
[b] Tecnun Escuela de Ingeniería, Universidad de Navarra, Manuel de Lardizábal 13, 20018 San Sebastián, Spain
[c] Department of Mathematical & Statistical Sciences, University of Alberta, Edmonton, Canada, T6G 2G1

ARTICLE INFO

*Article history:*
Received 17 April 2021
Received in revised form 17 October 2021
Accepted 22 January 2022
Available online 1 February 2022

*MSC:*
primary 62F35
secondary 62D99

*Keywords:*
Logistic regression
Loss function
Mislabelling
Optimal subsampling
Probit regression
Sequential sampling

ABSTRACT

We introduce a method of Robust Learning ('ROBL') for binary data, and propose its use in situations where Active Learning is appropriate, and where sampling the predictors is easy and cheap, but learning the responses is hard and expensive. We seek robustness against both modelling errors and the mislabelling of the binary responses. Thus we aim to sample effectively from the population of predictors, and learn the responses only for an 'influential' sub-population. This is carried out by probability weighted sampling, for which we derive optimal 'unbiased' sampling weights, and weighted likelihood estimation, for which we also derive optimal estimation weights. The robustness issues can lead to biased estimates and classifiers; it is somewhat remarkable that our weights eliminate the mean of the bias – which is a random variable as a result of the sampling – due to both types of errors mentioned above. These weights are then tailored to minimize the mean squared error of the predicted values. Simulation studies indicate that when bias is of significant concern, ROBL allows for substantial reductions, relative to Passive Learning, in the prediction errors. The methods are then illustrated in real-data analyses.

## 1. Introduction and summary

The case for robust methods of data analysis, especially when the data are processed without intelligent intervention, has been convincingly made by Huber (1981). That for robust methods of design, in the face of model uncertainties, was as well made by Wiens (2015). For binary models, Copas (1988) proposed methods robust against mislabelling of binary data; this work was extended by Carroll and Pederson (1993). Copas restricted his theory and examples to rather small rates of mislabelling; in his discussion of this paper D. Cox expressed surprise at this restriction. Indeed, Meyer and Mittag (2017) report rates exceeding 20% in economic data.

With the advent of 'Big Data', huge – indeed, astronomical – data sets are becoming increasingly prevalent. With them comes an increased need not only for robustness but for data reduction, prior to the analysis, in order to reduce not only the computational burden but also the costs of sampling. For this, data subsampling has been proposed (Wang et al., 2018b; Nie et al., 2018; Drovandi et al., 2017; Wang et al., 2018a; Nachtsheim and Stufken, 2019; Wang et al., 2019). This in turn leads to revisiting methods of Optimal Experimental Design (OED) – as Drovandi et al. (2017) and Nie et al. (2018) point out, the goals and techniques are not dissimilar. The major difference is perhaps that in OED we have complete control over the predictors, whereas in Sampling Design these, and perhaps the responses as well, are sampled. But the

[*] Corresponding author.
*E-mail addresses:* fidalgo@unav.es (J. López-Fidalgo), doug.wiens@ualberta.ca (D.P. Wiens).

[1] Supplementary material is at www.ualberta.ca/~dwiens/.

https://doi.org/10.1016/j.jspi.2022.01.004

sampling need not be done uniformly. Simple Random Sampling (SRS) is termed Passive Learning (PASSL) in the machine learning literature; statisticians tend to be drawn instead to Active Learning, whereby the sampling is done in a more controlled, supervised manner, akin to OED.

Wang et al. (2018b) proposed a method of subsampling in active learning, with the goal being classification. In Wang et al. (2019) this was done for linear regression instead. There are two major ways in which that work departs from what is presented here, in a classification context.

(i) The framework of Wang et al. (2018b) and Wang et al. (2019) was that one was to subsample from a population containing both predictors *and responses.* Here we instead envision a scenario in which predictors can be sampled quickly and inexpensively (e.g. through electronic medical databases), whereas the responses are much more difficult to obtain (e.g. medical outcomes). One hopes to have to query the responses only in a relatively small number of cases, although the number of predictors sampled might be much larger.

(ii) In Wang et al. (2018b) and Wang et al. (2019) the model fitted to the data was assumed to be of a correct form. Here we place only slight and tentative faith in the fitted model, and instead seek sampling designs which are *robust* against both model misspecifications, and possible labelling errors. The former goal again has been well-developed in the field of Robustness of Experimental Design — see Wiens (2015) for a comprehensive review.

In Nie et al. (2018) this programme was carried out for regression; the resulting sampling schemes were found, in comparison with competing methods and by application to real data sets, to be both robust and efficient. Our approach in this current work is similar. We assume that, once gathered, the data will be analysed by fitting a Generalized Linear Model (GLM) – we focus on logistic and probit regression – through solving a weighted likelihood equation. As a result of the robustness issues raised above, the asymptotically normal estimates and resulting classifiers are biased. We propose estimation weights whereby, for given sampling weights, the asymptotic mean bias is eliminated (we write 'mean bias' since, as a result of the sampling, the bias is a random variable). It seems somewhat remarkable that the contributions to the means of both sources of error – model misspecification and response mislabelling – can be simultaneously eliminated. We then propose sampling weights under which the mean squared prediction errors are minimized. Comparisons with Passive Learning ( PASSL) illustrate the gains to be made, in terms of decreased prediction errors – or, equivalently, in terms of the reduction in the number of responses which must be queried in order to attain the same precision – from our Robust Learning (ROBL) approach.

An outline of the rest of the article is as follows. In Section 2 we present a precise formulation of the robustness issues to be addressed. Section 3 puts this in an asymptotic context, followed by our proposed solutions and algorithm in Section 4. These are tested in a simulation study in Section 5, and applied to examples from the machine learning literature in Section 6 and §7. The article concludes with a discussion in the Section 8. Proofs are in Appendix. The computing code used is available online, at ualberta.ca/~dwiens/home page/pubs/robl code.zip.

## 2. Model and labelling robustness

We aim to develop a theory of robust classification, and a methodology for applying this to large data sets such as arise in machine learning. The general idea is that there is a (large) population $\mathcal{Q} \subset \mathbb{R}^k$ of explanatory variables, which can be easily sampled. We denote by $\chi = \{\boldsymbol{x}_i\}_{i=1}^N$ the *distinct* values of $\boldsymbol{x} \in \mathcal{Q}$. With a certain probability, possibly depending on unknown parameters $\boldsymbol{\theta}_{d\times 1}$, an item with covariates $\boldsymbol{x}$ belongs to group 'A' (indicated by $Y = 1$), otherwise it belongs to group 'B' ($Y = 0$). We suppose that the determination of the appropriate group, given $\boldsymbol{x}$, is difficult and expensive, so that the investigator wishes to sample from $\mathcal{Q}$ in a manner which is more efficient than random sampling (i.e. PASSL).

More precisely, we let $q(\cdot)$ be the probability mass function (pmf) of $\boldsymbol{x} \in \chi$, assumed strictly positive. This is the relative frequency with which $\boldsymbol{x}$ appears in $\mathcal{Q}$, and need not be known – SRS from $\mathcal{Q}$ yields an empirical pmf $\hat{q}$ on $\chi$, converging in probability to $q$. We wish to find an 'optimal' pmf $p(\cdot)$ on $\chi$, and sample from the sub-population with this distribution. This will involve probability weighted sampling, with weights $\pi(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$ for $\boldsymbol{x} \in \chi$. We derive $\pi(\cdot)$; for this $p(\cdot)$ need not be known.

We consider a generalized linear model framework. Let $G$ be a strictly increasing, absolutely continuous distribution function $G(\cdot)$ with a density $g(\cdot)$ possessing two bounded derivatives. We concentrate on the cases $G(t) = 1/\left(1 + e^{-\sigma t}\right)$ (logistic regression; $\sigma = \pi/\sqrt{3}$ so that $\text{VAR}_G = 1$) and $G(t) = \Phi(t)$ (probit regression). Set

$$\beta(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{\theta},$$

for regressors $\boldsymbol{f}(\boldsymbol{x})$ and an unknown, $d$-dimensional parameter $\boldsymbol{\theta}$ ranging over a compact, convex space $\Theta$. The experimenter will model $\Pr(Y = 1|\boldsymbol{x}) \overset{def}{=} P(1|\boldsymbol{x})$ as

$$P(1|\boldsymbol{x}) = G(\beta(\boldsymbol{x}; \boldsymbol{\theta})), \tag{1}$$

for some $\boldsymbol{\theta}$, to be estimated.

Our focus is on efficient and robust estimation and prediction of $P(1|\boldsymbol{x})$. To quantify the robustness against both modelling and labelling errors, first define

$$\beta_*\left(\boldsymbol{x}; \boldsymbol{\theta}_\gamma\right) = \beta\left(\boldsymbol{x}; \boldsymbol{\theta}_\gamma\right) + \psi(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{\theta}_\gamma$ is the 'true' parameter under the mislabelling model described below, and $\psi(\boldsymbol{x})$ represents unknown model error constrained by

$$E_q\left[\psi^2(\boldsymbol{x})\right] \le \tau_1^2/n, \tag{3}$$

for a constant $\tau_1$ controlling the magnitude of the error in the incorrectly fitted model (1), and an anticipated sample size $n$. That $\psi$ be of order $1/\sqrt{n}$ is akin to the notion of contiguity in hypothesis testing (see, e.g., Le Cam (1960)) – and ensures that bias and standard error are of the same asymptotic order. This allows us to phrase the discussion in terms of asymptotic mean squared error.

Define as well

$$H_\gamma(t) = (1-\gamma)\,G(t) + \gamma\bar{G}(t),$$

where $\bar{G} = 1 - G$ and $\gamma$ is a mislabelling probability — see Copas (1988) and Carroll and Pederson (1993). We assume that

$$0 \le \gamma \le \min\left(\tau_2/\sqrt{n}, .5\right), \tag{4}$$

for a user-specified $\tau_2$. Our asymptotic statements will then always assume that $n > 4\tau_2^2$. Conditions (3) and (4) are necessary for a sensible asymptotic treatment; the dependence on $n$ is moot if $n$ is fixed.

