TITLE

Suyoung Park[*] and Daniel Eck
University of Illinois at Urbana-Champaign

Month 2020

Abstract

Space for the Abstract.

Key Words: list of keywords

1 Introduction

1.1 Background and Literature

Papers we are going to discuss:
1. *Heinze, 2002, logistf*
2. *Kosmidis, 2007, brglm2*
3. *Gelman, 2008, arm (bayesglm)*
4. *Geyer, 2009, gdor*

(Possibly we can talk about separation very briefly - Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 1984; 71:1–10, very first paper talking about the terms, "separation.") (Also Heinze is a good example) Briefly mention the history of effort to solve the separation problem (when canonical statistic lies on the boundary of its convex support). Explain how each method works.

1.2 Motivations and Contributions

Presenting Endometrial example for logistic regression and catrec example for Poisson case (Possibly we want to find "representative (or famous)"

[*] Email: spark148@illinois.edu

example that can be more persuasive). POINT out aforementioned methods (except Geyer's work) used the penalized method and "transform" the problem itself to avoid model-based problem (one of our motivation).

Introduce Daniel's work and guide reader to see for detail (theoretically). With Daniel's method, we can achieve 1) the smallest CI (and similar CI comparing to gdor, gdor example has to come to the last so that in the next paragraph I can connect to the lower computational cost) 2) lower computational cost in comparison to rcdd (we need to re-run the test for this case). 3) Canonical statistics are meaningless but only probabilities and expectations => glmdr provides the mean-value parameters meanwhile other 3 methods provide the canonical statistics (I may need to rewrite this in clearer way).

2 Problem Formulation

2.1 Exponential Family

Let X be a random variable or vector with sample space $\mathcal{X} \subset \mathbb{R}^p$ and θ be a vector parameter with space Θ. An exponential family of probability distributions is a parametric statistical model whose density has the following general form:

$$f_\theta(x) = h(x) \exp(\langle T(x), \theta \rangle - c(\theta)) \tag{1}$$

where $h(x)$ is a underlying measure, $T(x)$ is a vector of sufficient statistics and $c(\theta)$ is the cumulant generating function of the family. $\langle \cdot, \cdot \rangle$ represents the inner product, which is defined as:

$$\langle T(x), \theta \rangle = \sum_{i=1}^{p} T(x_i)\theta_i.$$

The density here can mean either a probability mass function (PMF), a probability density function (PDF), or probability mass-density function (PMDF) depending on the distribution we are referring to.

In the context of general exponential families, the statistics $T(x)$ and parameter θ are called *canonical* or *natural*. An exponential family is said to be *full* if its canonical parameter space is in the effective domain:

$$\Theta = \{\theta : c(\theta) < +\infty\}. \tag{2}$$

A full exponential family is said to be *regular* if its canonical parameter space is a non-empty open set.

The log likelihood for the exponential family takes the form:

$$l(\theta) = \langle T(x), \theta \rangle - c(\theta) \tag{3}$$

Since it is the convention that terms that do not contain the parameter can be dropped from a log likelihood, $h(x)$ in (1) is dropped while the log density goes to the log likelihood.

Let Y be a canonical statistic in the exponential family (i.e. $Y = T(x)$) and θ, ψ be two different canonical parameters in the space Θ. A exponential family of probability distributions is identifiable if θ and ψ correspond to a different distribution. An exponential family fails to be identifiable if θ and ψ correspond to the same distribution. This is equivalent to

$$\langle Y, (\theta - \psi) \rangle = C \tag{4}$$

where C is constant. Geometrically, we call this the canonical statistic Y is concentrated on a hyperplane $H = \{y : y^T v = C'\}$ for some non-zero vector v, which implies some of Y are affine functions of H (i.e. $Y \in H$). Since both θ and ψ are in the same parameter space Θ, we can rewrite ψ as the function of θ. Let $\psi = \theta + sv$ for any scalar s. Then, we call any vector v that satisfies (4) a *direction of constancy* of the log likelihood. The set of all directions of constancy is called the *constancy space* of the log likelihood. If $\langle Y, (\theta - \psi) \rangle \leq C$ Then, any vector v that satisfies this called a *direction of recession* of the log likelihood. If every direction of recession is a direction of constancy, then the MLE exists in a full exponential family.

