Psychological Test and Assessment Modeling, Volume 54, 2012 (3), 285-292

Bayesian predictive Configural Frequency Analysis

Eduardo Gutiérrez-Peña[1]

Abstract

Configural Frequency Analysis is a method for cell-wise inspection of cross-classifications. CFA searches for patterns of variable categories that occur either more often or less often than expected from a given base model. In this paper, we propose and discuss an alternative notion of types and antitypes that focuses on the likely values of the cell frequencies in future experiments, as opposed to the average values of such frequencies. The idea is developed from a Bayesian point of view.

Key words: Bayesian methods, Configural Frequency Analysis, predictive distribution

[1] *Correspondence concerning this article should be addressed to:* Eduardo Gutiérrez-Peña, PhD, Departamento de Probabilidad y Estadística, IIMAS-UNAM, Apartado Postal 20-726, 01000 México, D.F., Mexico; email: eduardo@sigma.iimas.unam.mx

Introduction

Configural Frequency Analysis (Lienert, 1969) is a method for cell-wise inspection of cross-classifications. CFA searches for "types" (respectively, "antitypes"), that is, patterns of variable categories that occur more often (respectively, less often) than expected from a given base model. Bayesian CFA (Gutiérrez-Peña and von Eye, 2000) 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. On the other hand, the Bayesian predictive CFA introduced in this paper focuses on the likely values of the cell frequencies in future experiments, as opposed to the average values of such frequencies. Both Bayesian CFA and Bayesian predictive 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.

In the next section we review the Bayesian approach to CFA. We then introduce the Bayesian predictive CFA and illustrate the method using a data set previously analyzed in the literature. Finally, the last section contains some concluding remarks.

Bayesian CFA

In this section we review the work of Gutiérrez-Peña and von Eye (2000) and introduce some notation. Consider a cross classification of $d \geq 2$ categorical variables. Let π_i denote the population probability for cell i (i=1,2,...K), and let $\boldsymbol{\pi}$ be the vector of such probabilities. For the sake of simplicity, here we shall only be concerned with multinomial sampling, where $\boldsymbol{M}$ (the vector of observed frequencies) can be regarded as an observation from a $(K-1)$-dimensional multinomial distribution with index $N = \sum_i m_i$ and unknown parameter vector $\boldsymbol{\pi}$. Other sampling schemes can be dealt with in a similar fashion.

From a Bayesian point of view, beliefs concerning the value of $\boldsymbol{\pi}$ must be described in terms of a *prior distribution*. The usual conjugate prior for the multinomial parameter is the Dirichlet distribution. This distribution is characterized by a parameter vector $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)$ such that $E(\pi_i) = \beta_i / \beta_*$, where $\beta_* = \sum_i \beta_i$. Small values of β_* imply vague prior information; in particular, the well-known Jeffreys' rule corresponds to $\boldsymbol{\beta} = (1/2,\ldots,1/2)$. The *posterior distribution* of $\boldsymbol{\pi}$ is also Dirichlet, with parameter $\boldsymbol{\beta} = (m_1 + 1/2,\ldots,m_K + 1/2)$. This distribution contains all the available information about the population probabilities $\boldsymbol{\pi}$, conditional on the observed contingency table.

Any base model imposes constraints on the possible values of $\boldsymbol{\pi}$. In other words, under the base model the population probability of cell i is given by $\pi_i^* = f_i(\boldsymbol{\pi})$ for some functions f_i. As simple example, consider a 2 × 2 cross classification and a base model which states that the two variables are independent. Then

$$\pi_1^* = \tilde{\pi}_{11}^* = f_1(\boldsymbol{\pi}) = (\pi_1 + \pi_3)\times(\pi_1 + \pi_2)\,,\ \pi_2^* = \tilde{\pi}_{12}^* = f_2(\boldsymbol{\pi}) = (\pi_1 + \pi_2)\times(\pi_2 + \pi_4)\,,$$
$$\pi_3^* = \tilde{\pi}_{21}^* = f_3(\boldsymbol{\pi}) = (\pi_3 + \pi_4)\times(\pi_1 + \pi_3) \text{ and } \pi_4^* = \tilde{\pi}_{22}^* = f_4(\boldsymbol{\pi}) = (\pi_3 + \pi_4)\times(\pi_2 + \pi_4)\,.$$

The base model can be tested on the basis of the posterior distribution of

$$\delta = \sum_i \log\left(\frac{\pi_i}{\pi_i^*}\right)\pi_i \,.$$

This quantity is always nonnegative and is zero if and only if the base model is correct. Thus, posterior distributions of δ concentrated near zero support the base model, whereas posterior distributions located away from zero lead to rejection of the base model.

