Logistic Regression Models for Aggregated Data

T. Whitaker, B. Beranger & S. A. Sisson

Taylor & Francis
Taylor & Francis Group

Check for updates

Logistic Regression Models for Aggregated Data

T. Whitaker, B. Beranger ⓘ, and S. A. Sisson ⓘ

UNSW Data Science Hub & School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia

ABSTRACT

Logistic regression models are a popular and effective method to predict the probability of categorical response data. However, inference for these models can become computationally prohibitive for large datasets. Here we adapt ideas from symbolic data analysis to summarize the collection of predictor variables into histogram form, and perform inference on this summary dataset. We develop ideas based on composite likelihoods to derive an efficient one-versus-rest approximate composite likelihood model for histogram-based random variables, constructed from low-dimensional marginal histograms obtained from the full histogram. We demonstrate that this procedure can achieve comparable classification rates to the standard full data multinomial analysis and against state-of-the-art subsampling algorithms for logistic regression, but at a substantially lower computational cost. Performance is explored through simulated examples, and analyses of large supersymmetry and satellite crop classification datasets. Supplementary materials for this article are available online.

1. Introduction

There are many well developed statistical methods for classification, such as logistic regression, discriminant analysis, and clustering, which predict a categorical variable that can take one of K distinct values given an input vector of predictor variables (e.g., Pampel 2000; Hastie, Tibshirani, and Friedman 2008). While these methods are effective for the analysis of standard data, when the data take nonstandard forms, such as random interval- or random histogram-based predictors, existing methods are either underdeveloped or do not exist. Interval, histogram and other-distribution based data summaries can arise through measurement error, data quantization, expert elicitation and, motivating this work, the desire to summarize large and complex datasets in an appropriate way so that they can be analyzed more efficiently than the full dataset (e.g., Rahman et al. 2020; Zhang, Beranger, and Sisson 2020). The field of symbolic data analysis (SDA) was developed to analyze such distributional data (Diday 1989; Billard and Diday 2003; Billard 2011; Beranger, Lin, and Sisson 2018). However, with a few exceptions (discussed below), the parameters of existing SDA-based methods undesirably lose their interpretation as parameters of models of the underlying (standard) micro-data, which means that the resulting inferences are not directly comparable to the equivalent standard full-data analysis (Beranger, Lin, and Sisson 2018; Zhang, Beranger, and Sisson 2020).

Logistic regression (e.g., Cox 1958; Hosmer, Lemeshow, and Sturdivant 2013) is one method of performing regression for categorical response data, and has been used extensively in many fields, including finance (Kim and Gu 2010; Hauser and Booth 2011), epidemiology (Merlo et al. 2006), medicine (Min 2013; Hosmer, Lemeshow, and Sturdivant 2013), diagnostics

(Knottnerus 1992), and modeling income (Pavlopoulos, Muffels, and Vermunt 2010). Such models are typically fitted by numerical maximum likelihood estimation as there is no closed form maximum likelihood estimator (MLE), conditional on certain conditions on the response and predictors such that the MLE exists (Albert and Anderson 1984). As a result, the computational overheads for determining the MLE of logistic regression models can be high for very large datasets.

With this motivation, Wang, Zhu, and Ma (2018) developed a data-subsampling scheme for binary logistic regression (for $K = 2$ classes) whereby each observation is assigned a weight according to a function that minimizes the asymptotic mean squared error (MSE) of the MLE. A subsample is then taken from the full dataset using these weights, with parameter estimates then obtained by maximizing the likelihood for the subsample. To calculate the observation weights, an initial parameter estimate is obtained using r_0 uniformly sampled observations. Wang, Zhu, and Ma (2018) demonstrated good estimator performance from the "optimal" subsample using $r_0 = 1000$.

In contrast, de Souza, Queiroz, and Cysneiros (2011) presented SDA-motivated versions of logistic regression for interval-valued data (where the intervals are constructed from the minimum and maximum observed values of each univariate predictor in the micro-data), whereby the regressions are constructed using the interval centers or endpoints as covariates. de Souza et al. (2008) developed a similar approach for multi-class classification with logistic regression using interval endpoints. While these SDA methods are computationally simple, because they are only based on random intervals (that is, two quantiles of the data), much of the distributional information in the

predictor values is lost. In addition the implementations treat each interval equally, which can cause problems if there are unbalanced numbers of micro-data in each category. Finally, as discussed above, the fitted models cannot be compared to the equivalent models fitted to the full (nonsummarized) dataset.

In other work, Tranmer and Steel (1997) investigated the effect of data aggregation on logistic regression when both the predictor and response variables are observed as the total sum of each variable for each group. Bhowmik, Ghosh, and Koyejo (2016) proposed an iterative algorithm for estimating generalized linear models (GLMs) when the response variables are aggregated into histogram-valued random variables. The resulting training error was numerically shown to approach that of the full (nonaggregated) analysis as the number of histogram bins became large. Armstrong (1985) considered the case where a single covariate is measured with error, and the distribution of the coarsened (binned) value given the true latent observation is known. Estimates for GLM coefficients are then obtained through the utilization of this known density. Johnson (2006) assumed Gaussian distributions for coarsened covariates and included their likelihood within the GLM framework. Lipsitz et al. (2004) proposed a method for implementing a GLM when one of the covariates is coarsened for only a subset of observations, whereby the resulting likelihood is the integral over the likelihood of all possible values that this variable could have taken, weighted by a density dependent on the uncoarsened, fully observed variables. Johnson and Wiest (2014) proposed a Bayesian approach for coarsened covariates in GLMs whereby a distribution is assumed for the coarsened covariates, given the uncoarsened data.

The above models work well if the distribution of the latent data given the observed data is well-specified, but run into problems if the distributional assumptions are violated. Further, if many covariates are provided in distributional form, these approaches require large computational overheads due to the curse of dimensionality. To the best of our knowledge, little progress has been made in developing logistic regression models for fully histogram-valued predictor variables, which provide far more insight into the shape of the underlying covariate data than the interval case. Here, our primary motivation is in histograms constructed from very large datasets by the analyst to facilitate increased computational speed of an analysis. In this setting, data could arrive in the form of univariate histograms for each predictor for each of the K categories, instead of large ($N \times (K + 1)$)-dimensional tables (where N, the number of observed predictor and class label pairs, is very large), allowing savings in data transmission, storage and analysis.

In this article, we develop methods for implementing logistic regression models with K response categories, where the covariate data for each response category are in the form of marginal or multivariate histograms (or random rectangles, as a special case of random histograms with a single bin). The basic component of our approach adopts the likelihood-based SDA construction of Beranger, Lin, and Sisson (2018) (see also Whitaker, Beranger, and Sisson 2020; Zhang, Beranger, and Sisson 2020), which unlike other SDA-based methods, directly fits models for the underlying micro-data given the

distributional-based data summaries. For this approach, suppose we let the standard classical likelihood be denoted by $L(\mathbf{x}, y; \theta) \propto \prod_{n=1}^{N} g_{X,Y}(y_n | x_n, \theta)$ for N iid pairs (x_n, y_n), where $g_{X,Y}(y_n | x_n, \theta)$ is the assumed distribution of the categorical response Y, given a set of covariates X and a parameter θ. Suppose that a histogram of the covariate values $(x_1, \ldots, x_n)$ is observed for each response category, rather than the observed values (the x_k) themselves, so that it is only known how many covariate vectors, s_b, are contained within bin volume Υ_b, for each bin b in the histogram. The main idea is to obtain the aggregated likelihood by averaging the unknown location of the covariate vector in the classical likelihood uniformly over the histogram bin in which it resides. As a result, the likelihood contribution of a given bin b is proportional to $\left(\int_{\Upsilon_b} g_{X,Y}(y_n | x_n, \theta) dx_n \right)^{s_b}$.

We propose a mixed model in which the underlying data are analyzed using a mixture of standard and SDA likelihoods: histogram bins with low counts are discarded and the residing micro-data are used directly to contribute standard likelihood terms, while bins with high counts contribute via the SDA likelihood framework. However, due to the computational difficulties associated with the potentially high-dimensional integrations required within the SDA likelihood (i.e., $\left(\int_{\Upsilon_b} g_{X,Y}(y_n | x_n, \theta) dx_n \right)^{s_b}$), this mixed model is unsuitable for moderate numbers of predictor variables. For this reason, along with the benefits accompanying the potential additional savings in data storage and computational overheads, we develop models based on lower-dimensional marginal histograms.

We introduce an approximate marginal composite likelihood approach to obtain estimates for the parameter vector of the complete (micro-data) model, using lower-dimensional marginal histograms of the observed covariate data. Univariate and bivariate covariate histograms are considered for these models, with the predictions performed on test datasets shown to be comparable with the full standard likelihood models, but at a much lower computational cost. The resulting parameter estimates are directly comparable to the parameter estimates of the standard full-data analysis.

This article is structured as follows. In Section 2, we recap two types of established logistic regression models for regular data. In Section 3, we briefly outline the general SDA-based likelihood construction of Beranger, Lin, and Sisson (2018), and introduce the mixed standard/SDA likelihood model. We also develop the approximate composite likelihood approach based on lower-dimensional marginal histograms. In Section 4, we perform various simulation studies to demonstrate the effectiveness of each model. We show that each model is able to produce comparable prediction accuracy compared to the full model at a lower computational cost, for a certain sample size, N. We also show that our method performs comparably with the recently developed optimal subsampling method for binary logistic regression by Wang, Zhu, and Ma (2018), but at a cheaper computational cost. In Section 5, we analyze satellite crop-prediction data from Queensland, Australia, and simulated particle collision dataset from the Machine Learning Repository (Dua and Graff 2017), and show that our approach is highly competitive with much more

computationally expensive standard statistical methods. We conclude with a discussion.

2. Logistic Regression Methods for Classification

There are a variety of methods for classifying an instance into one of $K \geq 2$ classes. Let Y denote a discrete random variable taking a value in $\Omega = \{1, \ldots, K\}$ and $X \in \mathbb{R}^D$ an associated vector of explanatory variables. Using the information contained in the covariates X, the aim is to estimate the probability of the outcome of $Y = k$ for $k \in \Omega$. Logistic regression is a widely used statistical modeling technique for binary ($K = 2$) dependent variables. Multinomial logistic regression is a generalization of logistic regression to problems with possible outcomes taking values in Ω ($K \geq 2$). An alternative problem representation recasts multinomial classification as multiple binary classification problems. One-versus-rest (OvR) logistic regression implements a separate binary logistic regression for each class $k \in \Omega$, assuming that each classification model is independent. To establish notation, we briefly describe both multinomial and OvR logistic regression methods below.

2.1. Multinomial Logistic Regression

For each outcome $k \in \Omega \backslash \{K\}$, the multinomial logistic regression model defines the log pairwise odds ratios through a linear model

$$\log \left(\frac{P_{\mathrm{M}}(Y = k|X)}{P_{\mathrm{M}}(Y = K|X)} \right) = \beta_{k0} + \beta_k^\top X, \quad (1)$$

where $P_{\mathrm{M}}(Y = k|X)$ denotes the probability that $Y = k$, under the multinomial model (M), when X is observed, $\beta_{k0} \in \mathbb{R}$ is an intercept and $\beta_k = (\beta_{k1}, \ldots, \beta_{kD})^\top \in \mathbb{R}^D$ represents the vector of regression coefficients associated with the D explanatory variables and the outcome k. The outcome K is referred to as the pivot or reference category, and its corresponding parameter $(\beta_{K0}, \beta_K^\top)^\top$ is the zero vector. Thus, (1) can be rearranged as

$$P_{\mathrm{M}}(Y = k|X) = \frac{e^{\beta_{k0} + \beta_k^\top X}}{1 + \sum_{j \in \Omega \backslash \{K\}} e^{\beta_{j0} + \beta_j^\top X}}$$

for all $k \in \Omega$. This implies that the odds of preferring one class over another do not depend on the presence or absence of other "irrelevant" alternatives.

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_N)$ is a vector of D-dimensional random vectors composed of N iid replicates of X and $\boldsymbol{y} = (Y_1, \ldots, Y_N)^\top$ is a vector of N iid replicates of Y, with respective realizations $\boldsymbol{x} \in \mathbb{R}^{D \times N}$ and $\boldsymbol{y} \in \Omega^N$. The likelihood function under the multinomial model is given by

$$L_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta}) = \prod_{n=1}^{N} \prod_{k \in \Omega} P_{\mathrm{M}}(Y = k|X = x_n)^{\mathbb{1}\{y_n = k\}}, \quad (2)$$

where $\mathbb{1}\{\cdot\}$ represents the indicator function, and $\boldsymbol{\beta} = (\check{\beta}_1, \ldots, \check{\beta}_K) \in \mathbb{R}^{(D+1) \times K}$ with $\check{\beta}_k = (\beta_{k0}, \beta_k^\top)^\top \in \mathbb{R}^{D+1}$. We denote the maximum likelihood estimator for the multinomial logistic regression model as $\hat{\boldsymbol{\beta}}^{\mathrm{M}} = \operatorname{argmax}_{\boldsymbol{\beta}} \log L_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$. Existence of the MLE can be examined through the concept of data separation.

Definition 1 (Multinomial model separation). There is complete separation of the data if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x_n > 0 \text{ for all } n \text{ such that } y_n = k, j \neq k,$$
$$(b_k - b_j)^\top x_n < 0 \text{ for all } n \text{ such that } y_n \neq k, j \neq k.$$

There is quasi-complete separation of the data if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x_n \geq 0 \text{ for all } n \text{ such that } y_n = k, j \neq k,$$
$$(b_k - b_j)^\top x_n \leq 0 \text{ for all } n \text{ such that } y_n \neq k, j \neq k,$$

with equality for at least one observation x_n in each class k.

Albert and Anderson ([1984](#)) proved that if there is neither complete nor quasi-complete separation in the data, then the MLE $\hat{\boldsymbol{\beta}}^{\mathrm{M}}$ exists for all classes $k \in \Omega$.

2.2. OvR Logistic Regression

For each outcome $k \in \Omega$, the OvR logistic regression model defines the log odds ratio through a linear model

$$\log \left(\frac{P_{\mathrm{O}}(Y = k|X)}{P_{\mathrm{O}}(Y \neq k|X)} \right) = \beta_{k0} + \beta_k^\top X,$$

where $P_{\mathrm{O}}(Y = k|X)$ denotes the probability that $Y = k$, under the OvR regression model (O), when X is observed, and where β_{k0} and β_k are as defined previously. This ratio can be rearranged as

$$P_{\mathrm{O}}(Y = k|X) = \frac{e^{\beta_{k0} + \beta_k^\top X}}{1 + e^{\beta_{k0} + \beta_k^\top X}}$$

for all $k \in \Omega$. Note that here β_K is different from the zero vector as each individual binary model has an implied reference category and $\sum_{k \in \Omega} P_{\mathrm{O}}(Y = k|X) \neq 1$. The likelihood function can be written as

$$L_{\mathrm{O}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta}) = \prod_{n=1}^{N} \left(P_{\mathrm{O}}(Y = y_n|X = x_n) \right.$$
$$\left. \prod_{k \in \Omega \backslash \{y_n\}} P_{\mathrm{O}}(Y \neq k|X = x_n) \right), \quad (3)$$

which is expressed as the product of K binary logistic regressions for each observation (x_n, y_n). The parameters of the multivariate and OvR regression models are not directly comparable, but the performance of each model can be assessed by evaluating their prediction accuracy on a training dataset (e.g., Eichelberger and Sheng [2013](#)).

As before, the MLE under the OvR model, $\hat{\boldsymbol{\beta}}^{\mathrm{O}} = \operatorname{argmax}_{\boldsymbol{\beta}} \log L_{\mathrm{O}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$, exists for all classes $k \in \Omega$ if there is neither complete nor quasi-complete separation of the data for each $k \in \Omega$, but under slightly modified definitions of separation compared to the multinomial model (Albert and Anderson [1984](#)).

