JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 1, 1–13
http://dx.doi.org/10.1080/10618600.2016.1154063

Taylor & Francis
Taylor & Francis Group

Regression Models for Multivariate Count Data

Yiwen Zhang[a], Hua Zhou[b], Jin Zhou[c], and Wei Sun[d]

[a]Department of Statistics, North Carolina State University, Raleigh, North Carolina; [b]Department of Biostatistics, University of California, Los Angeles, California; [c]Division of Epidemiology and Biostatistics, University of Arizona, Tucson, Arizona; [d]Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, Washington

ABSTRACT

Data with multivariate count responses frequently occur in modern applications. The commonly used multinomial-logit model is limiting due to its restrictive mean-variance structure. For instance, analyzing count data from the recent RNA-seq technology by the multinomial-logit model leads to serious errors in hypothesis testing. The ubiquity of overdispersion and complicated correlation structures among multivariate counts calls for more flexible regression models. In this article, we study some generalized linear models that incorporate various correlation structures among the counts. Current literature lacks a treatment of these models, partly because they do not belong to the natural exponential family. We study the estimation, testing, and variable selection for these models in a unifying framework. The regression models are compared on both synthetic and real RNA-seq data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2015
Revised October 2015

KEYWORDS
Analysis of deviance; Categorical data analysis; Dirichlet-multinomial; Generalized Dirichlet-multinomial; Iteratively reweighted Poisson regression (IRPR); Negative multinomial; Reduced rank GLM; Regularization

1. Introduction

Multivariate count data abound in modern application areas such as genomics, sports, imaging analysis, and text mining. When the responses are continuous, it is natural to adopt the multivariate normal model. For multivariate count responses, a common choice is the multinomial-logit model (McCullagh and Nelder 1983). However, the multinomial model is limiting due to its specific mean-variance structure and the implicit assumption that individual counts in the response vector are negatively correlated. In this article, we examine regression models for multivariate counts with more flexible mean–covariance and correlation structure. Parameter estimation in these models is typically hard because they do not belong to the exponential family. We propose a unifying iteratively reweighted Poisson regression (IRPR) method for the maximum likelihood estimation. IRPR is stable, scalable to high-dimensional data, and simple to implement using existing software. Testing and regularization methods for these models are also studied. Our methods are implemented in the R package and Matlab toolbox `mglm` (Zhang and Zhou 2015).

The remaining of the article is organized as follows. Section 2 motivates our study with analysis of count data in modern genomics. Section 3 introduces a class of GLM models for multivariate count responses. A unifying maximum likelihood estimation procedure is proposed in Section 4.2. Testing and regularized estimation are treated in Sections 5 and 6 respectively, followed by numerical examples in Section 7.

2. Motivation

Our study is motivated by the analysis of high-throughput data in genomics. Next generation sequencing technology has become the primary choice for massive quantification of genomic features. The data obtained from sequencing technologies are often summarized by the counts of DNA or RNA fragments within a genomic interval. A prime example is the RNA-seq data.

In most mammalian genomes, one gene is composed of multiple exons and different combinations of exons lead to different protein products. One such exon combination is called an RNA isoform. The left panel of Figure 1 depicts a gene with 3 exons and all possible isoforms for that gene. Current RNA-seq platforms are able to deliver the expression counts of each exon set (Wang, Gerstein, and Snyder 2009). An exon set includes contiguous portions isoforms. Here, we use the number of RNA-seq reads from exon sets instead of exons because some RNA-seq fragments may overlap with multiple exons (Sun et al. 2015).

Data for one gene with d exon sets take the form

Subject	Exon Set 1	Exon Set 2	⋯	Exon Set d	Treatment	Gender	Age	⋯
1	15	0	⋯	3	Yes	M	43	⋯
2	0	52	⋯	0	Yes	F	35	⋯
⋮					⋮			
n	124	45	⋯	73	No	F	25	⋯

CONTACT Hua Zhou *huazhou@ucla.edu* Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA 90095-1772.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JCGS*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

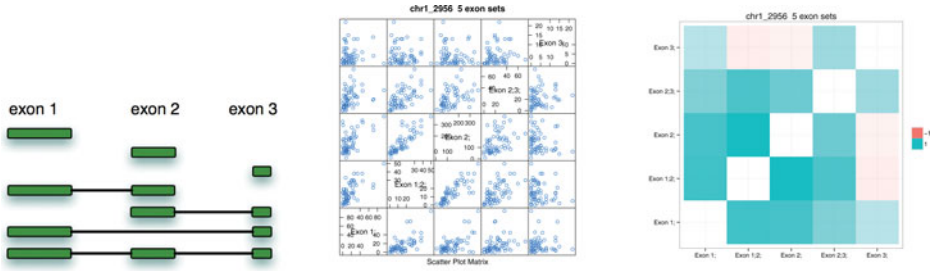

Figure 1. Left: a gene with 3 exons and 7 possible isoforms. Middle and right: pairwise scatterplots and correlations of exon counts of a gene with 5 exon sets.

The primary interest is to identify differential expression profiles and their relation to the covariates. The total expression of a gene in an individual can be obtained by summing the counts across exon sets. Negative binomial regression models have been developed to associate the total expression count with the covariates of interest (Anders and Huber 2010; Robinson, McCarthy, and Smyth 2010). This approach ignores the RNA isoform expression. More sophisticated methods have been developed to estimate RNA isoform expression and then assess association between isoform expression and covariates of interest. Nevertheless accounting for RNA isoform estimation uncertainty in the association step is a nontrivial task, since the responses in the association step are obtained from variable selection (Sun et al. 2015). An attractive alternative is to treat exon counts as multivariate responses and fit a generalized linear model (GLM). The multinomial-logit model is a popular choice, due to its wide availability in statistical software. However, it assumes negative correlation between counts. Two exon sets may belong to one or a few RNA isoforms, leading to complicated correlation structures among their counts. The middle panel of Figure 1 displays the pairwise scatterplots of exon counts for a pseudogene SPCS2P4 with 5 exon sets and the right panel depicts the corresponding correlations in a grayscale image. Notably the correlations can be both positive and negative.

We simulated RNA-seq read counts based on the mouse gene Phlda3 and the RNA-seq data collected from a mouse study (Sun et al. 2015). This gene has 4 exons and 6 exon sets that have nonzero observed read counts. The number of RNA-seq fragments per exon set was simulated by a negative binomial distribution with the mean equal to a linear combination of underlying RNA isoform expression and with the overdispersion estimated from RNA-seq data. In the generative model, the RNA isoform expression is associated with covariate `treatment`, but not with `age` and `gender`. $n = 200$ observations were generated. We fit a negative binomial regression (NegBin) using the total read counts from the 6 exon sets as responses; we also fit the multinomial-logit (MN) model using the multivariate count vectors as responses. The predictor `log(TotalReads)` is included as routinely done in RNA-seq data analysis. Based on 300 simulation replicates, the empirical rejection rates of the Wald test for testing each predictor are reported in Table 1. NegBin regression has well-controlled Type I error rate for the null predictors `age` and `gender`. However, it has almost no power for detecting the `treatment` effect. The multinomial-logit model has seriously inflated Type I error for the two null predictors `age` and `gender`. As a prelude, we also fit three other GLMs to the same datasets: Dirichlet-multinomial regression (DM), generalized Dirichlet-multinomial (GDM) regression, and negative multinomial (NM) regression. Details of these models are given in Section 3. We find that GDM shows both well-controlled Type I error for `age` and `gender` and high power for detecting the `treatment` effect. Model selection criteria AIC (Akaike information criterion) and BIC (Bayesian information criterion) also indicate appropriateness of GDM. AIC/BIC of negative binomial regression is not listed because it uses sum of counts and is incomparable to the multivariate models. We remark that the generative model has a marginal negative binomial distribution and it has nothing to do with the GDM model. Success of GDM results from its ability to learn the complex correlation between counts.

3. Models

Table 2 lists four regression models for multivariate count responses. They impose different correlation structures on the counts. Except for the multinomial-logit model, none of the other three belongs to the natural exponential family. We denote the data by $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{y}_i \in \mathbb{N}^d$ are d-dimensional count vectors and $\boldsymbol{x}_i \in \mathbb{R}^p$ are p-dimensional covariates. $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\mathsf{T} \in \mathbb{N}^{n\times d}$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T} \in \mathbb{R}^{n\times p}$ are called the response and design matrix respectively. For each model, we introduce the response distribution, propose a link function, and derive the score and information matrices (listed in Supplementary Materials S.2), which are essential for statistical estimation and inference.

Multinomial Regression (MN). To be self-contained, we start with the classical multinomial-logit model (McCullagh and Nelder 1983). The response $\boldsymbol{y}$ is modeled as multinomial with

Table 1. Empirical rejection rates for testing each predictor in the simulated RNA-seq data.

	NegBin	MN	DM	GDM	NM
Intercept	1.00(0.00)	0.98(0.01)	0.25(0.02)	0.22(0.02)	1.00(0.00)
log(TotalReads)	1.00(0.00)	0.98(0.01)	0.11(0.02)	0.10(0.02)	1.00(0.00)
treatment	0.06(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
gender	0.06(0.01)	0.97(0.01)	0.10(0.02)	0.07(0.01)	0.96(0.01)
age	0.03(0.01)	0.94(0.01)	0.10(0.02)	0.06(0.01)	0.91(0.02)
AIC		16142.42 (591.00)	8951.63 (64.28)	8775.60 (60.29)	18521.80 (1082.78)
BIC		16224.88 (591.00)	9050.58 (64.28)	8940.51 (60.29)	18624.05 (1082.78)
Type I error	✓	×	×	✓	×
Power	×	✓	✓	✓	✓

NOTES: The boldfaced predictor **treatment** has nonzero effect on the exon expression counts. Gender and age have no effects on the expression counts. Numbers in the brackets are standard errors based on 300 simulation replicates.

Table 2. Generalized linear models for multivariate categorical responses. d: dimension of response vector; p: number of predictors in regression models; m: the batch size $|\boldsymbol{y}| = \sum_j y_j$ of the response vector $\boldsymbol{y} = (y_1, \ldots, y_d)$; $|\boldsymbol{\alpha}| = \sum_j \alpha_j$.

