Journal of Statistics: Advances in Theory and Applications
Volume 29, Number 1, 2024, Pages 1-11
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100122307

ESTIMATION OF SPARSE MULTINOMIAL CELL PROBABILITIES: A REVIEW

LAHIRU WICKRAMASINGHE

Department of Mathematics and Statistics
University of Winnipeg
Winnipeg
Canada
e-mail: l.wickramasinghe@uwinnipeg.ca

Abstract

Sparse data, particularly in the form of sampling zeros or categories with very low counts, pose significant challenges to traditional estimation methods, often leading to biased parameter estimates, reduced statistical power, and unreliable conclusions. The pervasive nature of sparse multinomial data across various disciplines, including genetics, ecology, and the social sciences, underscores the urgent need for improved analytical techniques. This review paper highlights the critical importance of developing methods that can more accurately and robustly handle sparse data. By effectively managing zeros and low counts, these advanced techniques offer a more accurate representation of underlying distributions, thereby enhancing the validity of statistical inferences. Such improvements are crucial for informed decision-making and sound policy formulation across multiple fields of study.

2020 Mathematics Subject Classification: 62F15, 62G05.
Keywords and phrases: Dirichlet distribution, multinomial distribution, Bayesian.
Received August 13, 2024

1. Introduction

Categorical data is a type of data that consists of categories or groups (see Agresti [2]). Each observation in the data belongs to one of the categories. There are two types of categorical data: nominal and ordinal. Nominal data consists of categories that do not have a specific order or rank. Examples of nominal data include gender, race, and country of origin. Ordinal data consists of categories that have a specific order or rank. Examples of ordinal data include educational level (e.g., high school, college, graduate school) and income level (e.g., low, medium, high). In addition to these two main types, there are other types of categorical data, such as dichotomous and count data. Dichotomous data is a type of categorical data that has only two categories (see Hosmer Jr et al. [18]); on the other hand, count data is a type of categorical data where the categories represent counts of events or occurrences (see Cameron and Trivedi [9] and Hilbe [16]).

Count data often follows certain distributions that are appropriate for modelling purposes. The Poisson distribution is often used to model when we have data with a small range of possible counts, and when the average number of counts is equal to the variance (see Kleiber and Zeileis [21]). Another distribution that's used for count data is called the negative binomial distribution, which is used when the counts have a larger range of possible values than a Poisson distribution and when the variance is greater than the mean (see Long [25]). Hilbe [17] provided an overview of when to use Poisson and negative binomial distributions. Binomial and multinomial are another two distributions that can be used for count data, but rather represent categorical data where each observation falls into one of several distinct categories. The binomial distribution is used to model the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes (success or failure). This is a discrete probability distribution that has been extensively covered by Ross [28] and Walpole et al. [34]. On the other hand, the multinomial

distribution is used to model the probability of outcomes in a situation where there are more than two possible outcomes, and each outcome has a specific probability. This is a generalization of the binomial distribution and is also a discrete probability distribution. The multinomial distribution has been extensively covered by Johnson and Kotz [20] and Bishop [5]. The Likert scale was introduced by Likert [23]. The data collected using the Likert scale is a type of ordinal data that is commonly used in surveys and questionnaires to measure attitudes, opinions, and perceptions of individuals toward a particular topic or issue (see Carifio and Perla [10]).

There are several common issues that can arise when working with categorical data. Imbalanced categories occur when one or more categories have a much larger number of observations than others, which can also lead to biased statistical inference. To address imbalanced categories, Huang et al. [19] proposed techniques such as oversampling, undersampling, and weighting can be employed. Missing data, on the other hand, occur when some observations have incomplete information on the categorical variable of interest, which can reduce the sample size and affect statistical power. Various methods have been developed for handling missing data in categorical variables, such as multiple imputations, maximum likelihood estimation, and Bayesian methods (see Rubin [29] and Little and Rubin [24]). Another potential issue with categorical data is that it may be prone to measurement error, meaning that the categories used may not accurately reflect the underlying construct being measured. Carroll et al. [11] provided an in-depth analysis of the sources and consequences of measurement error and misclassification, and discusses various methods for dealing with these issues in statistical analysis. These issues can be addressed through careful data cleaning and preparation and using appropriate statistical analysis techniques.

2. Sparsity

This paper focuses on another issue that occurs with multinomial data called sparsity. Sparsity in multinomial data occurs when a dataset contains many variables or categories, resulting in a high number of possible outcomes but relatively few observations for each category. This situation is particularly common in scenarios with numerous classification variables or variables with many levels. As a result, many categories may have zero or very few observations, making statistical analysis challenging. There are two types of sparsity in multinomial data:

(1) When outcomes are not observed for one or more categories because these outcomes are unobservable (cell probabilities are zero).