Definition 2 (OvR model separation). There is complete separation of the data for the kth class if a $b_k \in \mathbb{R}^D$ exists such that

$$b_k^\top x_n > 0 \text{ for all } n \text{ such that } y_n = k,$$
$$b_k^\top x_n < 0 \text{ for all } n \text{ such that } y_n \neq k.$$

There is quasi-complete separation of the data for the kth class if a $b_k \in \mathbb{R}^D$ exists such that

$$b_k^\top x_n \geq 0 \text{ for all } n \text{ such that } y_n = k,$$
$$b_k^\top x_n \leq 0 \text{ for all } n \text{ such that } y_n \neq k,$$

with equality for at least one observation x_n in the kth class, and at least one observation $x_{n'}$, $n' \neq n$, not in the kth class.

3. Classification for Aggregated Data

For each class $k \in \Omega$, define $X^{(k)} = (X_n|Y_n = k, n = 1, \ldots, N) \in \mathbb{R}^{D \times N_k}$, where $N_k = \sum_{n=1}^{N} \mathbb{1}\{Y_n = k\}$, as the matrix of N_k covariate vectors associated with outcome k. In this way $X^{(1)}, \ldots, X^{(K)}$ partition the full covariate set X. When N is very large and $|\Omega| \ll N$ then it can be expected than one or more of the N_k will also be large. In this context, directly optimizing likelihood functions for logistic regression models could be computationally prohibitive. As an alternative, it might be appealing to aggregate the information contained in $X^{(k)}$ into distributional form (such as a histogram) and to implement a classification algorithm using these summaries only, which could be much more computationally efficient. The concept of performing statistical or inferential analyses on such distributional "datapoints" originated from Diday (1989), and has become known as SDA, where the "symbol" corresponds to the distributional summary (see also Bock and Diday 2000; Billard and Diday 2003, 2006).

A symbolic random variable $S_k \in \mathcal{D}_{S_k}$ can be viewed as the result of applying an aggregation function $\pi(\cdot)$ to $X^{(k)} \in \mathcal{D}_{X^{(k)}}$ (where $\mathcal{D}_{X^{(k)}} = \mathbb{R}^{D \times N_k}$), that is, $S_k = \pi(X^{(k)}) : \mathcal{D}_{X^{(k)}} \to \mathcal{D}_{S_k}$ so that $x^{(k)} \mapsto s_k$. In the present context, s_k corresponds to a vector of counts of the number of covariate vectors in $x^{(k)}$ that reside in each histogram bin (see Section 3.1 for more explicit detail). Various likelihood-based techniques for fitting statistical models given the information in distributional summaries have been developed (Le Rademacher and Billard 2011; Brito and Silva 2012; Lin, Caley, and Sisson 2017; Beranger, Lin, and Sisson 2018; Zhang, Beranger, and Sisson 2020). Here we follow the construction of Beranger, Lin, and Sisson (2018) and Zhang, Beranger, and Sisson (2020) who fully model the construction of the symbols from the generating process of the standard random variables $X^{(k)}$. Specifically, the likelihood of observing s_k is

$$L(s_k; \theta, \vartheta) \propto \int_{\mathcal{D}_{X^{(k)}}} f_{S_k|X^{(k)}=x^{(k)}}\left(s_k|x^{(k)}, \vartheta\right) g_{X^{(k)}}\left(x^{(k)}; \theta\right) dx^{(k)},$$

(4)

where $f_{S_k|X^{(k)}}(\cdot; \vartheta)$ is the conditional density of S_k given $X^{(k)}$ relating to the aggregation of $x^{(k)} \mapsto s_k$, $g_{X^{(k)}}(x^{(k)}; \theta)$ is the standard likelihood function of the model at the data level with parameter of interest θ, and $x^{(k)} = \left(x_1^{(k)}, \ldots, x_{N_k}^{(k)}\right)$, with $x_n^{(k)} \in \mathbb{R}^D$ denoting the covariate vector of the nth observation with outcome k. The likelihood (4) is a general expression, for which

the density $f_{S_k|X^{(k)}}(\cdot; \vartheta)$ takes different forms depending on the type of distributional summary considered (see Beranger, Lin, and Sisson 2018, for several examples using random intervals/rectangles and random histograms). In the following, we are interested in aggregating the covariates $X^{(k)}$ that have the same outcome k into histograms (with fixed or random bins), S_k, and to fit logistic regression type models $(g_{X^{(k)}}(x^{(k)}; \theta))$.

3.1. Logistic Regressions Using Histogram-Valued Data

For each class $k \in \Omega$, suppose that the dth margin of $\mathbb{R}^D$ is partitioned into B_k^d bins, so that $B_k^1 \times \cdots \times B_k^D$ bins are created in $\mathbb{R}^D$ through the D-dimensional intersections of each marginal bin. Index each bin by $b_k = (b_{1_k}, \ldots, b_{D_k})$, $b_{d_k} = 1, \ldots, B_k^d$, as the D-dimensional vector of co-ordinates of each bin in the histogram, and let $\mathcal{B}_k$ denote the set of bin indices for class k. The bin b_k is constructed over the space $\boldsymbol{\Upsilon}_{b_k} = \Upsilon_{b_k}^1 \times \cdots \times \Upsilon_{b_k}^D \subset \mathbb{R}^D$.

Let S_k represent a D-dimensional histogram constructed from $X^{(k)}$ through the aggregation function π where

$$S_k = \pi(X^{(k)}) : \mathbb{R}^{N_k \times D} \to \{0, \ldots, N_k\}^{B_k^1 \times \cdots \times B_k^D} \quad (5)$$

$$x^{(k)} \mapsto s_k = \left(s_{b_k} = \sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \boldsymbol{\Upsilon}_{b_k}\}, \forall b_k \in \mathcal{B}_k\right).$$

The quantity S_{b_k} denotes the random number of observed data points $X_1^{(k)}, \ldots, X_{N_k}^{(k)}$ that fall in the bin b_k and consequently, the histogram-valued random variable S_k represents the full $(B_k^1 \times \cdots \times B_k^D)$-dimensional vector of counts. In this manner, we can construct the collection of histograms $S = (S_1, \ldots, S_K)$, with one S_k for each outcome index $k \in \Omega$, that summarize the information contained in $X = (X^{(1)}, \ldots, X^{(K)})$.

Proposition 1. Suppose that $X^{(k)}$, the covariates associated with each outcome $k \in \Omega$, are aggregated via (5), and let $S = (S_1, \ldots, S_K)$ denote the resulting collection of histograms. For this summarized data S, using (4), the likelihood functions for the multinomial (2) and OvR (3) logistic regression models become

$$L_{\text{SM}}(s; \beta) \propto \prod_{k \in \Omega} \prod_{b_k \in \mathcal{B}_k} \left(\int_{\boldsymbol{\Upsilon}_{b_k}} P_{\text{M}}(Y = k|X = x)dx\right)^{s_{b_k}}, \quad (6)$$

$$L_{\text{SO}}(s; \beta) \propto \prod_{k \in \Omega} \prod_{b_k \in \mathcal{B}_k} \left(\int_{\Upsilon_{b_k}} P_{\text{O}}(Y = k|X = x)dx\right.$$
$$\left.\prod_{k' \in \Omega \setminus \{k\}} \int_{\Upsilon_{b_k}} P_{\text{O}}(Y \neq k'|X = x)dx\right)^{s_{b_k}},$$

(7)

where s is the observed value of S, and s_{b_k} denotes the number of observations in bin b_k in histogram k. For a derivation see the Appendix.

We refer to these models as the symbolic multinomial (SM) and symbolic OvR (SOvR) logistic models. In effect, the uncertainty of the location of each predictor X is averaged uniformly over its location in the histogram bin in which it resides. Note

that the likelihood functions (6) and (7) only implicitly depend on the vector of outcomes $\boldsymbol{y}$ since for each possible outcome $k \in \Omega$ the covariates $\boldsymbol{X}^{(k)}$ are summarized in a histogram S_k, and so the product of the N (Y_n, X_n) observations in the standard likelihoods (2) and (3) is replaced by a product over the K outcomes. Further, the parameter ϑ in (4) denotes quantities relevant to constructing the symbol (e.g., the number of bins and their locations), and so is fixed in this setting, and is therefore omitted in the notation.

Following similar arguments to Heitjan (1989) and Beranger, Lin, and Sisson (2018), the symbolic likelihoods $L_{\mathrm{SM}}(\boldsymbol{s}; \boldsymbol{\beta})$ and $L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$ can each be shown to approach their classical equivalent, $L_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$ and $L_{\mathrm{O}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$, as the number of bins in each histogram approaches infinity and the volume of each bin approaches zero. In this scenario, in the limit each bin will either be empty ($s_{b_k} = 0$) and so the bin will not contribute a likelihood term, or will contain exactly one point ($s_{b_k} = 1$) observed at each value of $x_n^{(k)}$, which then recovers the classical likelihood term for $x_n^{(k)}$. In this manner, the histogram-based likelihoods can be viewed as approximations to the standard likelihood functions for each model.

To establish conditions for the existence of the respective MLEs, $\hat{\boldsymbol{\beta}}^{\mathrm{SM}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{SM}}(\boldsymbol{s}; \boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}^{\mathrm{SO}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$, we need to consider modified definitions of complete and quasi-complete separation of the data, in analogy with Definitions 1 and 2, to account for the fact that the location of each covariate vector x_n is only known up to the histogram bin in which it resides.

Definition 3 (Histogram-based multinomial model separation). There is complete separation of the set of histograms $\boldsymbol{s}$ if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x > 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{b_k} > 0, j \neq k,$$
$$(b_k - b_j)^\top x < 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{b_{k'}} > 0, j \neq k$$
$$\text{and} \quad k' \neq k.$$

There is quasi-complete separation of the set of histograms $\boldsymbol{s}$ if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x \geq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{b_k} > 0, j \neq k,$$
$$(b_k - b_j)^\top x \leq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{b_{k'}} > 0, j \neq k$$
$$\text{and} \quad k' \neq k.$$

with equality for at least one point in the nonempty histogram bins for the kth class.

Definition 4 (Histogram-based OvR model separation). There is complete separation of the histogram for the kth class, s_k, if a vector b_k exists such that

$$b_k^\top x > 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{b_k} > 0,$$
$$b_k^\top x < 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{b_{k'}} > 0, k \neq k'.$$

There is quasi-complete separation of the histogram for the kth class if there exists a vector b_k such that

$$b_k^\top x \geq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{b_k} > 0,$$
$$b_k^\top x \leq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{b_{k'}} > 0, k \neq k',$$

with equality for at least one point in the nonempty histogram bins for the kth class, and equality for at least one point in any nonempty histogram bin not in the kth class.

Proposition 2. If the set of histograms $\boldsymbol{s} = (s_1, \ldots, s_K)$ does not exhibit complete or quasi-complete separation as described under Definition 3, then $L_{\mathrm{SM}}(\boldsymbol{s}; \boldsymbol{\beta})$ has a unique global maximum. If the set of histograms does not exhibit complete or quasi-complete separation for any class $k \in \Omega$ as described under Definition 4, then $L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$ has a unique global maximum. For a proof see the Appendix.

From Definitions 1–4, it can be seen that if there is separation of the histograms then there also has to be separation of the underlying data. In this sense, the definitions of complete and quasi-complete separation for random histograms (Definitions 3 and 4) are stronger conditions than those for standard random vectors (Definitions 1 and 2). As a result, from Proposition 2 this means that if no histogram-based MLE ($\hat{\boldsymbol{\beta}}^{\mathrm{SM}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{SO}}$) exists, then the equivalent standard logistic model MLE ($\hat{\boldsymbol{\beta}}^{\mathrm{M}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{O}}$) also does not exist. That is, the standard MLE cannot exist without the histogram-based MLE also existing. Conversely, however, it is possible to have separation of the underlying data but no separation in the derived histograms as a result of covariate vector's locations being "blurred" in the histogram bins. Accordingly, it is possible that the histogram-based MLE exists without the standard logistic model MLE existing. (In this particular setting, we therefore have the interesting case of the existence of an MLE for a given histogram converging to the nonexistence of the MLE in the limit as the number of histogram bins become large while the volume of each bin approaches zero.)

It is also possible for the histogram-based MLEs to exist for one histogram derived from an underlying dataset, but not exist for a different histogram (e.g., with different numbers of and/or locations of bins) derived from the same dataset. For example, a shift in the location of the bins can lead to low-count bins becoming empty (or empty bins becoming nonempty), meaning that a set of histograms previously exhibiting separation, could instead exhibit nonseparation (and vice versa). Practically, as a result, if the underlying classical data are available, the practitioner should consider verifying the separation conditions on the underlying dataset (Definitions 1 and 2) prior to aggregation, to ensure the appropriateness of the logistic model.

Assuming that the MLEs exist, the benefits of using histograms as data summaries for logistic regression modeling are obvious in the presence of large amounts of data. The effective number of likelihood terms in $L_{\mathrm{SM}}(\boldsymbol{s}; \boldsymbol{\beta})$ and $L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$ is the number of histogram bins multiplied by the number of classes. For very large datasets this can be much smaller than the N terms in $L_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$ and $L_{\mathrm{SO}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$, and so computing MLEs given the histogram summaries can be much more efficient. The tradeoff is the loss of some accuracy due to the loss of information in the binned data.

Despite its computational advantages, the above construction has some limitations. We now discuss these and propose some statistical and computational improvements.

3.2. Using Both Classical Data and Histograms

When constructing histograms, for example using the method described in Section 3.1, the number of observations s_{b_k} within each bin b_k will typically vary widely over bins, from very low to very high counts. Where bins have high counts, large computational efficiencies are obtained in the evaluation of $L_{\mathrm{SM}}(s; \beta)$ and $L_{\mathrm{SO}}(s; \beta)$ over the standard logistic regression likelihood functions. However, when a bin has low numbers of underlying data points, it may be that the computational cost in evaluating the bin-specific integrals in (6) and (7) (such as, e.g., $\int_{\Upsilon_{b_k}} P_{\mathrm{M}}(Y = k | X = x) dx$) is just as high or higher than evaluating the standard likelihood contributions (e.g., $P_{\mathrm{M}}(Y = k | X = x_n)$) for each of the underlying datapoints in that bin. Taken together with the loss of information in moving from the underlying data to a count of datapoints in a bin, in this case it is obviously worse statistically (and perhaps also computationally) to work with the histogram bin rather than the original data in that bin. Creating bins with low data counts becomes more likely as the dimension D of the predictors increases.

To avoid this situation, we introduce a lower bound, $\tau_k \in \{1, \dots, N_k\}$, on the number of underlying datapoints in a bin region Υ_{b_k} that is required before these data can be summarized into a histogram bin for their contribution to the likelihood function. When the number of underlying datapoints is lower than τ_k, the original data x_n are retained, and contribute to the likelihood in the standard way. Selection of the threshold τ_k can be based on the time needed to evaluate the likelihood for the binned data (of either the exact value or a numerical approximation of the integral in (6) or (7)) compared to the standard likelihood function, so that binned data are used when it is more computationally efficient to do so.

Under the assumption that the underlying data are available (which may not always be the case), and that the evaluation of the classical likelihood is computationally infeasible given the dataset size, we therefore propose to use the modified aggregation function

$$S_k = \tilde{\pi}(X^{(k)}) : \mathbb{R}^{N_k \times D} \to \{\tau_k, \dots, N_k\}^u \times \mathbb{R}^{v \times D}$$

$$x^{(k)} \mapsto \left(\begin{cases} s_{b_k} = \sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \Upsilon_{b_k}\} & \text{if } s_{b_k} \geq \tau_k \\ x_{b_k}^{(k)} = \{x_n^{(k)} : x_n^{(k)} \in \Upsilon_{b_k}\} & \text{otherwise} \end{cases}, \ b_k \in \mathcal{B}_k \right),$$

where $\tau_k \in \{1, \dots, N_k\}, u \in [0, \dots, B_k^1 \times \cdots \times B_k^D]$ is the number of histogram bins containing at least τ_k observations, and $v = N_k - \sum s_{b_k}$ is the number of retained classical datapoints in bins containing less than τ_k observations. That is, the resulting S_k is a mixture of those histogram bins that contain at least τ_k observations, combined with any remaining predictor vectors $X_n^{(k)}$ that would otherwise be put into bins with less than τ_k observations.