	Multinomial (MN)	Dirichlet-multinomial (DM)	Negative multinomial (NM)	Gen. Dir-Mult (GDM)
Response data $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$	negatively correlated	negatively correlated	positively correlated	general correlation
Parameter $\boldsymbol{\theta}$	$\boldsymbol{p} = (p_1, \ldots, p_d) \sum_j p_j = 1, p_j > 0$	$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \alpha_j > 0$	$(\boldsymbol{p}, \beta) = (p_1, \ldots, p_{d+1}, \beta)$ $\sum_j p_j = 1, p_j > 0, \beta > 0$	$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_1, \ldots, \alpha_{d-1},$ $\beta_1, \ldots, \beta_{d-1}), \alpha_j, \beta_j > 0$
Mean $\mu_j = \mathrm{E}(Y_j)$	mp_j	$m\alpha_j / \lvert\boldsymbol{\alpha}\rvert$	$\beta p_j / p_{d+1}$	see Lemma 1
Var. $\sigma_j = \mathrm{Var}(Y_j)$	$mp_j(1-p_j)$	$m \frac{\lvert\boldsymbol{\alpha}\rvert+m}{\lvert\boldsymbol{\alpha}\rvert+1} \frac{\alpha_j}{\lvert\boldsymbol{\alpha}\rvert}(1-\frac{\alpha_j}{\lvert\boldsymbol{\alpha}\rvert})$	$\beta \frac{p_j}{p_{d+1}}(1+\frac{p_j}{p_{d+1}})$	see Lemma 1
Cov. $\sigma_{jj'} = \mathrm{Cov}(Y_j, Y_{j'})$	$-mp_j(1-p_j)$	$-m \frac{\lvert\boldsymbol{\alpha}\rvert+m}{\lvert\boldsymbol{\alpha}\rvert+1} \frac{\alpha_j}{\lvert\boldsymbol{\alpha}\rvert} \frac{\alpha_{j'}}{\lvert\boldsymbol{\alpha}\rvert}$	$\beta \frac{p_j}{p_{d+1}} \frac{p_{j'}}{p_{d+1}}$	see Lemma 1
Regress. param. B	$(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d-1})$ $\boldsymbol{\beta}_j \in \mathbb{R}^p$	$(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ $\boldsymbol{\beta}_j \in \mathbb{R}^p$	$(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d, \boldsymbol{\beta})$ $\boldsymbol{\alpha}_j, \boldsymbol{\beta} \in \mathbb{R}^p$	$(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{d-1}), \boldsymbol{\alpha}_j \in \mathbb{R}^p$ $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d-1}), \boldsymbol{\beta}_j \in \mathbb{R}^p$
Systematic comp.	$\eta_j = \boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}_j$	$\eta_j = \boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}_j$	$\eta_j = \boldsymbol{x}^\mathsf{T} \boldsymbol{\alpha}_j, \zeta = \boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}$	$\eta_j = \boldsymbol{x}^\mathsf{T} \boldsymbol{\alpha}_j, \zeta_j = \boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}_j$
Link $\eta_j = g_j(\boldsymbol{\theta})$	$\eta_j = \ln(\frac{p_j}{1-\sum_{j'=1}^{d-1} p_{j'}})$	$\eta_j = \ln(\alpha_j)$	$\eta_j = \ln(\frac{p_j}{1-\sum_{j'=1}^{d} p_{j'}}), \zeta = \ln(\beta)$	$\eta_j = \ln \alpha_j, \zeta_j = \ln \beta_j$
Inverse link $\theta_j = g_j^{-1}(\boldsymbol{\eta})$	$p_j = \frac{\exp(\eta_j)}{1+\sum_{j'=1}^{d-1} \exp(\eta_{j'})}$	$\alpha_j = \exp(\eta_j)$	$p_j = \frac{\exp(\eta_j)}{1+\sum_{j'=1}^{d} \exp(\eta_{j'})}, \beta = \exp(\zeta)$	$\alpha_j = \exp(\eta_j), \beta_j = \exp(\zeta_j)$

$m = |\boldsymbol{y}| = \sum_{j=1}^d y_j$ trials and success probability parameter $\boldsymbol{p} = (p_1, \ldots, p_d)$, $p_j > 0$, $\sum_{j=1}^d p_j = 1$. The probability mass function is

$$f(\boldsymbol{y}|\boldsymbol{p}) = \binom{m}{\boldsymbol{y}} \prod_{j=1}^d p_j^{y_j}.$$

It is well known that the multinomial distribution belongs to the natural exponential family. Parameter $\boldsymbol{p}$ is linked to the covariates $\boldsymbol{x} \in \mathbb{R}^p$ via the multinomial-Poisson transformation (Baker 1994)

$$p_j = \frac{e^{\boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}_j}}{\sum_{j'} e^{\boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}_{j'}}}, \quad j = 1, \ldots, d,$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d \in \mathbb{R}^p$ are the regression coefficients. Because of the constraint $\sum_j p_j = 1$, we set $\boldsymbol{\beta}_d = \boldsymbol{0}_p$ for identifiability and only estimate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d-1}$, which are collected in the regression coefficient matrix $\boldsymbol{B} \in \mathbb{R}^{p \times (d-1)}$. Log-likelihood of n independent observations $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is

$$\ell_n(\boldsymbol{B}) = \sum_{i=1}^n \sum_{j=1}^d y_{ij} \left(\boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_j - \ln \sum_{j'=1}^d e^{\boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_{j'}} \right) + \sum_{i=1}^n \ln \binom{m_i}{\boldsymbol{y}_i}. \quad (1)$$

The mapping $(\eta_1, \ldots, \eta_d)^\mathsf{T} \mapsto -\ln \sum_j e^{\eta_j}$ is concave; thus the log-likelihood function (1) is concave.

Dirichlet-Multinomial Regression (DM). Multinomial model is not sufficient when there is observed over-dispersion. Dirichlet-multinomial distribution models the variation among the percentages $\boldsymbol{p}$ in the multinomial distribution by a Dirichlet distribution (Mosimann 1962). The probability mass of a d-category count vector $\boldsymbol{y}$ over $m = |\boldsymbol{y}| = \sum_j y_j$ trials under Dirichlet-multinomial with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d), \alpha_j > 0$, is

$$f(\boldsymbol{y}|\boldsymbol{\alpha}) = \int_{\Delta_d} \binom{m}{\boldsymbol{y}} \prod_j p_j^{y_j} \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_j \Gamma(\alpha_j)} \prod_j p_j^{\alpha_j - 1} \, d\boldsymbol{p}$$

$$= \binom{m}{\boldsymbol{y}} \prod_{j=1}^d \frac{\Gamma(\alpha_j + y_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\sum_j \alpha_j)}{\Gamma(\sum_j \alpha_j + \sum_j y_j)}$$

$$= \binom{m}{\boldsymbol{y}} \frac{\prod_{j=1}^d (\alpha_j)_{(y_j)}}{(|\boldsymbol{\alpha}|)_{(m)}}, \quad (2)$$

where $(a)_{(k)} = a(a+1)\ldots(a+k-1)$ denotes the rising factorial and $|\boldsymbol{\alpha}| = \sum_j \alpha_j$. Because the data y_j and parameter α_j are intertwined in the gamma terms and do not factorize, Dirichlet-multinomial distribution does *not* belong to the natural exponential family. The first two moments of DM are

$$\mathrm{E}(\boldsymbol{Y}) = m \frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|},$$

$$\mathrm{cov}(\boldsymbol{Y}) = m \frac{|\boldsymbol{\alpha}| + m}{|\boldsymbol{\alpha}| + 1} \left[\mathrm{diag}\left(\frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|} \right) - \left(\frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|} \right) \left(\frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|} \right)^\mathsf{T} \right].$$

It is clear that the counts are negatively correlated and the quantity $(|\boldsymbol{\alpha}| + m)/(|\boldsymbol{\alpha}| + 1)$ measures overdispersion. To incorporate covariates, the inverse link function $\alpha_j = e^{\boldsymbol{x}^\mathsf{T} \boldsymbol{\beta}_j}$ relates the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ of Dirichlet-multinomial distribution to the covariates $\boldsymbol{x}$. The log-likelihood for n independent data points $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ takes the form

$$\ell_n(\boldsymbol{B}) = \sum_{i=1}^n \sum_{j=1}^d \sum_{k=0}^{y_{ij}-1} \ln \left(e^{\boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_j} + k \right) - \sum_{i=1}^n \sum_{k=0}^{m_i - 1} \ln \left(\sum_{j=1}^d e^{\boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_j} + k \right) + \sum_{i=1}^n \ln \binom{m_i}{\boldsymbol{y}_i}, \quad (3)$$

where $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) \in \mathbb{R}^{p \times d}$ collects all regression coefficients. The log-likelihood, as a difference of two concave terms, is not concave in general.