(2) When outcomes are not observed for one or more categories due to the limited size of the sample and cell probabilities being small, but not actually zero.

The first type of sparsity is commonly known as structural zeros, which refers to outcomes that will always have a count of zero, regardless of the sample size n_i. If some or all structural zeros are known beforehand, they can be excluded from the model or assigned no probability mass. Additionally, some research has focused on identifying structural zeros, as discussed by Bishop et al. [6]. Xie et al. [39] proposed a Bayesian hierarchical model to identify structural zeros and impute dropouts in single-cell Hi-C data, which can be adapted to address similar challenges in multinomial data. Feng [13] provided a comparison of zero-inflated models and hurdle models that are frequently employed to handle excess zeros in count data, including structural zeros.

This paper's primary focus is on the second type of sparsity, exemplified by instances of sampling zeros (as well as extremely low observed counts). This issue can significantly hinder statistical inference. To mitigate this, augmenting the effective cell counts by integrating various data sources can be beneficial. Under standard conditions, the

Maximum Likelihood Estimator (MLE) for p_{ij} is known for its consistency and efficiency. Consistency implies that as the sample size increases, the estimates converge in probability to the true cell probabilities. Efficiency signifies that no other consistent estimator has a lower mean squared error (MSE) than $\hat{p}_{ij}^{\text{MLE}}$ as the sample size grows. However, the MLE can perform poorly by underestimating the true cell probabilities when dealing with sparse data (see Molenberghs and Verbeke [27]). For sparse multinomial datasets, the MLE often results in zero probability estimates that are difficult to interpret, and fail to meet the sparse asymptotic consistency criterion in certain scenarios, i.e.,

$$\sup_j \left[\frac{\hat{p}_{ij}^{\text{MLE}}}{p_{ij}} - 1 \right] \neq o_p(1)$$

(see Lambert [22]; Min and Agresti [26]). This inconsistency underscores the need for alternative estimation techniques that can handle sparsity more effectively.

3. Handling Sampling Zeros

The second type of sparsity, characterized by sampling zeros and very low observed counts, presents significant challenges in statistical analysis and inference. Very low observed counts contribute to instability in estimates and increase the variance of the estimators. This sparsity type is difficult to handle because it often leads to biased parameter estimates and increased uncertainty. This can result in misleading conclusions about the underlying population parameters. To address these challenges, I outline several approaches that effectively address the second type of sparsity.

The first approach involves an estimator, which is obtained by combining $\hat{p}_{ij}^{\text{MLE}}$ with an informed guess for p_{ij}. This method can provide improved performance over $\hat{p}_{ij}^{\text{MLE}}$ (see Fienberg and Holland

[14]). Different techniques to construct this informed guess lead to so-called shrinkage estimators, which borrow information across other multinomial populations and cell categories. The resulting estimator can have significantly improved performance in some contexts. Different methods for borrowing information from other available data can enhance the estimation of p_{ij}. One initial approach, particularly applicable to ordinal categories, involves borrowing information from adjacent cells within the same multinomial population to improve the estimation of cell probabilities. This method is also extensively used to gather information from neighbouring cells in a sparse contingency table. Various methodologies have been developed to leverage information from neighbouring cells. Simonoff [31] considered an estimator based on a maximum penalized likelihood criterion for sparse multinomial data. Burman [8], and Hall and Titterington [15] proposed kernel-type estimators for sparse multinomial data, both of which are sparsity asymptotic consistent under certain restrictive conditions on the true cell probabilities. Dong and Simonoff [12] used boundary kernels to relax some of these conditions. Aerts et al. [1] proposed an estimator based on a local polynomial approach for sparse contingency tables. Albert [4] demonstrated that the Bayesian paradigm offers significant flexibility in handling boundary bias and sparsity when analyzing sparse contingency tables. In this approach, information is borrowed within a "block," where the structure of these blocks and the data structure significantly affect the determination of neighbouring cells.

