Generalised Median Polish Based on Additive Generators

Balasubramaniam Jayaram[1] and Frank Klawonn[2,3]

Abstract Contingency tables often arise from collecting patient data and from lab experiments. A typical question to be answered based on a contingency table is whether the rows or the columns show a significant difference. Median Polish (MP) is fast becoming a prefered way to analyse contingency tables based on a simple additive model. Often, the data need to be transformed before applying the MP algorithm to get better results. A common transformation is the logarithm which essentially changes the underlying model to a multiplicative model. In this work, we propose a novel way of applying the MP algorithm with generalised transformations that still gives reasonable results. Our approach to the underlying model leads us to transformations that are similar to additive generators of some fuzzy logic connectives. In fact, we illustrate how to choose the best transformation that give meaningful results by proposing some modified additive generators of uninorms. In this way, MP is generalied from the simple additive model to more general nonlinear connectives. The recently proposed way of identifying a suitable power transformation based on IQRoQ plots [1] also plays a central role in this work.

1 Introduction

Contingency tables often arise from collecting patient data and from lab experiments. The rows and columns of a contingency table correspond to two

[1]Department of Mathematics, Indian Institute of Technology Hyderabad, Yeddumailaram - 502 205, India jbala@iith.ac.in ·[2]Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany f.klawonn@ostfalia.de ·[3] Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany frank.klawonn@helmholtz-hzi.de

	≤ 8	$9-11$	12	$13-15$	≥ 16
North-West	25.3	25.3	18.2	18.3	16.3
North-Central	32.1	29.0	18.8	24.3	19.0
South	38.8	31.0	19.3	15.7	16.8
West	25.4	21.1	20.3	24.0	17.5

Table 1 Infant Mortality vs Educational Qualification of the Parents in deaths per 1000 live births in the years 1964-1966 (Source: U.S. Dept. of Health, Education and Welfare)

different categorical attributes. Table 1 shows an example of a contingency table.

A typical question to be answered based on data from a contingency table is whether the rows or the columns show a significant difference. For the example of the contingency Table 1, one would be interested in finding out whether the education of the father or the regions have an influence on the infant mortality.

Hypothesis tests with non-parametric tests like the Wilcoxon-Mann-Whitney-U test, Analysis of variance (ANOVA) and the t-test are some of the common options. However, each of them has its own drawbacks. For more on this, please refer to [1] and the references therein.

Median polish [2] – a technique from robust statistics and exploratory data analysis – is another way to analyse contingency tables based on a simple additive model. We briefly review the idea of median polish in Section 2. Although the simplicity of median polish as an additive model is appealing, it is sometimes too simple to analyse contingency table. Very often, especially in the context of gene, protein or metabolite expression profile experiments, the measurements are not taken directly, but are transformed before further analysis. In the case of expression profile, it is common to apply a logarithmic transformation. The logarithmic transformation is a member of a more general family, the so-called power transformations which are explained in Section 3.

However, it is not clear whether the MP applied to the transformed data would still unearth the interesting characteristics of the data, since the logarithmic transformation essentially changes the underlying model to a multiplicative model. In this work, we propose a novel way of applying the MP algorithm that still gives reasonable results. Our approach to the underlying model leads us to transformations that are similar to additive generators of some fuzzy logic connectives. In fact, we illustrate how to choose the best transformation that give meaningful results by proposing some modified additive generators of uninorms. The recently proposed way of identifying a suitable power transformation based on IQRoQ plots [1] also plays a central role in this work.

Overall: 20.775						
	≤ 8	$9-11$	12	$13-15$	≥ 16	RE
NW	-1.475	0.075	0.0125	-1.075	0.625	-1.475 -
NC	1.475	-0.075	-3.2375	1.075	-0.525	2.375
S	10.900	4.650	-0.0125	-4.800	0.000	-0.350
W	-3.200	-5.950	0.2875	2.800	0.000	0.350
CE	7.4750	5.9250	-1.1125	0.0750	-3.6250	

Table 2 Median polish for the Infant Mortality data

2 Median Polish

The underlying additive model of median polish is that each entry x_{ij} in the contingency table can be written in the form

$$x_{ij} \;=\; g + r_i + c_j + \varepsilon_{ij}.$$

- g represents the overall or grand effect in the table. This can be interpreted as general value around which the data in the table are distributed.
- r_i is the row effect reflecting the influence of the corresponding row i on the values.
- c_j is the column effect reflecting the influence of the corresponding column j on the values.
- ε_{ij} is the residual or error in cell (i, j) that remains when the overall, the corresponding row and column effect are taken into account.