Negative Multinomial Regression (NM). Negative correlation of the multinomial and Dirichlet-multinomial models prevents

their use for positively correlated counts. The negative multinomial distribution provides a model for such data. The probability mass of a count vector $\boldsymbol{y} \in \mathbb{N}^d$ under a negative multinomial distribution with parameter $(p_1, \ldots, p_{d+1}, \beta)$, $\sum_{j=1}^{d+1} p_j = 1$, $p_j, \beta > 0$, is

$$f(\boldsymbol{y}|\boldsymbol{p}, \beta) = \binom{\beta + m - 1}{m} \binom{m}{\boldsymbol{y}} \prod_{j=1}^{d} p_j^{y_j} p_{d+1}^{\beta}$$
$$= \frac{(\beta)_{(m)}}{m!} \binom{m}{\boldsymbol{y}} \prod_{j=1}^{d} p_j^{y_j} p_{d+1}^{\beta}.$$

Parameter β and data m do not factorize; thus negative multinomial does *not* belong to the exponential family when β is unknown. Denote $\boldsymbol{p} = (p_1, \ldots, p_d)$. Then, the first two moments are

$$\mathrm{E}(\boldsymbol{Y}) = \beta \left(\frac{\boldsymbol{p}}{p_{d+1}} \right),$$
$$\mathrm{cov}(\boldsymbol{Y}) = \beta \left[\mathrm{diag} \left(\frac{\boldsymbol{p}}{p_{d+1}} \right) + \left(\frac{\boldsymbol{p}}{p_{d+1}} \right) \left(\frac{\boldsymbol{p}}{p_{d+1}} \right)^{\mathsf{T}} \right],$$

showing positive pairwise correlation between Y_j. We use the link functions

$$p_j = \frac{e^{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\alpha}_j}}{1 + \sum_{j=1}^{d} e^{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\alpha}_j}}, 1 \le j \le d,$$
$$p_{d+1} = \frac{1}{1 + \sum_{j=1}^{d} e^{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\alpha}_j}}, \quad \beta = e^{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\beta}},$$

to relate covariates $\boldsymbol{x} \in \mathbb{R}^p$ to distribution parameter $(p_1, \ldots, p_{d+1}, \beta)$. Let $\boldsymbol{B} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d, \boldsymbol{\beta}) \in \mathbb{R}^{p \times (d+1)}$ collect all the regression coefficients. Given n independent data points $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, the log-likelihood is

$$\ell_n(\boldsymbol{B}) = \sum_{i=1}^{n} \sum_{k=0}^{m_i - 1} \ln \left(e^{\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}} + k \right)$$
$$- \sum_{i=1}^{n} \left(e^{\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}} + m_i \right) \ln \left(\sum_{j=1}^{d} e^{\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_j} + 1 \right)$$
$$+ \sum_{i=1}^{n} \sum_{j=1}^{d} y_{ij} \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_j - \sum_{i=1}^{n} \sum_{j=1}^{d} \ln y_{ij}!. \tag{4}$$

When the overdispersion parameter β is not linked to covariates, the log-likelihood becomes

$$\ell_n(\boldsymbol{B}) = \sum_{i=1}^{n} \sum_{k=0}^{m_i - 1} \ln(\beta + k) - \sum_{i=1}^{n} (\beta + m_i) \ln \left(\sum_{j=1}^{d} e^{\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_j} + 1 \right)$$
$$+ \sum_{i=1}^{n} \sum_{j=1}^{d} y_{ij} \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_j - \sum_{i=1}^{n} \sum_{j=1}^{d} \ln y_{ij}!, \tag{5}$$

where $\boldsymbol{B} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_d^T, \beta)^T \in \mathbb{R}^{pd+1}$. Neither log-likelihood (4) nor (5) is necessarily concave.

Generalized Dirichlet-Multinomial Regression (GDM). It is possible to relax the restrictions of pairwise negative correlation in MN and DM or pairwise positive correlation in NM by choosing a more flexible mixing distribution as a prior for the multinomial. Connor and Mosimann (1969) suggested a generalized Dirichlet-multinomial distribution, which provides a flexible model for multivariate categorical responses with general correlation structure.

The probability mass of a count vector $\boldsymbol{y}$ over m trials under the generalized Dirichlet-multinomial model with parameter $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_1, \ldots, \alpha_{d-1}, \beta_1, \ldots, \beta_{d-1})$, $\alpha_j, \beta_j > 0$, is

$$f(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \binom{m}{\boldsymbol{y}} \prod_{j=1}^{d-1} \frac{\Gamma(\alpha_j + y_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\beta_j + z_{j+1})}{\Gamma(\beta_j)} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j + \beta_j + z_j)}$$
$$= \binom{m}{\boldsymbol{y}} \prod_{j=1}^{d-1} \frac{(\alpha_j)_{(y_j)} (\beta_j)_{(z_{j+1})}}{(\alpha_j + \beta_j)_{(z_j)}}, \tag{6}$$

where $z_j = \sum_{k=j}^{d} y_k$. The Dirichlet-multinomial (DM) distribution is a special case of GDM by taking $\beta_j = \alpha_{j+1} + \cdots + \alpha_d$. GDM does not belong to the exponential family. The moments of the GDM are recorded in Lemma 1 . A distinctive property of the GDM is that the correlations between counts can be simultaneously positive or negative, yielding the most modeling flexibility among the models in Table 2. Below $(a)_{[k]} = a(a-1)\ldots(a-k+1)$ denote the falling factorial.

Lemma 1. The falling factorial moments of GDM are

$$\mathrm{E} \prod_{j=1}^{d} (Y_j)_{[r_j]} = \mathrm{E} \prod_{j=1}^{d} Y_j (Y_j - 1) \ldots (Y_j - r_j + 1)$$
$$= (m)_{[\sum_j r_j]} \prod_{j=1}^{d-1} \frac{(\alpha_j)_{(r_j)} (\beta_j)_{(\delta_{j+1})}}{(\alpha_j + \beta_j)_{(\delta_j)}},$$

where $\delta_j = r_j + \cdots + r_d$ for $j = 1, \ldots, d$. Specifically the first two moments are

$$\mathrm{E}Y_j = m \begin{cases} \dfrac{\alpha_1}{\alpha_1 + \beta_1} & j = 1 \\ \dfrac{\alpha_j}{\alpha_j + \beta_j} \prod_{k=1}^{j-1} \dfrac{\beta_k}{\alpha_k + \beta_k} & j = 2, \ldots, d-1 \\ \prod_{k=1}^{d-1} \dfrac{\beta_j}{\alpha_j + \beta_j} & j = d \end{cases}$$

and

$$\mathrm{cov}(Y_j, Y_{j'}) = m \frac{\alpha_{j'}}{\alpha_{j'} + \beta_{j'}} \prod_{k=1}^{j'-1} \frac{\beta_k}{\alpha_k + \beta_k}$$
$$\times \left((m-1) \prod_{k=1}^{j-1} \frac{\beta_k + 1}{\alpha_k + \beta_k + 1} \frac{\alpha_j + 1_{\{j=j'\}}}{\alpha_j + \beta_j + 1} \right.$$
$$\left. - m \prod_{k=1}^{j-1} \frac{\beta_k}{\alpha_k + \beta_k} \frac{\alpha_j}{\alpha_j + \beta_j} + 1_{\{j=j'\}} \right)$$

for $j \le j'$.

We employ the link functions

$$\alpha_j = e^{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\alpha}_j}, \quad \beta_j = e^{\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\beta}_j}, \quad 1 \le j \le d-1,$$

to relate covariates $\boldsymbol{x}$ to distribution parameters $(\alpha_1, \ldots, \alpha_{d-1}, \beta_1, \ldots, \beta_{d-1})$ of the GDM model. Here we

slightly abuse notation and use scalar α_j and β_j to denote parameters of the GDM distribution. Boldfaced vectors $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$, both of dimension p, denote regression coefficients. Let $\boldsymbol{B} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{d-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d-1}) \in \mathbb{R}^{p\times 2(d-1)}$ collect all the regression coefficients. Given n independent data points $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, the log-likelihood is

$$\ell_n(\boldsymbol{B}) = \sum_{i=1}^{n}\sum_{j=1}^{d-1}\left[\sum_{k=0}^{y_{ij}-1}\ln\left(e^{\boldsymbol{x}_i^\top\boldsymbol{\alpha}_j}+k\right) + \sum_{k=0}^{z_{i,j+1}-1}\ln\left(e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}_j}+k\right) - \sum_{k=0}^{z_{ij}-1}\ln\left(e^{\boldsymbol{x}_i^\top\boldsymbol{\alpha}_j}+e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}_j}+k\right)\right] + \sum_{i=1}^{n}\ln\binom{m_i}{\boldsymbol{y}_i}, \quad (7)$$

which is not concave in general.

Model Choice. Multinomial is a limiting case of Dirichlet-multinomial by taking $\boldsymbol{\alpha}/|\boldsymbol{\alpha}| \to \boldsymbol{p}$ as $|\boldsymbol{\alpha}| \to \infty$. Dirichlet-multinomial is a special case of generalized Dirichlet-multinomial by taking $\beta_j = \alpha_{j+1} + \cdots + \alpha_d$. Therefore for distribution fitting, standard tests such as the likelihood ratio test (LRT) help choose the best one among the three nested models: MN $\subset$ DM $\subset$ GDM. However, this nesting structure is lost in regression models. For instance, in the presence of predictors, multinomial regression is *not* a submodel of the Dirichlet-multinomial regression model, and the latter is not a special case of the generalized Dirichlet-multinomial (GDM) regression model. Information criteria such as the AIC and BIC can be used to determine a best regression model for the data.

4. Estimation

In this section, we consider the MLE when sample size is greater than the number of parameters. In Section 6, we consider regularized estimation that is useful for model selection. MLE for the DM, NM, and GDM models is nontrivial as they do not belong to the natural exponential family and the classical IRWLS (iteratively reweighted least squares) machinery is difficult to apply, as explained in Section 4.1.

4.1 Difficulties

We illustrate the difficulties using Dirichlet-multinomial (DM) regression as an example. Given iid data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$ be the design matrix and $\boldsymbol{A} = (\alpha_{ij}) \in \mathbb{R}^{n\times d}$, where $\alpha_{ij} = e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}_j}$, be the matrix of distribution parameters at each observation. The observed information matrix is

$$-d^2\ell_n(\boldsymbol{B}) = (\boldsymbol{I}_d \otimes \boldsymbol{X})^\top \cdot \operatorname{diag}[\operatorname{vec}(\operatorname{diag}(\boldsymbol{v})\boldsymbol{A} - \boldsymbol{K})] \cdot (\boldsymbol{I}_d \otimes \boldsymbol{X}) - (\boldsymbol{A}^\top \odot \boldsymbol{X}^\top)\operatorname{diag}(\boldsymbol{w})(\boldsymbol{A}^\top \odot \boldsymbol{X}^\top)^\top,$$

where $\boldsymbol{v} \in \mathbb{R}^n$ has entries $v_i = \sum_{k=1}^{m_i-1}(|\boldsymbol{\alpha}_i| + k)^{-1}$, $\boldsymbol{w} \in \mathbb{R}^n$ has entries $w_i = \sum_{k=0}^{m_i-1}(|\boldsymbol{\alpha}_i| + k)^{-2}$, and $\boldsymbol{K} = (k_{ij}) \in \mathbb{R}^{n\times d}$ has entries

$$k_{ij} = \sum_{k=0}^{y_{ij}-1}\left(\frac{\alpha_{ij}}{\alpha_{ij}+k}\right)\left(\frac{k}{\alpha_{ij}+k}\right).$$

Here, $\otimes$ denotes the Kronecker product and $\odot$ denotes the Khatri–Rao product. See Supplementary Materials S.2.2 for the derivation.