We suppose that in fact

$$P(1|\boldsymbol{x}) = H_\gamma\left(\beta_*\left(\boldsymbol{x};\boldsymbol{\theta}_\gamma\right)\right), \tag{5}$$

so that in fitting (1) the experimenter, possibly incorrectly, takes $\gamma = 0$, $\psi \equiv 0$.

Neither $\boldsymbol{\theta}_\gamma$ nor $\psi$ are completely defined by (2) – for instance we can replace $\boldsymbol{\theta}_\gamma$ by $\boldsymbol{\theta}_\gamma + \boldsymbol{\theta}_\dagger$, and $\psi(\boldsymbol{x})$ by $\psi(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{\theta}_\dagger$. We address this by *defining* $\boldsymbol{\theta}_\gamma$ by

$$\boldsymbol{\theta}_\gamma = \arg\min_{\boldsymbol{\tau}} E_q\left[\left(H_\gamma^{-1}\{P(1|\boldsymbol{x})\} - \boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{\tau}\right)^2\right], \tag{6}$$

whence $\psi$ in (2) may be *defined* by $\psi(\boldsymbol{x}) = H_\gamma^{-1}\{P(1|\boldsymbol{x})\} - \boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{\theta}_\gamma$. A consequence of the minimization is the condition

$$E_q\left[\boldsymbol{f}(\boldsymbol{x})\,\psi(\boldsymbol{x})\right] = \boldsymbol{0}. \tag{7}$$

The following assumptions are made; the second is required for the asymptotics, to ensure the existence of a root of the likelihood equation.

**Assumption (A1)** The population $\chi$ is such that no non-trivial linear combination $\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{x})$, $\boldsymbol{c} \neq \boldsymbol{0}$ of the predictors is identically zero on $\chi$.

**Assumption (A2)** For all $\boldsymbol{x} \in \chi$, $G(\beta(\boldsymbol{x};\boldsymbol{\theta}_0)) < 1$.

Assumption (A1) implies in particular that $\boldsymbol{M}_q \overset{def}{=} E_q\left[\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})\right] \succ \boldsymbol{0}$ (i.e. is positive definite). For logistic and probit regression, (A2) requires only that $\|\boldsymbol{f}(\boldsymbol{x})\|$ be finite on $\chi$.

By (A1) the solution to the least squares problem (6) is

$$\boldsymbol{\theta}_\gamma = \boldsymbol{M}_q^{-1}E_q\left[\boldsymbol{f}(\boldsymbol{x})\,H_\gamma^{-1}\{P(1|\boldsymbol{x})\}\right].$$

The parameter $\boldsymbol{\theta}_\gamma$ furnishes the best $L_2$-approximation of $H_\gamma^{-1}\{P(1|\boldsymbol{x})\}$ by $\boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{\theta}$, in the presence of both modelling and labelling errors. The parameter $\boldsymbol{\theta}_0$ of interest assumes no mislabelling, and so is

$$\boldsymbol{\theta}_0 = \boldsymbol{M}_q^{-1}E_q\left[\boldsymbol{f}(\boldsymbol{x})\,G^{-1}\{P(1|\boldsymbol{x})\}\right], \tag{8}$$

whence

$$\boldsymbol{\theta}_\gamma - \boldsymbol{\theta}_0 = \boldsymbol{M}_q^{-1}E_q\left[\boldsymbol{f}(\boldsymbol{x})\left(H_\gamma^{-1}\{P(1|\boldsymbol{x})\} - G^{-1}\{P(1|\boldsymbol{x})\}\right)\right]. \tag{9}$$

From the definition of $H_\gamma$ we obtain the expansion

$$H_\gamma^{-1}(z) = G^{-1}\left(\frac{z-\gamma}{1-2\gamma}\right) = G^{-1}(z) + \frac{\gamma\,(2z-1)}{g\left(G^{-1}(z)\right)} + O\left(\gamma^2\right),$$

and then substituting $z = P(1|\boldsymbol{x}) = H_\gamma\left(\beta_*\left(\boldsymbol{x};\boldsymbol{\theta}_\gamma\right)\right) = G\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}_\gamma\right)\right) + O\left(n^{-1/2}\right)$ gives

$$H_\gamma^{-1}\{P(1|\boldsymbol{x})\} = G^{-1}(P(1|\boldsymbol{x})) + \gamma\,\frac{2G\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}_\gamma\right)\right) - 1}{g\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}_\gamma\right)\right)} + O\left(n^{-1}\right). \tag{10}$$

With

$$h\left(\boldsymbol{x};\boldsymbol{\theta}\right)\overset{def}{=}\frac{2G\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)-1}{g\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)},$$

the expansion (10) in (9) then gives $\boldsymbol{\theta}_{\gamma}-\boldsymbol{\theta}_0=\left(E_q\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\right]\right)^{-1}E_q\left[\boldsymbol{f}\left(\boldsymbol{x}\right)h\left(\boldsymbol{x};\boldsymbol{\theta}_{\gamma}\right)\right]\gamma+O\left(n^{-1}\right)$. In particular $\boldsymbol{\theta}_{\gamma}-\boldsymbol{\theta}_0=O\left(n^{-1/2}\right)$, whence (9) becomes

$$\boldsymbol{\theta}_{\gamma}-\boldsymbol{\theta}_0=\boldsymbol{M}_q^{-1}E_q\left[\boldsymbol{f}\left(\boldsymbol{x}\right)h\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\right]\gamma+O\left(n^{-1}\right).$$

## 3. Asymptotic theory

When sampling from $p\left(\cdot\right)$, the $n_i$ are observed values of a multinomial random vector

$$\boldsymbol{S}=\left(S_1,\cdots,S_N\right)'\sim multi\left(n,p_1=p\left(\boldsymbol{x}_1\right),\cdot\cdot\cdot,p_N=p\left(\boldsymbol{x}_N\right)\right).$$

With $\boldsymbol{p}=\left(p_1,\cdots,p_N\right)'$ and $\boldsymbol{D}_p=diag\left(p_1,\cdots,p_N\right)$ we have that

$$E\left[\boldsymbol{S}\right]=n\boldsymbol{p},\ \operatorname{cov}\left[\boldsymbol{S}\right]=n\left[\boldsymbol{D}_p-\boldsymbol{p}\boldsymbol{p}'\right],\ n^{-1}\boldsymbol{S}\overset{pr}{\to}\boldsymbol{p}. \tag{11}$$

We require the function

$$r\left(t\right)=\frac{d}{dt}\log\frac{G\left(t\right)}{\bar{G}\left(t\right)}=\frac{g\left(t\right)}{G\left(t\right)\bar{G}\left(t\right)},$$

with derivative $\dot{r}\left(t\right)$. If $G$ is the logistic distribution then $r\equiv\sigma$. Define

$$\alpha\left(\boldsymbol{x};\boldsymbol{\theta}\right)=r\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)g\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right).$$

Given a sample of size $n$ from $p\left(\boldsymbol{\cdot}\right)$, suppose that $\boldsymbol{x}_i$ appears $n_i\geq 0$ times ($\sum_{i=1}^{N}n_i=n$). Upon learning the group memberships, denote by $y_i$ the proportion of occurrences of group A, so that $n_iy_i$ is the number of such occurrences. The estimate is defined as the root of a (weighted, with weights $w\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)$) likelihood equation:

$$\boldsymbol{U}_n\left(\boldsymbol{\theta}\right)\overset{def}{=}\sum_{i=1}^{N}\frac{n_i}{n}w\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)r\left(\beta\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right)\left(y_i-G\left(\beta\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right)\right)\boldsymbol{f}\left(\boldsymbol{x}_i\right)=\boldsymbol{0}. \tag{12}$$

Note that $E\left[-\dot{\boldsymbol{U}}_n\left(\boldsymbol{\theta}_0\right)|\boldsymbol{S}\right]$ is the conditional weighted information matrix, given $\boldsymbol{S}=\left(n_1,\cdots,n_N\right)$.

Define

$$\begin{aligned}v\left(\boldsymbol{x};\boldsymbol{\theta}\right)&=G\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)\bar{G}\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right),\\ v_*\left(\boldsymbol{x};\boldsymbol{\theta}\right)&=H_{\gamma}\left(\beta_*\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)\bar{H}_{\gamma}\left(\beta_*\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right),\end{aligned}$$

where $\bar{H}_{\gamma}=1-H_{\gamma}$. Note that $v\left(\boldsymbol{x};\boldsymbol{\theta}\right)$ (resp. $v_*\left(\boldsymbol{x};\boldsymbol{\theta}\right)$) is the variance of $Y_{|\boldsymbol{x}}$ under (1) (resp. (5)). An occasionally useful identity is

$$\alpha\left(\boldsymbol{x};\boldsymbol{\theta}\right)=v\left(\boldsymbol{x};\boldsymbol{\theta}\right)r^2\left(\beta\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right).$$

Define $\boldsymbol{\psi}_{N\times 1}=\left(\cdots,\psi\left(\boldsymbol{x}_i\right),\cdots\right)'$, and

$$\begin{aligned}\boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right)&=E_p\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\alpha\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)w\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\right],\\ \boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)&=E_p\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\alpha\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)w^2\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\right],\\ \boldsymbol{b}_{\psi}\left(\boldsymbol{\theta}_0\right)&=E_p\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\alpha\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)w\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\psi\left(\boldsymbol{x}\right)\right],\\ \boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)&=E_q\left[\boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_q^{-1}\boldsymbol{f}\left(\boldsymbol{x}\right)h\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\right]-E_p\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\alpha\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)w\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)h\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\right],\\ \boldsymbol{b}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)&=\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)\gamma+\boldsymbol{b}_{\psi}\left(\boldsymbol{\theta}_0\right).\end{aligned}$$

The presence of $\boldsymbol{\psi}$ and $\gamma$ in the vector $\boldsymbol{b}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)$ determine the bias of the estimate. The matrix $\boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right)$ is the asymptotic unconditional weighted information matrix, i.e. the limit, as $n\to\infty$, of $E_{\boldsymbol{S}}E\left[-\dot{\boldsymbol{U}}_n\left(\boldsymbol{\theta}_0\right)|\boldsymbol{S}\right]$; it and $\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)$ appear in the 'sandwich' covariance matrix. For constant weights $w\equiv 1$, $\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)=\boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right)$.