For a detailed explanation of the MP algorithm please refer to [2]. Table 2 shows the result of median polish applied to Table 1.

The result of median polish can help to better understand the contingency table. In the ideal case, the residuals are zero or at least close to zero. Close to zero means in comparison to the row or column effects. If most of the residuals are close to zero, but only a few have a large absolute value, this is an indicator for outliers that might be of interest. Most of the residuals in Table 2 are small, but there is an obvious outlier in Southern region for fathers with the least number of years of education.

3 Median Polish on Transformed Data

Transformation of data is a very common step of data preprocessing (see for instance [3]). There can be various reasons for applying transformations before other analysis steps, like normalisation, making different attribute ranges comparable, achieving certain distribution properties of the data (symmetric, normal etc.) or gaining advantage for later steps of the analysis.

The logarithm is a special instance of parametric transformations, called power transformations (see for instance [2]) that are defined by

$$t_\lambda(x) \;=\; \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x) & \text{if } \lambda = 0. \end{cases}$$

It is assumed that the data values x to be transformed are positive. If this is not the case, a corresponding constant ensuring this property should be added to the data.

We restrict our considerations on power transformations that preserve the ordering of the values and therefore exclude negative values for λ.

3.1 The Non-additive Model

When we choose $\lambda = 0$, i.e. the logarithm for the power transformation, we obtain the following model.

$$\ln(x_{ij}) \;=\; g + r_i + c_j + \varepsilon_{ij}. \tag{1}$$

Transforming back to the original data yields the model

$$x_{ij} \;=\; e^g \cdot e^{r_i} \cdot e^{c_j} \cdot e^{\varepsilon_{ij}}.$$

So it is in principle a multiplicative model (instead of an additive model as in standard median polish) as follows:

$$x_{ij} \;=\; \tilde{g} \cdot \tilde{r}_i \cdot \tilde{c}_j \cdot \tilde{\varepsilon}_{ij}$$

where $\tilde{g} = e^g$, $\tilde{r}_i = e^{r_i}$, $\tilde{c}_j = e^{c_j}$, $\tilde{\varepsilon}_{ij} = e^{\varepsilon_{ij}}$. The part of the model which is not so nice is that the residuals also enter the equation by multiplication. Normally, residuals are always additive, no matter what the underlying model for the approximation of the data is.

Towards overcoming this drawback, we propose the following approach. We apply the median polish algorithm to the log-transformed data in order to compute g (or $\tilde{g}$), r_i (or $\tilde{r}_i$) and c_j (or $\tilde{c}_j$). The residuals are then defined at the very end as

$$\varepsilon_{ij} := x_{ij} - \tilde{g} \cdot \tilde{r}_i \cdot \tilde{c}_j. \tag{2}$$

Let us now rewrite Eq. (1) in the following form:

$$\ln(x_{ij}) \;=\; \ln(\tilde{g}) + \ln(\tilde{r}_i) + \ln(\tilde{c}_j) + \ln(\tilde{\varepsilon}_{ij}).$$

Assuming that the residuals are small, we have

$$\ln(x_{ij}) \;\approx\; \ln(\tilde{g}) + \ln(\tilde{r}_i) + \ln(\tilde{c}_j).$$

Transforming this back to the original data, we obtain

$$x_{ij} \approx \exp\left(\ln(\tilde{g}) + \ln(\tilde{r}_i) + \ln(\tilde{c}_j)\right).$$

A natural question that arises now is the following: *What happens with other power transformations, i.e., for* $\lambda > 0$*?* In principle the same, as we obtain

$$x_{ij} \approx t_\lambda^{-1}(t_\lambda(\tilde{g}) + t_\lambda(\tilde{r}_i) + t_\lambda(\tilde{c}_j)). \tag{3}$$

Let us denote by $\oplus_\lambda$ the corresponding, possibly associative, operator obtained as follows:

$$x \oplus_\lambda y = t_\lambda^{-1}\left(t_\lambda(x) + t_\lambda(y)\right) . \tag{4}$$

Now, we can interpret Eq. (3) as

$$x_{ij} \approx g \oplus_\lambda \tilde{r}_i \oplus_\lambda \tilde{c}_j \tag{5}$$

Thus the problem of determining a suitable transformation of the data before applying the median polish algorithm essentialy boils down to finding that operator $\oplus_\lambda$ which minimises the residuals in (2), viz.,

$$\varepsilon_{ij} = x_{ij} - g \oplus_\lambda \tilde{r}_i \oplus_\lambda \tilde{c}_j.$$

3.2 Finding a Suitable Transformation Based on IQRoQ Plots

As stated earlier, power transformations are the most commonly used transformations on data. Recently Klawonn *et al.* [1] have proposed a novel way of finding the particular λ of a power transformation to be applied on the data such that applying the Median Polish on that still reveals interesting characteristics of the data. In the following we briefly detail their technique.