Types and antitypes from a Bayesian perspective. If we knew the actual value of $\boldsymbol{\pi}$, 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 (respectively, antitype) unless $\pi_i - \pi_i^*$ was significantly greater than zero (respectively, less than zero). This suggests the following definition of types and antitypes: Cell *i* is regarded as a type if and only if $u_i < \pi_i - \pi_i^*$, and as an antitype if and only if $\pi_i - \pi_i^* < l_i$, where u_i and l_i are suitable threshold values (see von Eye and Gutiérrez-Peña, 2004). From the posterior distribution of $\boldsymbol{\pi}$ we can (for example) compute the posterior probability of Cell *i* being a type, namely, $\Pr(u_i < \pi_i - \pi_i^*)$.

Patterns of types and antitypes. An interesting feature of the Bayesian approach is that it allows us to calculate the joint posterior probability of several cells being all types simultaneously. More generally, we can 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 $\boldsymbol{\pi}$ induces a probability distribution on the set of all possible patterns. Consider, for example, a 2×2 cross classification. Then such possible patterns include

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

where *T* stands for 'type', *A* for 'antitype', and *N* for 'neither'.

A Bayesian solution to the Configural Frequency Analysis problem would then be to report the *most probable pattern*. However, even for problems of moderate size, the number of all possible patterns may be too large for a direct implementation of this approach to be feasible. In practice, we can dramatically reduce the numerical burden of the approach described above if we only look at patterns in a 'neighbourhood' of the particular pattern suggested by an exploratory analysis which looks at each cell individually. In this way, we can then compare two or more plausible patterns in terms of their relative posterior probability.

The Bayesian predictive approach

In this section, we introduce an alternative notion of types and antitypes that focuses on the likely values of the cell frequencies in future experiments, as opposed to the average values of such frequencies. This approach may be useful in developmental or any other research area in which repeated measurement designs are employed.

Types and antitypes. Let $\tilde{m}_i$ denote the (as yet unobserved) count in Cell i for a future experiment, and $\tilde{m}_i^*$ the corresponding count assuming the base model is true. The posterior predictive distribution of the $\tilde{m}_i$'s (*i.e.*, the conditional distribution of the $\tilde{m}_i$'s given the observed counts $\boldsymbol{M}$) can be readily obtained from the multinomial sampling model for $\boldsymbol{M}$ and the posterior distribution of the parameters $\boldsymbol{\pi}$. On the other hand, the $\tilde{m}_i^*$'s are a function of the $\tilde{m}_i$'s just as the π_i^*'s are a function of $\boldsymbol{\pi}$.

Thus, Cell i can be regarded as a type if $E(\tilde{m}_i^* | \boldsymbol{M}) < m_i$, and as an antitype if $m_i < E(\tilde{m}_i^* | \boldsymbol{M})$. As in the previous case, however, even if $m_i \neq E(\tilde{m}_i^* | \boldsymbol{M})$ we would be unwilling to classify Cell i as a type (respectively, antitype) unless $m_i - E(\tilde{m}_i^* | \boldsymbol{M})$ was significantly greater than zero (respectively, less than zero).

This suggests the following predictive definition of types and antitypes: Cell i is regarded as a type if the observed count m_i falls on the right tail of the posterior predictive distribution of $\tilde{m}_i^*$, and as an antitype if the observed count m_i falls on the left tail of that distribution. Specifically, in what follows, Cell i will be labelled as a type if $q_{0.95}(\tilde{m}_i^* | \boldsymbol{M}) < m_i$ and as an antitype if $m_i < q_{0.05}(\tilde{m}_i^* | \boldsymbol{M})$, where $q_\alpha(\tilde{m}_i^* | \boldsymbol{M})$ denotes the α-quantile of the posterior distribution of $\tilde{m}_i^*$. From the joint posterior predictive distribution of the $\tilde{m}_i$'s, we can now compute the corresponding posterior predictive probability of any pattern of type and antitypes.