Traditional optimization methods encounter difficulties when maximizing the DM log-likelihood. The Newton–Raphson method can be unstable because of the nonconcavity of the log-likelihood (3). The observed information matrix $-d^2\ell_n(\boldsymbol{B})$ is not necessarily positive definite and the Newton iterates may diverge. This issue is serious in regression problems because often there is no good starting point available. The Fisher scoring algorithm, also known as the iteratively reweighted least squares (IRWLS), replaces the observed information matrix by its expectation, that is, the Fisher information matrix $\mathrm{E}[-d^2\ell_n(\boldsymbol{B})]$. The Fisher information matrix is always positive semidefinite and, combined with line search, is guaranteed to generate nondecreasing iterates. However, evaluation of the Fisher information matrix involves computing

$$\mathrm{E}(k_{ij}) = \mathrm{E}\left[\sum_{k=0}^{y_{ij}-1}\left(\frac{\alpha_{ij}}{\alpha_{ij}+k}\right)\left(\frac{k}{\alpha_{ij}+k}\right)\right] = \sum_{k=0}^{m_i-1}\left(\frac{\alpha_{ij}}{\alpha_{ij}+k}\right)\left(\frac{k}{\alpha_{ij}+k}\right)\mathrm{P}(Y_{ij} > k),$$

where $(Y_{i1}, \ldots, Y_{id})$ is a Dirichlet-multinomial random vector with parameter $(\alpha_{i1}, \ldots, \alpha_{id})$ and batch size m_i. Marginal Beta-binomial tail probabilities $\mathrm{P}(Y_{ij} > k)$ have to be evaluated for each combination of $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$, and $k \in \{0, \ldots, m_i - 1\}$, making Fisher scoring method computationally expensive for even moderate sample size n and d. Even when the information matrix is positive definite and can be evaluated, a $p(d-1)$ by $p(d-1)$ linear system needs to be solved in each Newton or scoring iteration. Finally, quasi-Newton methods such as BFGS updates (Nocedal and Wright 2006) may alleviate some of the issues but suffer from slow convergence and instability in many examples.

Table 3 reports results from a simple numerical experiment. DM data are generated with $d = 3, 15, 20, 30$ response categories, $p = 6$ predictors, and sample size $n = 200$. The true parameter value is set at $\boldsymbol{B} = 3 \times \boldsymbol{1}_{p\times d}$. Entries of the design matrix $\boldsymbol{X}$ are drawn from independent standard normals. For each d, 100 replicates are simulated. For each replicate, we use the `nlminb` function in R to fit DM regression using Newton (with analytical gradient and Hessian) and quasi-Newton (BFGS with analytical gradient) methods. The starting point for each run is set at $\boldsymbol{B}^{(0)} = \boldsymbol{0}_{p\times d}$. For small dimensional problem $d = 3$, Newton method converges in 72% of replicates. As d increases, Newton's method fails more and more often. The quasi-Newton method apparently suffers from both instability and slow convergence.

The same difficulties, namely instability, high cost of evaluating Fisher information matrix, and high dimensionality of parameters, beset the MLE for the negative multinomial (NM) and generalized Dirichlet-multinomial (GDM) regressions.

4.2 MLE via IRPR

We propose a unifying framework, the iteratively reweighted Poisson regression (IRPR), for the MLE of the four regression models. The plain version of the IRPR scheme is summarized in Algorithm 1. At each iteration, we update some working

Table 3. Comparison of Newton, Quasi-Newton (BFGS), and IRPR methods for fitting Dirichlet-multinomial (DM) regression. Numbers in parentheses are standard errors based on 100 simulation replicates.

Dimension	Method	Converge%	Total time (sec.)	Norm of gradient	Iterations
$d = 3$	Newton	0.72	6.30	74.67 (281.97)	12.30 (2.32)
$p = 6$	Quasi	0.91	26.43	0.24 (1.07)	122.56 (8.02)
	IRPR	1.00	6.98	0.00 (0.00)	7.45 (1.42)
$d = 15$	Newton	0.36	26.03	14.12 (40.90)	15.10 (2.57)
$p = 6$	Quasi	0.00	91.50	9.98 (5.60)	146.86 (5.19)
	IRPR	1.00	39.89	0.00 (0.00)	9.26 (1.15)
$d = 20$	Newton	0.31	41.30	11.31 (44.13)	16.06 (2.48)
$p = 6$	Quasi	0.00	117.62	12.48 (11.07)	145.97 (6.10)
	IRPR	1.00	54.63	0.00 (0.00)	9.53 (1.00)
$d = 30$	Newton	0.25	81.75	14.30 (45.97)	17.13 (2.62)
$p = 6$	Quasi	0.00	176.66	8.93 (9.10)	147.53 (4.58)
	IRPR	1.00	81.41	0.00 (0.01)	10.02 (1.29)

responses $y_{ij}^{(t)}$ and weights $w_{ij}^{(t)}$ and then perform d_e weighted Poisson regressions to update the regression parameters. Here, $d_e = d - 1$ for MN, d for DM, $2(d-1)$ for GDM, and $d+1$ for NM. Therefore, IRPR is extremely simple to implement using existing software such as R and Matlab with a sound Poisson regression solver. Specific choice of the working responses and weights for each model is listed below.

Initialize $\boldsymbol{B}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_{d_e}^{(0)})$;
repeat
  Update working responses $y_{ij}^{(t)}$ and weights $w_{ij}^{(t)}$, $i = 1, \ldots, n$, $j = 1, \ldots, d_e$;
  for $j = 1, \ldots, d_e$ **do**
    Update $\beta_j^{(t+1)}$ by solving a weighted Poisson regression with working responses $y_{ij}^{(t)}$, covariates $\boldsymbol{x}_i$, and working weights $w_{ij}^{(t)}$, $i = 1, \ldots, n$;
  end
until *log-likelihood* ℓ *converges*;

Algorithm 1: Iterative reweighted Poisson regression (IRPR) scheme for MLE. Working responses $y_{ij}^{(t)}$ and weights $w_{ij}^{(t)}$ for each model are specified in Section 4.2.

- Multinomial-Logit Regression (MN).
  Given $\boldsymbol{B}^{(t)} = (\boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_{d-1}^{(t)})$, we update $\boldsymbol{\beta}_j$, $j = 1, \ldots, d-1$, by solving a Poisson regression with working weights and responses

$$w_{ij}^{(t)} = \frac{m_i}{\sum_{j'=1}^{d} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_{j'}^{(t)}}}, \quad y_{ij}^{(t)} = \frac{y_{ij}}{w_{ij}^{(t)}}.$$

- Dirichlet-Multinomial Regression (DM).
  Given $\boldsymbol{B}^{(t)} = (\boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_d^{(t)})$, we update $\boldsymbol{\beta}_j$, $j = 1, \ldots, d$, by solving a Poisson regression with working weights and responses

$$w_{ij}^{(t)} = \sum_{k=0}^{m_i - 1} \left(\sum_{j'} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_{j'}^{(t)}} + k \right)^{-1},$$
$$y_{ij}^{(t)} = (w_{ij}^{(t)})^{-1} \left(\sum_{k=0}^{y_{ij}-1} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(t)}}}{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(t)}} + k} \right).$$

- Generalized Dirichlet-Multinomial Regression (GDM).
  Given $\boldsymbol{B}^{(t)} = (\boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_{d-1}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_{d-1}^{(t)})$, we update $\boldsymbol{\alpha}_j$, $j = 1, \ldots, d-1$, by solving a weighted Poisson regression with working weights and responses

$$w_{ij}^{(t)} = \sum_{k=0}^{z_{ij}-1} \left(e^{\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j^{(t)}} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(t)}} + k \right)^{-1},$$
$$y_{ij}^{(t)} = (w_{ij}^{(t)})^{-1} \left(\sum_{k=0}^{y_{ij}-1} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j^{(t)}}}{e^{\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j^{(t)}} + k} \right),$$

  and update $\boldsymbol{\beta}_j$, $j = 1, \ldots, d-1$, by solving a weighted Poisson regression with working weights and responses

$$w_{ij}^{(t)} = \sum_{k=0}^{z_{ij}-1} \left(e^{\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j^{(t)}} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(t)}} + k \right)^{-1},$$
$$y_{ij}^{(t)} = (w_{ij}^{(t)})^{-1} \left(\sum_{k=0}^{y_{i,j+1}-1} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(t)}}}{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(t)}} + k} \right).$$

- Negative Multinomial Regression (NM).
  Given $\boldsymbol{B}^{(t)} = (\boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_d^{(t)}, \boldsymbol{\beta}^{(t)})$, we update $\boldsymbol{\beta}$ by solving a weighted Poisson regression with working weights and responses

$$w_i^{(t)} = \ln \left(\sum_{j=1}^{d} e^{\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j^{(t)}} + 1 \right),$$
$$y_i^{(t)} = (w_i^{(t)})^{-1} \left(\sum_{k=0}^{m_i - 1} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)}}}{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)}} + k} \right).$$

  Updating $\boldsymbol{\alpha}_j$, $j = 1, \ldots, d$, is a weighted Poisson regression with working weights and responses

$$w_{ij}^{(t)} = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t+1)}} + m_i}{\sum_{j=1}^{d} e^{\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j^{(t)}} + 1}, \quad y_{ij}^{(t)} = \frac{y_{ij}}{w_{ij}^{(t)}}.$$

IRPR is derived in the online Supplementary Materials S.3. Most importantly, IRPR iterations always increase the log-likelihood and thus enjoy superior stability.

