Application of Bayesian configural frequency analysis (BCFA) to determine characteristics user and non-user motor X

Muhamad Iqbal Mawardi¹, Septiadi Padmadisastra², Bertho Tantular³
Statistics Department, Faculty of Mathematics and Science, Universitas Padjadjaran, Indonesia
Email: 1992iqbal@gmail.com

Abstract. Configural Frequency Analysis is a method for cell-wise testing in contingency tables for exploratory search type and antitype, that can see the existence of discrepancy on the model by existence of a significant difference between the frequency of observation and frequency of expectation. This analysis focuses on whether or not the interaction among categories from different variables, and not the interaction among variables. One of the extensions of CFA method is Bayesian CFA, this alternative method pursue the same goal as frequentist version of CFA with the advantage that adjustment of the experiment-wise significance level α is not necessary and test whether groups of types and antitypes form composite types or composite antitypes. Hence, this research will present the concept of the Bayesian CFA and how it works for the real data. The data on this paper is based on case studies in a company about decrease Brand Awareness & Image motor X on Top Of Mind Unit indicator in Cirebon City for user 30.8% and non user 9.8%. From the result of B-CFA have four characteristics from deviation, one of the four characteristics above that is the configuration 2212 need more attention by company to determine promotion strategy to maintain and improve Top Of Mind Unit in Cirebon City.

1. Introduction

Configural Frequency Analysis (CFA) is a statistical method that allows one to determine whether unexpected events are significantly different than expected. This analysis focuses on whether or not the interaction among categories from different variables exist and not the interaction among the variables. The goal of CFA is to detect configuration pattern in the data whether experiencing deviation based on discrepancies from the model, which is marked by the existence of difference beetween the frequency of observation and expectation. In addition, CFA (Von Eye, 2002) is a method for identifying the configuration of categorical variables in multidimensional cross-tables. In CFA, deviation on the configurations will occur when an event more often or more rarely appears from what has been previously expected. An event occurred more often than what has been previously expected is called type, while an event rarely occurred from what has been previously expected is called antitype. The types of data used in CFAs are category pairs. This is based on the notion of configuration (Lienert in Von Eye 2002) which is a category pair that describes a cell of a cross-table.

CFA has been develop intensively since its first presentation in 1968 and is now among the more popular methods of data analysis. One of the reason for the increasing popularity of CFA is that the results of this method of data analysis are deemed easy to interpret. CFA is performed in five steps. The first step involves selecting a base model. The second step involves the selection of a concept of deviation from independence. The third step involves the selection of significance test. The fourth step involves estimation (or determining) the expected cell frequencies, performing the significance test,
and identifying those configurations that constitute types or antitypes. The fifth step involves interpreting types or antitypes using substantive background information.

Consider (Gutiérrez-Peña and von Eye, 2000) the cross-classification of \( d \geq 2 \) categorical variables. Let \( M \) be the vector of the model frequencies in this cross-classification, and \( m_i \) the observed frequency in cell \( i \) where \( i \) goes over all cells in the cross-classification. Let \( \hat{m}_i \) be the estimated expected cell frequency for cell \( i \) and \( \pi_i \) the population probability for cell \( i \). A log-frequency model for this cross-classification can be described as \( \log M = X\lambda \), where \( M \) is the vector of model cell frequencies, \( X \) is the design matrix and \( \lambda \) is the parameter vector. CFA asks whether the model prediction, \( \hat{m}_i \), deviates significantly from the observed value \( m_i \). Most popular are the \( z \)-test \( z = m_i - \hat{m}_i / \sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)} \), where the \( \pi_i \) are estimated by \( \hat{m}_i / n \), and Pearson \( X^2 \) component test, \( X_i^2 = (m_i - \hat{m}_i)^2 / \hat{m}_i \), \( (df = 1) \). The hypothesis \( H_0: E(m_i, \hat{m}_i) = 0 \) holds, one applies either of these or one of the other tests discussed in the context CFA under the appropriate measure for protection of the experiment-wise \( \alpha \). The most frequently used methods for \( \alpha \) protection are Bonferroni method, where an adjusted \( \alpha \) is calculated as \( \alpha' = \alpha / t \), where \( \alpha \) is the significance threshold, \( \alpha = 0.05 \) and \( t \) is the number of test. Cases in cell \( i \) are said to belong to type if \( E(m_i, \hat{m}_i) > 0 \) and antitype if \( E(m_i, \hat{m}_i) < 0 \). Bayesian CFA is one of the extensions of the CFA frequentist methods. Bayesian CFA (Gutiérrez-Peña, 2012) defines types and antitypes in terms of the true (unknown) values of the parameters, which can be estimated but will never be observed or fully known. One can (1) search for types and antitypes as before with the advantage that adjustment of the experiment-wise significance level \( \alpha \) is not necessary; and (2) test whether groups of types and antitypes form composite types or composite antitypes. Both CFA frequentist and Bayesian CFA are capable of assigning probabilities to patterns of types and antitypes, and thus allow for the comparison of such patterns by means of relative probabilities.