In the context of logistic regression modeling, this mixture histogram construction produces likelihood functions that are a mixture of the standard and histogram-based likelihood functions given in Sections 2.1, 2.2, and 3.1. For example, we can construct the likelihood function for a mixture of histogram and classical data under the multinomial logistic regression model as

$$L_{\mathrm{MM}}(s; \beta) \propto \prod_{k \in \Omega} \prod_{b_k \in \mathcal{B}_k} \left(\int_{\Upsilon_{b_k}} P_{\mathrm{M}}(Y = k | X = x) dx \right)^{s_{b_k} \mathbb{1}\{s_{b_k} \geq \tau_k\}}$$

$$\left(\prod_{x \in x_{b_k}^{(k)}} P_{\mathrm{M}}(Y = k | X = x) \right)^{\mathbb{1}\{s_{b_k} < \tau_k\}}, \qquad (8)$$

where MM denotes the multinomial mixture. A similar mixture-likelihood, $L_{\mathrm{MO}}(s; \beta)$, can be constructed for the OvR logistic regression model.

Because $L_{\mathrm{MM}}(s; \beta)$ can be considered as a special case of $L_{\mathrm{SM}}(s; \beta)$ (and $L_{\mathrm{MO}}(s; \beta)$ a special case of $L_{\mathrm{SO}}(s; \beta)$) in which the retained classical data vectors x_n can be thought of as residing in zero-volume bins, one for each retained vector, it is immediate that $L_{\mathrm{MM}}(s; \beta) \to L_M(x, y; \beta)$ also approaches that classical data likelihood function as the number of bins becomes large and the volume of each bin approaches zero (similarly $L_{\mathrm{MO}}(s; \beta) \to L_O(x, y; \beta)$).

Similarly, by considering the obvious definition of complete and quasi-complete separation for the mixture of histogram and retained classical data vectors as a combination of those in Definitions 1 and 3 (for the standard multinomial regression model) and Definitions 2 and 4 (for the OvR regression model), similar statements to Proposition 2 about the existence of the MLEs $\hat{\beta}^{\mathrm{MM}} = \arg\max_\beta \log L_{\mathrm{MM}}(s; \beta)$ and $\hat{\beta}^{\mathrm{MO}} = \arg\max_\beta \log L_{\mathrm{MO}}(s; \beta)$ can be made. For example, if no full-histogram MLE exists ($\hat{\beta}^{\mathrm{SM}}$ or $\hat{\beta}^{\mathrm{SO}}$) then no mixture-likelihood MLE exists ($\hat{\beta}^{\mathrm{MM}}$ or $\hat{\beta}^{\mathrm{MO}}$) and no standard likelihood MLE ($\hat{\beta}^{\mathrm{M}}$ or $\hat{\beta}^{\mathrm{O}}$) exists. That is, the standard MLE cannot exist without the mixture-likelihood MLE existing, which cannot itself exist without the full-histogram MLE existing. However, the full-histogram model MLE can exist without the mixture-likelihood MLE existing, and the mixture-likelihood MLE can exist without the standard MLE existing.

The choice of τ_k, for all $k \in \Omega$, controls the tradeoff between computational efficiency and information loss. On one hand, if τ_k is too large then we face the original issue of having a huge number of terms slowing down evaluation of the likelihood function. On the other hand if τ_k is too low then we risk a loss of efficiency (and perhaps higher computation) compared to higher τ_k. As a result, one option is to set τ_k to be the value such that evaluating the integrated bin likelihood term over over a bin b_k is less computationally expensive than evaluating the standard likelihood contribution τ_k times. Some strategies along these lines are explored in the simulations in Section 4.

3.3. Composite Likelihoods for Logistic Regression Models

Mixing histogram and micro data can lead to substantial statistical efficiency improvements, but it does not address the issue of grid-based multivariate histograms becoming highly inefficient as data summaries as the number of covariates (D) increases. In particular, the integrals required to compute the likelihood function $L_{\mathrm{SM}}(s; \beta)$ (6) have no analytical solution when the outcome has more than two possible classes ($K > 2$)

and there are more than two explanatory variables ($D > 2$). Similarly, the integrals in the likelihood function $L_{SO}(s; \boldsymbol{\beta})$ (7) have no analytical solution when more than two explanatory variables are considered ($D > 2$). In all nontrivial settings, then, these integrals must be computed numerically. This can be computationally costly when D is large, which can then defeat the purpose (i.e., improved computational efficiency) of using data aggregates.

To circumvent the issue of computing the probabilities of data falling in high-dimensional bins, Whitaker, Beranger, and Sisson (2020) proposed implementing a composite likelihood approach. This consisted of approximating the likelihood function of a high-dimensional histogram by the weighted product of likelihood functions for lower-dimensional *marginal* histograms, which yielded asymptotically consistent likelihood-based parameter estimates (Lindsay 1988; Varin, Reid, and Firth 2011). Assuming all weights are equal for simplicity, a j-wise composite likelihood function can be expressed as $L^{(j)}(\theta) \propto \prod_{i=1}^{m} L_i(\theta)$, where $L_i(\theta)$ is the likelihood function of one of m j-wise marginal events for a given parameter vector θ. In the current context, $L_i(\theta)$ corresponds to a likelihood contribution based on the subset of covariates represented by a j-dimensional marginal histogram derived from the full D-dimensional histogram, $\theta = \boldsymbol{\beta}$ and $m = \binom{D+1}{j}$.

However, omitting an important variable in probit and logistic regression analyses will depress the estimated vector of the remaining coefficients toward zero (Wooldridge 2002; Cramer 2007). It is therefore nonviable to directly apply a composite likelihood approach to logistic regression problems. However, in the OvR setting and under the assumption that all predictors are independent, Cramer (2007) showed that the nonomitted coefficients of a logistic regression can be written as functions of the regression coefficients in the scenario that no regressor is omitted. This result was primarily aimed at highlighting the effect of omitting variables in a regression analysis context, and had no practical use for, for example, compensating for zero-depressed parameter estimates, since the established correspondences required information about the variances of the omitted variables, which were unavailable. However, in the composite likelihood setting such information about each covariate is available, and the result of Cramer (2007) can therefore be implemented within each marginal likelihood contribution to compensate for the covariates that are omitted in that term. We implement this concept in Proposition 3.

In the remainder of this section, we construct composite likelihoods for the OvR and histogram-based OvR logistic regression model. We do not consider multinomial logistic regression here as the results of Cramer (2007) do not hold for multinomial logistic regression. Let $\boldsymbol{i} = (i_1, \ldots, i_I) \subseteq \{1, \ldots, D\}$, where for convenience $i_1 < \cdots < i_I$, and define by $\mathcal{I}_j = \{\boldsymbol{i} : |\boldsymbol{i}| = j\}$ the set of all j-dimensional subsets of $\{1, \ldots, D\}$. We adopt the notation that a vector with superscript $\boldsymbol{i}$ denotes the subvector containing those elements corresponding to the index set $\boldsymbol{i}$. For matrices with the superscript $\boldsymbol{i}$, the operation is replicated column-wise. For example, for $\boldsymbol{i} \in \mathcal{I}_j$, $\boldsymbol{X}^{(k)\boldsymbol{i}} = (X_1^{(k)\boldsymbol{i}}, \ldots, X_{N_k}^{(k)\boldsymbol{i}}) \in \mathbb{R}^{j \times N_k}$ where $X_n^{(k)} \in \mathbb{R}^j$ is a subvector of $X_n^{(k)}$, $n = 1, \ldots, N_k$. Then if $S_{\boldsymbol{b}_k^{\boldsymbol{i}}}^{\boldsymbol{i}}$ is the random number of observed data points in $\boldsymbol{X}^{(k)\boldsymbol{i}}$

that fall in bin $\boldsymbol{b}_k^{\boldsymbol{i}}$ and $\mathcal{B}_k^{\boldsymbol{i}}$ the set of all such bin indices, we may construct an I-dimensional random *marginal* histogram $\boldsymbol{S}_k^{\boldsymbol{i}}$ which satisfies $\sum_{\boldsymbol{b}_k^{\boldsymbol{i}}} S_{\boldsymbol{b}_k^{\boldsymbol{i}}}^{\boldsymbol{i}} = N_k$.

The following proposition establishes how to perform approximate composite likelihood estimation for the OvR and histogram-based OvR regressions models using the results in Cramer (2007). As independence between predictors, as assumed by Cramer (2007), is unrealistic, the proposition also extends the results of Cramer (2007) to account for the correlation between the included set of predictor variables within each composite likelihood term and the omitted variables. Without loss of generality, consider a random vector $X \in \mathbb{R}^D$ and for $\boldsymbol{i} \in \mathcal{I}_j$, let $X^{\boldsymbol{i}} \in \mathbb{R}^j$ represent the observed variables of X. Further let $\mathcal{I}_1^{-\boldsymbol{i}} = \{1, \ldots, D\} \backslash \{\boldsymbol{i}\}$ such that for all $i' \in \mathcal{I}_1^{-\boldsymbol{i}}$, $X^{i'}$ represents an omitted variable of X. Following Cramer (2007), we define the omitted variables $X^{i'}$ to be a linear function of the observed variables via $X^{i'} = \alpha_{\boldsymbol{i}i'}^\top X^{\boldsymbol{i}} + \epsilon_{\boldsymbol{i}i'}$, where $\alpha_{\boldsymbol{i}i'} = (\alpha_{i_1 i'}, \ldots, \alpha_{i_j i'})^\top \in \mathbb{R}^j$ and $\epsilon_{\boldsymbol{i}i'} \sim N(0, \lambda_{\boldsymbol{i}i'}^2)$. Denote $\text{cov}(\epsilon_{\boldsymbol{i}i'_1}, \epsilon_{\boldsymbol{i}i'_2}) = \lambda_{\boldsymbol{i}i'_1 i'_2}$.

Proposition 3. The j-wise approximate composite likelihood functions for the standard and the histogram-based D-dimensional OvR logistic regression models are, respectively, given by

$$L_O^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_O(\boldsymbol{x}^{\boldsymbol{i}}, y; \widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}) \quad \text{and}$$

$$L_{SO}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_{SO}(\boldsymbol{s}^{\boldsymbol{i}}, y; \widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}),$$

where the lower dimensional regression observed coefficients are given by $\widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}} = (\widetilde{\beta}_1^{\boldsymbol{i}}, \ldots, \widetilde{\beta}_K^{\boldsymbol{i}}) \in \mathbb{R}^{(j+1) \times K}$ where

$$\widetilde{\beta}_k^{\boldsymbol{i}} = \frac{\beta_k^{\boldsymbol{i}} + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta_k^{i'} \alpha_{\boldsymbol{i}i'} \right)^\top \right]^\top}{\sqrt{1 + \frac{\pi^2}{3} \sum_{i'_1 \in \mathcal{I}_1^{-\boldsymbol{i}}} \left[(\beta_k^{i'_1} \lambda_{\boldsymbol{i}i'_1})^2 + 2 \sum_{i'_2 \in \mathcal{I}_1^{-\boldsymbol{i}}, i'_2 \neq i'_1} \beta_k^{i'_1} \beta_k^{i'_2} \lambda_{\boldsymbol{i}i'_1 i'_2} \right]}}$$

$$\in \mathbb{R}^{(j+1)}. \tag{9}$$

If there is neither complete nor quasi-complete separation in the full D-dimensional dataset $\boldsymbol{x}$ (Definition 1) for any binary logistic model, then $L_O^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta})$ will have a unique global maxima. Similarly, if there is neither complete nor quasi-complete separation in any of the sets of marginal histograms $\boldsymbol{s}^{\boldsymbol{i}}$, $\boldsymbol{i} \in \mathcal{I}_j$, then $L_{SO}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ will have a unique global maxima. See the Appendix for a derivation and proof.

Note that $L_O^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta})$ and $L_{SO}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ are *approximate* composite likelihood functions rather than true composite likelihood functions. The resulting maximum composite likelihood estimators ($\hat{\boldsymbol{\beta}}_O^{(j)}$ and $\hat{\boldsymbol{\beta}}_{SO}^{(j)}$) are not unbiased or consistent. They are, however, reasonable parameter estimates if one is motivated to estimate logistic regression model parameters within the composite likelihood framework (as is the case here), that are more accurate than those estimated via a naive composite likelihood implementation (which we define here as simply $\tilde{L}_{SO}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta}) =$

$\prod_{\boldsymbol{i}\in\mathcal{I}_j} L_{\mathrm{SO}}(\boldsymbol{s}^{\boldsymbol{i}},\boldsymbol{y};\boldsymbol{\beta}^{\boldsymbol{i}})$). We numerically demonstrate the performance of these estimators against the naive implementation in the simulation study in Section 4.2. However, when our primary aim is model predictive accuracy, we will demonstrate that the performance of the fitted model using the approximate composite likelihood MLE ($\hat{\boldsymbol{\beta}}_{\mathrm{O}}^{(j)}$ or $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$) is highly competitive with using the full-data standard MLE, $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$, while, in the case of $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$, being much more computationally efficient to obtain.

Following similar arguments to before, it is clear that $L_{\mathrm{SO}}^{(j)}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\beta}) \rightarrow L_{\mathrm{O}}^{(j)}(\boldsymbol{s};\boldsymbol{\beta})$ as the number of histogram bins becomes large while the volume of each bin approaches zero. Whitaker, Beranger, and Sisson (2020) proved the asymptotic normality and consistency of the histogram-based (true) composite likelihood estimator, and highlighted that parameter variance consistency requires the number of bins in each histogram to become large and the number of histograms representing the data be close to N, the number of data points. While these results are not directly applicable under the approximate composite likelihood of Proposition 3, we intuitively expect that the estimated variances of $\hat{\boldsymbol{\beta}}_{\mathrm{O}}^{(j)}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$ will be inflated compared to that of $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$ unless the same conditions hold.

The estimation procedure for the histogram-based D-dimensional OvR logistic regression model is presented in Algorithm 1. It illustrates the use of the j-wise approximate composite likelihood given in Proposition 3, and considers the full likelihood equivalent of Proposition 1 as the special case of $j = D$. This algorithm can be adapted for the multinomial logistic regression model by considering the likelihood function $L_{\mathrm{SM}}(\boldsymbol{s};\boldsymbol{\beta})$ given in (6) rather than $L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$, but only for $j = D$. Estimation for the mixture models MM and MO is achieved in a similar manner.

Algorithm 1: Estimation procedure using the j-wise approximate composite likelihood for the histogram-based D-dimensional OvR logistic regression model.

Input data: $Y \in \{1,\dots,K\}^N, \boldsymbol{X} \in \mathbb{R}^{N\times D}$

1 **Divide $\boldsymbol{X}$ into $\boldsymbol{X}^{(1)},\dots,\boldsymbol{X}^{(K)}$ where**
  $\boldsymbol{X}^{(k)} = (X_n | Y_n = k, n = 1,\dots,N) \in \mathbb{R}^{D\times N_k}$

2 Define the locations $\Upsilon_{\boldsymbol{b}_k}$ of the bins $\boldsymbol{b}_k$

3 **if** $j = D$ **then**

4   Construct $\boldsymbol{S} = (S_1,\dots,S_K)$ by applying (5) to $\boldsymbol{X}^{(1)},\dots,\boldsymbol{X}^{(K)}$

5   Evaluate $L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$ using (7)

6   Compute $\hat{\boldsymbol{\beta}}^{\mathrm{SO}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$

  else

7   **for** $\boldsymbol{i} \in \mathcal{I}_j$ **do**

8   Construct $\boldsymbol{S}^{\boldsymbol{i}} = (S_1^{\boldsymbol{i}},\dots,S_k^{\boldsymbol{i}})$ by applying (5) to subvectors $\boldsymbol{X}^{(1)\boldsymbol{i}},\dots,\boldsymbol{X}^{(K)\boldsymbol{i}}$

9   Estimate $\alpha_{\boldsymbol{ii'_1}}, \lambda_{\boldsymbol{ii'_1}}$ and $\lambda_{\boldsymbol{ii'_1 i'_2}}$ for all $i'_1, i'_2 \in \mathcal{I}_1^{-\boldsymbol{i}}, i'_1 \neq i'_2$

10   Evaluate $\widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}$ using (9) and $L_{\mathrm{SO}}(\boldsymbol{s}^{\boldsymbol{i}},y;\widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}})$ using (7)

11 Compute $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)} = \arg\max_{\boldsymbol{\beta}} \sum_{\boldsymbol{i}\in\mathcal{I}_j} \log L_{\mathrm{SO}}(\boldsymbol{s}^{\boldsymbol{i}},y;\widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}})$

In specific cases, we can obtain a closed-form approximate composite likelihood function for a D-dimensional random histogram, $L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$. In the particular case of a binary outcome ($K = 2$) and using $j = 1$ to construct the composite likelihood from all univariate marginal events (the set $\mathcal{I}_1 = \{1,\dots,D\}$), we have

$$L_{\mathrm{SO}}^{(1)}(\boldsymbol{s};\boldsymbol{\beta}) = \prod_{i\in\mathcal{I}_1}\prod_{k=1}^{2}\prod_{b_{i_k}\in\mathcal{B}_k^i}\left[\left(\frac{1}{\widetilde{\beta}_{k1}^i}\right)^2 \log\left(\frac{1+e^{\widetilde{\beta}_{k0}^i+\widetilde{\beta}_{k1}^i y_{b_i}^i}}{1+e^{\widetilde{\beta}_{k0}^i+\widetilde{\beta}_{k1}^i y_{b_i-1}^i}}\right)\right.$$
$$\left.\log\left(\frac{1+e^{-\widetilde{\beta}_{k0}^i-\widetilde{\beta}_{k1}^i y_{b_i-1}^i}}{1+e^{-\widetilde{\beta}_{k0}^i-\widetilde{\beta}_{k1}^i y_{b_i}^i}}\right)\right]^{s_{b_{i_k}}^i},$$
(10)

where $\widetilde{\beta}_k^i = (\widetilde{\beta}_{k0}^i, \widetilde{\beta}_{k1}^i) \in \mathbb{R}^2$ and the bin indexed by $\boldsymbol{b}_k^i = b_{i_k}$ is constructed over the space $\Upsilon_{\boldsymbol{b}_k^i}^i = (y_{b_i-1}^i, y_{b_i}^i]$. In all other cases, the integrals in $L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$ (7) require numerical estimation.