**Lemma 1.** *Supposing that $\boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right)\succ\boldsymbol{0}$, and ignoring terms which are $O\left(n^{-1}\right)$, the unconditional mean squared error matrix $E_{\boldsymbol{S}}\left[E\left(\left\{\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right\}\left\{\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right\}'|\boldsymbol{S}\right)\right]$ is*

$$\mathrm{MSE}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)=\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\left\{\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)+n\boldsymbol{b}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)\boldsymbol{b}'_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)\right\}\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right), \tag{13}$$

*as $n\to\infty$.*

See Remark 1 in the next section, where the supposition of Lemma 1 is shown to be implied by (A1), for the estimation and sampling schemes considered in this article.

The proof of Lemma 1 is in the Appendix; we give a brief outline here. We first condition on the sample **S**, and apply results of Fahrmeir (1990) – here is where (A2) is used – to assert the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$. From this the conditional MSE matrix is obtained; averaging over **S** yields the final result.

## 4. Unbiased sampling

A sensible measure of the loss is the Mean Squared Prediction Error of $\boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_n$ as a predictor of $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}_0$:

$$\text{MSPE} = E_q\left[E\left[\left\{\sqrt{n}\left(\boldsymbol{f}'(\boldsymbol{x})\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right)\right\}^2\right]\right]. \tag{14}$$

Define $\pi(\boldsymbol{x};\boldsymbol{\theta}_0) = p(\boldsymbol{x})/q(\boldsymbol{x})$, so that

$$p(\boldsymbol{x}) = \pi(\boldsymbol{x};\boldsymbol{\theta}_0)\, q(\boldsymbol{x}).$$

We propose probability weighted sampling from $q(\cdot)$, under which $\boldsymbol{x}$ is chosen from $\chi$ with probability $q(\boldsymbol{x})$ (i.e., by SRS from $\mathcal{Q}$), and accepted into the sample with a procedure, explained below, according to the sampling weights $\pi(\boldsymbol{x};\boldsymbol{\theta}_0)$. We shall normalize $\pi$ by $E_q[\pi(\boldsymbol{x};\boldsymbol{\theta}_0)] = 1$, ensuring that $p(\boldsymbol{x})$ is a pmf on $\chi$. Under such a scheme,

$$\begin{aligned}
\boldsymbol{M}_1(\boldsymbol{\theta}_0) &= E_q\left[\boldsymbol{f}(\boldsymbol{x})\,\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\, w(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi(\boldsymbol{x};\boldsymbol{\theta}_0)\boldsymbol{f}'(\boldsymbol{x})\right],\\
\boldsymbol{M}_2(\boldsymbol{\theta}_0) &= E_q\left[\boldsymbol{f}(\boldsymbol{x})\,\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\, w^2(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi(\boldsymbol{x};\boldsymbol{\theta}_0)\boldsymbol{f}'(\boldsymbol{x})\right],\\
\boldsymbol{b}_\psi(\boldsymbol{\theta}_0) &= E_q\left[\boldsymbol{f}(\boldsymbol{x})\,\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\, w(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi(\boldsymbol{x};\boldsymbol{\theta}_0)\,\psi(\boldsymbol{x})\right],\\
\boldsymbol{b}_h(\boldsymbol{\theta}_0) &= E_q\left[\left\{\boldsymbol{M}_1(\boldsymbol{\theta}_0)\boldsymbol{M}_q^{-1} - \alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\, w(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi(\boldsymbol{x};\boldsymbol{\theta}_0)\boldsymbol{I}_d\right\}\boldsymbol{f}(\boldsymbol{x})\, h(\boldsymbol{x};\boldsymbol{\theta}_0)\right],
\end{aligned} \tag{15}$$

and $\text{MSPE} = E_q\left[\boldsymbol{f}'(\boldsymbol{x})\,\text{MSE}_{\psi,\gamma}(\boldsymbol{\theta}_0)\boldsymbol{f}(\boldsymbol{x})\right] = tr\left\{\boldsymbol{M}_q\text{MSE}_{\psi,\gamma}(\boldsymbol{\theta}_0)\right\}$ is, ignoring terms which are $O\left(n^{-1}\right)$, given by

$$\begin{aligned}
\text{MSPE} =\ & tr\left\{\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_2(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\right\}\\
&+ n\left(\boldsymbol{b}_h(\boldsymbol{\theta}_0)\gamma + \boldsymbol{b}_\psi(\boldsymbol{\theta}_0)\right)'\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\left(\boldsymbol{b}_h(\boldsymbol{\theta}_0)\gamma + \boldsymbol{b}_\psi(\boldsymbol{\theta}_0)\right).
\end{aligned}$$

A joint maximization of MSPE over both $\psi$ and $\gamma$ is possible but hopelessly complicated; we shall instead maximize over each individually, with the other set to zero, and then aim to minimize the sum of the first term above – arising solely from variation – and these two maxima, arising from model error and mislabelling respectively; viz. we minimize

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_0) =\ & tr\left\{\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_2(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\right\}\\
&+ \max_\psi n\boldsymbol{b}_\psi'(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{b}_\psi(\boldsymbol{\theta}_0)\\
&+ \max_\gamma n\boldsymbol{b}_h'(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{b}_h(\boldsymbol{\theta}_0)\gamma^2.
\end{aligned} \tag{16}$$

We carry this out under a side condition of *unbiasedness*, by which we mean the vanishing of the two maxima in (16). Our main theoretical results are Theorem 1 and its corollary.

**Theorem 1.** *(i) With $ch_{\max}(\cdot)$ denoting the maximum eigenvalue, the maximized loss (16) is*

$$\max_{\psi,\gamma}\mathcal{L}(\boldsymbol{\theta}_0) = \left(1+\tau_1^2+\tau_2^2\right)\mathcal{L}_{\nu_1,\nu_2}(w,\pi),$$

*where, with $\nu_1 \overset{def}{=} \tau_1^2/\left(1+\tau_1^2+\tau_2^2\right)$, and, $\nu_2 \overset{def}{=} \tau_2^2/\left(1+\tau_1^2+\tau_2^2\right)$,*

$$\begin{aligned}
&\mathcal{L}_{\nu_1,\nu_2}(w,\pi) =\\
&(1-\nu_1-\nu_2)\, tr\left\{\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_2(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\right\}\\
&+\nu_1\left[ch_{\max}\left\{\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)E_q\left[\boldsymbol{f}(\boldsymbol{x})\,\alpha^2(\boldsymbol{x};\boldsymbol{\theta}_0)\,w^2(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi^2(\boldsymbol{x};\boldsymbol{\theta}_0)\boldsymbol{f}'(\boldsymbol{x})\right]\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\right\} - 1\right]\\
&+\nu_2\boldsymbol{b}_h(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_q\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{b}_h(\boldsymbol{\theta}_0).
\end{aligned} \tag{17}$$

*(ii) For given sampling weights $\pi$, and with*

$$s(\boldsymbol{x}) \overset{def}{=} \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{M}_q^{-1}\boldsymbol{f}(\boldsymbol{x}),$$

*unbiased estimation weights are $w_*(\boldsymbol{x};\boldsymbol{\theta}_0) = 1/\left\{\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi(\boldsymbol{x};\boldsymbol{\theta}_0)\right\}$, for which*

$$\mathcal{L}_{\nu_1,\nu_2}(w_*,\pi) = (1-\nu_1-\nu_2)\,E_q\left[\frac{s(\boldsymbol{x})/\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)}{\pi(\boldsymbol{x};\boldsymbol{\theta}_0)}\right]. \tag{18}$$

*(iii) Sampling weights minimizing $\mathcal{L}_{\nu_1,\nu_2}(w_*,\pi)$ in (18), and corresponding estimation weights, are*

$$\pi_*(\boldsymbol{x};\boldsymbol{\theta}_0) = \frac{1}{c_*}\sqrt{\frac{s(\boldsymbol{x})}{\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)}}, \tag{19a}$$

$$w_*(\boldsymbol{x};\boldsymbol{\theta}_0) = \frac{c_*}{\sqrt{\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\,s(\boldsymbol{x})}}, \tag{19b}$$

*where*

$$c_* = E_q\left[\sqrt{\frac{s(\boldsymbol{x})}{\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)}}\right]. \tag{20}$$

*Minimum loss is*

$$\mathcal{L}_{\nu_1,\nu_2}(w_*,\pi_*) = (1-\nu_1-\nu_2)\,c_*^2. \tag{21}$$

*(iv) For passive learning ($\pi^{passive}\equiv 1$), unbiased estimation weights are $w_*^{passive}(\boldsymbol{x}) = 1/\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)$. Then*

$$\mathcal{L}_{\nu_1,\nu_2}\left(w_*^{passive},\pi^{passive}\right) = (1-\nu_1-\nu_2)\,E_q\left[\frac{s(\boldsymbol{x})}{\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)}\right] > \mathcal{L}_{\nu_1,\nu_2}(w_*,\pi_*). \tag{22}$$

*(v) For unweighted estimation ($w^{constant}\equiv 1$), unbiased sampling weights are $\pi^{unbiased}(\boldsymbol{x};\boldsymbol{\theta}_0) = 1/\{c_0\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\}$, where $c_0 = E_q[1/\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)]$. Loss is*

$$\mathcal{L}_{\nu_1,\nu_2}\left(w^{constant},\pi^{unbiased}\right) = (1-\nu_1-\nu_2)\,c_0 d > \mathcal{L}_{\nu_1,\nu_2}(w_*,\pi_*).$$

