Data Transformation and Its Validity in a Two-Sample Problem: An Illustration Based on Graphical Models

Data transformation and its validity in a two-sample problem: an illustration based on graphical models

Erika Banzato[1], Davide Risso[1], Monica Chiogna[2], and Vera Djordjilović[3]

[1] Department of Statistical Sciences, University of Padua, via C. Battisti 241, Padua, Italy, erika.banzato@unipd.it
[2] Department of Statistical Sciences, University of Bologna, Via Belle Arti, 41, Bologna, Italy
[3] Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, Venice, Italy

Abstract. The rapid expansion of technology in medical and biological domains has led to a surge in available data, accompanied by increased complexity. Classical statistical methods, primarily developed for analyzing data assumed to follow a normal distribution, may prove inadequate for handling this complexity. We focus our attention on transcriptomic data, in particular single-cell RNAseq data, which measure gene expression as counts rather than intensities. Through a simulation study, this study aims to demonstrate the limitations of common data transformations intended to fit data into a normal distribution, highlighting potential misinterpretations. Simulation results show that the transformation of the data might affect the results and their interpretation. In particular, we show that in a two-sample problem in a graphical framework, the set of nodes that differs in the two conditions is affected by the transformations. This behavior might be due to the fact that transformations can distort graphical structures reliant on conditional dependencies among variables, affecting final conclusions. Specifically, when analyzing RNA-seq data, methods tailored for count data are preferable over those designed for normally distributed data. The findings underscore the need for specialized approaches in handling count data in two-sample problems and advocate for further research into alternative methods for differential network analysis.

Keywords: two-sample problem, data transformation, graphical models, count data

Introduction

Data transformation represents a controversial tool for statistical analysis. It is often used as a pre-processing step, but little can be found in the literature about the risks of improper and naive use. This work aims to point out the increasing need for *ad hoc* methods to deal with the new complex types of data now available in the biological and medical fields. New tools should reflect the complex nature

of data, instead of trying to transform data to match the assumptions behind the classical statistical methodologies, usually developed for normally distributed data.

Originally, when transcriptomics data started to become available, the technology of choice for measuring gene expression was the microarray assay. This technology optically scans fluorophore intensities, providing data on a continuous scale [1]. This kind of data is commonly assumed to be normally distributed after being transformed in the logarithmic scale, and many statistical techniques have been developed to deal with such data. Later, a new technology allowed for the high-throughput sequencing of RNA molecules (i.e., RNA-seq), and now represents the reference technology for the study of genome-wide transcription levels [2]. One of the main advantages of RNA-seq over microarrays is that it allows the study of gene expression even at single-cell resolution [3]. However, as measures of gene expression, RNA-seq yields counts rather than intensities, which typically show skewed distributions with a high variance. Moreover, they very often show a large number of zero counts, typically larger than expected in a Poisson model [4].

Network analysis is one of the most prominent fields of research, with a notable emphasis on uncovering differences within networks in multiple conditions. The aim is to identify potential differences in the network by comparing data arising from different conditions [5]. However, statistical techniques have been mainly developed for data assumed to be normally distributed, but there is little information about how this applies to count data, such as RNA-seq.

When dealing with such data, it is common to assume normality on the logarithmic scale or on other transformations [6]. Some transformations show good performance when the mean and variance of the variables are relatively high and the presence of ties does not have a strong impact on the analysis. However, the impact of these transformations on more skewed data is not clear. In fact, the main issue with single-cell data is the overabundance of zeros that leads to a high presence of ties, which is a big hurdle when trying to normalize the data. Moreover, in this particular setting, an important aspect is that the adopted transformation should preserve the conditional independence of the variables and potential differences. The primary interest is the characterization of a transformation that preserves the relations between variables while maintaining possible changes in the relationships.

Through a simulation study, this work aims to compare different transformation approaches that are used in practical analyses, to highlight the main advantages and drawbacks when data are transformed. The purpose is to emphasize how the data transformation process could change the conditional independence relationships between variables and lead to misleading results.

Simulation study

We ran a simulation study to compare the performances of different data transformations applied to data from a Poisson graphical model. Let $G = (V, E)$ be

the underlying undirected graph, where V is the set of nodes and $E = \{(u, v) : u, v \in V\}$ is the set of edges. A graph is said to be undirected if two vertices a and b are connected by an edge (a, b) in addition to (b, a).