Lemma 2 (Monotonicity). The IRPR algorithmic iterates $\boldsymbol{B}^{(t)}$ satisfy $\ell(\boldsymbol{B}^{(t+1)}) \geq \ell(\boldsymbol{B}^{(t)})$ for $t \geq 0$.

Monotonicity and simplicity of IRPR are reminiscent of the celebrated expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). Derivation of the IRPR hinges upon the minorization–maximization (MM) principle (Lange, Hunter, and Yang 2000), a powerful generalization of the EM algorithm. The same principle has been successfully applied to distribution fitting (Zhou and Lange 2010; Zhou and Zhang 2012). Monotonicity of the algorithm does not guarantee the convergence of the iterates. The following proposition specifies conditions for the global convergence of the IPRP algorithm.

Proposition 1. Assume that (i) the design matrix $\boldsymbol{X}$ has full column rank, (ii) the log-likelihood ℓ is bounded above, (iii) the set $\{\boldsymbol{B} : \ell(\boldsymbol{B}) \geq \ell(\boldsymbol{B}^{(0)})\}$ is compact, and (iv) the set of stationary points of ℓ are isolated. Then, the IRPR iterates $\boldsymbol{B}^{(t)}$ converge to a stationary point of ℓ.

In practice, EM and MM algorithms may suffer from slow convergence. Their convergence rate is linear at best. We combine stability of IRPR and fast convergence of the Newton method by using a mix-and-match strategy. During each iteration, we choose either the IRPR or the Newton update that yields a higher log-likelihood. This strategy works very well in practice, as exemplified by the exceptional stability and competitive run times in Table 3. Another pleasant observation is that the d_e Poisson regressions within each iteration are separated, making IRPR amenable to parallel computing. This is a common feature of many EM or MM algorithms that are able to divide a high-dimensional optimization problem into many small ones (Zhou, Lange, and Suchard 2010).

5. Testing

Scientific applications often involve testing the significance of covariate(s). Although the regression models in Table 2, except for the multinomial, do not belong to the exponential family, they are regular in many statistical senses as their densities are quadratically mean differentiable (qmd) (Lehmann and Romano 2005). Hence, the standard asymptotic tests (LRT, score, Wald) apply.

With p covariates, each covariate involves d_e regression parameters, leading to a total of pd_e parameters. The likelihood ratio test (LRT) for two nested models is asymptotically distributed as a chi-square distribution with $d_e \Delta p$ degrees of freedom, where Δp is the difference in the numbers of covariates. To apply the LRT, optimizations have to be performed under the null and alternative hypotheses separately. The score test avoids computing the MLE at the alternative hypothesis and the Wald test avoids optimization at the null hypothesis.

6. Regularization

The asymptotics fail when the sample size n is only moderately larger than or even less than the number of regression parameters pd_e. In such cases, regularization is a powerful tool for reducing the variance of estimate and improving its mean squared error. In general, we consider regularized problem

$$\min_{\boldsymbol{B}} \; h(\boldsymbol{B}) = -\ell(\boldsymbol{B}) + J(\boldsymbol{B}), \tag{8}$$

where ℓ is the log-likelihood function and J is a regularization functional. Choice of J depends on specific applications.

- Different covariates may be associated with different category. Sparsity in terms of $\|\text{vec}\,\boldsymbol{B}\|_0$ is sought in this situation. For instance, in a sparse Dirichlet-multinomial regression model proposed by Chen and Li (2013), β_{kj} being nonzero indicates association of predictor k with bacterial taxon j. In general, elementwise shrinkage and sparsity can be achieved by the regularization term

$$J(\boldsymbol{B}) = \sum_{k \in \mathcal{S}} \sum_{j=1}^{d} P_\eta(|\beta_{kj}|, \lambda),$$

  where the set $\mathcal{S}$ indexes the covariates subject to regularization, $P_\eta(|x|, \lambda)$ is a scalar penalty function, λ is the penalty tuning parameter, and η is an index for the penalty family. Widely used penalties include: power family (Frank and Friedman 1993), where $P_\eta(|x|, \lambda) = \lambda|x|^\eta$, $\eta \in (0, 2]$, and in particular lasso (Tibshirani 1996) ($\eta = 1$) and ridge ($\eta = 2$); elastic net (Zou and Hastie 2005), where $P_\eta(|x|, \lambda) = \lambda[(\eta - 1)x^2/2 + (2 - \eta)|x|]$, $\eta \in [1, 2]$; SCAD (Fan and Li 2001), where $\partial/\partial|x| P_\eta(|x|, \lambda) = \lambda\{1_{\{|x| \leq \lambda\}} + [(\eta\lambda - |x|)_+/(\eta - 1)\lambda]1_{\{|x| > \lambda\}}\}$, $\eta > 2$; and MC+ penalty (Zhang 2010), where $P_\eta(|x|, \lambda) = \{\lambda|x| - x^2/(2\eta)\}1_{\{|x| < \eta\lambda\}} + 0.5\lambda^2\eta 1_{\{|x| \geq \eta\lambda\}}$, among many others.
- Predictor selection can be achieved by the group penalty (Yuan and Lin 2006; Meier, van de Geer, and Bühlmann 2008)

$$J(\boldsymbol{B}) = \lambda \sum_{k \in \mathcal{S}} \|\boldsymbol{\beta}_{[k]}\|_2,$$

  where λ is the penalty tuning constant, $\boldsymbol{\beta}_{[k]}$ is the vector of regression coefficients associated with the kth covariate, and $\|\boldsymbol{v}\|_2$ is the ℓ_2 norm of a vector $\boldsymbol{v}$. In other words, $\boldsymbol{\beta}_{[k]}$ is the kth row of the regression parameter matrix $\boldsymbol{B} \in \mathbb{R}^{p \times d_e}$.
- Sparsity at both the predictor level and within predictors can be achieved by the $\ell_{2,1}$ penalty (Zhao, Rocha, and Yu 2009)

$$J(\boldsymbol{B}) = \lambda \sum_{k \in \mathcal{S}} \left(\sum_{j=1}^{d_e} |\beta_{kj}| \right)^{1/2}.$$

- Shrinkage and sparsity in terms of the rank of $\boldsymbol{B}$ is achieved by regularization term

$$J(\boldsymbol{B}) = \lambda \|\boldsymbol{B}\|_* = \lambda \sum_i \sigma_i(\boldsymbol{B}),$$

  where the nuclear norm $\|\boldsymbol{B}\|_* = \sum_j \sigma_j(\boldsymbol{B})$, and $\sigma_j(\boldsymbol{B})$'s are the singular values of the matrix $\boldsymbol{B}$. The nuclear norm $\|\boldsymbol{B}\|_*$ is a suitable measure of the "size" of a matrix parameter, and is a convex relaxation of $\text{rank}(\boldsymbol{B}) = \|\sigma(\boldsymbol{B})\|_0$. This extends the reduced rank multivariate linear regression (Yuan et al. 2007) to multivariate GLM.

The nonsmooth minimization problem (8) is nontrivial. The MM principle underlying the IRPR algorithm in Section 4.2 is able to separate the columns of parameter matrix $\boldsymbol{B}$. Unfortunately both the group and nuclear norm regularization terms are not separable in columns. The success of the coordinate descent algorithm, which is efficient for ℓ_1 regularization in univariate GLM models (Friedman, Hastie, and Tibshirani

2010), suggests the block descent algorithm for minimizing (8). However, updating each block is a possibly nonconvex problem and nontrivial.

1 Initialize $S^{(1)} = B^{(0)} = B^{(1)}$, $\delta > 0$, $\alpha^{(0)} = 0$, $\alpha^{(1)} = 1$;
2 **repeat**
3 **repeat**
4 $A_{\text{temp}} \leftarrow S^{(t)} + \delta \nabla \ell(S^{(t)})$;
5 $B_{\text{temp}} \leftarrow \text{argmin}(2\delta)^{-1}\|B - A_{\text{temp}}\|_{\text{F}}^2 + J(B)$;
6 $\delta \leftarrow \delta/2$;
7 **until** $h(B_{\text{temp}}) \le g(B_{\text{temp}} \mid S^{(t)}, \delta)$;
8 $\alpha^{(t+1)} \leftarrow (1 + \sqrt{1 + (2\alpha^{(t)})^2})/2$;
9 **if** $h(B_{\text{temp}}) \le h(B^{(t)})$ **then**
10 $B^{(t+1)} \leftarrow B_{\text{temp}}$;
11 $S^{(t+1)} \leftarrow B^{(t+1)} + \left(\frac{\alpha^{(t)}-1}{\alpha^{(t+1)}}\right)(B^{(t+1)} - B^{(t)})$;
12 **else**
13 $B^{(t+1)} \leftarrow B^{(t)}$;
14 $S^{(t+1)} \leftarrow B_{\text{temp}} + \frac{\alpha^{(t)}-1}{\alpha^{(t+1)}}(B^{(t)} - B_{\text{temp}})$;
15 **end**
16 **until** *objective value converges*;

Algorithm 2: Accelerated proximal gradient method for regularized estimation (8).

We use the accelerated proximal gradient method that has been successful in solving various regularization problems (Beck and Teboulle 2009). The accelerated proximal gradient algorithm as summarized in Algorithm 2 consists of two steps per iteration: (a) predicting a search point S based on the previous two iterates (line 2) and (b) performing gradient descent from the search point S, possibly with a line search (lines 2-2). We first describe step (b). The gradient descent step effectively minimizes the surrogate function

$$\begin{aligned} g(B|S^{(t)}, \delta) &= -\ell(S^{(t)}) - \langle \nabla \ell(S^{(t)}), B - S^{(t)} \rangle \\ &\quad + \frac{1}{2\delta}\|B - S^{(t)}\|_{\text{F}}^2 + J(B) \\ &= \frac{1}{2\delta}\|B - [S^{(t)} + \delta \nabla \ell(S^{(t)})]\|_{\text{F}}^2 + J(B) + c^{(t)}, \quad (9) \end{aligned}$$

where the constant $c^{(t)}$ collects terms irrelevant to the optimization. Here, we abuse the notation to use $\nabla \ell(S^{(t)}) \in \mathbb{R}^{p \times d_e}$ to denote the *matrix* of first derivatives $\partial \ell / \partial s_{kj}^{(t)}$. These gradients are given in Propositions S.1–S.3 of supplementary materials as the score functions for the models considered. The ridge term $(2\delta)^{-1}\|B - S^{(t)}\|_{\text{F}}^2$ in the surrogate function (9) shrinks the next iterate toward $S^{(t)}$, which is desirable since the first-order approximation is good only within the neighborhood of current search point. Minimizing the surrogate function $g(B|S^{(t)}, \delta)$ is achieved by simple thresholding. Let $A_{\text{temp}} = S^{(t)} + \delta \nabla \ell(S^{(t)})$ be the intermediate matrix with rows $a_{[k]}^{(t)}$. The minimizer of g, denoted by B_{temp}, for various J are listed below (line 2).