**Remarks:**

1. In the sequel we compare four sampling and estimation scenarios: Robust Weighted (ROBL/W), with estimation weights $w_*$ and sampling weights $\pi_*$, Robust Unweighted (ROBL/U) as in (v), using $\left(w^{constant},\pi^{unbiased}\right)$, Passive Weighted (PASSL/W) with constant sampling weights $\pi^{passive}$ and unbiased estimation weights $w_*^{passive}$, and Passive Unweighted (PASSL/U), with constant sampling $\pi^{passive}$ and constant estimation weights $w^{constant}$. For the first three of these, the product $\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)\,w(\boldsymbol{x};\boldsymbol{\theta}_0)\,\pi(\boldsymbol{x};\boldsymbol{\theta}_0)$ in (15) is constant and positive, so that $\boldsymbol{M}_1(\boldsymbol{\theta}_0)$ is proportional to $\boldsymbol{M}_q$ and the supposition of Lemma 1 is vacuous in light of (A1). For PASSL/U, this product reduces to $\alpha(\boldsymbol{x};\boldsymbol{\theta}_0)$, and it suffices that $g$ be such that this is bounded away from zero. For logistic and probit regression the verification of this is straightforward.
2. By (iv) and then (v) of Theorem 1, when evaluated at $\boldsymbol{\theta}_0$ ROBL/W is necessarily a strict improvement on PASSL/W and on ROBL/U when $\mathcal{L}_{\nu_1,\nu_2}$ – which in these three cases reduces to (a multiple of) the mean variance of the predictions – is used as the measure of performance. These orderings need not hold when the parameters are estimated.
3. The $\{s(\boldsymbol{x}_i)\}$ are analogous to the influence measures in regression diagnostics. In the special case $q(\boldsymbol{x}_i)\equiv 1/N$, we have that $s(\boldsymbol{x}_i) = Nh_{ii}$, where
$$\boldsymbol{F} \overset{def}{=} \left[\boldsymbol{f}(\boldsymbol{x}_1)\cdots\boldsymbol{f}(\boldsymbol{x}_N)\right],$$
and $\boldsymbol{H} = \boldsymbol{F}\left(\boldsymbol{F}'\boldsymbol{F}\right)^{-1}\boldsymbol{F}'$ is the 'hat' matrix formed from all of $\chi$, with diagonal elements $\{h_{ii}\}$. For this reason we will call $s(\boldsymbol{x}_i)$ an 'influence measure'. We note that Wang et al. (2018b) and Nie et al. (2018) also obtained sampling weights based on $\{h_{ii}\}$.
4. The (most powerful) Neyman–Pearson test, to choose between $p$ and $q$, decides in favour of $p$ if $p(\boldsymbol{x})/q(\boldsymbol{x}) = \pi(\boldsymbol{x};\boldsymbol{\theta}_0)$ is sufficiently large. Then from (19) we have the interpretation that the sampling favours points with large influence or at which the response variance is small, whereas the regression downweights influential points while favouring those with large sampling weights.
5. For logistic regression, $\alpha(\boldsymbol{x};\boldsymbol{\theta}_0) = v(\boldsymbol{x};\boldsymbol{\theta}_0)$ and the unbiased estimation weights for passive learning are the inverses of the response variances.

### *4.1. Mean squared classification error*

As an alternative to MSPE, one might instead seek to minimize the Mean Squared Classification Error (MSCE) of $\boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_n$ as a predictor of $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}_0$. This is $E_q\left[E\left[\left\{\sqrt{n}\left(G\left(\boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_n\right) - G\left(\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}_0\right)\right)\right\}^2\right]\right]$; apart from terms which are $o(1)$ it is

$$\text{MSCE} = E_q\left[g^2\left(\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}_0\right)E\left[\left\{\sqrt{n}\boldsymbol{f}'(\boldsymbol{x})\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right\}^2\right]\right]. \tag{23}$$

Then with

$$\boldsymbol{M}_g(\boldsymbol{\theta}_0) \overset{def}{=} E_q\left[\boldsymbol{f}(\boldsymbol{x})\,g^2\left(\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}_0\right)\boldsymbol{f}'(\boldsymbol{x})\right],$$

we have that $\textsc{msce} = E_q\left[g^2\left(\boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{\theta}_0\right)\left\{\boldsymbol{f}'\left(\boldsymbol{x}\right)\text{MSE}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)\boldsymbol{f}\left(\boldsymbol{x}\right)\right\}\right] = tr\left\{\boldsymbol{M}_g\ \text{MSE}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right)\right\}$ is, ignoring terms which are $O\left(n^{-1}\right)$, given by

$$\begin{aligned}\textsc{msce} = {} & tr\left\{\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\right\} \\ & + n\left(\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)\gamma + \boldsymbol{b}_\psi\left(\boldsymbol{\theta}_0\right)\right)'\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\left(\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)\gamma + \boldsymbol{b}_\psi\left(\boldsymbol{\theta}_0\right)\right).\end{aligned}$$

Note that this agrees with $\textsc{mspe}$, with $\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)$ replacing $\boldsymbol{M}_q$. The proof of the following corollary is then almost identical to that of Theorem 1; part (i) is proved in the Appendix.

**Corollary 1.** *(i) Replace* $\boldsymbol{M}_q$ *by* $\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)$ *in* (16) *and in* (17) *(but not in the definition of* $\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)$*). Then (i) of* Theorem 1 *holds, with*

$$\begin{aligned}&\mathcal{L}_{\nu_1,\nu_2}\left(w,\pi\right) = \\ &\left(1-\nu_1-\nu_2\right) tr\left[\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\right] \\ &+ \nu_1 ch_{\max}\left[\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\left\{\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)E_q\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\alpha^2\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)w^2\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\pi^2\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\right]\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right) - \boldsymbol{M}_q^{-1}\right\}\right] \\ &+ \nu_2\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right),\end{aligned} \tag{24}$$

*and gives the maximum of* $\textsc{msce}$ *over* $\psi$ *and* $\gamma$*. (ii) Statements (ii)–(v) of* Theorem 1 *hold, with* $s\left(\boldsymbol{x}\right) = \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{M}_q^{-1}\boldsymbol{f}\left(\boldsymbol{x}\right)$ *replaced by* $s\left(\boldsymbol{x};\boldsymbol{\theta}_0\right) \overset{def}{=} \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{M}_q^{-1}\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_q^{-1}\boldsymbol{f}\left(\boldsymbol{x}\right)$ *and* $d$ *in (v) replaced by* $E_q\left[s\right] = tr\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_q^{-1}$.

### *4.2. Sampling and estimation methods*

The following algorithm will be used in the simulation and examples which follow.

**Initialization** Our methods require an estimate of $\boldsymbol{M}_q$ and, if $\textsc{msce}$ is to be minimized, of $\boldsymbol{M}_g\left(\boldsymbol{\theta}_0\right)$. For this, take a $\textsc{srs}$ $\mathcal{S}_0 = \left\{\boldsymbol{x}_{(i)}\right\}_{i=1}^{n_0}$ from $\mathcal{Q}$. Learn the corresponding class memberships and compute the unweighted $\textsc{mle}$ $\hat{\boldsymbol{\theta}}$ by solving (12) with $w \equiv 1$. Set

$$\hat{\boldsymbol{M}}_q = \frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{f}\left(\boldsymbol{x}_{(i)}\right)\boldsymbol{f}'\left(\boldsymbol{x}_{(i)}\right),$$

$$\hat{\boldsymbol{M}}_g = \frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{f}\left(\boldsymbol{x}_{(i)}\right)g^2\left(\boldsymbol{f}'\left(\boldsymbol{x}_{(i)}\right)\hat{\boldsymbol{\theta}}\right)\boldsymbol{f}'\left(\boldsymbol{x}_{(i)}\right).$$

For $\boldsymbol{x} \in \chi$ estimate $s\left(\boldsymbol{x}\right)$ by $\hat{s}\left(\boldsymbol{x}\right) = \boldsymbol{f}'\left(\boldsymbol{x}\right)\hat{\boldsymbol{M}}_q^{-1}\boldsymbol{f}\left(\boldsymbol{x}\right)$, and $s\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)$ by $\boldsymbol{f}'\left(\boldsymbol{x}\right)\hat{\boldsymbol{M}}_q^{-1}\hat{\boldsymbol{M}}_g\hat{\boldsymbol{M}}_q^{-1}\boldsymbol{f}\left(\boldsymbol{x}\right)$.

Carry out one of the following two methods:

**(i) Group sampling** For a desired study size of $n$, carry out probability weighted sampling ($\textsc{pws}$) with weights $\pi_*\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}\right)$ as at (19a) to choose $k = n - n_0$ new sample members. To do $\textsc{pws}$, we first take a $\textsc{srs}$ of size $n_*$ (large; we use $n_* = 10\left(n-n_0\right)$) from the population, and let $\mathcal{S}_1 = \left\{\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(n_*)}\right\}$ be the corresponding, and not necessarily unique, members of $\chi$, which thus have pmf $q$. We then use the implementation, in $\textsc{matlab}$, of the method of Wang and Easton (1980) to draw a sample (without replacement) from $\mathcal{S}_1$, with pmf $\pi$. The unconditional pmf of this sample $\left\{\boldsymbol{x}_{i_*}\right\}_{i=1}^{k}$ is then $q\pi = p$.

Update $\hat{\boldsymbol{\theta}}$ by solving (12) with the desired weights.

**(ii) Sequential sampling** This is group sampling with $k = 1$ carried out $n-n_0$ times, and with $\hat{\boldsymbol{\theta}}$ and $\left\{\pi_*\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}\right)\right\}$ updated after each new point has been added to the sample.

## 5. A simulated example

We first generated a set $\chi \subset \mathbb{R}^2$, consisting of $N = 20000$ independent random vectors $\boldsymbol{x}_i$ with independent nonnegative elements, distributed as the absolute values of N(0,1) variables. Each $\boldsymbol{x}_i$ was then repeated an additional $T_i$ times, where the $T_i$ were independent Poisson random variables, each with unit mean. This yielded a bivariate population $\mathcal{Q}$ of expected size $2N$. See Fig. 1(a) - (c).