Single-cell RNAseq data can be described using a truncated Poisson graphical model (TPGM) [7, 8]. The joint distribution of the p-dimensional random vector $\boldsymbol{Y}$ with domain $Y_i \in \{0, 1, \ldots, D_i\}$ for $i = 1, \ldots, p$ can be written as

$$p(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) \propto \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{\phi} \boldsymbol{y} + \sum_{j=1}^{p} \log(y_j!) \right\} \tag{1}$$

where the matrix ϕ represents the graphical structure of the model.

For illustrative purposes, we consider a simple graph with four nodes and four edges. Data were simulated from the graph in Figure 1, using the code in [8]. We assumed that the samples come from two TPGM distributions in (1), differing only for the parameter θ_3. Then four transformations were applied to the generated data: logarithm, square root, deviance residuals, and randomized quantile residuals [6]. To avoid computational issues caused by zero counts generated in the logarithmic transformation, a unit was added to each observation when log-transforming the data. A total of 500 simulations were run for each sample size $n \in (500, 2500, 5000, 10000, 25000, 40000)$.

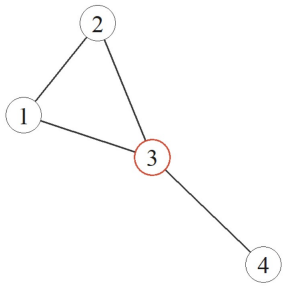

Fig. 1: Graph used for the simulations.

To address the issue of identifying differences in the network, we applied to the transformed data the *SourceSet* method (based on the normality assumption). The statistical guarantees of the tool are described in [9] and the algorithm is implemented in the `R` package `sourceSet` [10]. To improve the asymptotic approximation of the test statistics we applied the correction in [11]. Given that $\theta_3^{(1)} \neq \theta_3^{(2)}$, and following the rules in [9] we would expect that the procedure finds the node 3 as the source of the difference in the two conditions. In fact, this algorithm identifies the smallest set containing the difference, either a clique or a separator, where a clique is a maximal subset of nodes all connected among each other, while a separator is the intersection of two cliques; we refer the reader to the book in [12] for a detailed explanation of these terms. The identification

of node 3 is possible because it is a separator of the two cliques: $\{1, 2, 3\}$ and $\{3, 4\}$. Finally, FWER was controlled at 0.05 level.

Results

The results of the simulations are shown in Figure 2.

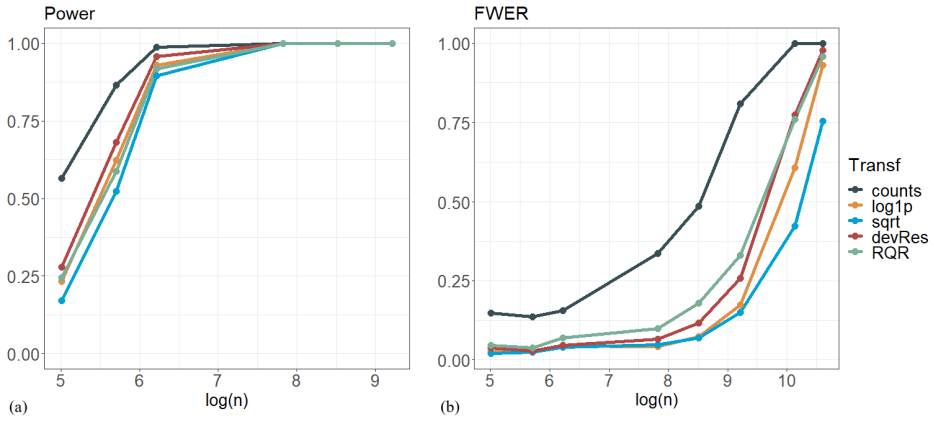

Fig. 2: (a) Power: fraction of times the node 3 is identified as different in the two conditions. (b) FWER: fraction of times at least one of the nodes 1, 2, and 4 has been falsely discovered. The black line is for reference and refers to the *raw*-counts.

Figure 2(a) shows the power, namely the fraction of times the node 3 was identified as different in the two conditions, when the procedure is applied to all the transformed data. We can observe that as the sample size n increases, the power also increases as expected. However, Figure 2(b) shows that as n increases, also the number of false rejections increases. This behavior is common to all the transformations, with a small advantage of the square-root transformation. Nevertheless, this transformation also show an increasing trend.

Further simulations showed that, on the contrary, when the difference affects a node which is not a separator, the method has comparable results in terms of power and the FWER is controlled at the nominal level. Moreover, when considering bigger graphs with a larger number of cliques, the false rejections considering these cliques only are controlled at the nominal level. On the other hand, the same results observed in the previous paragraph are obtained for the cliques affected by the difference.