Evaluating (9) within the approximate composite likelihood requires knowledge of $\alpha_{\boldsymbol{ii'_1}}, \lambda_{\boldsymbol{ii'_1}}$, and $\lambda_{\boldsymbol{ii'_1 i'_2}}$ for all $i'_1, i'_2 \in \mathcal{I}_1^{-\boldsymbol{i}}, i'_1 \neq i'_2$. The $\alpha_{\boldsymbol{ii'_1}}$ terms are the coefficients explaining the variations of an unobserved variable $X^{i'_1}$ as a linear function of the observed variables $X^{\boldsymbol{i}}$ (i.e., $X^{i'} = \alpha_{\boldsymbol{ii'}}^\top X^{\boldsymbol{i}} + \epsilon_{\boldsymbol{ii'}}$). In the case where $\boldsymbol{i} = i \in \mathcal{I}_1$ (i.e., a simple linear regression with $j = 1$ as in (10)), an estimate of $\alpha_{\boldsymbol{ii'_1}}$ is $\mathrm{cov}(X^i, X^{i'_1})/\mathrm{var}(X^i)$. In this context, rewriting $\epsilon_{\boldsymbol{ii'}} = X^{i'_1} - \alpha_{\boldsymbol{ii'}} X^i$, we also have that

$$\hat{\lambda}_{\boldsymbol{ii'_1}}^2 = \mathrm{var}(X^{i'_1}) - \frac{\mathrm{cov}(X^i, X^{i'_1})^2}{\mathrm{var}(X^i)} \quad \text{and}$$

$$\hat{\lambda}_{\boldsymbol{ii'_1 i'_2}} = \mathrm{cov}(X^{i'_1}, X^{i'_2}) - \frac{\mathrm{cov}(X^i, X^{i'_1})\mathrm{cov}(X^i, X^{i'_2})}{\mathrm{var}(X^i)}.$$

Knowledge of variances and covariances between the covariates X is similarly required when two or more predictors are considered in each approximate composite likelihood contribution, that is, when $\boldsymbol{i} \in \mathcal{I}_j$, for $j \geq 2$. Ideally these variances and covariances should be computed and stored prior to the data aggregation process, that is, on the full dataset, but if this information is unavailable, variance and covariance estimates can be derived directly from the histograms either with (Beranger, Lin, and Sisson 2018) or without (Billard and Diday 2003; Billard 2011) parametric assumptions. Clearly, the assumption that a missing variable can be written as a linear combination of observed variables may not hold. Where viable, transformations (e.g., Box–Cox) or more flexible regression models can be applied to provide a more realistic model, either regressing on the transformed covariates or modifying the form of $\widetilde{\beta}_k^{\boldsymbol{i}}$ (9) as required.

Finally, suppose that we again consider the case $j = 1$ so that the approximate composite likelihood is constructed with each term comprising each covariate separately (10). This is the most computationally efficient histogram-based likelihood as it is based solely on univariate histograms (and known covariances with the other covariates). This construction implies that only univariate marginal histograms are required. Beranger, Lin, and Sisson (2018) introduced two likelihood constructions

for histogram-valued variables. The first, which we have used until now, assumes that histogram bins are fixed and the corresponding counts are random, which works straightforwardly in D-dimensions. The second construction is a quantile-based approach for univariate variables only, where the bin locations are assumed random and the counts fixed. Defining bin locations using quantiles can better describe the behavior of the underlying data, and also has the advantage of retaining some of the micro-data (at the observed quantiles), which resembles the mixture of histogram and standard likelihood approach of Section 3.2.

In this setting, for each $k \in \Omega$ and each marginal component $i \in \mathcal{I}_1$, define a vector of order statistics $t = (t_1, \ldots, t_B)^\top$ where $1 \leq t_1 \leq \cdots \leq t_B \leq N_k$, such that a quantile-based histogram-valued random variable is obtained through the aggregation function

$$S_k^i = \dot{\pi}(X^{(k)i}) : \mathbb{R} \to \mathcal{S} = \{(a_1, \ldots, a_B) \in \mathbb{R}^B : a_1 < \cdots < a_B\} \times \mathbb{N} \tag{11}$$

$$x^{(k)i} \mapsto s_k^i = \left(x_{(t_1)}^{(k)i}, \ldots, x_{(t_B)}^{(k)i}, N_k \right),$$

where $x_{(t_b)}^{(k)i}$ denotes the t_bth order statistic of $x^{(k)i} \in \mathbb{R}$. (Note that to ease notation we have omitted superscripts and subscripts related to i and k in the order statistics t.) The bth histogram bin is then defined over the range $(s_{kb-1}^i, s_{kb}^i]$, with fixed counts of underlying datapoints $t_b - t_{b-1}$, for $b = 1, \ldots, B+1$, where $s_0 = -\infty$, $s_{B+1} = +\infty$, $t_0 = 0$ and $t_{B+1} = N_k+1$. If each covariate is aggregated via (11), then the resulting approximate composite likelihood function is

$$L_{\mathrm{OO}}^{(1)}(s; \beta) = L_{\mathrm{O}}^{(1)}(\{x^{(k)i}\}; \beta) L_{\mathrm{SO}}^{(1)}(s; \beta),$$

where $s = (s_1, \ldots, s_K)$ and $s_k = (s_k^1, \ldots, s_k^D)$ with s_k^i defined in (11), and where $L_{\mathrm{SO}}^{(1)}$ is the likelihood shown in (10).

4. Simulation Studies

We now examine the parameter estimation and classification capabilities of the methods developed in Section 3 based on simulated data. We consider both the statistical and computational performance of the histogram-based analyses compared to the standard full-data approach.

In the following we set the number of possible outcomes K to define $\Omega = \{1, \ldots, K\}$, the domain of the response variable Y. We obtain the $(D \times 2N)$ matrix of covariates X by generating $2N$ observations from a specified D-dimensional distribution. Given a fixed matrix of regression coefficients $\beta \in \mathbb{R}^{(D+1) \times K}$, for each $n = 1, \ldots, 2N$ we compute the probability of every outcome in Ω using (1). These probabilities are then used to generate $Y \in \Omega^{2N}$ from a multinomial distribution. The dataset (Y, X) is then split into equally sized training and test datasets, and the estimates $\hat{\beta}^{\mathrm{M}}, \hat{\beta}^{\mathrm{O}}, \hat{\beta}^{\mathrm{MM}}, \hat{\beta}^{\mathrm{O}(j)}$, and $\hat{\beta}^{\mathrm{SO}(j)}$ are obtained by maximizing their respective likelihood functions on the training dataset. Using the test dataset we compute the prediction accuracy (PA) of a model (and estimation procedure) as

$$\mathrm{PA} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Y_n^{\mathrm{Pred}} = Y_n\},$$

where $Y_n^{\mathrm{Pred}} = \mathrm{argmax}_{k \in \Omega} P(Y = k | X = X_n), n = 1, \ldots, N$, denotes the predicted outcome under the model (multinomial or OvR) with estimated coefficients $\hat{\beta}^{\mathrm{M}}, \hat{\beta}^{\mathrm{O}}, \hat{\beta}^{\mathrm{MM}}, \hat{\beta}^{\mathrm{O}(j)}$, or $\hat{\beta}^{\mathrm{SO}(j)}$. After replicating the above 1000 times, we report the mean squared errors (MSE) and mean prediction accuracies of each estimator.

In Section 3.2, we proposed to improve the statistical and computational efficiency of the MLE by retaining the data underlying a histogram bin, rather than using the bin itself, if the number of underlying observations in the bin s_{b_k} is less than τ_k. We use quadrature methods to perform the integrations in (6), for which the number of function evaluations used to approximate the integral needs to be specified. We specified a value for which adding more function evaluations only yielded negligible changes in the estimated value of the integral. (A similar approaches could be adopted to determine the most appropriate number of marginal bins B.) Through experimentation we determined that using 2^j function evaluations for each integral globally produced small enough approximation errors when integrating over j-dimensional bins to obtain comparable results to the classical model. As this implies the minimum number of evaluations necessary for a reasonable approximation of the integrals across all bins, we set $\tau_k = 2^j$.

4.1. Varying the Number of Bins, B

We specify (training and test) datasets each comprising $N = 25{,}000$ observations for which the response variable Y can take values in $\Omega = \{1, 2, 3\}$ ($K = 3$) conditional on $D = 5$ explanatory variables. The true vector of regression coefficients β_{true} has entries randomly drawn from a $U[-5, 5]$ distribution. The explanatory variables are drawn from D-dimensional normal and skew-normal distributions, with correlation matrices containing zero correlations (the identity matrix) or correlations drawn from $U[-0.75, 0.75]$. The elements of the skew-normal slant vector are drawn from $U[-7, 7]$. While the correlation parameter of the skew-normal distribution is not equivalent to the correlation matrix of the associated random variable, skew-normal data simulated using the identity matrix as the correlation parameter typically have low correlations. When aggregating the design matrix X into a histogram through (5), an equal number B of bins is set for each margin and each outcome k, that is, $B_k = (B, B, B, B, B, B)$ for all $k \in \Omega$.

We use both the full multinomial regression (M) model fit using (2) and the OvR (O) model fit using the full likelihood (3) as reference fits. We also fit the multinomial mixture (MM) model of histogram and underlying classical data (8), and the univariate approximate composite likelihood $L_{\mathrm{O}}^{(1)}$ (see Proposition 3). For the histogram-based, univariate approximate composite likelihood (10) we make the assumption that the covariates are either independent ($\alpha_{ii'} = 0$ and $\lambda_{ii'_1 i'_2} = 0$ in (9)) or that the variance-covariances of the covariates are to be estimated.

Figure 1 illustrates the mean prediction accuracies (over 1000 replicates) as a function of the number of bins B, obtained for each of the above models and estimation procedures. For the full dataset, using an approximate composite likelihood approach to fit the OvR model (dashed gray line) yields comparable
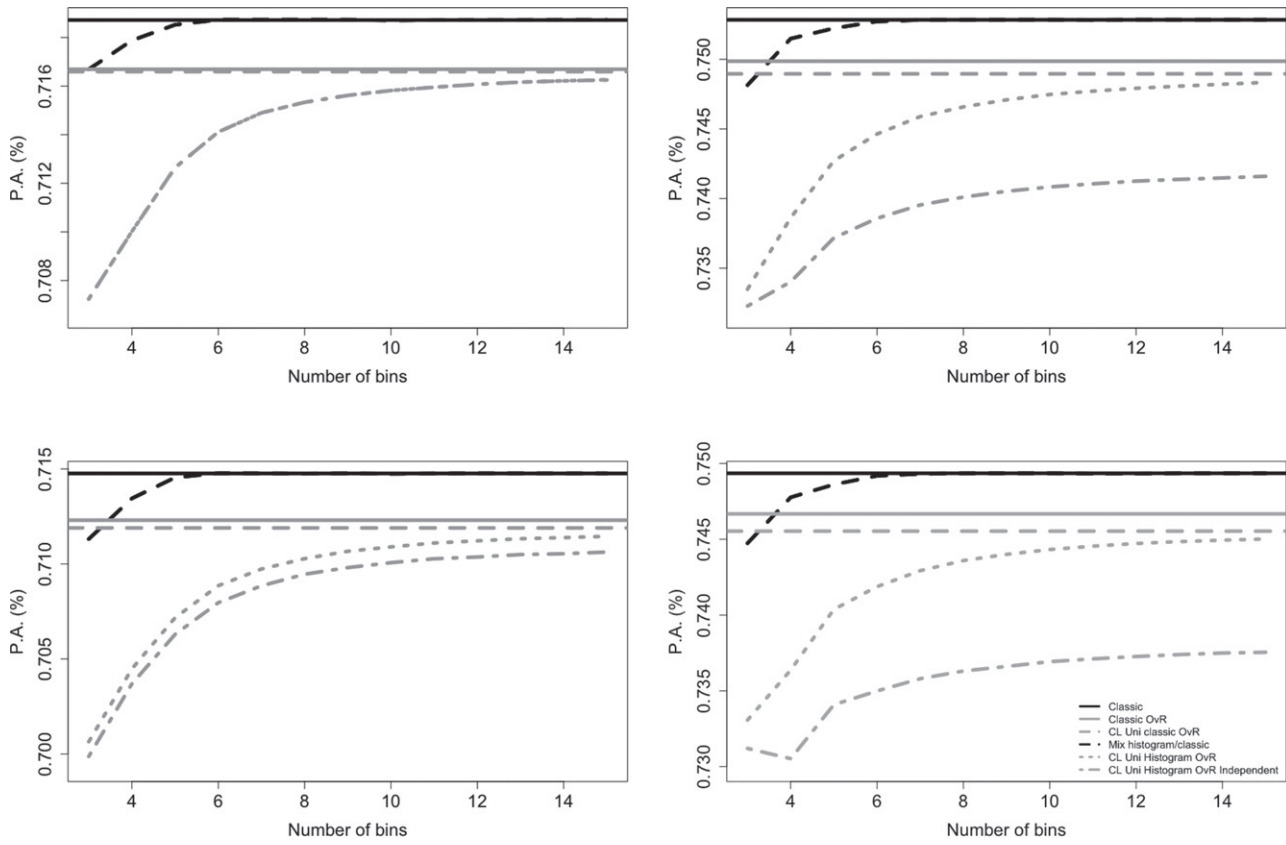
Figure 1. Average prediction accuracy (PA) computed over 1000 replications for the multinomial model using full likelihood (solid black), mixture multinomial model (dashed black), OvR model using full likelihood (solid gray) and approximate composite likelihood (dashed gray), histogram-based OvR model using approximate composite likelihood assuming independence of the covariates (dot dashed gray) and using additional covariate assumption (dotted gray). Top panels consider covariates simulated from the multivariate normal distribution and bottom panels using the skew normal distribution. Left panels assume the covariates have zero correlation parameter, and right panels use nonzero correlations.

prediction accuracies to the full likelihood approach (solid gray line), in particular when the covariates are independent of each other (left panels).

When the empirical variance-covariance matrix is used, the histogram-based OvR model fitted using the $L_{SO}^{(1)}$ approximate composite likelihood (gray dotted line) is able to obtain comparable prediction accuracies to the $L_O^{(1)}$ full-data approximate composite likelihood OvR model (solid gray line), for a reasonable ($\approx$ 10 marginal bins) level of data aggregation, and for any covariate distribution (rows). This clearly demonstrates that $L_{SO}^{(j)}(x, y; \beta) \rightarrow L_O^{(j)}(s; \beta)$ as discussed in Section 3.3. It also performs well compared to the analysis on the full data (solid black line) when the covariates are independent (left panels; note the small y-axis scale).