A parameter vector $\boldsymbol{\theta}$ was randomly generated. For each $\boldsymbol{x} \in \mathcal{Q}$, with value $\boldsymbol{x}_i \in \chi$ and $\boldsymbol{f}\left(\boldsymbol{x}_i\right) = \left(1, x_{i1}, x_{i2}, x_{i1}\cdot x_{i2}\right)'$, set $\beta_*\left(\boldsymbol{x};\boldsymbol{\theta}\right) = \boldsymbol{f}'\left(\boldsymbol{x}_i\right)\boldsymbol{\theta} + \boldsymbol{\psi}\left(\boldsymbol{x}\right)$, with $\boldsymbol{\psi}$ computed as described below. Then $\boldsymbol{\theta}_\gamma = \boldsymbol{\theta}$, as can be seen by noting that $\boldsymbol{\theta}$ satisfies

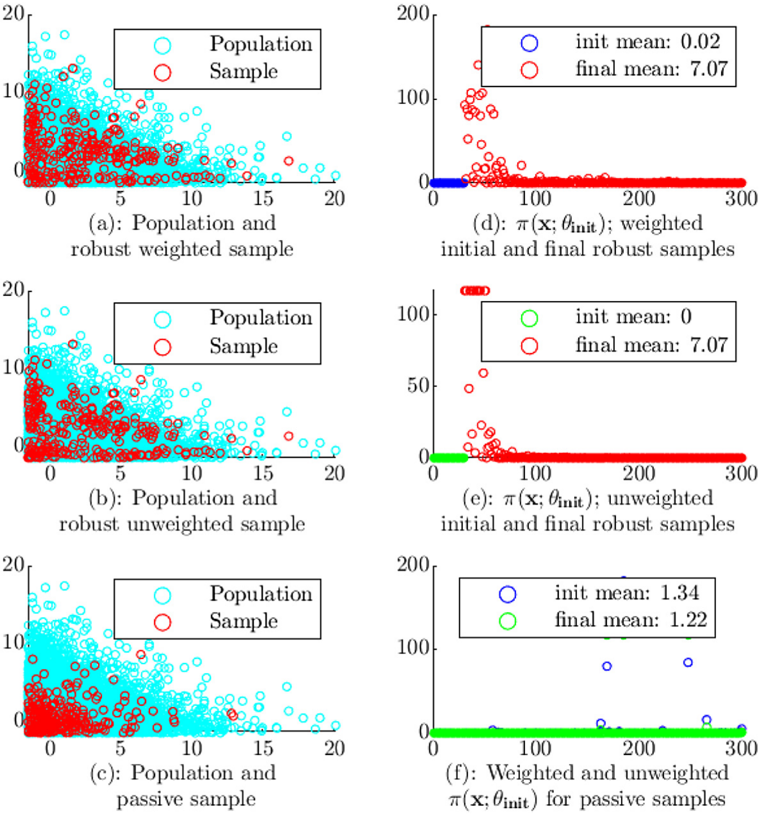

**Fig. 1.** Population values and training samples; $\tau_1 = 3$, $\tau_2 = 1$.

(6), whose rhs is

$$\begin{aligned}
&\arg\min_{\boldsymbol{t}} E_q\left[\left(H_\gamma^{-1}\left\{P\left(1|\boldsymbol{x}\right)\right\} - \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{t}\right)^2\right] \\
&= \arg\min_{\boldsymbol{t}} E_q\left[\left(\beta_*\left(\boldsymbol{x};\boldsymbol{\theta}\right) - \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{t}\right)^2\right] \\
&= \arg\min_{\boldsymbol{t}} \left(\boldsymbol{\theta}-\boldsymbol{t}\right)'\boldsymbol{M}_q\left(\boldsymbol{\theta}-\boldsymbol{t}\right) + E_q\left[\psi^2\left(\boldsymbol{x}\right)\right],
\end{aligned}$$

using (7). The minimizer is clearly $\boldsymbol{t} = \boldsymbol{\theta}$.

To compute a least favourable $\boldsymbol{\psi} = \boldsymbol{\psi}\left(\boldsymbol{\theta}_0\right)$, and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0\left(\boldsymbol{\psi}\right)$, we iterated between computing $\boldsymbol{\psi}$ from (A.1) and (A.2), with constant sampling and estimation weights and with $\boldsymbol{c}$ the maximizing eigenvector, and then inserting $\boldsymbol{\psi}$ into (5) and computing $\boldsymbol{\theta}_0$ from (8). This process generally converged very quickly. See Fig. 2 for typical examples.

Fig. 1 refers to the case $\tau_1 = 3$, $\tau_2 = 1$. We chose $\gamma = \min\left(\tau_2/\sqrt{n}, .5\right)$, where $n = 300$ is the final sample size. Finally, for each $\boldsymbol{x} \in \mathcal{Q}$, with value $\boldsymbol{x}_i \in \chi$, we generated $Y\left(\boldsymbol{x}\right) \sim bin\left(1, G\left(\beta_*\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right)\right)$, and then an 'observed' value $Y_{obs}$, with $Y_{obs}\left(\boldsymbol{x}\right)|Y\left(\boldsymbol{x}\right) \sim bin\left(1, \left(1-\gamma\right)Y\left(\boldsymbol{x}\right) + \gamma\left(1 - Y\left(\boldsymbol{x}\right)\right)\right)$.

Having simulated $Y$-values for the entire population we then compared the four sampling and estimation scenarios of Remark 1 above. The computations were carried out in MATLAB; Eq. (12) was solved using their nonlinear solver *fsolve*. For ROBL we carried out both group and sequential sampling as described above, with $n_0 = 30$ and $n = 300$. For the passive estimates the results were based on a SRS of the same size, hence the same number of queried responses, as were used by ROBL.

Recall Remark 3 and see Fig. 1(d)–(f) — our sampling scheme seems to be performing as intended, in selecting points likely to be from $p\left(\cdot\right)$.

After obtaining final estimates $\hat{\boldsymbol{\theta}}$ we classified all of the unqueried population values, as $\hat{Y} = 1$ if $G\left(\beta\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}\right)\right)$ exceeded a cutoff, equal to that value minimizing the misclassification rate (the proportion of times that $\hat{Y}\left(\boldsymbol{x}\right) \neq Y\left(\boldsymbol{x}\right)$) in the training sample. As a check on the effectiveness of this method, we also computed the 'optimal' – but not available in practice – misclassification rates, using a cutoff equal to that based on $\hat{\boldsymbol{\theta}}$ but minimizing the misclassification rate in the entire population. In all cases the two error rates were very close.

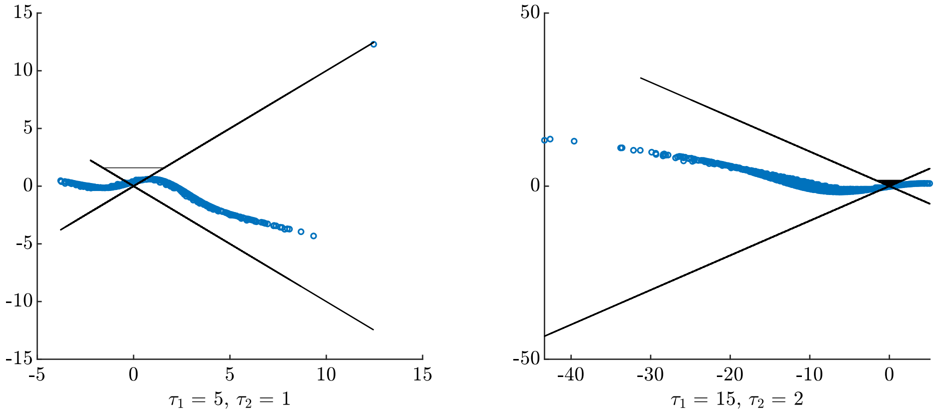

**Fig. 2.** $\psi(x)$ vs. true means $\boldsymbol{F}\boldsymbol{\theta}_\gamma$, lines at $y = \pm x$.

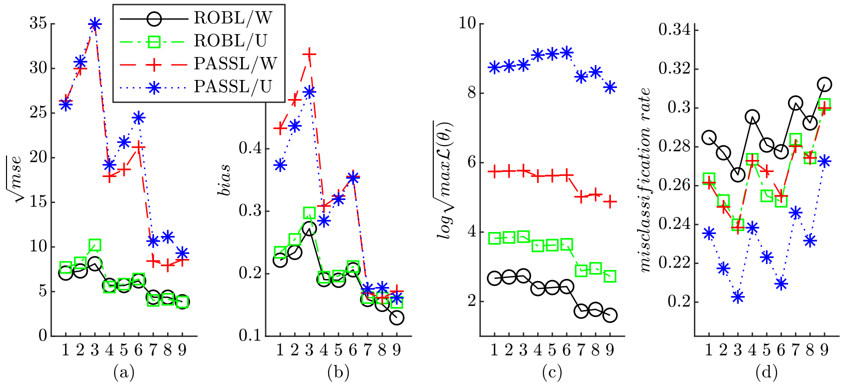

**Fig. 3.** Simulated data. Loss measures from a logistic regression (group sampling, MSPE loss, $n = 300$) for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $(\tau_1, \tau_2) = (0, .5), (3, .5), (6, .5), (0, 1), (3, 1), (6, 1), (0, 2), (3, 2), (6, 2)$ (1–9, respectively). Means over 100 simulations of (a) $\sqrt{\text{MSPE}}$ as at (14), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) $\log\sqrt{\max_{\psi,\gamma}\mathcal{L}(\boldsymbol{\theta}_0)}$, (d) misclassification rates.

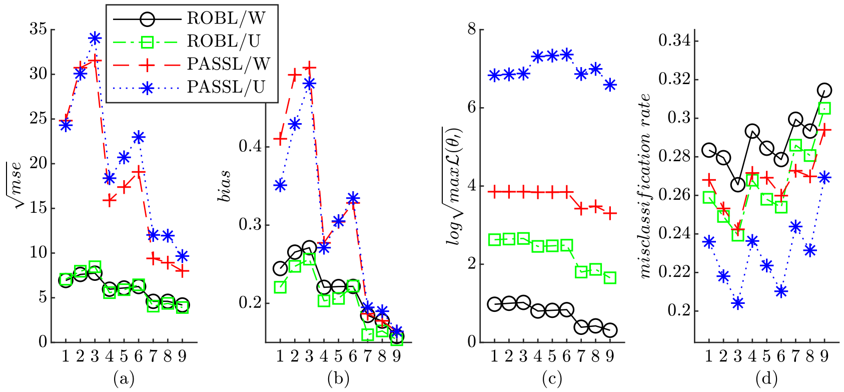

**Fig. 4.** Simulated data. Loss measures from a logistic regression (group sampling, MSCE loss, $n = 300$) for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $(\tau_1, \tau_2) = (0, .5), (3, .5), (6, .5), (0, 1), (3, 1), (6, 1), (0, 2), (3, 2), (6, 2)$ (1–9, respectively). Means over 100 simulations of (a) $\sqrt{\text{MSCE}}$ as at (23), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) $\log\sqrt{\max_{\psi,\gamma}\mathcal{L}(\boldsymbol{\theta}_0)}$, (d) misclassification rates.