Another approach to handle this second type of sparsity by borrowing information from other multinomial populations rather than from neighbouring cells within the same population to enhance the estimation of p_{ij}. This method identifies other multinomial populations that are similar in p to the target population and borrows information from these similar populations for each category separately. Ahmed [3] demonstrated that shrinkage estimators outperform maximum likelihood estimators for

p_{ij}. This approach is conceptually different from the first one, as it focuses on finding methodologies that leverage information from other multinomial populations. Wickramasinghe et al. [37] introduced a semi-parametric Bayesian estimator utilizing the Dirichlet process (DP) through the stick-breaking construction, originally proposed by Sethuraman [30]. This Dirichlet process method effectively borrows information across similar populations by clustering them during the MCMC posterior simulation iterations. Additionally, Wickramasinghe et al. [35] developed an approach to model batting outcomes in baseball using weighted likelihood concepts. This weighted likelihood method for estimating multinomial probabilities also facilitates sharing relevant information among different batters (populations). Both techniques improve the estimation of cell probabilities by borrowing information exclusively across other multinomial populations. Importantly, neither method enables the sharing of information between cells within the same population, which is particularly valuable in the context of ordered categories.

A third method involves integrating the first and second approaches I discussed before by borrowing information across both multinomial populations and cell categories. This combined approach enhances the estimation of multinomial cell probabilities. It is particularly useful when the categories are ordinal, meaning they have a natural order. In such cases, borrowing information from neighbouring cell categories is conceptually sound and beneficial. Wickramasinghe et al. [36] introduced a Bayesian estimator, based on a smoothed Dirichlet prior (see Wickramasinghe et al. [38]), which acts as a scaled shrinkage estimator. This method allows for simultaneous inference across numerous multinomial populations by borrowing information between populations and cell categories. This introduced estimator can be viewed as a double (or two-way) shrinkage estimator that enhances the overall inference for sparse multinomial data.

4. Future Research Areas

I outline a few future research areas for handling this second type of sparsity. One potential method is using hierarchical Bayesian models. Hierarchical Bayesian models offer a robust framework for enhancing the estimation of sparse multinomial data by pooling information across different levels of a hierarchical structure. Developing new priors that balance flexibility and informativeness can help in better capturing the underlying structure of sparse data. Priors such as hierarchical Dirichlet processes (see Teh et al. [32]) or nonparametric priors can adaptively borrow strength across categories while preserving local variation. Also, innovations and efficient computational algorithms in scalable inference techniques, including variational inference (see Blei et al. [7]) and advanced Markov Chain Monte Carlo (MCMC) methods, are crucial for handling sparse multinomial data. Using penalized likelihood methods for improving the estimation of sparse multinomial data is a promising research area. Penalized likelihood methods (see Tibshirani [33]), such as those incorporating adaptive penalties, can improve parameter estimation by controlling the complexity of the model. Selecting an appropriate penalty term is crucial to handle sparsity effectively.

References

[1] M. Aerts, I. Augustyns and P. Janssen, Smoothing sparse multinomial data using local polynomial fitting, Journal of Nonparametric Statistics 8(2) (1997), 127-147.

DOI: https://doi.org/10.1080/10485259708832717

[2] A. Agresti, Categorical Data Analysis, John Wiley & Sons, 2018.

[3] S. E. Ahmed, Construction of improved estimators of multinomial proportions, Communications in Statistics - Theory and Methods 29(5-6) (2000), 1273-1291.

DOI: https://doi.org/10.1080/03610920008832544

[4] J. H. Albert, Bayesian Methods for Contingency Tables, Encylopedia of Biostatistics, 2004.

DOI: https://doi.org/10.1002/9781118445112.stat04849

[5] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[6] Y. M. M. Bishop, S. E. Fienberg and P. W. Holland, Discrete Multivariate Analysis: Theory and Practice, Cambridge, Mass: MIT Press, 1975.

[7] D. M. Blei, A. Kucukelbir and J. D. McAuliffe, Variational inference: A review for statisticians, Journal of the American Statistical Association 112(518) (2017), 859-877.

DOI: https://doi.org/10.1080/01621459.2017.1285773

[8] P. Burman, Smoothing sparse contingency tables, Sankhya: The Indian Journal of Statistics, Series A 49(1) (1987), 24-36.

[9] A. C. Cameron and P. K. Trivedi, Regression Analysis of Count Data, Cambridge University Press, 2013.

[10] J. Carifio and R. J. Perla, Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes, Journal of Social Sciences 3(3) (2007), 106-116.

DOI: https://doi.org/10.3844/jssp.2007.106.116

[11] R. J. Carroll, D. Ruppert and L. A. Stefanski, Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments, CRC Press, 2006.

[12] J. Dong and J. Simonoff, The construction and properties of boundary kernels for smoothing sparse multinomials, Journal of Computational and Graphical Statistics 3(1) (1994), 57-66.

DOI: https://doi.org/10.2307/1390795

[13] C. X. Feng, A comparison of zero-inflated and hurdle models for modeling zero-inflated count data, Journal of Statistical Distributions and Applications 8(1) (2021); Article 8.