An interesting feature of the Bayesian approach is that it allows us to calculate the posterior probability of any specific pattern of types and antitypes in a cross-classification. Given a particular base model, the posterior distribution of \( \pi \) induces a probability distribution on the set of all possible patterns. In practice pattern in a neighbourhood of the particular pattern suggested by an exploratory analysis which looks at each cell individually. Different to CFA frequentist, from Bayesian viewpoint the posterior distribution of \( \pi \) produces joint probability statements concerning possible patterns of types and antitypes in a given group of cells. There is no need for Bonferroni-type corrections within the Bayesian framework.

2. Bayesian CFA

2.1 Prior and Posterior Distribution

In this section we introduce some notation (Gutiérrez-Peña and von Eye, 2000). Let \( R \) denote the total number of cells in the cross-classification and let \( \pi_i \) denote the population probability for cell \( i \) \((i = 1, \ldots, R) \). Finally let \( \pi \) be the vector of such probabilities. Under multinomial sampling, the vector of observed frequencies, \( m = (m_1, \ldots, m_R) \) can be regarded as an observation from a \((R-1)\) dimensional multinomial distribution with index \( N = \sum m_i \) and unknown parameter vector \( \pi \). From Bayesian point of view, all prior belief concerning the value of \( \pi \) must be described in terms of a prior distribution. The usual conjugate prior for the multinomial parameter Dirichlet distribution. This distribution is characterized by a parameter vector \( \beta = (\beta_1, \ldots, \beta_R) \) such that \( E(\pi_i) = \beta_i / \beta_* \), where \( \beta_* = \sum \beta_i \). The value of \( \beta_i \) is interpreted as a hypothetical prior sample size and determines the strength of the information contained the prior: a small \( \beta_i \) implies vague prior information whereas a large \( \beta_* \) suggests strong prior belief about \( \pi \). The posterior distribution is also Dirichlet with parameter \( \beta_m = (m_1 + \beta_1, \ldots, m_R + \beta_R) \). This distribution contains all the available information about the population proportion \( \pi \) conditional on the observed frequencies.

In the absence of prior information, also called a non informative prior will typically be used. Of the most widely used methods to derive non informative priors is the called Jeffreys’ rule, which is in this case yields for the multinomial parameter is precisely the Dirichlet distribution with parameter \( \beta = (1/2, \ldots, 1/2) \). One of the nice features of this prior is that it is conjugate (close under sampling).
Keeping in mind the interpretation of $\beta_i$ as a prior sample size, the quantity $i = \beta_i/(N + \beta_i)$ can be regarded as the proportion of the total information that is contributed by the prior. Since $E(\pi_i) = \beta_i/\beta_*$, the individual values of $\beta_i$ should be chosen according to the prior belief concerning $E(\pi_i)$.

The Dirichlet prior provides a simple way to elicit prior information concerning prior population probabilities, but it may not be flexible enough to describe specific prior information about other parameter such as interactions (Knuiiman and Speed, 1998).