We ran this and similar examples 100 times, with new samples drawn in each run, and a new population simulated for each choice of $(\tau_1, \tau_2)$. Figs. 3 and 4 are typical — logistic regression and group sampling generally resulted in much superior parameter estimates and predictions, but systematically somewhat poorer classifications overall, when compared

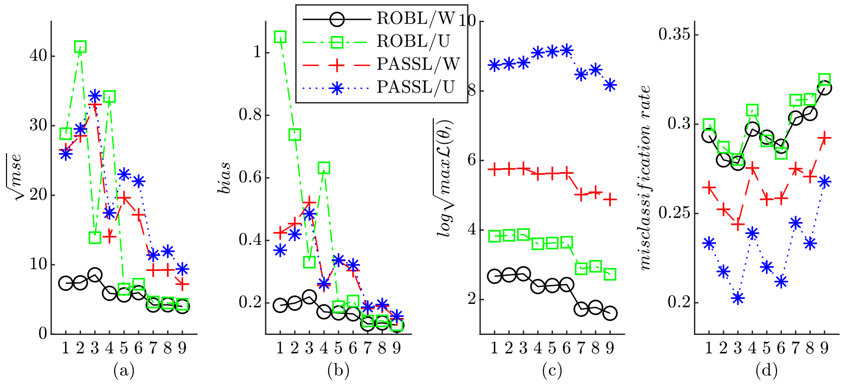

**Fig. 5.** Simulated data. Loss measures from a logistic regression (sequential sampling, MSPE loss, $n = 300$) for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $(\tau_1, \tau_2) = (0, .5), (3, .5), (6, .5), (0, 1), (3, 1), (6, 1), (0, 2), (3, 2), (6, 2)$ (1–9, respectively). Means over 100 simulations of (a) $\sqrt{\text{MSPE}}$ as at (14), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) $\log\sqrt{\max_{\psi,\gamma}\mathcal{L}(\boldsymbol{\theta}_0)}$, (d) misclassification rates.

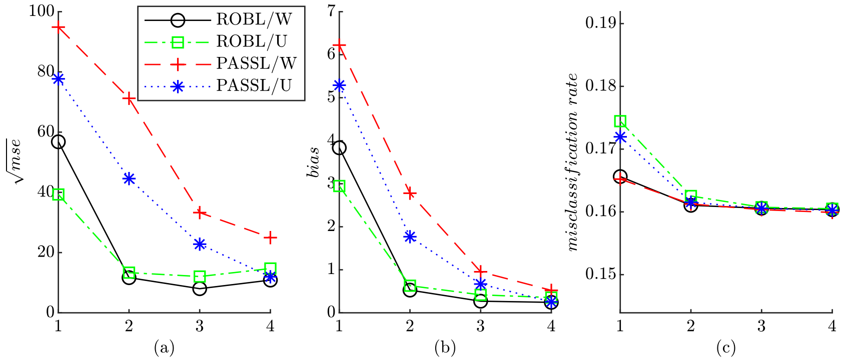

**Fig. 6.** Tweet data. Loss measures from a probit regression for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $n = 200, 500, 1000, 2000$ (1–4, respectively). Means over 100 simulations of (a) $\sqrt{\text{MSPE}}$ as at (14), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) misclassification rates.

to the passive methods. The latter is perhaps to be expected, since the ROBL estimates are tailored to a more influential, but not necessarily representative, subpopulation.

In these simulations the weighted ROBL was uniformly more efficient than the unweighted version, with respect to the maximized loss (see panels (c) in Figs. 3–5), and almost uniformly so with respect to BIAS and MSE. This is in contrast to the results of Wang (2019), where subsampling schemes in correctly specified binary models are derived. These schemes depend on the responses as well as the predictors, and the weighted versions are seen to be typically less efficient than their unweighted counterparts.

Sequential sampling is much slower, and – see Fig. 5 – yields very similar results.

In the 'tweet' example which follows, robust sampling is more efficient than passive sampling, with neither being uniformly superior with respect to classification.

## 6. Example: Tweet sources

We have a set, which for this example is viewed as the population, of 48, 748 tweets that we want to classify as 'coming from a human' or 'coming from a bot' according to the explanatory variables 'friends', 'listed', 'favourites', 'status' and 'profile'. The final variable is binary while the rest are counts. These were the training data in "NYU Tandon Spring 2017 Machine Learning Competition: Twitter Bot classification" at *www.kaggle.com/c/twitter-bot-classification/data*.

We carried out an analysis as in Section 4.2. The transformation $x \to \log(1 + x)$ was first applied to the count variables. Since $\boldsymbol{\theta}_0$ cannot be computed as it was in Section 5, we approximated it by the result of an unweighted regression using all values in the population. Using the probit link for a variety of values of $n$ with $n_0 = \max(40, n/10)$ gave the means, over 100 runs, of the performance measures as illustrated in Fig. 6 for MSPE loss and Fig. 7 for MSCE loss. For these figures we used group sampling. In both cases ROBL led to significantly smaller losses than, and misclassification rates that were nearly identical to, those from PASSL.

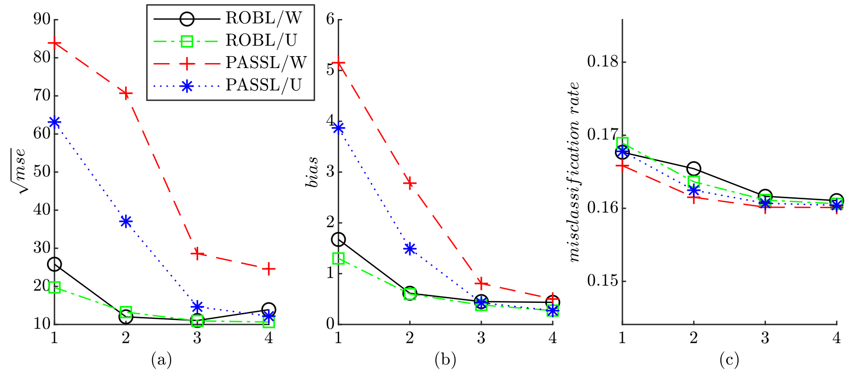

**Fig. 7.** Tweet data. Loss measures from a probit regression for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $n = 200, 500, 1000, 2000$ (1–4, respectively). Means over 100 simulations of (a) $\sqrt{\text{MSCE}}$ as at (14), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) misclassification rates.

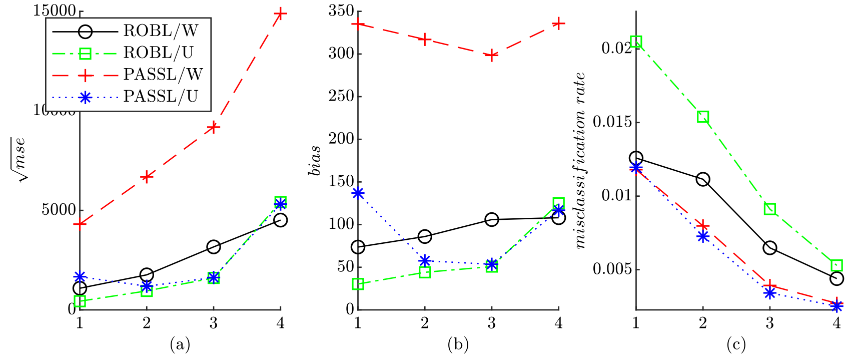

**Fig. 8.** Skin data. Loss measures from a logistic regression for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $n = 200, 500, 1000, 2000$ (1–4, respectively). Means over 20 simulations of (a) $\sqrt{\text{MSPE}}$ as at (14), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) misclassification rates.

## 7. Example: Skin segmentation

The Skin Segmentation dataset (Bhatt et al., 2009) is collected by randomly sampling blue, green and red values from face images of various age groups (young, middle, and old), race groups (White, Black, and Asian), and genders. The ‘population’ size is 245057, out of which 50859 are the skin samples and 194198 are non-skin samples. The variable to be classified is ‘skin’ or ‘non-skin’; this is meant to be used as a prelude to face and body detection and tracking.

See Figs. 8 and 9 for the output from 20 runs, using logistic regression and group sampling. In both cases the unweighted methods ROBL/U and PASSL/U led to estimates with smaller losses; all misclassification rates were very small, with those from PASSL being slightly smaller.

## 8. Discussion

We have proposed estimation and classification methods for binary data, robust against model misspecification and response mislabelling. The simulation studies indicate that when these are of significant concern, the proposed method ROBL allows for substantial reductions, relative to Passive Learning, in the prediction errors.

A defining feature of ROBL is that the responses need be observed only as they correspond to the sampled predictors — there is no need to have available the entire population of responses. This can of course be a huge advantage when learning the responses is expensive.

However, as a reviewer has pointed out, our methods might perform poorly in ‘rare events’ situations, where response imbalance is common (Wang, 2020). We obtain the responses only for the ‘influential’ subsample which is gathered; if this subsample is not representative of the general population – as one might expect in the case of rare events – then the classifier built from the subsample might be poor. An example is the ‘2009 Mortgage Default’ data, available at *packages.revolutioinanalytics.com/datasets/* and with data on one million mortgages, about 2.5% of which resulted in defaults. Drovandi et al. (2017) studied designs for these data, and took subsamples of size 5000. The main goal was parameter estimation; classification was not considered. For brevity we have not presented our analysis here, but found that even with subsamples of size 10,000 the resulting classifiers performed poorly.