When the empirical variance-covariance matrix is unavailable and the correlations between explanatory variables are assumed to be independent ($\alpha_{ii'} = 0$ and $\lambda_{ii'_1 i'_2} = 0$), the histogram-based approximate composite likelihood model (dot-dashed gray line) still produces reasonable prediction accuracies. However, as should be expected, there is a clear loss in performance compared to when the covariances are known, for densities with high covariate correlations (right panels) and asymmetric distributions (bottom panels).

The prediction accuracies obtained from the multinomial model (dashed black line) have converged relatively quickly to its classical equivalent (solid black line) requiring only around

$B = 6$ bins per margin: To understand this, note that the histogram comprises $6^5 = 7776$ bins for 25,000 raw data points and 3 possible outcomes. As a result there will be numerous empty bins, and bins with low covariate vector counts. In addition, because $\tau_k = 2^5 = 32$ (see above), data are only aggregated if there are more than 32 observations in each bin. Consequently, for increasing B, the summarized dataset can quickly retain the most important information from the original dataset. The multinomial model gives higher overall prediction accuracies than the OvR model in each case, however, recall that the data were simulated according to the multinomial model, giving it a natural advantage. While the multinomial model is generally preferred over the OvR model here, in practice the OvR approach can outperform the multinomial model for some datasets, and can give almost as good results in many other cases (e.g., Eichelberger and Sheng 2013).

Figure 2 illustrates the mean computational efficiency of fitting each model. It highlights the computational superiority of the histogram-based OvR model when univariate components are included in the likelihood ($L_{SO}^{(1)}(s; \beta)$; gray lines) against the full data multinomial model (black lines). The computation time increases as the number of bins increases when $L_{SO}^{(1)}(s; \beta)$ is used, however, comparable predictions to the full multinomial model are achieved for $B \approx 10$ (see Figure 1) at a significantly cheaper computational cost. For the mixture multinomial model (black dashed lines), the computation time increases strongly
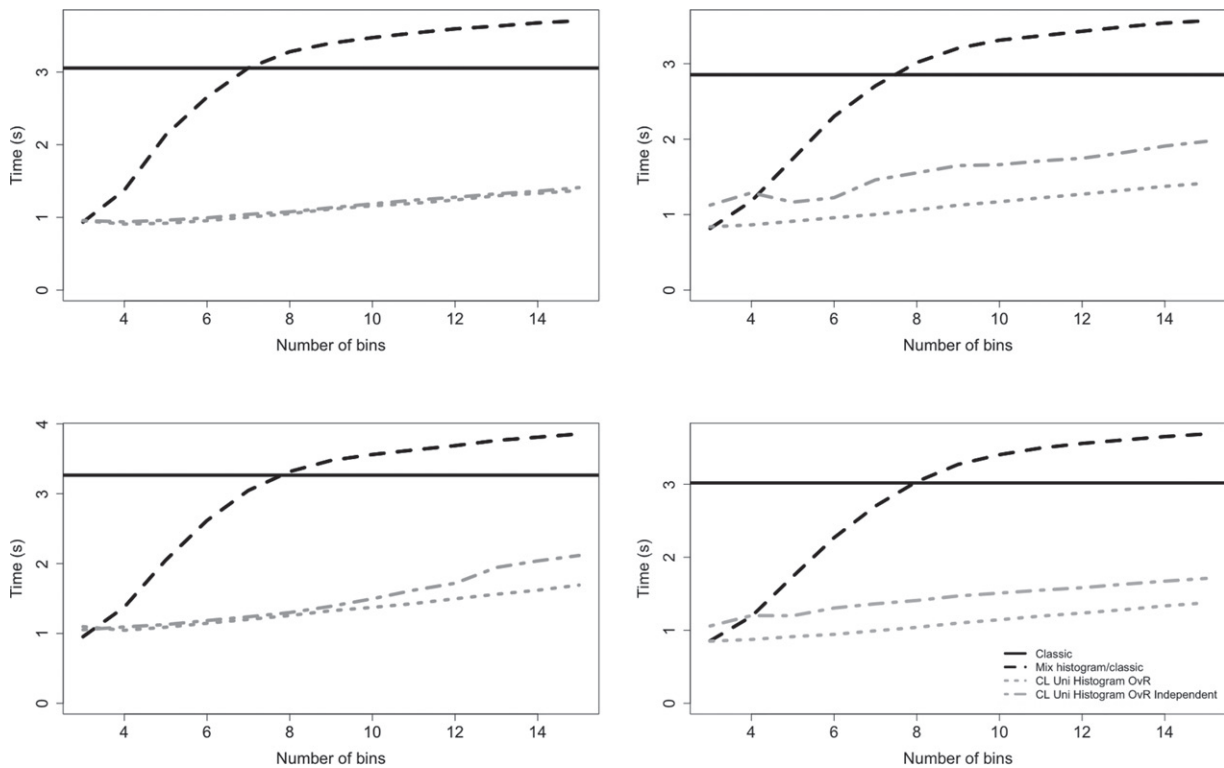
Figure 2. Average computation time (in CPU seconds) over 1000 replications for the multinomial model using full likelihood (solid black), mixture multinomial model (dashed black), histogram-based OvR model using approximate composite likelihood assuming independence of the covariates (dot dashed gray) and using additional covariate assumption (dotted gray). Top panels consider covariates simulated from the multivariate normal distribution and bottom panels using the skew normal distribution. Left panels use covariates from a zero correlation parameter, and right panels use nonzero correlations.

with increasing B. This phenomena is due to the intractability of the integrands in (6) requiring numerical integration, and our choice of τ_k. However, for this setup $B \approx 5$ is sufficient to provide comparable predictions to the multinomial model (Figure 1), and with lower computational overheads than the complete data case.

Even though the $L_{SO}^{(1)}(s; \boldsymbol{\beta})$ and $L_{MM}(s; \boldsymbol{\beta})$ likelihood approaches for histogram-valued data increase in computational intensity with increasing B, relative to the classical M model, we need to keep in mind that $N = 25{,}000$ observations are considered. Increasing N will directly increase the computational time of the classical approach (solid black line), while computational overheads will remain relatively unchanged for the histogram-based methods (where computation is proportional to the number of bins, not datapoints within bins). Consequently, there are clear computational benefits to employing a histogram likelihood approach when analyzing extremely large datasets. We explore this in Section 4.2.

4.2. Varying the Number of Underlying Observations, N and Comparison With Subsampling

Aggregating data into summaries and performing an analysis on these new "datapoints" seems a good strategy when the sample size is large. It is natural to compare this approach to other popular techniques for downsizing data volume, such as subsampling algorithms. We use the two-step subsampling scheme given by Wang, Zhu, and Ma (2018) for logistic regression modeling, which first uniformly draws a subsample of size $r_0 = 1000$ from

the dataset to produce a provisional MLE $\hat{\boldsymbol{\beta}}_0$, and uses this to produce an optimal weight for each datapoint. The second step then draws with replacement a subsample of size $r = 1000$ using the optimal weights, and then determines the final estimate of $\boldsymbol{\beta}$ using the total subsample of size $r_0 + r$.

In the following binary response ($K = 2$) experiment each element in the true vector of regression coefficients is generated from $U[-1, 1]$, and the number of observations, N, varies between 5000 and 100,000. The explanatory variables are drawn from eight-dimensional skew-normal distributions ($D = 8$), with either zero correlations (identity matrix) or correlations drawn uniformly on $[0, 0.75]$. The slant vector of the skew-normal distribution is drawn from $U[-10, 10]$. (Recall that the identity correlation matrix for the skew-normal distribution does not lead to independent covariates, but rather low correlations between the covariates.) Note that in the case of binary responses the multinomial and OvR models are identical. After aggregating the covariates into histograms with $B = 15$ bins for each margin, the histogram-based OvR model is fitted using $L_{SO}^{(1)}$ and $L_{SO}^{(2)}$ (i.e., univariate and bivariate marginal histograms), including covariate correlations. The $L_{SO}^{(1)}$ model is also fitted assuming independence between covariates.

Figure 3 shows that the mean prediction accuracies obtained by each method are increasing functions of the sample size N. Overall the symbolic based methods yield higher prediction accuracies than the subsampling approach when the covariate correlations are incorporated, and the more informative bivariate histogram setup $L_{SO}^{(2)}$ will outperform the univariate histogram-based $L_{SO}^{(1)}$. When the covariates exhibit low
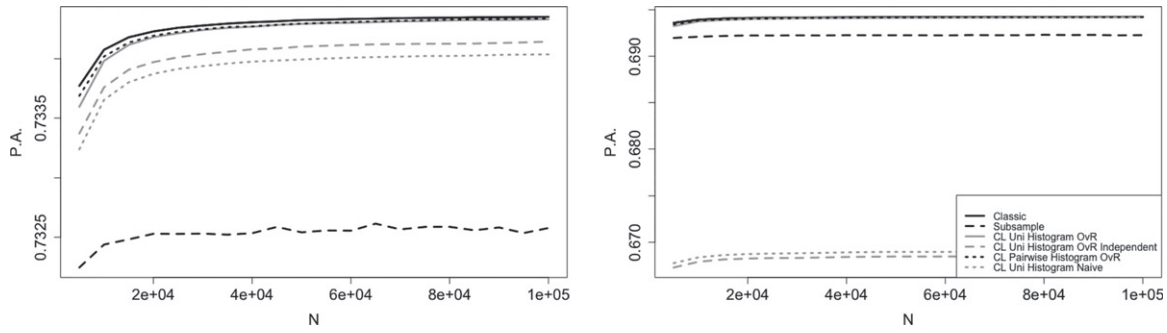
Figure 3. Mean prediction accuracies (P.A.) using the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{\mathrm{SO}}^{(1)}$ with independence assumption (dashed gray line), $L_{\mathrm{SO}}^{(1)}$ with correlations (solid gray line), $L_{\mathrm{SO}}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted gray line) as a function of the number of datapoints N. The covariates are generated from eight-dimensional skew-normal distributions, with zero correlations (left), or correlations drawn uniformly on $[0, 0.75]$ (right). The responses have two possible outcomes ($K = 2$). Results are based on 1000 replicate analyses.
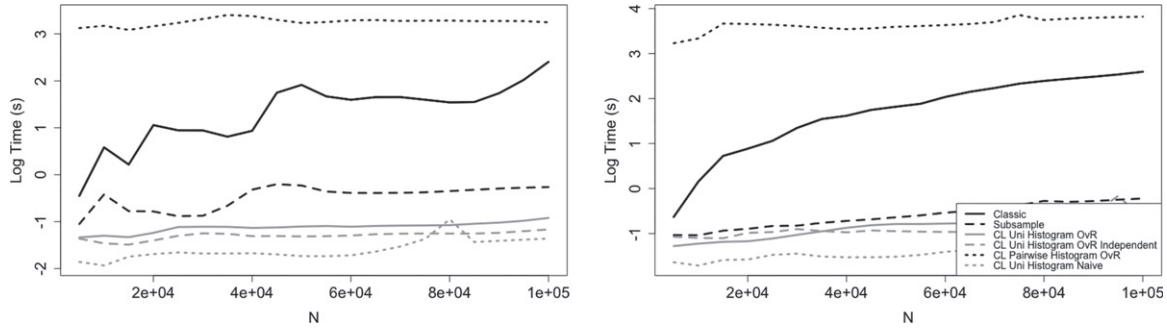

Figure 4. Mean total computation times (in CPU seconds) for the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{\mathrm{SO}}^{(1)}$ with independence assumption (dashed gray line), $L_{\mathrm{SO}}^{(1)}$ with correlations (solid gray line), $L_{\mathrm{SO}}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted gray line) as a function of the number of datapoints N. The covariates are generated from eight-dimensional skew-normal distributions, considering zero (left) and nonzero (right) correlation parameters. Results are based on 1000 replicate analyses.

correlations, the $L_{\mathrm{SO}}^{(1)}$ model provides significant improvements over the naive composite likelihood analysis, regardless of whether correlations are incorporated. When there are correlations between the covariates, the $L_{\mathrm{SO}}^{(1)}$ model only provides significant improvements over the naive composite likelihood analysis if the covariate correlations are included. The magnitude of the variations in the prediction accuracy confirms that the extra efforts to use $L_{\mathrm{SO}}^{(2)}$ are not justified in this case, and that univariate marginal histograms provide enough information and produce comparable results to a classical full data analysis.

Figure 4 supports these conclusions by providing the overall computation times (including aggregation and optimization) for each model in Figure 3. We observe that the mean computational time required for the univariate symbolic model is significantly lower than for the multivariate model with full data, with an increasing disparity as the sample size N increases, as the number of terms in the histogram-based OvR model depends on the histogram construction and not N (and so is approximately constant in these plots). In addition to its prediction superiority, $L_{\mathrm{SO}}^{(1)}$ also computationally outperforms the subsampling approach of Wang, Zhu, and Ma (2018). Note that Figure 4 indicates that $L_{\mathrm{SO}}^{(2)}$ (dotted black line) is computationally more demanding than using the full data (solid black line), making it superfluous in this setting. However, note that as computation for $L_{\mathrm{SO}}^{(2)}$ is constant in N, there will be some value N_0 such that if $N > N_0$ then the computational overheads for $L_{\mathrm{SO}}^{(2)}$ will be more efficient than for the full data analysis.

Figure 5 explores parameter estimator performance via the mean squared error (MMSE) of a model's MLE, $\hat{\theta}^{\mathrm{Model}}$, defined as $\mathrm{MMSE}(\hat{\theta}^{\mathrm{Model}}) = S^{-1} \sum_{s=1}^{S} \|\hat{\theta}_s^{\mathrm{Model}} - \theta_s^{\mathrm{true}}\|^2$, where $\| \cdot \|$ denotes the Euclidean norm, θ_s^{true} is the true parameter vector, and $S = 1000$ the number of replicate analyses. Figure 5 demonstrates that subsampling methods perform better than the histogram-based methods if a low MMSE is desired. Using $L_{\mathrm{SO}}^{(1)}$ instead of the naive composite likelihood analysis leads to a lower MMSE, with the results further improving if the covariate correlations are included.