DOI: https://doi.org/10.1186/s40488-021-00121-4

[14] S. E. Fienberg and P. W. Holland, Simultaneous estimation of multinomial cell probabilities, Journal of the American Statistical Association 68(343) (1973), 683-691.

DOI: https://doi.org/10.2307/2284799

[15] P. Hall and D. M. Titterington, On smoothing sparse multinomial data, Australian Journal of Statistics 29(1) (1987), 19-37.

DOI: https://doi.org/10.1111/j.1467-842X.1987.tb00717.x

[16] J. M. Hilbe, Modeling Count Data, Cambridge University Press, 2014a.

DOI: https://doi.org/10.1017/CBO9781139236065

[17] J. M. Hilbe, Negative Binomial Regression, Cambridge University Press, 2014b.

[18] D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, 2013.

DOI: https://doi.org/10.1002/9781118548387

[19] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang and W. Xu, Applications of support vector machine (svm) learning in cancer genomics, Cancer Genomics & Proteomics 15(1) (2018), 41-51.

DOI: https://doi.org/10.21873/cgp.20063

[20] N. L. Johnson and S. Kotz, Continuous Univariate Distributions, John Wiley & Sons, 1970.

[21] C. Kleiber and A. Zeileis, Applied Econometrics with R, Springer Science & Business Media, 2008.

DOI: https://doi.org/10.1007/978-0-387-77318-6

[22] D. Lambert, Zero-inflated poisson regression, with an application to defects in manufacturing, Technometrics 34(1) (1992), 1-14.

DOI: https://doi.org/10.2307/1269547

[23] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 140(22) (1932), 1-55.

[24] R. J. Little and D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, 2019.

DOI: https://doi.org/10.1002/9781119482260

[25] J. S. Long, Regression Models for Categorical and Limited Dependent Variables, Sage Publications, 1997.

[26] Y. Min and A. Agresti, Random effect models for repeated measures of zero-inflated count data, Statistical Modelling 5(1) (2005), 1-19.

DOI: https://doi.org/10.1191/1471082X05st084oa

[27] G. Molenberghs and G. Verbeke, Models for Discrete Longitudinal Data, Springer, 2007.

[28] S. M. Ross, A First Course in Probability, Pearson Education, 2010.

[29] D. B. Rubin, Multiple imputation for nonresponse in surveys, John Wiley & Sons, 1987.

DOI: https://doi.org/10.1002/9780470316696

[30] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica 4(2) (1994), 639-650.

[31] J. S. Simonoff, A penalty function approach to smoothing large sparse contingency tables, The Annals of Statistics 11(1) (1983), 208-218.

DOI: https://doi.org/10.1214/aos/1176346071

[32] Y. Teh, M. Jordan, M. Beal and D. Blei, Hierarchical Dirichlet processes, Journal of the American Statistical Association 101(476) (2006), 1566-1581.

DOI: https://doi.org/10.1198/016214506000000302

[33] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58(1) (1996), 267-288.

DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[34] R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye, Probability & Statistics for Engineers & Scientists, Pearson Education, 2011.

[35] L. Wickramasinghe, A. Leblanc and S. Muthukumarana, Model-based estimation of baseball batting metrics, Journal of Applied Statistics 48(10) (2020), 1775-1797.

DOI: https://doi.org/10.1080/02664763.2020.1775792

[36] L. Wickramasinghe, A. Leblanc and S. Muthukumarana, Bayesian inference on sparse multinomial data using smoothed Dirichlet distribution with an application to covid-19 data, Model Assisted Statistics and Applications 18(3) (2023), 207-226.

DOI: https://doi.org/10.3233/MAS-221411

[37] L. Wickramasinghe, A. Leblanc and S. Muthukumarana, Semi-parametric Bayesian estimation of sparse multinomial probabilities with an application to the modelling of bowling performance in T20I cricket, Annals of Biostatistics and Biometric Applications 5(1) (2023), 1-13.

[38] L. Wickramasinghe, A. Leblanc and S. Muthukumarana, Smoothed Dirichlet distribution, Journal of Statistical Theory and Applications 22(4) (2023), 237-261.

DOI: https://doi.org/10.1007/s44199-023-00062-8

[39] Q. Xie, C. Han, V. Jin and S. Lin, HiCimpute: A Bayesian hierarchical model for identifying structural zeros and enhancing single cell Hi-C data, PLOS Computational Biology 18(6) (2022), 1-19.

DOI: https://doi.org/10.1371/journal.pcbi.1010129

■