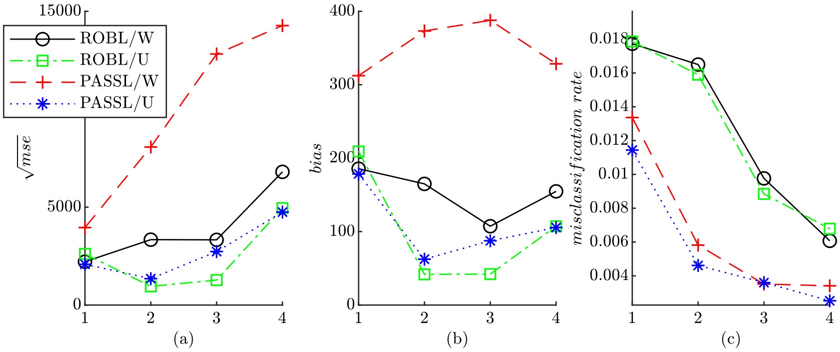

**Fig. 9.** Skin data. Loss measures from a logistic regression for ROBL/W, ROBL/U, PASSL/W and PASSL/U vs. $n = 200, 500, 1000, 2000$ (1–4, respectively). Means over 20 simulations of (a) $\sqrt{\text{MSCE}}$ as at (14), (b) BIAS $\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\|$, (c) misclassification rates.

## Acknowledgements

Jesús López–Fidalgo is sponsored by Ministerio de Economía y Competitividad, Spain MTM2016-80539-C2-1-R and by Program Salvador de Madariaga PRX18/00339. The research of Douglas P. Wiens is supported by the Natural Sciences and Engineering Council of Canada. The authors thank Linglong Kong, José Moler, Haiying Wang and Zhichun Zhai for helpful comments. In particular, we are grateful for the insightful comments of several anonymous reviewers.

## Appendix. Derivations

**Proof of Lemma 1.** We first condition on $\boldsymbol{S}$. For this, define

$$\boldsymbol{M}_{n,0}(\boldsymbol{\theta}_0) = \sum_{i=1}^{N} \frac{n_i}{n} \alpha(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, w(\boldsymbol{x}_i;\boldsymbol{\theta}_0) \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}'(\boldsymbol{x}_i),$$

$$\boldsymbol{V}_{n,0}(\boldsymbol{\theta}_0) = \sum_{i=1}^{N} \frac{n_i}{n} \alpha(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, w^2(\boldsymbol{x}_i;\boldsymbol{\theta}_0) \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}'(\boldsymbol{x}_i),$$

$$\begin{aligned}\boldsymbol{c}_{n,0}(\boldsymbol{\theta}_0) = {} & \sum_{i=1}^{N} \frac{n_i}{n} \alpha(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, w(\boldsymbol{x}_i;\boldsymbol{\theta}_0) \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}'(\boldsymbol{x}_i) \left(E_q\left[\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})\right]\right)^{-1} E_q\left[\boldsymbol{f}(\boldsymbol{x})\, h(\boldsymbol{x};\boldsymbol{\theta}_0)\right]\gamma \\ & - \sum_{i=1}^{N} \frac{n_i}{n} \boldsymbol{f}(\boldsymbol{x}_i)\, \alpha(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, w(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, h(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, \gamma + \sum_{i=1}^{N} \frac{n_i}{n} \boldsymbol{f}(\boldsymbol{x}_i)\, \alpha(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, w(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, \psi(\boldsymbol{x}_i).\end{aligned}$$

Then in this notation, and under mild conditions as given along with asymptotic theory for misspecified models in Fahrmeir (1990), the MLE $\hat{\boldsymbol{\theta}}_n$ is asymptotically normally distributed:

$$\boldsymbol{M}_{n,0}(\boldsymbol{\theta}_0) \cdot \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \sim AN\left(\sqrt{n}\boldsymbol{c}_{n,0}(\boldsymbol{\theta}_0), \boldsymbol{V}_{n,0}(\boldsymbol{\theta}_0)\right).$$

See Wiens (2021) for additional details. From (11), $\boldsymbol{M}_{n,0}(\boldsymbol{\theta}_0) = \boldsymbol{M}_1(\boldsymbol{\theta}_0) + o_P(1)$ and $\boldsymbol{V}_{n,0}(\boldsymbol{\theta}_0) = \boldsymbol{M}_2(\boldsymbol{\theta}_0) + o_P(1)$. It follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \sim AN\left(\sqrt{n}\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{c}_{n,0}(\boldsymbol{\theta}_0),\ \boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{M}_2(\boldsymbol{\theta}_0)\, \boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\right).$$

Ignoring terms which are $O\left(n^{-1}\right)$, the unconditional MSE matrix is

$$\text{MSE}_{\psi,\gamma}(\boldsymbol{\theta}_0) = \boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\left\{nE_{\boldsymbol{S}}\left[\boldsymbol{c}_{n,0}(\boldsymbol{\theta}_0)\, \boldsymbol{c}'_{n,0}(\boldsymbol{\theta}_0)\right] + M_2(\boldsymbol{\theta}_0)\right\} \boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0).$$

With

$$\boldsymbol{t}(\boldsymbol{x}_i) = \alpha(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, w(\boldsymbol{x}_i;\boldsymbol{\theta}_0) \cdot \left\{\begin{array}{c} \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}'(\boldsymbol{x}_i)\left(E_q\left[\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})\right]\right)^{-1} E_q\left[\boldsymbol{f}(\boldsymbol{x})\, h(\boldsymbol{x};\boldsymbol{\theta}_0)\right]\gamma \\ -\boldsymbol{f}(\boldsymbol{x}_i)\, h(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\, \gamma + \boldsymbol{f}(\boldsymbol{x}_i)\, \psi(\boldsymbol{x}_i) \end{array}\right\},$$

and $\boldsymbol{T}_{d\times N} = (\boldsymbol{t}(\boldsymbol{x}_1), \ldots, \boldsymbol{t}(\boldsymbol{x}_N))$ we see that $\boldsymbol{c}_{n,0}(\boldsymbol{\theta}_0)$ is the realized value of $n^{-1}\boldsymbol{TS}$, so that

$$E_{\boldsymbol{S}}\left[\boldsymbol{c}_{n,0}(\boldsymbol{\theta}_0)\right] = E_{\boldsymbol{S}}\left[\frac{1}{n}\boldsymbol{TS}\right] = \boldsymbol{Tp},$$

and

$$\begin{aligned} E_{\boldsymbol{S}}\left[\boldsymbol{c}_{n,0}\left(\boldsymbol{\theta}_0\right)\boldsymbol{c}'_{n,0}\left(\boldsymbol{\theta}_0\right)\right] &= \mathrm{cov}_{\boldsymbol{S}}\left[\boldsymbol{c}_{n,0}\left(\boldsymbol{\theta}_0\right)\right] + E_{\boldsymbol{S}}\left[\boldsymbol{c}_{n,0}\left(\boldsymbol{\theta}_0\right)\right]E_{\boldsymbol{S}}\left[\boldsymbol{c}'_{n,0}\left(\boldsymbol{\theta}_0\right)\right] \\ &= \mathrm{cov}_{\boldsymbol{S}}\left[\frac{1}{n}\boldsymbol{TS}\right] + \left\{E_{\boldsymbol{S}}\left[\frac{1}{n}\boldsymbol{TS}\right]\right\}\left\{E_{\boldsymbol{S}}\left[\frac{1}{n}\boldsymbol{TS}\right]\right\}' \\ &= \frac{1}{n}\boldsymbol{T}\left[\boldsymbol{D}_p - \boldsymbol{pp}'\right]\boldsymbol{T}' + \{\boldsymbol{Tp}\}\{\boldsymbol{Tp}\}'. \end{aligned}$$

Since $\boldsymbol{T}$ is $O\left(n^{-1/2}\right)$, we obtain

$$E_{\boldsymbol{S}}\left[\boldsymbol{c}_{n,0}\left(\boldsymbol{\theta}_0\right)\boldsymbol{c}'_{n,0}\left(\boldsymbol{\theta}_0\right)\right] = \{\boldsymbol{Tp}\}\{\boldsymbol{Tp}\}' + O\left(n^{-2}\right) = E_p\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right]E_p\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right]' + O\left(n^{-2}\right).$$

Now (13) follows, with $\boldsymbol{b}_{\psi,\gamma}\left(\boldsymbol{\theta}_0\right) \overset{def}{=} E_p\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right]$. □

**Proof of Theorem 1.** (i) That $\max_\gamma n\boldsymbol{b}'_h\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_q\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right)\gamma^2$ is as stated is obvious. For the maximization over $\psi$ it is convenient to adopt a canonical form. Let $\boldsymbol{F}_q = \boldsymbol{D}_q^{1/2}\boldsymbol{F}$, where $\boldsymbol{D}_q = diag(\ldots, q\left(\boldsymbol{x}_i\right), \ldots)$, and let

$$\boldsymbol{F}_q = \left(\boldsymbol{Q}_1 \vdots \boldsymbol{Q}_2\right)\left(\begin{array}{c}\boldsymbol{R}\\ \boldsymbol{0}\end{array}\right)$$

be the qr-decomposition, with $\boldsymbol{Q}_1 : N \times d$. With $\boldsymbol{\psi} = (\ldots, \psi\left(\boldsymbol{x}_i\right), \ldots)'$, conditions (3) and (7) force

$$\boldsymbol{\psi} = \frac{\tau_1}{\sqrt{n}}\boldsymbol{D}_q^{-1/2}\boldsymbol{Q}_2\boldsymbol{c}, \tag{A.1}$$

for $\boldsymbol{c} \in \mathbb{R}^{N-d}$ and $\|\boldsymbol{c}\| \leq 1$. In these terms, and with $\boldsymbol{D}_\alpha = diag(\ldots, \alpha\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right), \ldots)$, $\boldsymbol{D}_w = diag(\ldots, w\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right), \ldots)$, $\boldsymbol{D}_\pi = diag(\ldots, \pi\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right), \ldots)$ we have

$$\begin{aligned} \boldsymbol{M}_q &= \boldsymbol{F}'_q\boldsymbol{F}_q = \boldsymbol{R}'\boldsymbol{R}, \\ \boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right) &= \boldsymbol{R}'\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\boldsymbol{R}, \\ \boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right) &= \boldsymbol{R}'\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w^2\boldsymbol{D}_\pi\boldsymbol{Q}_1\boldsymbol{R}, \\ \boldsymbol{b}_\psi\left(\boldsymbol{\theta}_0\right) &= \frac{\tau_1}{\sqrt{n}}\boldsymbol{R}'\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2\boldsymbol{c}, \end{aligned}$$