Any base model imposes constraints on the space of possible values of $\pi$, in other words under the base model, the population proportion for cell $i$ is given by $\pi_i = f_i(\pi)$ for some function $f_i$. Consider a 2×12 cross-classification and base model which state that the three variables are independent, then:

$$f_1(\pi) = (\pi_1 + \pi_2) \cdot (\pi_3 + \pi_5 + \pi_7 + \pi_9 + \pi_{11} + \pi_{15} + \pi_{17} + \pi_{19} + \pi_{21} + \pi_{23})$$

$$f_2(\pi) = (\pi_1 + \pi_2) \cdot (\pi_3 + \pi_4 + \pi_6 + \pi_8 + \pi_{10} + \pi_{12} + \pi_{14} + \pi_{16} + \pi_{18} + \pi_{20} + \pi_{22} + \pi_{24})$$

$$f_3(\pi) = (\pi_3 + \pi_4) \cdot (\pi_1 + \pi_3 + \pi_5 + \pi_7 + \pi_9 + \pi_{11} + \pi_{13} + \pi_{15} + \pi_{17} + \pi_{19} + \pi_{21} + \pi_{23})$$

$$\vdots$$

$$f_{24}(\pi) = (\pi_{23} + \pi_{24}) \cdot (\pi_2 + \pi_4 + \pi_6 + \pi_8 + \pi_{10} + \pi_{12} + \pi_{14} + \pi_{16} + \pi_{18} + \pi_{20} + \pi_{22} + \pi_{24})$$

We can calculate the posterior probability of any event involving the population proportions, from posterior distribution of $\pi$. In particular, that posterior distribution induces a posterior distribution on the vector $\pi$ of model based proportion. Even if $\pi_i \neq \pi_i^*$, we would be unwilling to classify cell $i$ as type or antitype unless $\pi_i - \pi_i^*$ was significantly greater than zero or less than zero.

Any base model imposes constraints on the possible values of $\pi$. In other words, under the base model the population probability of cell $i$ is given by $\pi_i = f_i(\pi)$ for some functions $f_i$. The base model can be tested on the basis of the posterior distribution of $\delta = \sum_i \log(\pi_i/\pi_i^*) \pi_i$. This quantity is always nonnegative and is zero if and only if the base model is correct. Posterior distributions of $\delta$ is not available in close form, it can easily be obtained from that $\pi$ using Monte Carlo techniques, that posterior distributions concentrated near zero support the base model, where as posterior distributions located away from zero lead to rejection of the base model.

2.2 Types and Antitypes from Bayesian Perspective

If we knew (Gutiérrez-Peña, 2012) the actual value of $\pi$, then cell $i$ could be regarded as a type if $\pi_i > \pi_i^*$, and as an antitype if $\pi_i < \pi_i^*$. However, even if $\pi_i \neq \pi_i^*$, we would be unwilling to classify cell $i$ as a type and antitype unless $\pi_i - \pi_i^*$ was significantly greater and less than zero. This suggests the following definition of types and antitypes: cell $i$ is regarded as a type if and only if $\pi_i < \pi_i^*$ and antitype in and only if $\pi_i - \pi_i^* < 1$, where $\pi_i$ and $\pi_i^*$ are suitable threshold values (von Eye and Gutiérrez-Peña, 2004). From the posterior distribution of $\pi$ we can compute the posterior probability of cell $i$ being a type namely, $Pr(\pi_i < \pi_i^*)$.

2.3 Patterns Types and Antitypes

We now turn to the Bayesian analysis (von Eye and Gutiérrez-Peña, 2004) of cross classification under product multinomial sampling. In order to simplify the exposition, we shall only consider discrimination between two group on the basis of two explanatory variables arranged in a $a \times b$ cross classification for each of the groups. More general situations can be handled in a similar fashion. For example a $3 \times 2$ cross classification, then such possible patterns include (T for type, A for antitype, and N for neither)

$$\begin{pmatrix} T & T \\ T & T \end{pmatrix}, \begin{pmatrix} T & A \\ T & A \end{pmatrix}, \begin{pmatrix} N & T \\ T & T \end{pmatrix}, \begin{pmatrix} N & A \\ T & T \end{pmatrix}, \begin{pmatrix} A & A \\ A & A \end{pmatrix}$$