2.2 Limiting Conditional Model

For a full exponential family having canonical statistic Y, canonical parameter $\theta \in \Theta$, and $H = \{y : y^T v = C\}$ for some constant C, and $\text{pr}(Y \in H) > 0$ for some distribution in the family,

$$\lim_{s \to \infty} f_{\theta + sv}(x) = \begin{cases} 0 & \langle Y, (\theta - \psi) \rangle < C \\ f_\theta(x \mid Y \in H) & \langle Y, (\theta - \psi) \rangle = C \\ +\infty & \langle Y, (\theta - \psi) \rangle > C \end{cases} \tag{5}$$

Notice that the family

$$\{f_\theta(\cdot \mid Y \in H) : \theta \in \Theta\} \tag{6}$$

is also an exponential family with the same canonical statistic and parameter with the original family (1). We now will call this family the limiting conditional model (LCM). The canonical parameter space of the family is at least

$$\Theta + \Gamma_{\text{lim}} = \{\theta + \gamma : \theta \in \Theta \text{ and } \gamma \in \Gamma_{\text{lim}}\}, \tag{7}$$

where Θ is the canonical parameter space of the original family and Γ_{lim} is the constancy space of the family (6).

The log likelihood for family (6) is

$$l_{\text{LCM}}(\theta) = l(\theta) - \log(\text{pr}_\theta(Y \in H)) = \langle Y, \theta \rangle - c(\theta) - \log(\text{pr}_\theta(Y \in H)).$$

Suppose the MLE exists for the LCM. Then, we call this as a MLE in Barndorff-Nielsen completion of the original family if 1) it maximizes the likelihood in the union of the LCM and the original family, 2) it maximizes the likelihood in the family that that is the set of all limits of sequences of distributions in the original family.

2.3 Null Space of the Fisher Information

The null space of the Fisher information matrix (also, the variance-covariance matrix of the canonical statistic or null eigenvector of the Fisher information matrix) is crucial for the statistical inference in our method because we approximate the null space of the Fisher information matrix for an exponential family to find the MLE in the Barndorff-Nielsen completion. Since it is the support of the canonical statistic under the MLE distribution in the completion, it must contain the mean value vector of the canonical statistic. Thus, we can see Γ_{lim} in (7) is the null space of the Fisher information matrix. In our method, we can estimate the null space of the Fisher information matrix from its eigenvalues and eigenvectors using inexact computer arithmetic.

2.4 One-Sided Confidence Interval

Let $\theta = M\beta$ denote the saturated model canonical parameter (also called "linear predictor" in the GLM) where β denotes the vector of submodel canonical parameters (often called "coefficients" in statistical software) and M denotes a model matrix. Let $\hat{\beta}$ be a MLE in the LCM. Let I denote the index set of the observed values in the response vector on which we condition the original model to get the LCM. Hence, Y_I and y_I denote a random vector of these observed values and its realizations, respectively. Then endpoints for a $100(1-\alpha)\%$ confidence interval for a scalar parameter $g(\beta)$ are

$$\min_{\substack{\gamma \in \Gamma_{\text{lim}} \\ \text{pr}_{\hat{\beta}+\gamma}(Y_I = y_I) \geq \alpha}} g(\hat{\beta} + \gamma) \quad \text{and} \quad \max_{\substack{\gamma \in \Gamma_{\text{lim}} \\ \text{pr}_{\hat{\beta}+\gamma}(Y_I = y_I) \geq \alpha}} g(\hat{\beta} + \gamma) \tag{8}$$

where Γ_{lim} is the null space of the Fisher information matrix and α is the significance level. Since we always know if the observed values in the response vector is at the upper or lower of end of its rangem we only need to compute the other end of the confidence interval.