An ideal result for median polish would be when all residuals are zero or at least small. The residuals get smaller automatically when the values in the contingency table are smaller. This would mean that we tend to put a high preference on the logarithmic transformation ($\lambda = 0$), at least when the values in the contingency table are greater than 1.

Neither single outliers of the residuals nor of the row or column effects should have an influence on the choice of the transformation. What we are interested in is being able to distinguish between significant row or column effects and residuals. Therefore, the spread of the row or column effects should be large whereas at least most of the absolute values of the residuals should be small.

To measure the spread of the row or column effects, [1] uses the interquartile range which is a robust measure of spread and not sensitive to outliers like the variance. The interquartile range is the difference between the 75%- and the 25%-quantile, i.e. the range that contains 50% percent of the data in the middle. They use the 80% quantile of the absolute values of all residuals to judge whether most of the residuals are small. One should not expect all residuals to be small. There might still be single outliers that are of high interest.

Finally, they compute the quotient of the interquartile range of the row or column effects and divide it by the 80% quantile of the absolute values of all residuals. They call this quotient the IQRoQ value (InterQuartile Range over the 80% Quantile of the absolute residuals). The higher the IQRoQ value, the better is the result of median polish. For each value of λ, the corresponding power transformation is applied to the contingency table and calculate the IQRoQ value. In this way, we obtain an IQRoQ plot, plotting the IQRoQ value depending on λ.

4 Transformations and Additive Generators of Fuzzy Logic Connectives

It is very interesting to note the similarity between the operator $\oplus_\lambda$ and t-norms / t-conorms [4] in fuzzy logic.

On the one hand, the above family of power transformations closely resemble the Schweizer-Sklar family of additive generators of t-norms. In fact, the power transformations are nothing but the negative of the additive generator of the Schweizer-Sklar t-norms. Note that additive generators of t-norms are non-increasing, and in the case of continuous t-norms they are strictly decreasing, which explains the need for a negative sign to make the function decreasing.

On the other hand, given continuous and strict additive generators, one constructs t-norms / t-conorms precisely by using Eq. (4).

However, it should be emphasised that additive generators of t-norms or t-conorms cannot be directly used here. The additive generator of a t-norm is non-increasing while one requires a transformation to maintain the monotonicity in the arguments. In the case of the additive generator of a t-conorm, though monotonicity can be ensured, their domain is restricted to just $[0, 1]$. This can be partially overcome by normalising the data to fall in this range. However, this type of normalisation may not be reasonable always. Further, the median polish algorithm applied to the transformed data do not always remain positive and hence determining the inverse with the original generator is not possible.

The above discussion leads us to consider a suitable modification of the additive generators of t-norms / t-conorms that can accommodate a far larger

range of values both in their domain and co-domain. Representable uninorms are another class of fuzzy logic connectives that are obtained by the additive generators of both a t-norm and a t-conorm. In this work, we construct new transformations by suitably modifying the underlying generators of these representable uninorms [4].

4.1 Modified Additive Generators of Uninorms : An Example

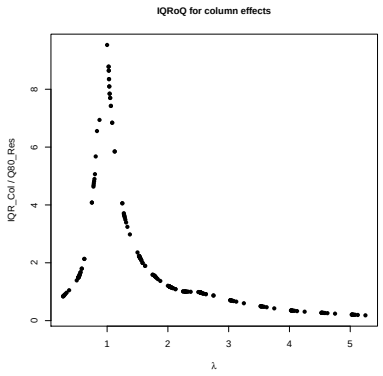
Let us assume that the data x are coming from $(-M, M)$. Consider the following modified generator of the uninorm obtained from the additive generators of the Schweizer-Sklar family of t-norms and t-conorms. Let $e \in (-M, M)$ be any arbitrary value. Then the following is a valid transformation with $h_\lambda : [-M, M] \to \left[\frac{(-M)^\lambda - e^\lambda}{\lambda}, \frac{1}{\lambda}\right]$, for all $\lambda \in [-\infty, 0[\cup]0, \infty]$.