A Bayesian solution to the CFA problem would be to report the most probability pattern. For problems of moderate size, the number of all possible patterns may be too large for direct implementation of this approach to be feasible. The calculations of all the posterior probabilities reported based on the simulation of 10000 samples from the Dirichlet distribution of $\pi$ (von Eye and Gutiérrez-Peña, 2000).
3. Bayesian Analysis of the ‘Top of Mind Motor X’ Data

In this section, we present the data used are XYZ Company survey on Brand Awareness & Image characteristics in Cirebon City as many as 190 respondents. There are 4 factors to look at:

1. Top Of Mind
   Top of Mind X motors are divided into two categories: X (1) and Competitor (2).
2. Jobs
   The Jobs is divided into three categories: Student (1), Employee (2), Others (3).
3. Spending per month
   Spending per month is divided into two categories: ≤ Rp. 1.250.000 (1) and > Rp. 1.250.000 (2).
4. Motor cycle Users
   Motor cycle users are divided into two categories: Users (motor cycle users) (1) and Non User (potential motor cycle users) (2).

Now consider the base model:

\[
\log \mu_{ijkl} = \lambda_0 + \lambda_A^i + \lambda_B^j + \lambda_C^k + \lambda_D^l + \lambda_{AB} + \lambda_{AC} + \lambda_{AD} + \lambda_{BC} + \lambda_{BD} + \lambda_{CD} + \lambda_{ABC} + \lambda_{ABD} \\
+ \lambda_{A^CD} + \lambda_{B^CD} + \lambda_{A^BCD} 
\]

Where \( \mu_{ijkl} \) is the model frequency for cell \( ijk l \), with \( \lambda_0 = \) intercept, \( \lambda_A = \) job, \( \lambda_B = \) spending per month, \( \lambda_C = \) motor cycle users, \( \lambda_D = \) top of mind and the other is interaction between variables. Suppose now we have independent (non informative) Dirichlet priors so that corresponding posterior distributions are also Dirichlet distribution. In order to calculate the posterior probability that cell \( ijk l \) is a type, we use a Monte Carlo approach. The base model can be tested on the basis of the posterior distribution of

\[
\delta = \sum_i \sum_j \sum_k \sum_l \log \left( \frac{\pi_i}{\pi_j} \right) \pi_j 
\]

Table 1 shows the results of the Bayesian CFA of the top of mind data, we use R Software and in Figure 1, we show a histogram of the posterior distribution of \( \delta \). This distribution puts most of its mass away from zero, indicating significant discrepancies between the data and the expectancies under the assumption of independence. Notice that both the classical and the Bayesian CFA lead to the same results in this case. The posterior probability of the pattern provided by the two analyses is 0.126. The posterior distribution over the space of all possible patterns is much more concentrated around the pattern obtained from the preliminary analysis.