Figure 6 explores parameter estimate accuracy further, displaying the mean estimates for a selection of the regression parameter over the 1000 replicate analyses. The estimates obtained from $L_{\mathrm{SO}}^{(1)}$ are much closer to those of the full model analysis than that of the naive composite likelihood analysis, with accuracy improving if covariate correlations are incorporated into the model. While the subsampling method provides more accurate parameter estimates, the better performance of the histogram-based models for predictions can potentially be explained by the fact that the histogram-based models incorporate the entire dataset, whereas a subsampling scheme can still potentially omit important observations. For predictions and the binary model, an observation x_n is assigned to class 1 if $\beta^{\top} x_n < 0$, and class 2 otherwise. Consequently, the same predictions are obtained from any parameter vector $m\beta$, $m > 0$ for any dataset. In this case, the histogram-based models are more accurately estimating the model parameters to
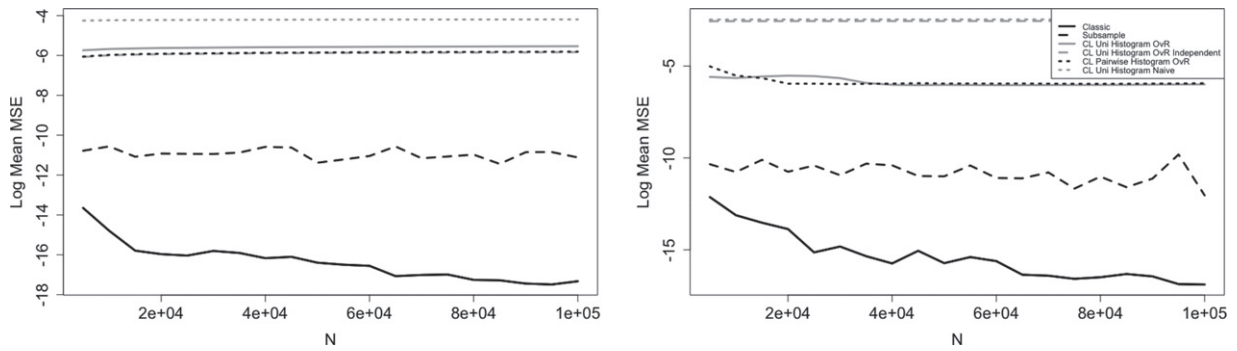
Figure 5. MMSE for the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{SO}^{(1)}$ with independence assumption (dashed gray line), $L_{SO}^{(1)}$ with correlations (solid gray line), $L_{SO}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted gray line) as a function of the number of datapoints N. The covariates are generated from eight-dimensional skew-normal distributions, considering zero (left) and nonzero (right) correlation parameters. Results are based on 1000 replicate analyses.
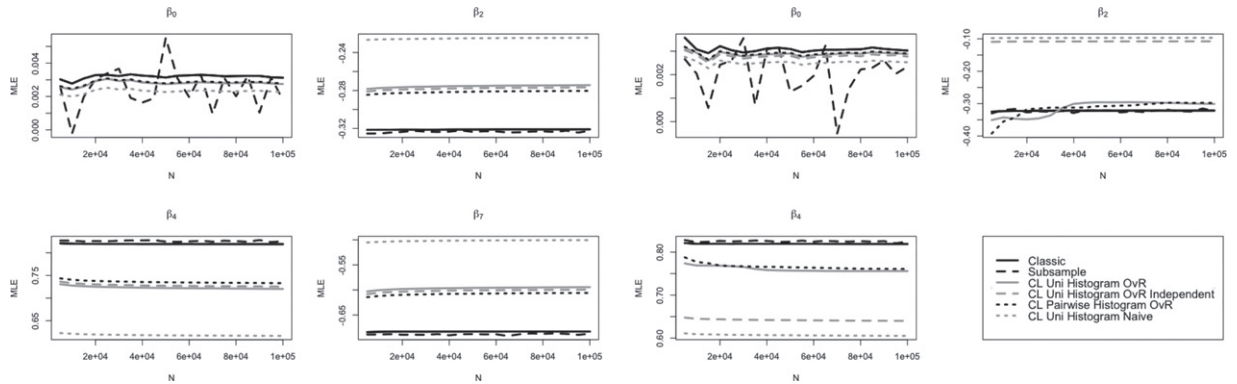

Figure 6. Mean MLEs using the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{SO}^{(1)}$ with independence assumption (dashed gray line), $L_{SO}^{(1)}$ with correlations (solid gray line), $L_{SO}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted gray line) as a function of the number of replicates N. The covariates are generated from eight-dimensional skew-normal distributions, considering zero (left two columns) and nonzero (right two columns) correlation parameters. Results are based on 1000 replicate analyses.

proportionality compared to the subsampling scheme, despite having a larger MSE.

In summary, this experiment suggests that if predictions are desired for logistic regression models, histogram-based solutions can be more accurate and computationally more efficient than subsampling-based methods, such as that in Wang, Zhu, and Ma (2018). The use of bivariate histograms to represent the covariate information improves the prediction of the response outcomes, but at an often impractical computational cost compared to univariate histograms. This simulation study suggests that marginally aggregating covariates into univariate histograms, in combination with knowledge of covariate correlations, provides the best tradeoff between accuracy and speed.

Methods based on, for example, Beranger, Lin, and Sisson (2018) can be used to estimate correlations from multidimensional histograms ($j \geq 2$). However, only variance estimation (not correlations) can be obtained from univariate marginal histograms, meaning that correlations need to be obtained prior to data aggregation. Given the small computational costs of estimating covariance matrices, even for large N datasets, this does not affect the efficiency of the proposed method. However, if an estimate of the observed covariance matrix is unavailable, this would force the analyst to assume independence between covariates, leading to a decrease in performance (if large correlations are present).
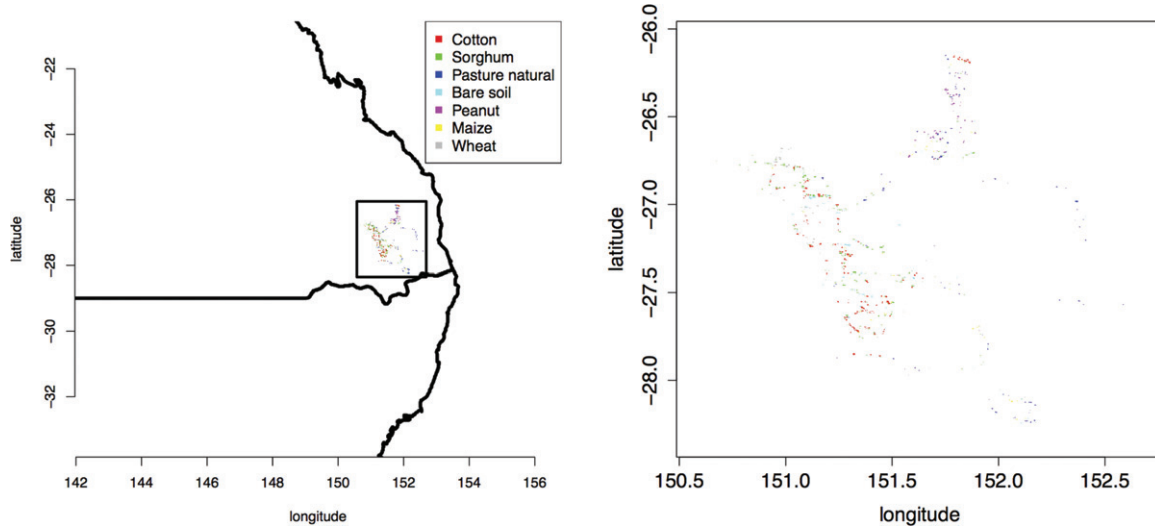
5. Real Data Analyses

We illustrate the applicability of our proposed methodology to two real data problems. We first consider a logistic regression problem where the goal is to distinguish between a process where new supersymmetric particles are produced and a background process. Second, we tackle a multinomial regression problem which consists of predicting crop types based on satellite-based pixel observations.

5.1. Supersymmetric Benchmark Dataset

The Supersymmetry dataset (SUSY) is available from the Machine Learning Repository (Dua and Graff 2017) and comprises 5 million Monte Carlo observations generated by Baldi, Sadowski, and Whiteson (2014). The binary response variable ($K = 2$) discriminates a signal process which produces supersymmetric particles from a background process which does not. There are 18 features ($D = 18$); the first 8 are low-level features representing the kinematic properties measured by the particle detectors, while the remaining 10 are high-level features derived as function of the previous 8 by physicists to help discriminate between the two outcomes. This dataset was analyzed by Wang, Zhu, and Ma (2018) to test their optimal subsampling scheme for logistic regression. Following Wang, Zhu, and Ma (2018), we consider a training dataset of 4,500,000

Table 1. Percentage prediction accuracy with computing time (in sec) for the Supersymmetry dataset, using histograms with B bins per margins, compared to the subsampling approach of Wang, Zhu, and Ma (2018).

Likelihood	Bins						
	6	8	10	12	15	20	25
$L_{OO}^{(1)}$	75.9	77.8	77.9	78.0	78.0	78.1	78.1
	(19.2)	(15.3)	(18.9)	(19.4)	(20.5)	(20.1)	(21.6)
$L_{SO}^{(1)}$	76.2	77.4	78.0	78.2	78.2	78.2	78.3
	(10.3)	(10.5)	(12.0)	(12.9)	(14.1)	(15.2)	(15.8)
Subsampling Wang, Zhu, and Ma (2018)							78.2
							(70.2)

Figure 7. The crop type dataset with different colors for each crop. Left: Location of the study area in the state of Queensland on the east coast of Australia. Right: farm location and crop type detail.

randomly chosen observations, and a test dataset with the remaining 500,000 observations.

Following the conclusions from Section 4, we fit the histogram-based OvR model using univariate marginal histogram aggregates ($L_{SO}^{(1)}$). While until now the focus has been on histograms with random counts (fixed bins), here we fit $L_{OO}^{(1)}$ to explore the performance of using random bin (fixed counts) histograms. For an integer B, we construct histograms for each covariate by partitioning the data into B bins with roughly equal counts, while avoiding the highest and lowest 1% of covariate values to reduce bin sensitivity to outliers. For example, for $B = 4$ we would use the 0.01, 0.25, 0.5, and 0.99 empirical quantiles to construct the histogram for each covariate, and create additional bins on the endpoints of the range of the data to accommodate the outliers.

The likelihood functions are optimized using Lasso regularization with 10-fold cross-validation and, for $L_{OO}^{(1)}$, for a range of values of B. Prediction accuracies obtained on the test dataset and the optimization times (in sec) on the training dataset are reported in Table 1. For the histogram-based models, there is an increase in prediction accuracy as the number of bins increases, which is as expected since these are more informative summaries. The improvement in prediction accuracy slows down at around the $B = 12$ bin mark, whereas the computation time naturally increases with the number of bins. When implementing the subsampling methodology of Wang, Zhu, and Ma (2018) we obtain a prediction accuracy of 78.2% with a computation time of 70.2 sec, that is, about 3–4 times more computation

than for the histogram-based models with $B = 25$. That is, the histogram-based methods offer as good prediction accuracy with much smaller computational overheads compared to state-of-the-art subsampling approaches.

5.2. Crop Type Dataset

We examine a crop type dataset (QUT 2016) which consists of 247,210 observations, each representing a $25 \times 25\,m^2$ pixel located over farmland across the state of Queensland, on the east coast of Australia (Figure 7). For each pixel the ground-truth crop type is available (observed at one of three possible times) as well as numerous vegetation indices, based on reflectance data taken from a LANDSAT 7 satellite. The aim of this analysis is to predict the crop type based on the vegetation indices. After selecting the most meaningful covariates by iteratively removing variables with correlations greater than 0.85, we retained $D = 7$ variables corresponding to various color reflectances measured by the satellite and functions of these indicators.

As poor prediction accuracy of classes with low numbers of observations is a well-known issue, we only retain crop types that are observed more than 10,000 times, reducing the dataset to 234,485 observations. The set of possible outcomes of our multinomial response variable $Y =$ "Crop type" is then $\Omega = $ {Bare Soil, Cotton, Maize, Pasture Natural, Peanut, Sorghum, Wheat} and thus $K = 7$. The resulting dataset is identical to the one used in a previous analysis in QUT (2016) which used the standard multinomial model $L_M(x, y; \beta)$.

As the approximate composite likelihood relies on the assumption of a linear relationship between the predictor variables, we use the R package `bestNormalize` to select the best transformation to achieve approximate predictor Gaussianity, according to the Pearson P-test statistic. The dataset is randomly partitioned into a training dataset of size 200,000 used for parameter estimation and a test dataset with the remaining 34,485 observations to evaluate the prediction accuracy. We perform constrained likelihood optimization with a Lasso regularization, and use 10-fold cross-validation to determine the best regularization parameter.

Table 2 presents the prediction accuracies for the $L_{SO}^{(1)}$ model for each crop type and the overall prediction accuracy when the covariate information is collapsed into univariate marginal histograms with $B = 6, 8, 10, 12, 15$, and 20 bins. The last column of Table 2 provides a comparison with the full data multinomial likelihood ($L_M(x, y; \beta)$) using the same Lasso regularization, as implemented in the original analysis by QUT (2016). The overall and crop-specific prediction accuracies have achieved good predictive performance compared to the full data multinomial model analysis using only $B \approx 10$–12 bins. Two particular crops produce notable results. The prediction accuracies for Maize are around 7.3% for the histogram-based analysis compared to the $\sim$14.2% accuracy of the full-data analysis. While both of these are low due to this crop being the least well represented of all crops in the study ($<6\%$ of all observations), and perhaps lowly informative vegetation indices for this crop, the $\sim$50% predictive underperformance for the histogram-based analysis suggests that categories with less data in a model are more sensitive to the degree of binning than those categories with larger representation in the dataset (although this is less apparent for Wheat).

In the case of Pasture Natural, the histogram-based prediction accuracies are even higher (at $\sim$82%) than for the full data analysis (77.6%). While this is not immediately understandable intuitively, in that by constructing histograms information in the dataset is certainly being lost and so performance should perhaps always be worse, the difference is only $\sim$4%, and moreover the likelihoods are not directly comparable in the sense that the limit of the approximate composite likelihood $L_{SO}^{(j)}(s; \beta)$ as $B \to \infty$ is not the full data multinomial model $L_M(x, y; \beta)$ as used in QUT (2016). So here the discrepancy is that the two likelihoods simply have different performances for these data. This argument notwithstanding, proponents of SDA sometimes ascribe to the idea that inference using the "shape" of the data (i.e., the histogram summary) may sometimes be more useful in an analysis than the full underlying dataset (Edwin Diday, personal communication).

Overall, while the histogram-based analysis gives comparable prediction accuracies to the full data analysis, the real gains are in the computational overheads required for each model. The full multinomial analysis takes considerably longer (more than $13\times$ the $B = 12$ analysis) to implement than the histogram based analysis. Finally, note that while the computational savings here are substantial, this dataset only contains $N = 234,485$ observations. For larger datasets the computational overheads will skyrocket for the standard multinomial model analysis (where computation is proportional to N), and yet will remain roughly constant for the histogram-based approach (where computation is proportional to the number of bins).

6. Discussion

In this article, we have developed a novel approach for classifying binary and multinomial random variables that alleviates the computational bottleneck that arises with very large datasets. The strategy relies on concepts from the field of SDA (Beranger, Lin, and Sisson 2018), aggregating the covariate data into histogram-valued random variables which have lower computational overheads to analyze and store, albeit with some loss of information. When computation for any histogram bin is larger than that for the standard likelihood contribution of the datapoints within that bin, the standard likelihood contribution for these datapoints can be used instead. However, because high-dimensional histograms are not efficient distributional summaries, we additionally introduced an approximate composite likelihood methodology, which quantitatively builds on the qualitative results of Cramer (2007). The individual components of the approximate composite likelihood are constructed from marginal histograms derived from the full D-dimensional histogram. This concept of approximate composite likelihoods for logistic regression does not solely apply to aggregated data and can be used in more general settings.

We have demonstrated through simulation studies and real data analyses that these histogram-based strategies can produce fitted models that have comparable prediction accuracies to the standard full data analysis, but at a much lower computational cost, even compared to state-of-the-art computational

Table 2. Crop specific and overall prediction accuracies (%) using univariate marginal histograms with B bins.

Crop type	N_k	Bins						$L_M(x, y; \beta)$
		6	8	10	12	15	20	
Cotton	72,450	90.9	93.6	92.4	91.4	91.3	91.3	92.6
Sorghum	66,751	78.8	76.2	78.6	79.0	78.4	77.7	80.3
Pasture Natural	27,479	76.1	81.2	81.4	81.3	81.8	82.2	77.6
Bare Soil	26,173	89.5	90.2	90.4	90.0	89.8	90.1	91.0
Peanut	17,868	80.2	79.7	79.9	79.7	80.2	79.8	82.9
Maize	12,986	0.6	1.3	5.0	11.9	5.5	7.3	14.2
Wheat	10,778	6.7	10.4	9.9	8.7	9.9	10.4	10.3
Overall	234,485	75.8	76.8	77.3	77.2	77.3	77.4	78.1
		(49)	(70)	(73)	(94)	(97)	(117)	(1263)

NOTE: The likelihood optimization times (in sec) are reported in the last row. The full model is the standard multinomial likelihood $L_M(x, y; \beta)$ (2) with Lasso regularization, as implemented by QUT (2016).

techniques for logistic regression such as subsampling (Wang, Zhu, and Ma 2018). On the down side, the resulting parameter estimates are biased, though not as much as for naive composite likelihood-based approaches.

One aspect of implementing histogram-based inference that we have not explored is identifying a principled method of constructing histograms for subsequent analysis. If the number of bins is too low, then important information in the data will be lost and model predictions may be poor (see, e.g., Figure 1). In contrast, as the number of bins becomes large, the inferential accuracy can approach the level of the full data analysis (within the context of the inferential model being used). However, the price of more accurate inference is an increase in the computational costs. A simple diagnostic to choose the number of bins approach is apparent from the real data analyses in Section 5. Here the number of bins could be increased until some quantity of interest—such as the model prediction accuracy, or the MLE, or its standard error—does not change drastically when increasing the number of bins further. Of course, this procedure is very ad-hoc, and so developing a more theoretically justified and principled approach would be highly beneficial. Such an approach could therefore consider balancing computational complexity and inferential accuracy, or alternatively by minimizing a loss function constructed over some useful criterion.