An example (Alcohol abuse). This example concerns a sample of $N = 108$ adult men who were diagnosed as alcohol abusers (Zucker, 1994). The diagnostic scale had the following four levels:

1 = 'Alcohol user'; 2 = 'Mild abuser'; 3 = 'Severe abuser'; and 4 = 'Alcohol dependent'

Three years later the same individuals were diagnosed again. Diagnostic categories were the same as at Time 1. However, in addition, the diagnosis 0 = 'No user of alcohol' was included. Table 1 displays the 4×5 cross-classification of the diagnoses at the two occasions.

The base model is a log-linear main effects model (first-order CFA). Figure 1 shows the posterior distribution of δ. This distribution is located away from zero, thus suggesting that the base model should be rejected (see the previous section). On the other hand, Table 2 compares the results of the Bayesian CFA and the classical CFA. According to the posterior distribution of $\boldsymbol{\pi}$, the pattern suggested by the Bayesian CFA is more than 30 times as likely as the pattern suggested by the classical CFA. See Gutiérrez-Peña and von Eye (2000) for further details.

Table 1:
Cross-Classification of Alcohol Diagnoses at two Occasions ($N = 108$)

Alcohol Abuse Diagnoses, Categories		Diagnoses at Time 2				
		0	1	2	3	4
Diagnoses at Time 1	1	10	8	1	0	0
	2	8	2	11	3	3
	3	4	2	7	10	3
	4	9	3	4	6	14

Figure 1:
Posterior distribution of δ for the alcohol abuse data

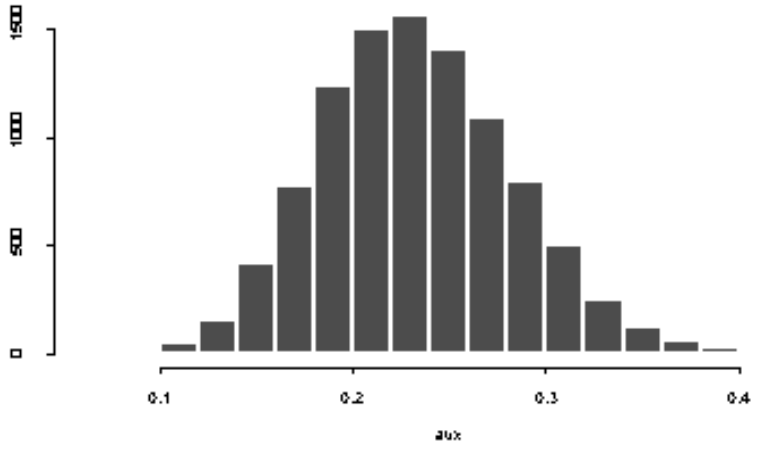

Table 3 shows the results of the Bayesian predictive CFA and compares them with those obtained from the usual Bayesian CFA as displayed in Table 2. In this case, according to the posterior predictive distribution of $\tilde{m}_i^*$'s, the pattern suggested by the Bayesian predictive CFA is about 10 times as likely as the pattern suggested by the classical CFA, and considerably more likely than the pattern suggested by the usual Bayesian CFA. Perhaps more interestingly, even under the original Bayesian CFA (*i.e.*, with respect to the posterior distribution of $\boldsymbol{\pi}$), the pattern suggested by the Bayesian predictive CFA is about 10 times as likely as the pattern suggested by the usual Bayesian CFA.

We have also analysed a data set concerning sleep behaviour and previously discussed in Gutiérrez-Peña and von Eye (2000). In that case, all three approaches lead to the same pattern of types and antitypes.