(in the proof of Corollary 1, $\boldsymbol{M}_q$ is replaced by $\boldsymbol{M}_g = \boldsymbol{R}'\boldsymbol{Q}'_1\boldsymbol{D}_g^2\boldsymbol{Q}_1\boldsymbol{R}$) and

$$\begin{aligned} &\max_\psi n\boldsymbol{b}'_\psi\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_q\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{b}_\psi\left(\boldsymbol{\theta}_0\right) \\ &= \tau_1^2\max_{\|\boldsymbol{c}\|\leq 1}\boldsymbol{c}'\boldsymbol{Q}'_2\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-2}\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2\boldsymbol{c} \qquad \text{(A.2)} \\ &= \tau_1^2 ch_{\max}\boldsymbol{Q}'_2\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-2}\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2, \end{aligned}$$

attained at the eigenvector of unit norm. By virtue of the fact that matrices $\boldsymbol{PQ}$ and $\boldsymbol{QP}$ have the same nonzero eigenvalues, the maximum eigenvalue is

$$\begin{aligned} &ch_{\max}\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2\boldsymbol{Q}'_2\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \\ &= ch_{\max}\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\left(\boldsymbol{I}_N - \boldsymbol{Q}_1\boldsymbol{Q}'_1\right)\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \\ &= ch_{\max}\left\{\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}'_1\boldsymbol{D}_\alpha^2\boldsymbol{D}_w^2\boldsymbol{D}_\pi^2\boldsymbol{Q}_1\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} - \boldsymbol{I}_d\right\} \\ &= ch_{\max}\left\{\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}'_1\boldsymbol{D}_\alpha^2\boldsymbol{D}_w^2\boldsymbol{D}_\pi^2\boldsymbol{Q}_1\left(\boldsymbol{Q}'_1\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\right\} - 1. \end{aligned}$$

Writing this in the original terms gives (17).

(ii) With weights $w_*$,

$$\boldsymbol{M}_1\left(\boldsymbol{\theta}_0\right) = E_q\left[\boldsymbol{f}\left(\boldsymbol{x}\right)\alpha^2\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)w^2\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\pi^2\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\right] = \boldsymbol{M}_q,$$

and so the maximum eigenvalue is that of the identity matrix, and $\boldsymbol{b}_h\left(\boldsymbol{\theta}_0\right) = \boldsymbol{0}$; thus these terms both vanish.

Now $\mathcal{L}_{\nu_1,\nu_2}\left(w_*, \pi\right) = \left(1 - \nu_1 - \nu_2\right)tr\left\{\boldsymbol{M}_q\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\right\}$, and we calculate that

$$tr\left\{\boldsymbol{M}_q\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_2\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\right\} = tr\left\{E\left[\boldsymbol{f}\left(\boldsymbol{x}\right)w_*\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\right]\right\},$$

which reduces to $E_q\left[s\left(\boldsymbol{x}\right)w_*\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\right]$.

(iii) We minimize (18), subject to $E_q\left[\pi\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\right] = 1$, by introducing a Lagrange multiplier and then minimizing pointwise. This gives (19a) , where $c_*$ is the normalizing constant. Then (21) and (19b) are immediate.

(iv) The form of $w_*^{passive}$, and the equality in (22) are (ii) of this theorem with $\pi \equiv 1$. The inequality is the Cauchy–Schwarz inequality applied to (20): $\left(E\left[\sqrt{s/\alpha}\right]\right)^2 < E\left[s/\alpha\right]$.

(v) The unbiasedness is a simple calculation, and then, using $E_q\left[s\right] = d$, the inequality is again the Cauchy–Schwarz Inequality: $\left(E\left[\sqrt{s}\sqrt{1/\alpha}\right]\right)^2 < E\left[s\right]E\left[1/\alpha\right].$ □

**Proof of (i) of Corollary 1.** With notation as in the proof of the theorem we have

$$\boldsymbol{M}_g = \boldsymbol{R}'\boldsymbol{Q}_1'\boldsymbol{D}_g^2\boldsymbol{Q}_1\boldsymbol{R},$$

for $\boldsymbol{D}_g = diag\left(\cdots, g\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right), \cdots\right)$, and

$$\begin{aligned}
&\max_{\psi} n\boldsymbol{b}_\psi'\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{M}_g\boldsymbol{M}_1^{-1}\left(\boldsymbol{\theta}_0\right)\boldsymbol{b}_\psi\left(\boldsymbol{\theta}_0\right)\\
&= \tau_1^2 \max_{\|\boldsymbol{c}\|\le 1} \boldsymbol{c}'\boldsymbol{Q}_2'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \cdot \boldsymbol{Q}_1'\boldsymbol{D}_g^2\boldsymbol{Q}_1\\
&\quad \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \cdot \boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2\boldsymbol{c}\\
&= \tau_1^2 ch_{\max}\boldsymbol{Q}_2'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\\
&\quad \cdot\boldsymbol{Q}_1'\boldsymbol{D}_g^2\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \cdot \boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2,
\end{aligned}$$

attained at the eigenvector of unit norm. By virtue of the fact that matrices $\boldsymbol{PQ}$ and $\boldsymbol{QP}$ have the same nonzero eigenvalues, the maximum eigenvalue is

$$\begin{aligned}
&ch_{\max}\boldsymbol{D}_g\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \cdot \boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_2\boldsymbol{Q}_2'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \cdot \boldsymbol{Q}_1'\boldsymbol{D}_g\\
&= ch_{\max}\boldsymbol{D}_g\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\\
&\quad \cdot\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\left(\boldsymbol{I}_N - \boldsymbol{Q}_1\boldsymbol{Q}_1'\right)\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} \cdot \boldsymbol{Q}_1'\boldsymbol{D}_g\\
&= ch_{\max}\left\{\boldsymbol{D}_g\boldsymbol{Q}_1 \cdot \left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}_1'\boldsymbol{D}_\alpha^2\boldsymbol{D}_w^2\boldsymbol{D}_\pi^2\boldsymbol{Q}_1\left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}_1'\boldsymbol{D}_g - \boldsymbol{D}_g\boldsymbol{Q}_1\boldsymbol{Q}_1'\boldsymbol{D}_g\right\}\\
&= ch_{\max}\left\{\boldsymbol{D}_g\boldsymbol{Q}_1\left[\left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1}\boldsymbol{Q}_1'\boldsymbol{D}_\alpha^2\boldsymbol{D}_w^2\boldsymbol{D}_\pi^2\boldsymbol{Q}_1\left(\boldsymbol{Q}_1'\boldsymbol{D}_\alpha\boldsymbol{D}_w\boldsymbol{D}_\pi\boldsymbol{Q}_1\right)^{-1} - \boldsymbol{I}_d\right]\boldsymbol{Q}_1'\boldsymbol{D}_g\right\}.
\end{aligned}$$

Writing this in the original terms gives (24). □

## References

Bhatt, R.B., Dhall, A., Sharma, G., Chaudhury, S., 2009. Efficient skin region segmentation using low complexity fuzzy decision tree model. IEEE-INDICON 2009 Conf., Ahmedabad, India 1–4.

Carroll, R., Pederson, S., 1993. On robustness in the logistic regression model. J. R. Stat. Soc. Ser. B 55, 693–706.

Copas, J.P., 1988. Binary regression models for contaminated data. J. R. Stat. Soc. Ser. B 50, 225–265.

Drovandi, C.C., Holmes, C.C., McGree, J.M., Mengersen, K., Richardson, S., Ryan, E.G., 2017. Principles of experimental design for big data analysis. Statist. Sci. 32, 385–404.

Fahrmeir, L., 1990. Maximum likelihood estimation in misspecified generalized linear models. Statistics 21, 487–502.

Huber, P.J., 1981. Robust Statistics. Wiley.

Le Cam, L., 1960. Locally asymptotically normal families of distributions. Univ. Calif. Publ. Stat. 3, 37–98.

Meyer, B.D., Mittag, N., 2017. Misclassification in binary choice models. J. Econometrics 200, 295–311.

Nachtsheim, A.C., Stufken, J., 2019. Comments on: Data science, big data and statistics. TEST 28, 345–348.

Nie, R., Wiens, D.P., Zhai, Z., 2018. Minimax robust active learning for approximately specified regression models. Canad. J. Statist. 46, 104–122.

Wang, H.-Y., 2019. More efficient estimation for logistic regression with optimal subsamples. J. Mach. Learn. Res. 20, 1–59.

Wang, H.-Y., 2020. Logistic regression for massive data with rare events. In: Proceedings Of The 37th International Conference On Machine Learning, Vol. 119. pp. 9829–9836.

Wang, C., Chen, M.-H., Wu, J., Yan, J., Zhang, Y., Schifano, E., 2018a. Online updating method with new variables for big data streams. Canad. J. Statist. 46, 123–146.

Wang, C.K., Easton, M.C., 1980. An efficient method for weighted sampling without replacement. SIAM J. Comput. 9, 111–113.

Wang, H.-Y., Yang, M., Stufken, J., 2019. Information-based optimal subdata selection for big data linear regression. J. Am. Stat. Assoc. 114, 393–405.

Wang, H.-Y., Zhu, R., Ma, P., 2018b. Optimal subsampling for large sample logistic regression. J. Am. Stat. Assoc. 113, 829–844.

Wiens, D.P., 2015. Robustness Of Design, In ‘Handbook Of Design And Analysis Of Experiments’. Chapman & Hall/CRC.

Wiens, D.P., 2021. Robust designs for dose-response studies: model and misclassification robustness. Comput. Statist. Data Anal. 158, 107189.