Reducing Response Categories in Multinomial Logistic Regression

Brad Price

University of Miami
Department of Management Science

April 2, 2015

Joint work with Adam Rothman and Charles Geyer (University of Minnesota School of Statistics)

Multinomial Logistic Regression

Let $x_i = (1, x_{i1}, \ldots, x_{ip})^T$, where $x_{i1}, \ldots, x_{ip}$ are the values of the p predictors ($i = 1, \ldots, N$)

Let $y_i = (y_{i1}, \ldots, y_{iC})$ be a vector of C category counts resulting from n_i independent multinomial trials that result in one of C categories ($i = 1, \ldots, N$)

y_i is a realization of $Y_i \sim \text{Multinom}\,(n_i, \pi_1(x_i), \ldots, \pi_C(x_i))$ where,

$$\pi_c(x_i) = \frac{\exp(x_i^T \beta_c)}{\sum_{m \in \mathcal{C}} \exp(x_i^T \beta_m)}, \; c \in \mathcal{C}$$

$Y_1, \ldots, Y_N$ are independent random vectors

Baseline Category Parameterization

To make the model identifiable we set $\beta_C = \vec{0}$

We call response category C the baseline category

$$\log\left(\frac{\pi_c(x_i)}{\pi_C(x_i)}\right) = x_i^T \beta_c$$

To compare response category c and m

$$\log\left(\frac{\pi_c(x_i)}{\pi_m(x_i)}\right) = x_i^T (\beta_c - \beta_m)$$

Group Fused Multinomial Logistic Regression

Goal: Reduce the number of response categories by minimizing

$$-\sum_{i=1}^{N}\left(\sum_{c\in\mathcal{C}} y_{ic}x_i^T\beta_c - n_i\log(\sum_{r\in\mathcal{C}}\exp\{x_i^T\beta_r\})\right) + \lambda\sum_{(m,c)\in\mathcal{C}\times\mathcal{C}}|\beta_c-\beta_m|_2$$

- $\sum_{(m,c)\in\mathcal{C}\times\mathcal{C}}|\beta_c-\beta_m|_2$ is the *group fused penalty*
- Why use the group fused penalty?

Good Question

Why use the group fused penalty?

- Promotes vector-wise similarity of the β's
- If $\beta_c = \beta_m$ then $\pi_c(x) = \pi_m(x)$
- Since the probabilities of an observation coming from category c and category m are always the same we combine the category

Reformulation

We reformulate the penalized negative log-likelihood as

$$-\sum_{i=1}^{N}\left(\sum_{c\in\mathcal{C}} y_{ic}x_i^T\beta_c - n_i\log(\sum_{r\in\mathcal{C}}\exp\{x_i^T\beta_r\})\right)+\lambda\sum_{(c,m)\in\mathcal{C}\times\mathcal{C}}|Z_{cm}|_2$$

where $Z_{cm}=\beta_c-\beta_m$ for all $c,m\in\mathcal{C}$

This reformulation allows us to use the Alternating Direction Method of Multipliers (ADMM) Algorithm

The ADMM Algorithm

The ADMM algorithm minimizes the penalized negative log-likelihood

$$-\sum_{i=1}^{N}\left(\sum_{c\in\mathcal{C}} y_{ic}x_i^T\beta_c - n_i\log(\sum_{r\in\mathcal{C}}\exp\{x_i^T\beta_r\})\right)+\lambda\sum_{(c,m)\in\mathcal{C}\times\mathcal{C}}|Z_{cm}|_2$$

with respect to β and Z subject to the constraint that $Z_{cm}=\beta_c-\beta_m$

- Developed in the 1970's
- Combines dual ascent and method of multipliers algorithm
- Great review of statistical applications in Foundations and Trends in Machine Learning Research Boyd et al (2011)

Iterative Procedure

The scaled augment Lagrangian is

$$-\sum_{i=1}^{N}\left(\sum_{c\in\mathcal{C}} y_{ic}x_i^T\beta_c - n_i\log(\sum_{r\in\mathcal{C}}\exp\{x_i^T\beta_r\})\right)$$
$$+\sum_{(c,m)\in\mathcal{C}\times\mathcal{C}}\left(\lambda|Z_{cm}|_2+\frac{\rho}{2}|\beta_c-\beta_m-Z_{cm}+U_{cm}|_2^2\right)$$

Minimize w.r.t β

- Ridge fusion penalized multinomial logistic regression
- We use a coordinate descent method that uses Newton-Raphson method

Minimize w.r.t. Z

- Analogous to group penalized least squares solution

Update U with

$$U_{cm}^{(k+1)} = U_{cm}^{(k)} + \widehat{\beta}_c - \widehat{\beta}_m - \widehat{Z}_{cm}$$

Algorithm Convergence

Theorem

The ADMM Algorithm that solves group fused multinomial logistic regression converges to the optimal objective function value, converges to the optimal values of (β, Z)*, and the dual variable converges to the optimal dual variable.*

Computational Issues

Let $\widehat{\beta}$, $\widehat{Z}$ and $\widehat{U}$, be the solutions found using the ADMM algorithm

Ridge fusion penalized solution

- $\widehat{\beta}$ never completely fused
- Use $\widehat{Z}$ as an indicator of the categories that should be combined

What happens if not all pairs of categories are penalized?