$$h_\lambda(x) = \begin{cases} \dfrac{x^\lambda - e^\lambda}{\lambda}, & x \in [-M, e] \\ \dfrac{1 - \left(\dfrac{M-x}{M-e}\right)^\lambda}{\lambda}, & x \in [e, M] \end{cases};$$

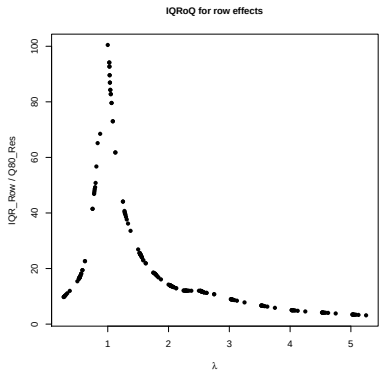
$$(h_\lambda)^{-1}(x) = \begin{cases} (x\lambda + e^\lambda)^{\frac{1}{\lambda}}, & x \leq 0 \\ M - (M-e)\left[(1 - x\lambda)\right]^{\frac{1}{\lambda}}, & x \geq 0 \end{cases}.$$

Note that h_λ is monotonic for all $\lambda \in [-\infty, 0[\cup]0, \infty]$ and increases with decreasing λ.

That this modified generator is a reasonable transform can be seen by applying on the following data. Consider the 10×10 table generated by the following additive model. The overall effect is 0, the row effects are 10, 20, 30, . . . , 100, the column effects are 1, 2, 3, . . . , 10. To each of these entries is added a noise from a uniform distribution over the interval [-0.5, 0.5]. From the IQRoQ plots for this data given in Figure 1, it can be seen that the global maxima occur at $\lambda = 1$. So the IQRoQ plots propose to apply the above transformation with $\lambda = 1$ which is a linear transformation of the data.



(a) Artificial Data, $e = 5, L = 110$, IQRoQ Column Plot



(b) Artificial Data, $e = 5, L = 110$, IQRoQ Row Plot

Fig. 1 IQRoQ plots for the column and row effects of the Artificial data with Modified Schweizer-Sklar generator

4.2 A Novel Way of Finding a Suitable Transformation

In this section we present the algorithm to find a suitable transformation of the given data such that the MP algorithm performs well to elucidate the underlying structures in the data. We only consider a one parameter family of operators with the parameter denoted by λ.

The proposed algorithm is as follows. Let $\oplus_\lambda$ denote the one parameter family of operators whose domain and range allow it to be operated on the data given in the contingency table. Then for each λ the following steps are performed:

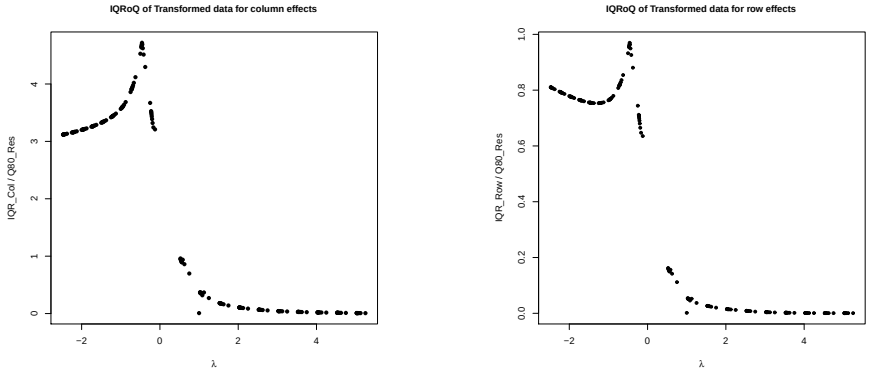
1. Apply the transformation $\oplus_\lambda$ to the contingency table.
2. Apply the median polish algorithm to find the overall, row and column effects, viz., $\tilde{g}, \tilde{r_i}, \tilde{c_j}$ for each i, j.
3. Find the residuals $\varepsilon_{ij} = x_{ij} - g \oplus_\lambda \tilde{r_i} \oplus_\lambda \tilde{c_j}$ for each i, j.
4. Determine the IQRoQ values of the above residuals.

Finally, we plot λ versus the above IQRoQ values to get the IQRoQ plots for the column and row effects.

Clearly, the operator corresponding to the λ at which the above IQRoQ plots peak is a plausible transformation for the given contingency table. Though, a rigorous mathematical analysis and support for the above statement is not immediately available, an intuitive explanation is clear from the earlier work of Klawonn *et al.* [1]. Further, we illustrate the same by applying the above h_λ transformations on some real data sets and present our results in the next section.