| Configuration | Obs | Pr(Type) | Pr(Neither) | Pr(Antype) | Type/Antitype |
|---------------|-----|----------|-------------|------------|---------------|
| 1111          | 14  | 0.0559   | 0.9341      | 0.0100     |               |
| 1121          | 2   | 0.0100   | 0.9341      | 0.0559     |               |
| 1211          | 4   | 0.0163   | 0.9420      | 0.0417     |               |
| 1221          | 3   | 0.0417   | 0.9420      | 0.0163     |               |
| 2111          | 12  | 0.0324   | 0.9469      | 0.0207     |               |
| 2121          | 2   | 0.0207   | 0.9469      | 0.0324     |               |
| 2211          | 18  | 0.1012   | 0.8926      | 0.0062     |               |
| 2221          | 7   | 0.0062   | 0.8926      | 0.1012     |               |
| 3111          | 4   | 0.2700   | 0.7293      | 0.0007     |               |
| 3121          | 13  | 0.0007   | 0.7293      | 0.2700     |               |
| 3211          | 16  | 0.1600   | 0.8379      | 0.0021     |               |
|    |   |     |     |     |
|----|---|-----|-----|-----|
| 3221 | 8 | 0.0021 | 0.8379 | 0.1600 |
| 1112 | 10 | 0.0000 | 0.2064 | 0.7936 |
| 1122 | 2 | 0.7936 | 0.2064 | 0.0000 |
| 1212 | 3 | 0.0002 | 0.7682 | 0.2316 |
| 1222 | 1 | 0.2316 | 0.7682 | 0.0002 |
| 2112 | 5 | 0.0061 | 0.9175 | 0.0764 |
| 2122 | 0 | 0.0764 | 0.9175 | 0.0061 |
| 2212 | 31 | 0.6389 | 0.3611 | 0.0000 |
| 2222 | 11 | 0.0000 | 0.3611 | 0.6389 |
| 3112 | 5 | 0.0317 | 0.9512 | 0.0171 |
| 3122 | 3 | 0.0171 | 0.9512 | 0.0317 |
| 3212 | 13 | 0.2175 | 0.7816 | 0.0009 |
| 3222 | 3 | 0.0009 | 0.7816 | 0.2175 |

**Figure 1.** Posterior distribution of $\delta$ for the top of mind data

### 4. Conclusion

Bayesian Configural Frequency Analysis (B-CFA) is capable of assigning probabilities to patterns of types and antitypes, and thus allows for the comparison of such patterns by means of relative probabilities. Recommendations can be derived as to which approach to select for data analysis, when single types and antitypes are of interest and the observed cells frequencies are deemed valid for the estimation of expected cell frequencies. The Bayesian approach presented allows for joint inferential statements about groups of cells and does not require Bonferroni type arguments.

There are eight characteristics that have been determined from deviation, based on the analysis type means observed value is more often than expected value appear on:

- **1122:** respondents with job as student, spending per month \( \leq Rp. 1.250.000.00 \), as a non user, and Top of Mind competitor

Whereas, antitype means observed value is less often than expected value appear on:

- **2212:** respondents with job as employee, spending per month \( > Rp. 1.250.000.00 \), as a user, and Top of Mind competitor.
1112: respondents with job as student, spending per month ≤ Rp. 1.250.000,00, as a user, and Top of Mind competitor.

2222: respondents with job as employee, spending per month > Rp. 1.250.000,00, as a non user, and Top of Mind competitor.

From that descriptions, recommendations for XYZ Company is need to pay special attention to this characteristics especially 2212, this will be very useful to determine strategy to maintain and improve Top Of Mind Unit in Cirebon City.

5. References

[1] Gutiérrez-Peña, E. and Von Eye, A. (2000). A Bayesian Approach to Configural Frequency Analysis. Journal of Mathematical Sociology, 24, 151-174.

[2] Gutiérrez-Peña, E. (2012). Bayesian predictive Configural Frequency Analysis. Journal Psychological Test and Assessment Modeling, 54, 285-292.

[3] Knuiman, M. W. and Speed, T. P. (1988) Incorporating prior information into the analysis of contingency tables, Biometrics, 44, pp. 1061-1071.

[4] Mawardi, M. I. 2014. Application of Configural Frequency Analysis (CFA) to Determine Characteristics User dan Non User Motor X in Cirebon City (Case Study PT XYZ). Undergraduate Thesis. Statistics Department, Faculty of Mathematics and Science, State University of Padjadjaran, Bandung, Indonesia.

[4] Von Eye, A. 2002. Configural Frequency Analysis: Methods, Models, and Applications. Lawrence Erlbaum Associates: London.

[5] Von Eye, A. and Brandstätter, J. (1997). Configural Frequency Analysis as a searching device for possible causal relationships. Methods of Psychological Research - Online, 2, 1-23

[6] Von Eye, A. And Gutiérrez-Peña, E (2004). Configural frequency Analysis – the search for extreme cells. Journal of Applied Statistics, 31, 981–997.