- Lasso penalty $J(B) = \lambda \|\text{vec}(B)\|_1$. B_{temp} has entries $(1 - \delta\lambda/|a_{ij}|)_+ a_{ij}$.
- Group penalty $J(B) = \lambda \sum_k \|\beta_{[k]}\|_2$. The rows of B_{temp} are given by $(1 - \delta\lambda/\|a_{[k]}\|_2)_+ a_{[k]}$.
- $\ell_{2,1}$ penalty does not have analytic solution for the minimizer g. B_{temp} can be solved by weighted lasso regression (Xu et al. 2010).
- Nuclear norm $J(B) = \lambda \|B\|_*$. Suppose A_{temp} admits singular value decomposition $U\text{diag}(a)V^{\top}$. Then $B_{\text{temp}} = U\text{diag}[(a - \delta\lambda)_+]V$.

Suppose the loss $-\ell(B)$ has gradient Lipschitz constant $\mathcal{L}$, that is, $\|\nabla \ell(B_1) - \nabla \ell(B_2)\| \le \mathcal{L}\|B_1 - B_2\|_{\text{F}}$ for all B_1, B_2. Then, we can fix $\delta = \mathcal{L}^{-1}$ and the line search described in Algorithm 2 terminates in a single step. Using a larger δ leads to a bigger gradient descent step (line 2), which sometimes must be contracted to send the penalized loss downhill.

In Step (a) of Algorithm 2. The search point S is found by an extrapolation based on the previous two iterates $B^{(t)}$ and $B^{(t-1)}$. This trick accelerates ordinary gradient descent by making this extrapolation. Without extrapolation, Nesterov's method collapses to a gradient method with the slow nonasymptotic convergence rate of $O(k^{-1})$ rather than $O(k^{-2})$.

Assume the loss $-\ell$ has gradient Lipschitz constant $\mathcal{L}(\ell)$. For convex loss such as in the multinomial-logit model, if the global minimum of the penalized loss $h(B)$ occurs at the point B^*, then the following nonasymptotic bound for the convergence of the objective values

$$h(B^{(t)}) - h(B^*) \le \frac{4\mathcal{L}(f)\|B^{(0)} - B^*\|_{\text{F}}^2}{(t+1)^2}$$

applies (Beck and Teboulle 2009). For nonconvex losses such as in the DM, NM and GDM regressions, convergence theory is hard. In general, it is only guaranteed that $\|B^{(t+1)} - B^{(t)}\|_{\text{F}}$ converges to 0. In practice, the algorithm almost always converges to at least a local minimum of the objective function.

7. Numerical Examples

We conducted extensive simulation studies to assess the finite sample performance of four regression models. Specifically, we demonstrate that misspecification of the model can cause serious errors in hypothesis testing and variable selection. This is of practical importance as practitioners routinely rely on the multinomial-logit (MN) model to analyze data with multiple categorical responses.

Table 4. Empirical rejection rates of the Wald test by the multinomial (MN), Dirichlet-multinomial (DM), generalized Dirichlet-multinomial (GDM), and negative multinomial (NM) regression models, based on 300 simulation replicates. The column for $\alpha_0 = 0$ corresponds to empirical Type I error rates and the other columns correspond to empirical power. Responses are generated from the generalized Dirichlet-multinomial (GDM) model. All standard errors are less than 0.02 thus not displayed here.

		Effect size (α_0)						
n	Model	0	0.05	0.1	0.5	1	2	5
50	MN	0.99	0.98	0.99	0.99	1.00	1.00	1.00
	DM	0.63	0.59	0.61	0.97	1.00	1.00	1.00
	GDM	0.08	0.10	0.10	0.84	1.00	1.00	1.00
	NM	0.99	0.98	0.99	0.99	1.00	1.00	1.00
100	MN	0.99	0.99	0.99	0.99	1.00	1.00	1.00
	DM	0.58	0.52	0.63	1.00	1.00	1.00	1.00
	GDM	0.05	0.07	0.15	1.00	1.00	1.00	1.00
	NM	0.99	0.99	1.00	0.97	1.00	1.00	0.99
200	MN	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	DM	0.48	0.57	0.66	1.00	1.00	1.00	1.00
	GDM	0.05	0.10	0.20	1.00	1.00	1.00	1.00
	NM	0.98	0.99	0.99	1.00	0.99	0.98	1.00
500	MN	0.99	1.00	0.99	1.00	0.99	0.99	1.00
	DM	0.45	0.62	0.90	1.00	1.00	1.00	1.00
	GDM	0.05	0.14	0.56	1.00	1.00	1.00	1.00
	NM	1.00	0.99	1.00	0.99	1.00	1.00	1.00

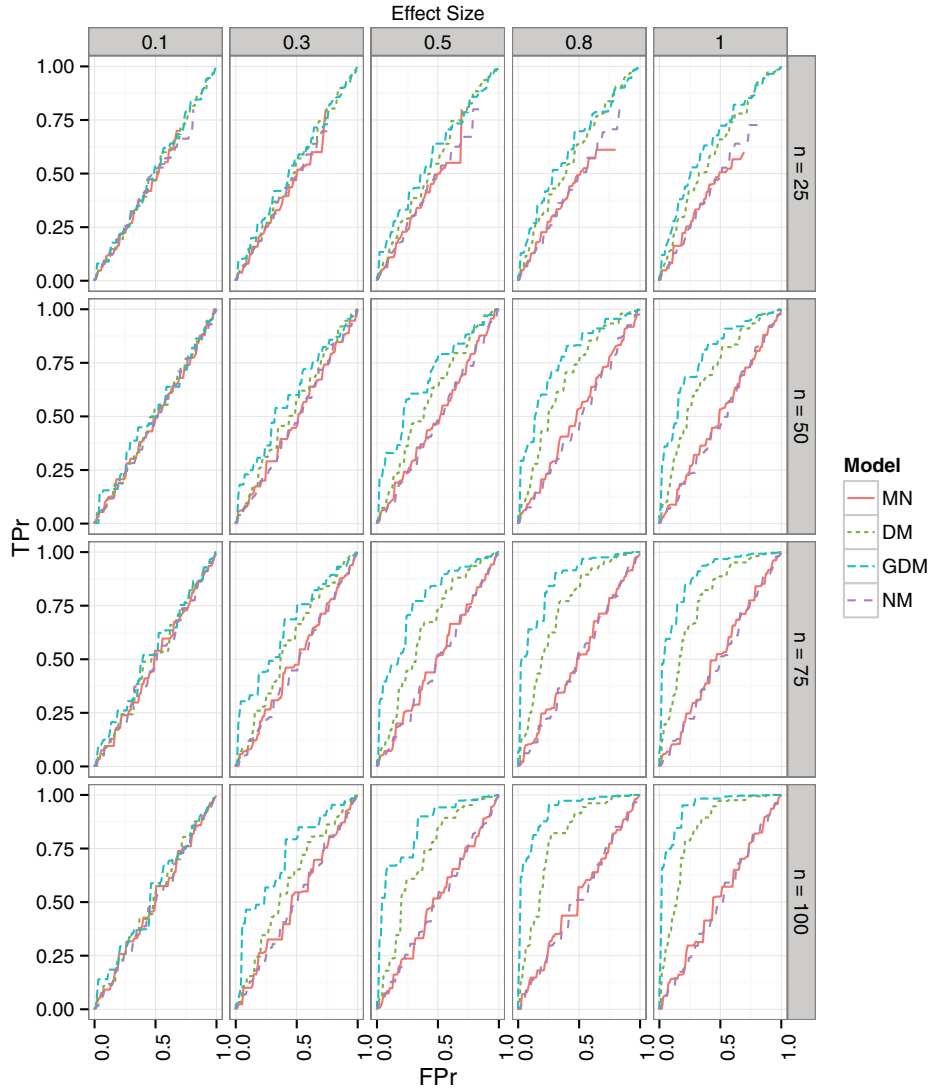

Figure 2. ROC curves from the group penalized estimation by the multinomial (MN), Dirichlet-multinomial (DM), and generalized Dirichlet-multinomial (GDM) regression models. ROC curves are summarized from 300 simulation replicates. Responses are generated from the generalized Dirichlet-multinomial (GDM) model.

7.1 Hypothesis Testing

We generate polytomous responses from each of the four models, MN, DM, GDM, and NM, and then fit the data with all four regression models.

In the generative model, there are $d = 5$ categories and $p = 6$ predictors. The first three predictors have nonzero effects size α_0 and the last three are null predictors. Therefore, the true parameter matrix is $\boldsymbol{B} = [\alpha_0, \alpha_0, \alpha_0, 0, 0, 0]^{\mathsf{T}} \mathbf{1}_{d_e}^{\mathsf{T}} \in \mathbb{R}^{p \times d_e}$. The number of parameters to estimate is 24, 30, 48, and 31 for MN, DM, GDM, and NM, respectively. Entries of the covariate matrix $\boldsymbol{X}$ are generated from independent standard normal. We vary the effect size α_0 at values 0, 0.05, 0.1, 0.5, 1, 2, and 5 and the sample size n at 50, 100, 200, and 500. The batch size m_i of MN, DM, GDM response vectors are generated from Binomial(200, 0.8). We simulate 300 replicates at each combination of effect size and sample size. Empirical Type I error and power of the Wald test for testing the significance of the first predictor are reported.