We can see solving (8) is the optimization problem. In our method, we used the sequential quadratic programming (SQP) method to solve the constrained nonlinear problem.

2.5 Calculating the MLE for discrete GLM

In a regular full discrete exponential family, the MLE falls into one of three cases; 1) a MLE exists in the original model, 2) a MLE in the Barndorff-Nielsen completion is completely degenerate, and 3) a MLE in the Barndorff-Nielsen completion is not completely degenerate. When the MLE exists in the conventional sense (the first case), we can simply use any statistical software to fit the model. In the latter two cases, we can use our method for statistical inference.

2.5.1 Completely Degenerate

Suppose Y is concentrated to some hyperplane H, then there exists a non-zero vector v in (4) that Y has no variability. In other words, the variance-covariance matrix for Y is not full rank and the model is not identifiable. If Y repeats to be concentrated to all possible hyperplanes that are nested within each other, the model is completely degenerated. The eigenvalues of the Fisher information then are all zeros and every parameter values $\theta \in \Theta$ corresponds to the same distribution. In this case, every vector of the canonical parameters is an MLE. It implies Γ_{lim} is the whole parameter space, Θ and I is the whole index for the reponse vector in (8).

2.5.2 Not Completely Degenerate

The MLE exists in the Barndorff-Nielsen completion is not completely degenerate, the limiting conditional model conditions on the non-problematic points (linearity = FALSE).

2.6 Logistic Regression

Let $p = \text{logit}^{-1}(\theta)$ denote the mean value parameter vector. Then the probabilities in (8) are

$$\text{pr}_\beta(Y_I = y_I) = \prod_{i \in I} p_i^{y_i}(1 - p_i)^{n_i - y_i}$$

where the n_i are the binomial sample sizes.

We could take the confidence interval problem to be

$$\begin{array}{ll} \text{maximize} & p_k \\ \text{subject to} & \displaystyle\prod_{i \in I} p_i^{y_i}(1 - p_i)^{n_i - y_i} \geq \alpha \end{array} \tag{9}$$

where p is taken to be the function of γ described above. And this can be done for any $k \in I$.

Notice that the mean value parameter contains exponential term in the numerator. Thus, when our canonical parameter is extremely large, we will lose precision. Thus, we transform our problems to make it computationally stable. Firstly, we convert maximize problem to minimize problem, optimize canonical parameter rather than mean value parameter and rewrite mean value parameter.

Firstly, solving minimizing problem is much more preferable in the general sense because it is much easier to solve computationally. Secondly, canonical parameter, $\theta_k = \text{logit}(p_k)$ is a monotone transformation, so if we find the solution, it will be the solution for the other. To avoid a catastrophic cancellation, we carefully construct the mean value parameters:

$$p_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \frac{1}{1 + \exp(-\theta_i)}$$

$$1 - p_i = \frac{1}{1 + \exp(-\theta_i)} = \frac{\exp(-\theta_i)}{1 + \exp(-\theta_i)}$$

Thus, our problem (9) is now

$$\begin{array}{ll} \text{minimize} & -\theta_k \\ \text{subject to} & \displaystyle\sum_{i \in I} [y_i \log{(p_i)} + (n_i - y_i) \log{(1 - p_i)}] - \log{(\alpha)}. \end{array} \tag{10}$$

2.7 Complete Separation

2.7.1 Bradley-Terry model

2.7.2 Poisson Model

3 Solution

3.1 Methodology

3.1.1 Computational Cost

3.2 Analysis and Discussion

3.2.1 Results

3.2.2 Comparison with other methods

When comparing the computational cost, check two cases when n increases and p increases and both.

4 Conclusion

Covering the drawback in the computational cost as n increases, we is more likely to have non-problematic points and need to iterate more.

- (completely degenerate case) Hence, our method is not practically useful because the computational cost is proportional to the number of the sample.

References

5 Appendix

Will be correctly labeled later...