Discussion

When dealing with count data, assumptions about the data structures are of great importance, especially if the analysis aims to study the conditional rela-

tionships between variables. In this setting, we have highlighted the fact that the data transformation does not seem to meet the goal of the analysis. In particular, when the target object of the analysis is the detection of a difference involving a separator, as shown in figure 2, the *SourceSet* approach has a power increasing with the sample size, but at the same time, the number of false rejections is increasing as well. This is not observed in situations where the node affected by a change in parameter belongs to a clique only. In the same way, the number of false rejections is also controlled for all those nodes in cliques not connected to the source of the difference.

This behavior might be the consequence of the conditional dependence structure. In fact, the *SourceSet* method highly relies on the graphical structure to identify the subsets of nodes affected by a change in the two groups. Hence, the fact that data transformation might act on the structure of the graph or on the relationship between the parameters of the model, could act as a propagator of the differences in node 3, such that, in the end, all the nodes connected to it are affected. A method that is specifically designed for data coming from a normal distribution may therefore result in an unsuitable approach when applied to data of a different nature. If, indeed, the nature of the difference between the two conditions lies in a parameter θ, relative to a single node, the transformation may not necessarily maintain the relationship with the corresponding parameter of the normal distribution, in this case the mean. Additionally, in the case of count data, it is well-known that there is a dependency between the mean and variance, and undoubtedly this may affect the transformed data. Therefore, the main problem is not the ability of the test to detect the differences, but rather not knowing exactly how the difference in a certain parameter of the initial distribution is then propagated to the other parameters in the distribution of the transformed data.

Conclusion

Thanks to the expansion of new technologies in medical and biological fields, the amount of data available is increasing, as is their complexity. The classical statistical methods were mostly developed for analyzing data that can be assumed to come from a normal distribution. However, this work demonstrates the need to go beyond the normality assumption to chase the complexity of the data. Through a simulation study, we showed how common data transformations, whose purpose is to make data fit a normal distribution, may lead to ambiguous or even wrong conclusions. When dealing with RNA-seq data, it is important to select methods that are specifically developed for handling count data rather than methods designed to handle normally distributed data. When the focus of the analysis involves the graphical structure, which is highly reliant on conditional dependencies among variables, the use of data transformation may alter the relationships and thus the final conclusions. We showed how the use of the transformations should be carefully considered within the context of the application.

These results highlight the need for *ad hoc* methods for handling count data in a two-sample problem. There is a distinct lack, in the literature, of alternative methods suitable for handling differential network analysis, representing an important area for future research. In particular, the KLIEP approach suggested in [13] offers promising development opportunities.

This paper concludes that future statistical research in the transcriptomic field should be focused on the development of tools designed specifically for tackling the complexity of this kind of data, while the hypothesis of transforming the data in order to use methods meant for other kinds of data should be carefully considered.

References

[1] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2):249–264.

[2] Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1):57–63.

[3] Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. Molecular cell, 58(4):610–620.

[4] Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics, 14(1):113–128.

[5] Shojaie, A. (2021). Differential network analysis: A statistical perspective. Wiley Interdisciplinary Reviews: Computational Statistics, 13(2):e1508.

[6] Ahlmann-Eltze, C., and Huber, W. (2023). Comparison of transformations for single-cell RNA-seq data. Nat Methods, 20(5):665-672.

[7] Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdisciplinary Reviews: Computational Statistics, 9(3):e1398.

[8] Zhang, R., Ren, Z., Celedón, J. C., and Chen, W. (2021). Inference of large modified Poisson-type graphical models: Application to RNA-seq data in childhood atopic asthma studies. The Annals of Applied Statistics, 15(2), 831-855.

[9] Djordjilović, V. and Chiogna, M. (2022). Searching for a source of difference in graphical models. Journal of Multivariate Analysis, 190:1–13.

[10] Salviato, E., Djordjilović, V., Chiogna, M., and Romualdi, C. (2019). SourceSet: A graphical model approach to identify primary genes in perturbed biological pathways. PLOS Computational Biology, 15(10):e1007357.

[11] Banzato, E., Chiogna, M., Djordjilović, V., and Risso, D. (2023). A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models. Statistics & Probability Letters, 193, 109732.

[12] Lauritzen, S. L. (1996). Graphical models. Clarendon Press.

[13] Liu, S., Quinn, J. A., Gutmann, M. U., Suzuki, T., and Sugiyama, M. (2014). Direct learning of sparse changes in Markov networks by density ratio estimation. Neural Computation, 26(6):1169–1197.