Similarly, the assumption of equal bin widths considered here could be relaxed by developing selection criteria for bin locations. As suggested by the Associate Editor, regression trees over covariate space offer one possibility to efficiently represent the data by identifying easily integrable (axis-parallel) regions over covariate space. One advantage of using equally spaced bins, however, is that bin locations are computationally cheap to obtain, requiring just the minimum and maximum values for each variable. Conversely, any optimal bin-choice method will be more computationally expensive to implement, and so any gains in accuracy must be balanced with the increase in computation. This is clearly an important component of the current methodology, and is the focus of current research.

Appendices

A.1. Proof of Proposition 1

We utilize the arguments presented in Beranger, Lin, and Sisson (2018) to derive of Proposition 1. Note that if the underlying microdata $\mathbf{X} = (X_1, \ldots, X_N)$ are iid then the K subsets $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$ are similarly iid. For histogram-valued data,

$$f(\mathbf{s}|\mathbf{x}, \mathbf{y}, \vartheta) = \prod_{k=1}^{K} f(s_k|\mathbf{x}^{(k)}, \vartheta),$$

where

$$f(s_k|\mathbf{x}^{(k)}, \vartheta) = \prod_{\mathbf{b}_k \in \mathcal{B}_k} \mathbb{1}\left\{\sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \mathbf{\Upsilon}_{\mathbf{b}_k}\} = s_{\mathbf{b}_k}\right\}.$$

For the symbolic, histogram-based model, we therefore obtain

$$L_{\text{SM}}(\mathbf{s}; \beta) = \int_{\mathcal{D}_{\mathbf{X}}} L_M(\mathbf{x}, \mathbf{y}; \beta) f(\mathbf{s}|\mathbf{x}, \mathbf{y}, \vartheta) d\mathbf{x}$$

$$= \prod_{k \in \Omega} \int_{\mathcal{D}_{\mathbf{X}^{(k)}}} L_M(\mathbf{x}^{(k)}, \mathbf{y}; \beta) f(s_k|\mathbf{x}^{(k)}, \vartheta) d\mathbf{x}^{(k)}$$

$$= \prod_{k \in \Omega} \prod_{n=1}^{N_k} \prod_{\mathbf{b}_k \in \mathcal{B}_k} \left(\int_{\mathcal{D}_{x_n^{(k)}}} L_M(x_n^{(k)}, y_n; \beta) dx_n^{(k)}\right)^{\mathbb{1}\{x_n^{(k)} \in \mathbf{\Upsilon}_{\mathbf{b}_k}\}}$$

$$\propto \prod_{k \in \Omega} \prod_{\mathbf{b}_k \in \mathcal{B}_k} \left(\int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} P_M(y = k|X = x) dx\right)^{s_{\mathbf{b}_k}}.$$

Similarly, for the histogram-based OvR model, we obtain

$$L_{\text{SO}}(\mathbf{s}; \beta) = \int_{\mathcal{D}_{\mathbf{X}}} L_O(\mathbf{x}, \mathbf{y}; \beta) f(\mathbf{s}|\mathbf{x}, \mathbf{y}, \vartheta) d\mathbf{x}$$

$$= \prod_{k \in \Omega} \int_{\mathcal{D}_{\mathbf{X}^{(k)}}} L_O(\mathbf{x}^{(k)}, \mathbf{y}; \beta) f(s_k|\mathbf{x}^{(k)}, \vartheta) d\mathbf{x}^{(k)}$$

$$= \prod_{k \in \Omega} \prod_{n=1}^{N_k} \prod_{\mathbf{b}_k \in \mathcal{B}_k} \left(\int_{\mathcal{D}_{x_n^{(k)}}} L_O(x_n^{(k)}, y_n; \beta) dx_n^{(k)}\right)^{\mathbb{1}\{x_n^{(k)} \in \mathbf{\Upsilon}_{\mathbf{b}_k}\}}$$

$$\propto \prod_{k \in \Omega} \prod_{\mathbf{b}_k \in \mathcal{B}_k} \left(\int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} P_O(Y = k|X = x) dx\right.$$

$$\left. \prod_{k' \in \Omega \setminus \{k\}} \int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} P_O(Y \neq k'|X = x) dx\right)^{s_{\mathbf{b}_k}}.$$

A.2. Proof of Proposition 2

We now show that if there is neither complete nor quasi-complete separation of the set of histograms $\mathbf{s}$, then $L_{\text{SO}}(\mathbf{s}; \beta)$ and $L_{\text{SM}}(\mathbf{s}; \beta)$ have unique global maxima. The following arguments are analogous to those proposed by Albert and Anderson (1984). Suppose that there is complete separation exhibited by the histogram dataset for the kth class, according to Definition 4.

As a result, the complete separation property holds for all vectors $\beta_k = a_k b_k$ for $a_k > 0$. We now examine the behaviors of the integrals of the $P_O(Y = k|X)$ and $P_O(Y \neq k|X)$ terms in the likelihood functions (6) and (7). Using the mean value theorem, and given $a_k b_k^\top x_{\mathbf{b}_j}^* > 0$ if $j = k$ and $a_k b_k^\top x_{\mathbf{b}_j}^* < 0$ if $j \neq k$ for all nonempty bins and $a_k > 0$, we obtain

$$\lim_{a_k \to \infty} \int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} P_O(Y = k|x) dx = \lim_{a_k \to \infty} \int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} \frac{e^{a_k b_k^\top x}}{1 + e^{a_k b_k^\top x}} dx$$

$$\propto \lim_{a_k \to \infty} \frac{e^{a_k b_k^\top x_{\mathbf{b}_k}^*}}{1 + e^{a_k b_k^\top x_{\mathbf{b}_k}^*}} = 1,$$

$$\lim_{a_k \to \infty} \int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} P_O(Y \neq k|x) dx = \lim_{a_k \to \infty} \int_{\mathbf{\Upsilon}_{\mathbf{b}_k}} \frac{1}{1 + e^{a_k b_k^\top x}} dx$$

$$\propto \lim_{a_k \to \infty} \frac{1}{1 + e^{a_k b_k^\top x_{\mathbf{b}_k}^*}} = 1,$$

where $x_{\mathbf{b}_k}^*$ is some point inside $\mathbf{\Upsilon}_{\mathbf{b}_k}$. Each integral therefore approaches a constant for all bins as a_k increases, meaning the maximum value of each likelihood function is attained at the boundary of the parameter space, that is, $\hat{\beta}_k = \infty$.

Now suppose there is quasi-complete separation exhibited by the histogram dataset, according to Definition 4. Continuing with the

previous notation, denote A_k^{D+1} as the set of all vectors b_k that satisfy the complete separation condition, meaning that A_k^{D+1} is a convex set. Denote the parameter vector $\alpha_k(a_k) = c_k + a_k b_k$, where $a_k > 0$ and $a_k \in A^{D+1}$. Consequently,

$$P_O(Y = k|x) = \frac{e^{\alpha_k(a)^\top x}}{1 + e^{\alpha_k(a_k)^\top x}}.$$

The log-likelihood for the component of the OvR model (7) estimating the parameters for the kth class, β_k, can therefore be expressed as

$$\log L_{SO}^k(\mathbf{s}; \beta_k) = \sum_{\boldsymbol{b}_k \in \mathcal{B}_k} s_{\boldsymbol{b}_k} \log \int_{\Upsilon_{\boldsymbol{b}_k}} P_O(Y = k|x) dx$$
$$+ \sum_{k' \in \Omega \setminus \{k\}} \sum_{\boldsymbol{b}_{k'} \in \mathcal{B}'_k} s_{\boldsymbol{b}_{k'}} \log \int_{\Upsilon_{\boldsymbol{b}_{k'}}} P_O(Y \neq k|x) dx$$
$$= \sum_{\boldsymbol{b}_k \in \mathcal{B}_k} s_{\boldsymbol{b}_k} \log \int_{\Upsilon_{\boldsymbol{b}_k}} \frac{e^{c_k^\top x + a_k b_k^\top x}}{1 + e^{c_k^\top x + a_k b_k^\top x}} dx$$
$$+ \sum_{k' \in \Omega \setminus \{k\}} \sum_{\boldsymbol{b}_{k'} \in \mathcal{B}'_k} s_{\boldsymbol{b}_{k'}} \log \int_{\Upsilon_{\boldsymbol{b}_{k'}}} \frac{1}{1 + e^{c_k^\top x + a_k b_k^\top x}} dx.$$

Given that $b_k^\top x > 0$ for all $x \in \Upsilon_{\boldsymbol{b}_k}$ such that $s_{\boldsymbol{b}_k} > 0$, the function $\frac{e^{c_k^\top x + a_k b_k^\top x}}{1 + e^{c_k^\top x + a_k b_k^\top x}}$ is monotonically increasing with a_k for all $x \in \Upsilon_{\boldsymbol{b}_k}$ such that $s_{\boldsymbol{b}_k} > 0$. Consequently,

$$\sum_{\boldsymbol{b}_k \in \mathcal{B}_k} s_{\boldsymbol{b}_k} \log \int_{\Upsilon_{\boldsymbol{b}_k}} \frac{e^{c_k^\top x + a_k b_k^\top x}}{1 + e^{c_k^\top x + a_k b_k^\top x}} dx$$

is monotonically increasing with increasing a_k. Similarly, given that $b_k^\top x < 0$ for all $x \in \Upsilon_{\boldsymbol{b}_{k'}}$ such that $s_{\boldsymbol{b}_{k'}} > 0$ and $k' \neq k$

$$\sum_{k' \in \Omega \setminus \{k\}} \sum_{\boldsymbol{b}_{k'} \in \mathcal{B}'_k} s_{\boldsymbol{b}_{k'}} \log \int_{\Upsilon_{\boldsymbol{b}_{k'}}} \frac{1}{1 + e^{c_k^\top x + a_k b_k^\top x}} dx$$

is monotonically increasing with increasing a_k. Therefore the log-likelihood function $\log L_{SO}^k(\mathbf{s}; \beta)$ is monotonically increasing with a_k, and the maximum value is attained at the boundary of the parameter domain, that is, $\hat{\beta}_k = \infty$.

The above arguments show that if there is complete or quasi-complete separation for any class $k \in \Omega$, then there is no unique MLE for the symbolic OvR model. Using similar arguments, it can be shown that the maximum value for the symbolic multinomial likelihood $\log L_{SM}(\mathbf{s}; \boldsymbol{\beta})$ is attained at $\hat{\boldsymbol{\beta}} = (\infty, \ldots, \infty)^\top$ if there is either complete or quasi-complete separation in the data.

By the mean value theorem,

$$L_{SM}(\mathbf{s}; \boldsymbol{\beta}) \propto \prod_{k \in \Omega} \prod_{\boldsymbol{b}_k \in \mathcal{B}_k} \left(\int_{\Upsilon_{\boldsymbol{b}_k}} P_M(y = k|X = x) dx \right)^{s_{\boldsymbol{b}_k}}$$
$$\propto \prod_{k \in \Omega} \prod_{\boldsymbol{b}_k \in \mathcal{B}_k} P_M(y = k|X = x^*_{\boldsymbol{b}_k})^{s_{\boldsymbol{b}_k}},$$

where $x^*_{\boldsymbol{b}_k} \in \Upsilon_{\boldsymbol{b}_k}$ is some point located inside the $\boldsymbol{b}_k$th bin. The symbolic multinomial likelihood $L_{SM}(\mathbf{s}; \boldsymbol{\beta})$ is therefore proportional to the classical likelihood for some dataset $\boldsymbol{x}^*$, consisting of $\sum_{k \in \Omega} \sum_{\boldsymbol{b}_k \in \mathcal{B}_k} \mathbb{1}\{s_{\boldsymbol{b}_k} > 0\}$ distinct values $x^*_{\boldsymbol{b}_k}$, each appearing $s_{\boldsymbol{b}_k}$ times, $\boldsymbol{b}_k \in \mathcal{B}_k$, $k \in \Omega$. Albert and Anderson (1984) proved that $L_M(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\beta})$ is a closed convex function. Therefore, $L_M(\boldsymbol{x}^*, \boldsymbol{y}; \boldsymbol{\beta})$ and subsequently $L_{SM}(\mathbf{s}; \boldsymbol{\beta})$ are closed convex functions. Similar arguments can be used to show the closed convex nature of $L_{SO}(\mathbf{s}; \boldsymbol{\beta})$. As a result, if the histogram-valued data do not exhibit complete separation or

quasi-complete separation (Definition 4), then there is a unique global maximum of $L_{SM}(\mathbf{s}; \boldsymbol{\beta})$. Similarly, if the histogram-valued data do not exhibit complete separation or quasi-complete separation for any class $k \in \Omega$ (Definition 3), then there is a unique global maximum of $L_{SO}(\mathbf{s}; \boldsymbol{\beta})$.

A.3. Proof of Proposition 3

We first use the arguments in Cramer (2007) and Wooldridge (2002) to derive $L_O^{(j)}$ and $L_{SO}^{(j)}$ using the latent variable formulation of the logistic regression model (equivalent to the log odds formulation in Section 2). Consider a binary ($K = 2$) logistic regression model, such that the OvR and multinomial model are equivalent. The latent variable formulation for the logistic regression model based on the $\boldsymbol{i}$th subset of variables can be written as

$$Y_n^* = \beta^{\boldsymbol{i}\top} X_n^{\boldsymbol{i}} + e_n^{\boldsymbol{i}}, \qquad n = 1, \ldots, N, \tag{.12}$$

where Y_n^* is an unseen latent variable and $e_n^{\boldsymbol{i}}$ is an error term following a logistic distribution. Classification is then achieved by setting $Y_n = 1$ if $Y_n^* < 0$ and $Y_n = 2$ otherwise. In the full model

$$Y_n^* = \beta^\top X_n + u_n,$$

where u_n follows a logistic distribution with zero mean and unit variance. In the smaller model (.12) indexed by $\boldsymbol{i}$, the omitted terms are absorbed into the error term. That is

$$e_n^{\boldsymbol{i}} = u_n + \sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta^{i'} X_{ni'}.$$

Suppose that there are correlations between included and omitted variables for the model based on subset $\boldsymbol{i}$, and we can express each omitted variable as a linear function of the included variables, that is, $X^{i'} = \alpha_{\boldsymbol{i}i'}^\top X^{\boldsymbol{i}} + \epsilon_{\boldsymbol{i}i'}$, as described in Section 3.3. W.l.o.g. we can assume the ith variable has zero mean and variance given by σ_i^2, and that the covariance between variables i and i' is given by $\sigma_{ii'}$. The error term $e_n^{\boldsymbol{i}}$ can therefore be expressed as

$$e_n^{\boldsymbol{i}} = u_n + \sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta^{i'} \left(\alpha_{\boldsymbol{i}i'}^\top X_n^{\boldsymbol{i}} + \epsilon_{ni i'} \right).$$

We now rewrite (.12) by absorbing the terms in $e_n^{\boldsymbol{i}}$ that are dependent on the included variables into the model. That is

$$Y_n^* = \sum_{i'_1 \in \boldsymbol{i}} \left(\beta^{i'_1} + \sum_{i'_2 \notin \boldsymbol{i}} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1} + \tilde{e}_n^{\boldsymbol{i}},$$

where

$$\tilde{e}_n^{\boldsymbol{i}} = u_n + \sum_{i' \notin \boldsymbol{i}} \beta^{i'} \epsilon_{ni i'}.$$