Table 4 shows the results when responses are generated from the GDM model. The fact that using a wrong model, MN, DM, or NM in this case, causes highly inflated Type I error is cautionary, as practitioners routinely rely on the multinomial-logit model to analyze count data. Similar patterns are observed when the responses are generated from the MN, DM, or NM models. Their results are presented in Tables 1–3 of the supplementary materials.

7.2 Variable Selection by Regularization

The simulation design for sparse regression is similar to the previous section, except that the response matrix has $d = 10$ categories and there are $p = 100$ predictors. Only the first 5 predictors are associated with the responses with effect size $\alpha_0 \in \{0.1, 0.3, 0.5, 0.8, 1\}$. Sample sizes are $n = 25, 50, 75, 100$. Three hundred replicates are simulated at each combination of effect size and sample size. For each data replicate, we perform predictor selection by fitting the group penalized estimation.

The variable selection performance is summarized by the receiver operating characteristic (ROC) curves that result from the solution path. At each tuning parameter value λ, we calculate the true positive rate (TPr) and false positive rate (FPr) of

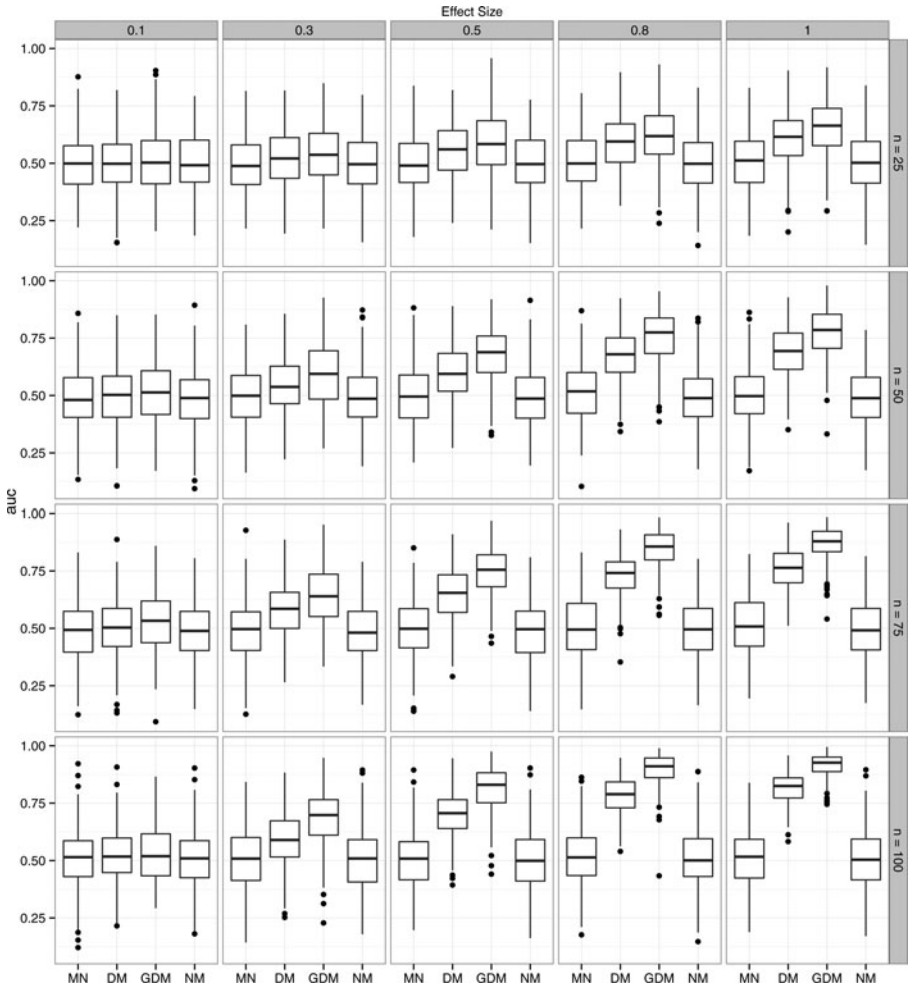

Figure 3. Box plots of AUCs from the group penalized estimation by the multinomial (MN), Dirichlet-multinomial (DM), generalized Dirichlet-multinomial (GDM) and negative multinomial (NM) regression models, based on 300 simulation replicates. Responses are generated from the GDM model.

the regularized estimate $\widehat{\boldsymbol{B}}(\lambda)$:

$$\text{True positive rate (TPr)} = \frac{\#\text{ true positives}}{\#\text{ true positive} + \#\text{ false negative}}$$

$$\text{False positive rate (FPr)} = \frac{\#\text{ false positive}}{\#\text{ false positive} + \#\text{ true negative}}.$$

At $\lambda = \infty$, TPr and FPr are 0. Both increase as λ decreases, approaching 1 at $\lambda = 0$. Thus, each solution path produces an ROC curve. For each regression model, the 300 ROC curves are summarized by an average ROC curve obtained from fitting an isotonic regression.

Figure 2 displays the summary ROC curves of the four regression models, when the responses are generated from the

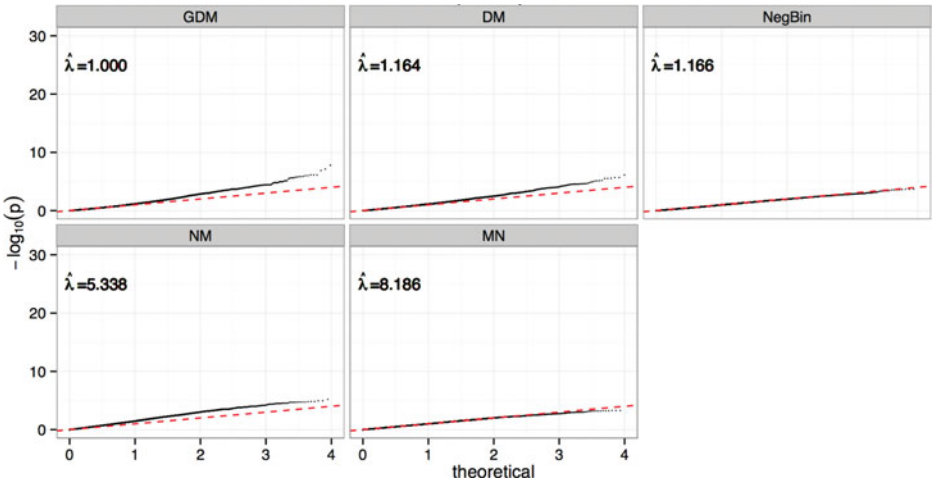

Figure 4. QQ plots of eQTL analysis of ST13P6 using MN, DM, GDM, and NM regressions. $\hat{\lambda}$'s are the estimated genomic control (GC) inflation factor.

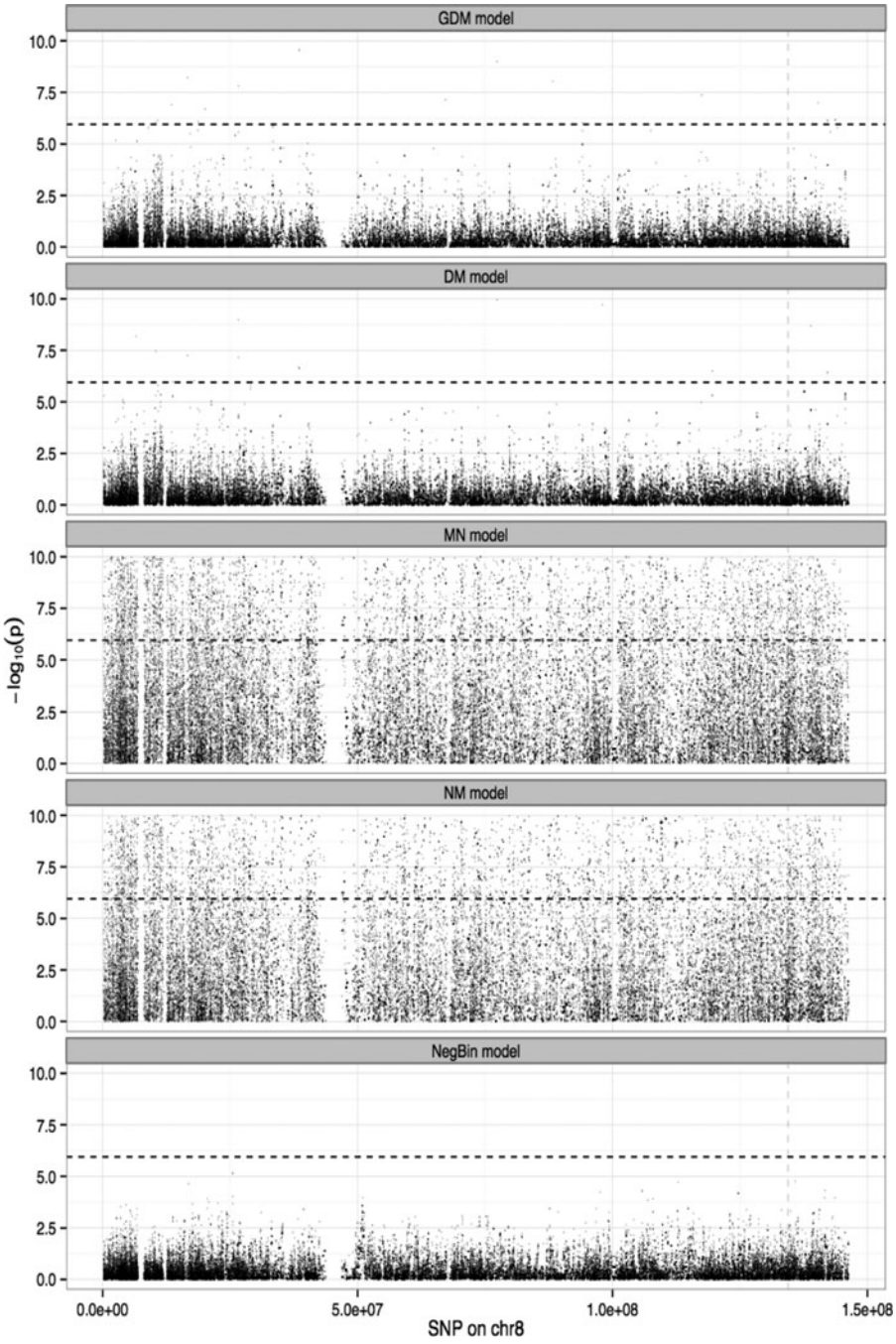

Figure 5. Manhattan plots of GDM, DM, MN, NM, and NegBin regressions for eQTL analysis of ST13P6. Dashed lines are chromosome-8-wide significance level.

generalized Dirichlet-multinomial distribution (GDM). GDM regression demonstrates the best variable selection performance, followed by the DM regression. The multinomial (MN) and negative multinomial (NM) model have little discriminatory power at various effect sizes and sample sizes. To appreciate the Monte Carlo errors, we also calculate the area under curve (AUC) of the ROC curve from each replicate. Larger AUC indicates better variable selection performance. The box plots of AUCs are displayed in Figure 3. Again, the correct model GDM shows superior performance, while the MN and NM models are no better than random guess.