- Size of Z and U change
- Algorithm converges under certain regularity conditions
- Adaptive Penalties

Tuning Parameter Selection

Combining Categories

The estimates produce a new *group structure*
$\widehat{\mathcal{G}} = (\hat{g}_1, \ldots, \hat{g}_G)$, $G < C$

- $\widehat{\mathcal{G}}$ is a partition of the set of response categories
- If $(c, m) \in \hat{g}_j$ then $\widehat{\beta}_m = \widehat{\beta}_c$

Response categories that are in the same group are combined, we call it $\tilde{y}$

Tuning Parameter Selection

Tuning parameter selection in this problem is equivalent to selecting the group structure

Two Step Procedure

- For a given λ use group fused multinomial logistic regression to find the estimated group structure $\widehat{\mathcal{G}}_\lambda$
- Refit the model using the reduced response categories given by $\widehat{\mathcal{G}}_\lambda$ we will call these estimates $\widehat{\eta}_\lambda$

Need to compare models with a different number of response categories

Comparing Models with different number of response categories

Exploit the fact fusion of two categories means the probabilities are equal for every value of the predictors

In the reduced category model $\tilde{y}_i = (\tilde{y}_{ig_1}, \ldots, \tilde{y}_{ig_G})$ is a realization of the distribution

$$\tilde{Y}_i \sim \text{Multinom}(n_i, \theta_{g_1}(x_i), \ldots, \theta_{g_G}(x_i))$$

$$AIC(\eta_{\widehat{\mathcal{G}}}) = -2\left(l_{\widehat{\mathcal{G}}}(\tilde{\theta}) - \sum_{j=1}^{G} n_{g_j} \log(\text{card}(g_j))\right) + 2(p+1)(G-1),$$

- $\tilde{\theta}$ are the estimated probabilities associated with $\widehat{\eta}_\lambda$
- $l_{\widehat{\mathcal{G}}}(\tilde{\theta})$ is the likelihood generated from the reduced categories indicated by $\widehat{\mathcal{G}}$

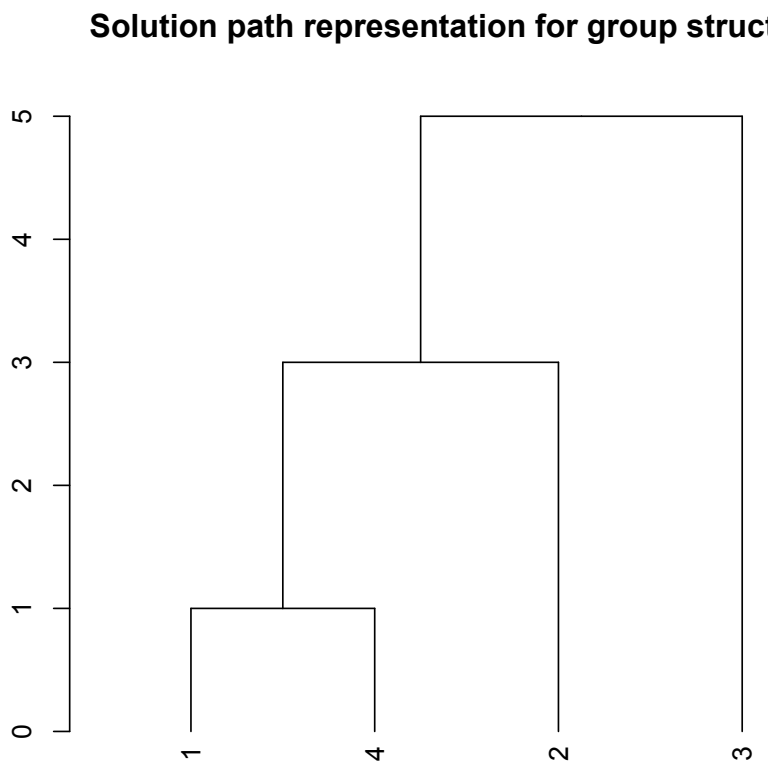
Selecting group structures for comparison