Table 2:
Bayesian CFA for the alcohol abuse data

Configuration	Obs. Freq.	Pr(Type)	Pr(Neither)	Pr(Antitype)	B-CFA	Classical CFA
1 - 0	10	0.5427	0.4573	0.0000	Type	
1 - 1	8	0.7351	0.2649	0.0000	Type	Type
1 - 2	1	0.0000	0.3585	0.6409	Antitype	
1 - 3	0	0.0000	0.1017	0.8983	Antitype	
1 - 4	0	0.0001	0.0799	0.9200	Antitype	
2 - 0	8	0.0419	0.9439	0.0142		
2 - 1	2	0.0026	0.7968	0.2006		
2 - 2	11	0.6472	0.3528	0.0000	Type	Type
2 - 3	3	0.0041	0.8062	0.1897		
2 - 4	3	0.0027	0.7648	0.2325		
3 - 0	4	0.0004	0.5099	0.4897		
3 - 1	2	0.0053	0.8021	0.1926		
3 - 2	7	0.1133	0.8854	0.0013		
3 - 3	10	0.7047	0.2953	0.0000	Type	Type
3 - 4	3	0.0041	0.8038	0.1921		
4 - 0	9	0.0076	0.9065	0.0859		
4 - 1	3	0.0020	0.7637	0.2343		
4 - 2	4	0.0003	0.4925	0.5072		
4 - 3	6	0.0202	0.9480	0.0318		
4 - 4	14	0.9238	0.0762	0.0000	Type	Type

Table 3:
Bayesian predictive CFA for the alcohol abuse data

Configuration	Obs. Freq.	Pr(Type)	Pr(Neither)	Pr(Antitype)	B-CFA	BP-CFA
1 - 0	10	0.2560	0.7434	0.0006	Type	
1 - 1	8	0.5085	0.4915	0.0000	Type	Type
1 - 2	1	0.0009	0.9991	0.0000	Antitype	
1 - 3	0	0.0000	1.0000	0.0000	Antitype	
1 - 4	0	0.0000	1.0000	0.0000	Antitype	
2 - 0	8	0.0421	0.9403	0.0176		
2 - 1	2	0.0050	0.9950	0.0000		
2 - 2	11	0.3360	0.6637	0.0003	Type	
2 - 3	3	0.0118	0.8640	0.1242		
2 - 4	3	0.0112	0.8674	0.1214		
3 - 0	4	0.0035	0.8110	0.1855		
3 - 1	2	0.0055	0.9945	0.0000		
3 - 2	7	0.0706	0.9215	0.0079		
3 - 3	10	0.4022	0.5971	0.0007	Type	
3 - 4	3	0.0098	0.8712	0.1190		
4 - 0	9	0.0213	0.9154	0.0633		
4 - 1	3	0.0062	0.8724	0.1214		
4 - 2	4	0.0018	0.8232	0.1750		
4 - 3	6	0.0280	0.9574	0.0146		
4 - 4	14	0.4157	0.5842	0.0001	Type	

Discussion

Bayesian 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. However, it 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. On the other hand, Bayesian predictive CFA focuses on the likely values of the cell frequencies in future experiments. These can in principle be observed and compared with the values predicted by the model on the basis of previous experiments. This approach may be more appropriate when sample sizes are small. All in all, Bayesian predictive CFA seems to be more conservative than the original Bayesian CFA. This could be due to the fact that there is more uncertainty involved in the prediction of new observations than in the estimation of their corresponding expected values (*i.e.* the parameters).

Author notes

Eduardo Gutiérrez-Peña, Department of Probability and Statistics, National University, Mexico. This work was partially supported by Sistema Nacional de Investigadores, Mexico. The author would like to thank Prof. Alexander von Eye for several comments and suggestions that greatly improved this paper.

References

Gutiérrez-Peña, E. and von Eye, A. (2000). A Bayesian approach to Configural Frequency Analysis. *Journal of Mathematical Sociology*, 24, 151-174.

Lienert, G.A. (1969). Die “Konfigurationsfrequenzanalyse” als Klassifikationsmethode in der klinischen Psychologie. In Irle, M. (Ed.), *Bericht über den 26. Kongress der Deutschen Gesellschaft für Psychologie in Tübingen 1968*, 244–253. Göttingen: Hogrefe.

von Eye, A. and Gutiérrez-Peña, E. (2004). Configural Frequency Analysis: the search for extreme cells. *Journal of Applied Statistics,* 31, 981-997.

Zucker, R.A. (1994). Pathways to alcohol problems and alcoholism: A developmental account of the evidence for multiple alcoholisms and contextual contributions to risk. In R.A. Zucker, J. Howard, & G. Boyd (Eds.), *The development of alcohol problems: Exploring the biopsychosocial matrix of risk*. Rockville, MD: National Institute on Alcohol Abuse and Alcoholism.