The supplementary materials contain the corresponding summery ROC curves and box plots for AUCs when data were generated from multinomial (MN), Dirichlet-multinomial (DM), and negative multinomial (NM) models. In general, we observe a similar pattern: using an incorrectly specified model leads to poor performance in variable selection.

7.3 Real Data

We apply the regression models to an RNA-seq dataset of 60 HapMap CEU samples (Montgomery et al. 2010). Genotype data are also available for these samples and we aim to assess the association between the genotype of each SNP (single nucleotide polymorphism) with the counts of multiple exon sets of each gene.

We demonstrate results by gene ST13P6 (suppression of tumorigenicity 13 pseudogene 6), which has been related to

B-Cell leukemia, multiple myeloma, and prostate cancer in previous studies (Sossey-Alaoui et al. 2002). ST13P6 has three exon sets and is located on chromosome 8. Expression counts of the 3 exon sets are regressed on covariates, namely three principle components, `log(TotalReads)`, and each of the 45,587 SNPs on chromosome 8, using the four models in Table 2. The total expression counts are also regressed on the same covariates using the negative binomial (NegBin) model. Thus, each model generates 45,587 p-values, which are summarized by the QQ plot and Manhattan plot. In reality, association of expression levels and SNPs is rare. Therefore, we expect none or at most a few SNPs that are significant after adjusting for multiple testing.

Figure 4 shows the QQ plots of the quantiles of the observed $-\log_{10}(p\text{-value})$ versus the theoretical quantiles under the null hypothesis of no association. Departure from the theoretical quantiles implies systematic bias in the data or statistical methods (Laird and Lange 2011). Specifically bending upward indicates there are too many false positives. Also reported is the genomic control (GC) inflation factor λ (Laird and Lange 2011), which is the ratio of the median of the observed test statistics to the median of chi-square distribution with d_e degrees of freedom. $\lambda > 1$ indicates inflation of Type I error. QQ plots and inflation factors in Figure 4 show serious inflation of type I errors under MN and NM models and moderately inflated Type I error under DM and NegBin.

Manhattan plots in Figure 5 plot $-\log_{10}(p\text{-value})$ of each SNP under different models (Ziegler and König 2006). The dashed lines in Manhattan plots indicate the chromosome-wide significance level after Bonferroni correction for multiple testing. MN and NM have numerous SNPs that pass the chromosome-wide significance level, indicating inflated Type I error. GDM and DM seem to have well-controlled Type I error and identify some signals on chromosome 8, while NegBin identifies none. Table 8 of the supplementary materials tabulates the names, positions, minor allele frequency (MAF), and functional annotation of the identified SNPs under the GDM model. Most of the detected SNPs are located in or close to the candidate genes for obesity, cardiovascular diseases, or cancers.

It is hard to draw conclusions based on a sample size of 60. However, the results seem to conform with our findings in the simulation study in Section 2: (1) choosing a limiting multivariate count model such as multinomial (MN) and negative multinomial (NM) models may inflate Type I errors and (2) collapsing counts by simple sums such as in the negative binomial (NegBin) model compromises power for detecting differential expression profiles.

8. Discussion

We have investigated GLMs with multivariate categorical responses that, compared to the multinomial-logit model, admit more flexible correlation structures among counts. The RNA-seq simulation example exposes the limitation of the widely used multinomial-logit model. Then, we examine three more flexible models for count data: Dirichlet-multinomial, generalized Dirichlet-multinomial, and negative multinomial. Although they do not belong to the exponential family, we show that MLE and regularized estimation can be treated in a unifying framework. The IRPR scheme for MLE is stable, efficient, simple to implement, and enjoys favorable convergence properties. The accelerated proximal gradient algorithm for regularized estimation incorporates various penalties, permitting variable selection, low rank regularization, and entrywise selection arising from different applications. These regression models provide practitioners more flexible tools for analyzing complex, multivariate count data.

The MLE, testing, and regularized estimation for all four models in Table 2 are implemented in the R package `mglm`, which is available on CRAN, and a Matlab toolbox, which is available at *http://hua-zhou.github.io/softwares/mglm/*. We refer readers to the companion article (Zhang and Zhou 2015) for more implementation and usage details.

Supplementary Materials

Supplementary materials contain technical details of proofs, extra simulation results, and detailed real data analysis results.

Acknowledgments

The work is partially supported by National Science Foundation (NSF) grant DMS-1310319 and National Institutes of Health (NIH) grants HG006139, GM105785, and GM53275.

References

Anders, S., and Huber, W. (2010), "Differential Expression Analysis for Sequence Count Data," *Genome Biology*, 11, R106. [2]

Baker, S.G. (1994), "The Multinomial-Poisson Transformation," *Journal of the Royal Statistical Society*, Series D, 43, 495–504. [3]

Beck, A., and Teboulle, M. (2009), "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, 2, 183–202. [8]

Chen, J., and Li, H. (2013), "Variable Selection for Sparse Dirichlet-Multinomial Regression With an Application to Microbiome Data Analysis," *The Annals of Applied Statistics*, 7, 418–442. [7]

Connor, R.J. , and Mosimann, J.E. (1969), "Concepts of Independence for Proportions With a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 64, 194–206. [4]

Dempster, A.P. , Laird, N.M. , and Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [7]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [7]

Frank, I.E. , and Friedman, J.H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [7]

Friedman, J.H. , Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [8]

Laird, N.M. , and Lange, C. (2011), *The Fundamentals of Modern Statistical Genetics*, Statistics for Biology and Health, New York: Springer. [12]

Lange, K., Hunter, D.R. , and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions" (with discussion, and a rejoinder by Hunter and Lange), *Journal of Computational and Graphical Statistics*, 9, 1–59. [7]

Lehmann, E.L. , and Romano, J.P. (2005), *Testing Statistical Hypotheses* (3rd ed.), Springer Texts in Statistics, New York: Springer. [7]

McCullagh, P., and Nelder, J.A. (1983), *Generalized Linear Models*, Monographs on Statistics and Applied Probability, London: Chapman & Hall. [1,2]

Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society*, Series B, 70, 53–71. [7]

Montgomery, S., Sammeth, M., Gutierrez-Arcelus, M., Lach, R., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. (2010), "Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population," *Nature*, 464, 773–777. [11]

Mosimann, J.E. (1962), "On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions," *Biometrika*, 49, 65–82. [3]

Nocedal, J., and Wright, S.J. (2006), *Numerical Optimization* (2nd ed.), Springer Series in Operations Research and Financial Engineering, New York: Springer. [5]

Robinson, M.D. , McCarthy, D.J. , and Smyth, G.K. (2010), "edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data," *Bioinformatics*, 26, 139–140. [2]

Sossey-Alaoui, K., Kitamura, E., Head, K., and Cowell, J.K. (2002), "Characterization of FAM10A4, a Member of the ST13 Tumor Suppressor Gene Family That Maps to the 13q14.3 Region Associated With B-Cell Leukemia, Multiple Myeloma, and Prostate Cancer," *Genomics*, 80, 5–7. [12]

Sun, W., Liu, Y., Crowley, J.J. , Chen, T.-H., Zhou, H., Chu, H., Huang, S., Kuan, P.-F., Li, Y., Miller, D., Shaw, G., Wu, Y., Zhabotynsky, V., McMillan, L., Zou, F., Sullivan, P.F. , and Pardo-Manuel de Villena, F. (2015), "IsoDOT Detects Differential RNA-isoform Usage With Respect to a Categorical or Continuous Covariate With High Sensitivity and Specificity," *Journal of the American Statistical Association*, 110, 975–986. [1,2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [7]

Wang, Z., Gerstein, M., and Snyder, M. (2009), "RNA-Seq: A Revolutionary Tool for Transcriptomics," *Nature Reviews Genetics*, 10, 57–63. [1]

Xu, Z., Zhang, H., Wang, Y., Chang, X., and Liang, Y. (2010), "$L_{1,2}$ Regularization," *Science China Information Sciences*, 53, 1159–1169. [8]

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression," *Journal of the Royal Statistical Society*, Series B, 69, 329–346. [7]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [7]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *Annals of Statistics*, 38, 894–942. [7]

Zhang, Y., and Zhou, H. (2015), "MGLM: R Package and Matlab Toolbox for Multivariate Categorical Data Analysis," under revision. [1,12]

Zhao, P., Rocha, G., and Yu, B. (2009), "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *Annals of Statistics*, 37, 3468–3497. [7]

Zhou, H., and Lange, K. (2010), "MM Algorithms for Some Discrete Multivariate Distributions," *Journal of Computational and Graphical Statistics*, 19, 645–665. [7]

Zhou, H., Lange, K., and Suchard, M. (2010), "Graphical Processing Units and High-Dimensional Optimization," *Statistical Science*, 25, 311–324. [7]

Zhou, H., and Zhang, Y. (2012), "EM vs. MM: A Case Study," *Computational Statistics & Data Analysis*, 56, 3909–3920. [7]

Ziegler, A., and König, I.R. (2006), *A Statistical Approach to Genetic Epidemiology*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, concepts and applications, With a foreword by Robert C. Elston. [12]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [7]