Overall: 0.2919985						
	≤ 8	$9-11$	12	$13-15$	≥ 16	RE
NW	0.00025312	0.0027983	-0.00025004	-0.010879	0.0000000	-0.010113225
NC	-0.00025312	-0.0027983	-0.01200293	0.010879	0.0078014	0.006694490
S	0.01098492	0.0091121	0.00025004	-0.044525	-0.0035433	-0.001558958
W	-0.01102793	-0.0305895	0.00456985	0.014641	0.0000000	0.001558958
CE	0.0318984143	0.0293532152	-0.0112376220	0.0002531186	-0.0294192135	

Table 3 Median polish on the h_λ-transformed Infant Mortality data with $\lambda = -0.5$



(a) $e = 2, M = 40$, IQRoQ Column Plot (b) $e = 2, M = 40$, IQRoQ Row Plot

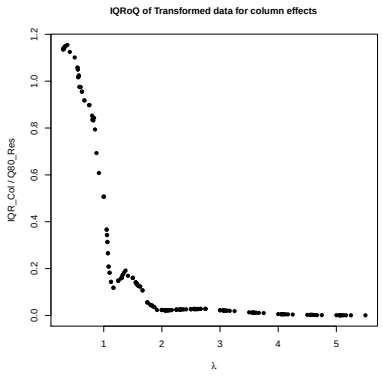
Fig. 2 IQRoQ plots for the column and row effects of the Infant Mortality data

4.3 Some Illustrative Examples

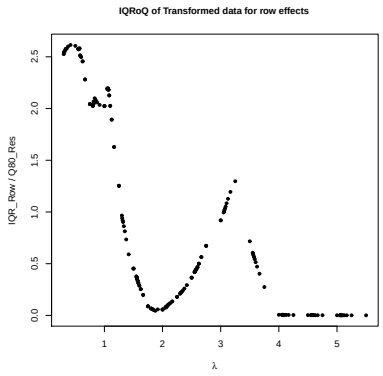
Let us consider the data given in the Contingency table Table 1. Applying the above algorithm with the transformation h_λ we obtain the IQRoQ plots - Figures 2(a) and (b) - which suggest a value of around $\lambda = -0.5$. The 'median polished' contingency table for $\lambda = -0.5$ is given in Table 3.

We finally consider two larger contingency tables with 14 rows and 97 columns that are far too large to be included in this paper. The tables consist of a data set displaying the metabolic profile of a bacterial strain after isolation from different tissues of a mouse. The columns reflect the various substrates whereas the rows consist of repetitions for the isolates from tumor and spleen tissue. The aim of the analysis is to identify those substrates that can be utilized by active enzymes and to find differences in the metabolic profile after growth in different organs.

The corresponding IQRoQ plots shown in Figures 3(a) and (b) suggest that we choose a value of around $\lambda = 0.4$. The 'median polished' contingency table for $\lambda = 0.4$ shows that the number of residuals that are larger than the absolute value of most of the row or column effects is roughly 50%.



(a) $e = 10, M = 20000$, IQRoQ Column Plot



(b) $e = 10, M = 20000$, IQRoQ Row Plot

Fig. 3 IQRoQ plots for the column and row effects of the Spleen data

5 Conclusions

In this work, we have shown that that the Median Polish algorithm does not always give interpretable results when applied to raw contingency tables. This necessitates a transformation of the data. However, both the choice of the transformation and the fact that the transformation leads to changing the underlying model of the data from a simple additive to a multiplicative model become an issue. We have proposed a novel way of applying the MP algorithm even in this case that still gives reasonable results. Our approach to the underlying model leads us to transformations that are similar to additive generators of some fuzzy logic connectives. Further, we have illustrated how to choose a suitable transformation that gives meaningful results.

References

1. Klawonn, F., Crull, K., Kukita, A., Pessler, F.: Median polish with power transformations as an alternative for the analysis of contingency tables with patient data. In He, J., Liu, X., Krupinski, E., Xu, G., eds.: Health Information Science, Berlin, Springer (2012) 25–35
2. Hoaglin, D., Mosteller, F., Tukey, J.: Understanding Robust and Exploratory Data Analysis. Wiley, New York (2000)
3. Berthold, M., Borgelt, C., Höppner, F., Klawonn, F.: Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data. Springer, London (2010)
4. Klement, E., Mesiar, R., Pap, E.: Triangular Norms. Kluwer Academic Publishers, Dodrecht (2000)