Different values of λ will produce different $\widehat{\mathcal{G}}$

Results in a set of candidate models

Candidate Models Example

Example where $C = 4$, $G = 3$, $g_1 = \{1, 4\}$, $g_2 = \{2\}$, $g_3 = \{3\}$



How to select the group structure

Use the line search on the solution path produced by group fused multinomial regression to find the group of candidate group structures

Refit the multinomial logistic regression models using the combined categories indicated by the estimated group structures

Use AIC to select the model

Simulation Setup

Evaluation on the group structure AIC selects when compared to the data generating model

Report the fraction of replications that return group structures of interest

$x_1, \ldots, x_N$ are generated from a $N_9(0, I)$

$\tilde{x}_i = (1, x_i)^T$

y_i is a realization of Multinom$(1, \pi_1(\tilde{x}_i), \ldots, \pi_4(\tilde{x}_i))$ where

$$\pi_c(x_i) = \frac{\exp(\tilde{x}_i^T \beta_c)}{\sum_{r=1}^{4} \exp(\tilde{x}_i^T \beta_r)}$$

100 replications of each setting with category 4 used as the baseline

Simulation 1

4 category problem where there are 2 groups

Group 1: $\beta_1 = -\vec{\delta}$

Group 2: $\beta_2 = \beta_3 = \beta_4 = \vec{0}$

Investigated the cases there $N = 50, 75$ and $\delta = 1$ and 3

Simulation 1 Results: Our Method

	$N = 50$		$N = 75$	
	$\delta = 1$	$\delta = 3$	$\delta = 1$	$\delta = 3$
1 Group	19/100	0/100	3/100	0/100
2 Groups (Correct)	58/100	71/100	80/100	83/100
2 Groups (Incorrect)	0/100	0/100	0/100	0/100
3 Groups (One-Step)	20/100	21/100	16/100	15/100
3 Groups (Incorrect)	0/100	0/100	0/100	0/100
4 Groups	3/100	8/100	1/100	2/100

For each N and δ the correct group structure is chosen the most

One-Step indicates that a partially correct group structure was found

Simulation 1 Results: Exhaustive Search

	$N = 50$		$N = 75$	
	$\delta = 1$	$\delta = 3$	$\delta = 1$	$\delta = 3$
1 Group	0/100	0/100	0/100	0/100
2 Groups (Correct)	46/100	59/100	62/100	82/100
2 Groups (Incorrect)	25/100	4/100	14/100	0/100
3 Groups (One-Step)	28/100	29/100	24/100	17/100
3 Groups (Incorrect)	0/100	0/100	0/100	0/100
4 Groups	1/100	8/100	0/100	1/100

Our method selects the correct group structure more often than the exhaustive search

The exhaustive search never selects one group, and is competitive when $N = 75$ and $\delta = 3$

Simulation 2

4 category problem with 3 groups

Group 1: $\beta_1 = \beta_4 = \vec{0}$

Group 2: $\beta_2 = -\vec{\delta}$

Group 3: $\beta_3 = \vec{\delta}$

Investigated the cases of $N = 50$, $\delta = 2$ and 3

Simulation 2 Results

	$\delta = 2$	$\delta = 3$
1 Group	0/100	0/100
2 Groups	0/100	0/100
3 Groups (Correct)	93/100	98/100
3 Groups (Incorrect)	0/100	0/100
4 Groups	7/100	2/100

For both values of δ the correct group structure is selected with the highest proportion

In the 100 replications for both values of δ 1 or 2 groups is never chosen

Agrees perfectly with exhaustive search

1996 Election Data

Understand a self identification of political party based of 944 voters based on education (7 levels), income (24 levels), and age (continuous)

Response categories are political party

- Strong, weak, independent democrat
- Strong, weak, independent republican
- Independent

Fit ordered and unordered response model

- Ordered response respects the relationship of the categories
- Unordered allows for any combinations of categories to be fused
- Exhaustive search also used

1996 Election Data: Results

Group	Unordered Responses	Ordered Responses
1	Strong Republican Weak Republican Independent Republican Independent Democrat	Strong Republican Weak Republican
2	Independent	Independent Republican Independent Independent Democrat
3	Strong Democrat Weak Democrat	Strong Democrat Weak Democrat

Exhaustive search agrees with the ordered response model, and use of model in Faraway (2002) Unordered response

model fits political science models

What did we do?

Propose group fusion penalty to reduce response categories in multinomial logistic regression

Propose an ADMM alogrithm with convergence properties based on minimal constraints

- Propose an AIC to compare multinomial logistic regression models with combined categories