Pingel (2014) showed that if X is distributed according to a logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$, then a standard normal distribution also fits the distribution of X reasonably well. The similarities between the standard logistic density and a rescaled normal density have also been investigated by Jeffress (1973), Bowling and Khasawneh (2009), and Pingel (2014), whereby different values for the rescaling factor C are proposed based on the criteria used to match the logistic and normal distributions. In practice, any of these values can be used here, but we proceed with $\frac{\pi^2}{3}$ in the simulations and real data analyses in Sections 4 and 5, respectively. As a result, $\sum_{i' \notin \boldsymbol{i}} \beta^{i'} \epsilon_{ni i'}$ is approximately normally distributed, and the error term $\tilde{e}_n^{\boldsymbol{i}}$ is approximately

logistically distributed with mean zero (i.e., $\mathbb{E}(\tilde{e}_n^i) = 0$) and variance given by

$$\operatorname{var}(\tilde{e}_n^i) = 1 + \frac{\sum_{i'_1 \in \mathcal{I}_1^{-i}} \left(\beta^{i'_1 2} \lambda_{ii'_1}^2 + 2 \sum_{i'_2 \in \mathcal{I}_1^{-i}, i'_2 \neq i'_1} \beta^{i'_1} \beta^{i'_2} \lambda_{ii'_1 i'_2} \right)}{\pi^2/3}.$$

In the full model,

$$P(Y_n = 2|X_n) = P(Y_n^* > 0|X_n) = P(u_n < \beta^\top X_n)$$
$$= \frac{\exp\{\beta^\top X_n\}}{1 + \exp\{\beta^\top X_n\}},$$

and we obtain MLEs $\hat{\beta}$ for β by maximizing over the sum of this quantity over all observations. Note that $P(u_n < \beta^\top X_n) = P_O(Y_n = 2|X_n)$, resulting in the equivalency between the latent and log odds formulations of the logistic regression model. For the omitted variable model,

$$P(Y_n = 2|X_n^i) = P(Y_n^{i*} > 0|X_n^i)$$
$$= P\left(\tilde{e}_n^i < \sum_{i'_1 \in i} \left(\beta^{i'_1} + \sum_{i'_2 \notin i} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1} \right).$$

Rescaling $\tilde{e}_n^i$ by its standard deviation gives us a random variable with approximately the same distribution as u_n, that is, $\tilde{u}_n = \frac{\tilde{e}_n^i}{\sqrt{\operatorname{var}(\tilde{e}_n^i)}}$ will approximately follow a logistic distribution with zero mean and unit variance. As a consequence,

$$P(Y_n = 2|X_n^i) = P\left(\frac{\tilde{e}_n^i}{\sqrt{\operatorname{var}(\tilde{e}_n^i)}} < \frac{\sum_{i'_1 \in i} \left(\beta^{i'_1} + \sum_{i'_2 \notin i} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1}}{\sqrt{\operatorname{var}(\tilde{e}_n^i)}} \right)$$
$$\approx P\left(\tilde{u}_n < \frac{\sum_{i'_1 \in i} \left(\beta^{i'_1} + \sum_{i'_2 \notin i} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1}}{\sqrt{\operatorname{var}(\tilde{e}_n^i)}} \right)$$
$$= \frac{\exp\{\tilde{\beta}^{i\top} X_n^i\}}{1 + \exp\{\tilde{\beta}^{i\top} X_n^i\}},$$

where $\tilde{\beta}^i = \frac{\beta^i + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-i}} \beta^{i'} \alpha_{ii'} \right)^\top \right]^\top}{\sqrt{\operatorname{var}(\tilde{e}_n^i)}} \in \mathbb{R}^{(j+1)}$. Therefore, we see that the regression parameters $\tilde{\beta}^i$ for the OvR model fit to the data indexed by i can be expressed as functions of the regression parameters β for the complete D-dimensional OvR model. The value for $\tilde{\beta}^i$ that maximizes the binary logistic likelihood for the ith subset is therefore a rescaled version of the value for β^i that maximizes the complete binary logistic likelihood.

Albert and Anderson (1984) proved that if the dataset x does not exhibit complete or quasi-complete separation, then there is a unique value for the MLE $\hat{\beta}$ for the complete model regression parameter β. As a result, unique values for the MLE $\hat{\beta}^i$ for $\tilde{\beta}^i$ exist if the data do not exhibit complete separation in the variables indexed by i. It is trivial to show that if the complete dataset x does not exhibit complete or quasi-complete separation, then there is no $i \in I_j$ for which x^i exhibits complete or quasi-complete separation. If there was, then there would exist a b^i such that

$$b^{i\top} x_n^i > 0 \text{ for all } n \text{ such that } y_n = 2,$$
$$b^{i\top} x_n^i < 0 \text{ for all } n \text{ such that } y_n = 1.$$

Setting $b_d = b_d^i$ if $d \in i$ and 0 otherwise yields the vector $b = (b_1, \ldots, b_D)$ such that

$$b^\top x_n = b^{i\top} x_n^i > 0 \text{ for all } n \text{ such that } y_n = 2,$$
$$b^\top x_n = b^{i\top} x_n^i < 0 \text{ for all } n \text{ such that } y_n = 1,$$

which is a contradiction. As a result, a sufficient condition for the existence and uniqueness of all MLEs $\hat{\beta}^i$ for $\tilde{\beta}^i$, $i \in i$, is that the complete data x does not exhibit complete or quasi-complete separation. Now, given $\hat{\beta}^i$ is the value for $\tilde{\beta}^i$ that minimizes the log-likelihood for the ith model

$$\log L(x^i, y; \tilde{\beta}^i) = \sum_{n=1}^N \mathbb{1}\{y_n = 1\} \log P(Y = 1|x_n^i, y_n, \tilde{\beta}^i) + \mathbb{1}\{y_n = 2\}$$
$$\log P(Y = 2|x_n^i, y_n, \tilde{\beta}^i),$$

that is, $\log L(x^i, y; \hat{\beta}^i) < \log L(x^i, y; \tilde{\beta}^i)$ for all $\tilde{\beta}^i \in \mathcal{D}_{\tilde{\beta}^i}$, the parameter $\hat{\beta}^{(j)} = \left\{ \hat{\beta}^i \right\}_{i \in I_j}$ therefore minimizes the log-likelihood

$$\log L_O^{(j)}(x, y; \tilde{\beta}^{(j)}) = \sum_{i \in I_j} \log L(x^i, y; \tilde{\beta}^i),$$

that is, $\log L_O^{(j)}(x, y; \hat{\beta}^{(j)}) < \log L_O^{(j)}(x, y; \tilde{\beta}^{(j)})$ for all $\tilde{\beta}^{(j)} \in \mathcal{D}_{\tilde{\beta}^{(j)}}$. Therefore an estimate $\hat{\beta}$ for β such that

$$\hat{\beta}^i = \frac{\beta^i + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-i}} \hat{\beta}^{i'} \alpha_{ii'} \right)^\top \right]^\top}{\sqrt{1 + \frac{\sum_{i'_1 \in \mathcal{I}_1^{-i}} \left(\beta^{i'_1 2} \lambda_{ii'_1}^2 + 2 \sum_{i'_2 \in \mathcal{I}_1^{-i}, i'_2 \neq i'_1} \beta^{i'_1} \beta^{i'_2} \lambda_{ii'_1 i'_2} \right)}{\pi^2/3}}},$$

(i.e., an estimate for β that yields the MLEs for each of the smaller-dimensional logistic models based on the variables indexed by i) will minimize $\log L_O^{(j)}(x, y; \tilde{\beta}^{(j)})$. By the definition of the model this $\hat{\beta}$ will exist. Given that a logistic OvR model is just the product of K binary logistic regression models, the above results hold for the OvR model for $K > 2$ classes. Consequently, the j-dimensional OvR model can therefore be written as

$$L_O^{(j)}(x, y; \beta) = \prod_{i \in \mathcal{I}_j} L_O(x^i, y; \tilde{\beta}^i),$$

where

$$\tilde{\beta}_k^i = \frac{\beta_k^i + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-i}} \beta_k^{i'} \alpha_{ii'} \right)^\top \right]^\top}{\sqrt{\operatorname{var}(\tilde{e}_{nk}^i)}} \in \mathbb{R}^{(j+1)}.$$

Through similar arguments, we can find an expression for the symbolic j-dimensional OvR model as

$$L_{SO}^{(j)}(s; \beta) = \prod_{i \in \mathcal{I}_j} L_{SO}(s^i, y; \tilde{\beta}^i).$$

ORCID

B. Beranger 🔘 http://orcid.org/0000-0002-7944-3925
S. A. Sisson 🔘 http://orcid.org/0000-0001-8943-067X

References

Albert, A., and Anderson, J. A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 71, 1–10. [1049,1051,1064,1065,1066]

Armstrong, B. (1985), "Measurement Error in the Generalized Linear Model," *Communication in Statistics—Simulation and Computation*, 14, 529–544. [1050]

Baldi, P., Sadowski, P., and Whiteson, D. (2014), "Searching for Exotic Particles in High-Energy Physics With Deep Learning," *Nature Communications*, 5, 1–9. [1061]

Beranger, B., Lin, H., and Sisson, S. A. (2018), "New Models for Symbolic Data Analysis," arXiv no. 1809.03659. [1049,1050,1052,1053,1056,1061,1063,1064]

Bhowmik, A., Ghosh, J., and Koyejo, O. (2016), "Generalized Linear Models for Aggregated Data," arXiv no. 1605.04466. [1050]

Billard, L. (2011), "Brief Overview of Symbolic Data and Analytic Issues," *Statistical Analysis and Data Mining*, 4, 149–156. [1049,1056]

Billard, L., and Diday, E. (2003), "From the Statistics of Data to the Statistics of Knowledge," *Journal of the American Statistical Association*, 98, 470–487. [1049,1052,1056]

—— (2006), *Symbolic Data Analysis*, Wiley Series in Computational Statistics, Chichester: Wiley. [1052]

Bock, H.-H., and Diday, E., eds. (2000), *Analysis of Symbolic Data: Exploratory Methods For Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization, Berlin: Springer-Verlag. [1052]

Bowling, S. R., and Khasawneh, M. T. (2009), "A Logistic Approximation to the Cumulative Normal Distribution," *Journal of Industrial Engineering and Management*, 2, 114–127. [1065]

Brito, P., and Silva, A. P. D. (2012), "Modelling Interval Data With Normal and Skew-Normal Distributions," *Journal of Applied Statistics*, 39, 3–20. [1052]

Cox, D. (1958), "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society*, Series B, 20, 215–232. [1049]

Cramer, J. S. (2007), "Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-Specified Disturbances," *Oxford Bulletin of Economics and Statistics*, 69, 545–555. [1055,1063,1065]

de Souza, R. M., Cysneiros, F. J. A., Queiroz, D. C., and Fagundes, R. A. de A. (2008), "A Multi-Class Logistic Regression Model for Interval Data," in *Conference: Systems, Man and Cybernetics*. [1049]

de Souza, R. M. C. R., Queiroz, D. C. F., and Cysneiros, F. J. A. (2011), "Logistic Regression-Based Pattern Classifiers for Symbolic Interval Data," *Pattern Analysis and Applications*, 14, 273–282. [1049]

Diday, E. (1989), "Introduction a l'approche symbolique en analyse des données," *RAIRO-Operations Research-Recherche Opérationnelle*, 23, 193–236. [1049,1052]

Dua, D., and Graff, C. (2017), "UCI Machine Learning Repository." [1050,1061]

Eichelberger, R. K., and Sheng, V. S. (2013), "An Empirical Study of Reducing Multiclass Classification Methodologies," in *Machine Learning and Data Mining in Pattern Recognition. MLDM 2013*, Lecture Notes in Computer Science (Vol. 7988), ed. P. Perner, Berlin, Heidelberg: Springer, pp. 505–519. [1051,1058]

Hastie, T., Tibshirani, R., and Friedman, J. (2008), *The Elements of Statistical Learning* (2nd ed.), New York: Springer. [1049]

Hauser, R. P., and Booth, D. (2011), "Predicting Bankruptcy With Robust Logistic Regression," *Journal of Data Science*, 9, 565–584. [1049]

Heitjan, D. F. (1989), "Inference From Grouped Continuous Data: A Review," *Statistical Science*, 4, 164–183. [1053]

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013), *Applied Logistic Regression* (3rd ed.), Hoboken, NJ: Wiley. [1049]

Jeffress, L. A. (1973), "The Logistic Distribution as an Approximation to the Normal Curve," *The Journal of the Acoustical Society of America*, 53, 1296. [1065]

Johnson, T. R. (2006), "Generalized Linear Models With Ordinally-Observed Covariates," *British Journal of Mathematical and Statistical Psychology*, 59, 275–300. [1050]

Johnson, T. R., and Wiest, M. M. (2014), "Generalized Linear Models With Coarsened Covariates: A Practical Bayesian Approach," *Psychological Methods*, 19, 281–299. [1050]

Kim, H., and Gu, Z. (2010), "A Logistic Regression Analysis for Predicting Bankruptcy in the Hospitality Industry," *Journal of Hospitality Financial Management*, 14, 24. [1049]

Knottnerus, J. A. (1992), "Application of Logistic Regression to the Analysis of Diagnostic Data: Exact Modeling of a Probability Tree of Multiple Binary Variables," *Medical Decision Making*, 12, 93–108. [1049]

Le Rademacher, J., and Billard, L. (2011), "Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data," *Journal of Statistical Planning and Inference*, 141, 1593–1602. [1052]

Lin, H., Caley, M. J., and Sisson, S. A. (2017), "Estimating Global Species Richness Using Symbolic Data Meta-Analysis," arXiv no. 1711.03202. [1052]

Lindsay, B. G. (1988), "Composite Likelihood Methods," in *Statistical Inference From Stochastic Processes (Ithaca, NY, 1987)*, Contemporary Mathematics (Vol. 80), Providence, RI: American Mathematical Society, pp. 221–239. [1055]

Lipsitz, S., Parzen, M., Natarajan, S., Ibrahim, J., and Fitzmaurice, G. (2004), "Generalized Linear Models With a Coarsened Covariate," *Journal of the Royal Statistical Society*, Series C, 53, 279–292. [1050]

Merlo, J., Chaix, B., Ohlsson, H., Beekman, A., Johnell, K., Hjerpe, P., Rstam, L., and Larsen, K. (2006), "A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Using Measures of Clustering in Multilevel Logistic Regression to Investigate Contextual Phenomena," *Journal of Epidemiology and Community Health*, 60, 290–297. [1049]

Min, H. (2013), "Ordered Logit Regression Modeling of the Self-Rated Health in Hawaii, With Comparisons to the OLS Model," *The Journal of Modern Applied Statistical Methods*, 12, 371–380. [1049]

Pampel, F. C. (2000), *Logistic Regression: A Primer* (Vol. 1), Thousand, CA: SAGE Publications. [1049]

Pavlopoulos, D., Muffels, R., and Vermunt, J. (2010), "Wage Mobility in Europe. A Comparative Analysis Using Restricted Multinomial Logit Regression," *Quality and Quantity*, 44, 115–129. [1049]

Pingel, R. (2014), "Some Approximations of the Logistic Distribution With Application to the Covariance Matrix of Logistic Regression," *Statistics and Probability Letters*, 85, 63–68. [1065]

QUT (2016), "Technical Report on Classification Methods for Crop Types," Technical Report, Queensland University of Technology. [1062,1063]

Rahman, P., Beranger, B., Roughan, M., and Sisson, S. A. (2020), "Likelihood-Based Inference for Modelling Packet Transit From Thinned Flow Summaries," arXiv no. 2008.13424. [1049]

Tranmer, M. H., and Steel, D. (1997), "Logistic Regression Analysis With Aggregate Data: Tackling the Ecological Fallacy," *Paper Presented at the American Statistical Association Conference*. [1050]

Varin, C., Reid, N., and Firth, D. (2011), "An Overview of Composite Likelihood Methods," *Statistica Sinica*, 21, 5–42. [1055]

Wang, H., Zhu, R., and Ma, P. (2018), "Optimal Subsampling for Large Sample Logistic Regression," *Journal of the American Statistical Association*, 113, 829–844. [1049,1050,1059,1060,1061,1062,1064]

Whitaker, T., Beranger, B., and Sisson, S. A. (2020), "Composite Likelihood Methods for Histogram-Valued Random Variables," *Statistics and Computing*, 30, 1459–1477. [1050,1055,1056]

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press. [1055,1065]

Zhang, X., Beranger, B., and Sisson, S. A. (2020), "Constructing Likelihood Functions for Interval-Valued Random Variables," *Scandinavian Journal of Statistics*, 47, 1–35. [1049,1050,1052]