COMPOSITIONAL IDEAS IN THE BAYESIAN ANALYSIS OF CATEGORICAL DATA WITH APPLICATION TO DOSE FINDING CLINICAL TRIALS

M. Gasparini[1]**and J. Eisele**[2]
[1]Politecnico di Torino, Torino, Italy; *mauro.gasparini@polito.it*
[2] Novartis Pharma, Basel, Switzerland

1. Introduction

Compositional random vectors are fundamental tools in the Bayesian analysis of categorical data. Many of the issues that are discussed with reference to the statistical analysis of compositional data have a natural counterpart in the construction of a Bayesian statistical model for categorical data.

This note builds on the idea of cross-fertilization of the two areas recommended by Aitchison (1986) in his seminal book on compositional data. Particular emphasis is put on the problem of what parameterization to use.

2. Bayesian analysis of categorical data

Categorical data record the allocation of n statistical units $u_1, \ldots, u_n$ into several mutually exclusive categories $C_1, \ldots C_D$. If $D = 2$ the data are called binary, or dichotomous, and they often code the presence or absence of a phenomenon or an attribute. The counts

$$n_i = \#\{j : u_j \in C_i\} \tag{1}$$

can be collected in a vector of statistics $(n_1, \ldots n_D)$ such that $\sum n_i = n$.

Categorical data in this form reflect a very basic stage of data collection. If it can be assumed that $u_1, \ldots, u_n$ is a random sample from a population of interest, then the counts (1) are a sufficient statistic and the parameter of interest becomes the vector of probabilities

$$(p_1, \ldots p_D), \tag{2}$$

such that $\sum p_i = 1$, where

$$p_i = \mathrm{P}(u \in C_i).$$

Categorical data and their inference are covered by many books and textbooks. A standard reference is for example Agresti (1990), where due importance is given to the fundamental problem of studying the behavior of $(p_1, \ldots p_D)$ in the presence of explanatory covariates.

If it is the case that categories $C_1, \ldots C_D$ are ordered, then we speak of ordinal categorical data. It is then meaningful to consider the (cumulative) probability of all categories less than or equal to the i-th category:

$$P_i = \sum_{k=1}^{i} p_k = \mathrm{P}(u \in C_i \text{ or smaller}), \quad i = 1, \ldots, D-1, \quad P_D = 1. \tag{3}$$

Cumulative probabilities are a parameterization equivalent to parameterization (2) but they may be preferred when they make it easier for the researcher to express certain assumptions and hypotheses. For example, as illustrated in Agresti (1990), chapter 9, cumulative logit models and continuation-ratio logits are naturally introduced within the framework of cumulative probabilities.

Notice that probabilities $(p_1, \ldots p_d)$, where $d = D - 1$, live in the simplex $\mathcal{S}^d$, whereas cumulative probabilities $(P_1, \ldots P_d)$ belong to the following set

$$\mathcal{OS}^d = \{(P_1, \ldots P_d) : 0 \leq P_1 \leq P_2 \leq \ldots \leq P_d \leq 1\}$$

which can be called the $d-$dimensional *ordered* simplex.

From a Bayesian point of view, the vector of probabilities given by formula (2), the parameter of interest, can be treated as a random variable having a distribution, prior to sampling, which reflects the opinions and the information available to the researcher. That is precisely where the connection with compositional data comes from, since $(p_1, \dots p_D)$ is a composition, although not one arising from the physical description of several parts of a whole.

Depending on the situation, it may be preferable, especially with ordinal data, to build a Bayesian model on parameterization (3), in which case the problem becomes the construction a prior distribution on the ordered simplex.

3. Different parameterizations and Bayesian priors

When it comes to assigning distributions to vectors of probabilities (2) or (3), the same issues arise as in the choice of a likelihood for compositional data.

To start with the Dirichlet distribution on the probabilities $(p_1, \dots p_d)$, it is a Bayesian textbook example which illustrates a property called *conjugacy*: it is reproducible under random sampling, that is, the posterior distribution on $(p_1, \dots p_d)$ given a Dirichlet prior is again a Dirichlet distribution. Although that is a computationally convenient feature, the shortcomings of the Dirichlet distribution are well known to Bayesian researchers, who realize that the Dirichlet distribution can seldom represent prior information about the interdependencies between the different levels of categorical variables.

When the probabilities $(p_1, \dots p_d)$ have a Dirichlet distribution, their partial sums $(P_1, \dots P_d)$ have an ordered Dirichlet distribution which has been the basis for the construction, since a seminal paper by Ferguson (1973), of the Dirichlet stochastic process, to be used in a nonparametric setting.

As it is the case for compositional data, alternatives to the Dirichlet distribution have been looked for in the direction of transferring the parameter space to the whole of R^d by making use, for example, of the additive logistic transformation in the case of unordered categorical data. This line of reasoning has a large Bayesian literature dating back at least to Leonard (1972).

As for the Dirichlet distribution, the distribution induced on the partial sums $(P_1, \dots P_d)$ by the additive logistic distribution on $(p_1, \dots p_d)$ also has a stochastic process generalization useful in a nonparametric setting and studied by Lenk (1988) among others.

The multiplicative logistic trasformation is more useful for ordered categorical data. Since such transformation is less popular than the additive logistic, its definition is recalled here.

Definition 1. *Let $(y_1, \dots, y_d)'$ be a vector in R^d. Following Aitchison (1986), the multiplicative logistic transformation is defined as*

$$p_i = \frac{e^{y_i}}{\prod_{j=1}^{i}(1+e^{y_j})}, \quad i = 1, \dots, d \tag{4}$$
$$p_D = \frac{1}{\prod_{l=1}^{d}(1+e^{y_l})}.$$

For the reasons explained in the previous section, in the ordered categorical case it is often more convenient to work with the partial sums, hence the following definition.

Definition 2. *Under the same assumptions as Definition 1, define the* cumulative *multiplicative logistic transformation as*

$$P_i = \sum_{j=1}^{i} p_j, i = 1, \dots, d.$$

Lemma 1. *The following identities hold:*

$$P_i = 1 - \frac{1}{\prod_{j=1}^{i}(1+e^{y_j})}, \quad i = 1, \dots, d \tag{5}$$

Proof. By definition,

$$P_i = \sum_{j=1}^{i} p_j = (1 - \frac{1}{(1+e^{y_1})}) + \sum_{j=2}^{i} (\frac{1}{\prod_{k=1}^{j-1}(1+e^{y_k})} - \frac{1}{\prod_{l=1}^{j}(1+e^{y_l})})$$

from which (5) follows by a series of telescopic cancellations. ∎

Unlike the Dirichlet and the additive logistic distributions, the multiplicative logistic distribution does not have a continuous-time generalization as a stochastic process to be used in nonparametric statistics.

The inverse transformation of (5) is

$$y_i = \log(\frac{P_i - P_{i-1}}{1 - P_i}), \quad i = 2, \ldots, d,$$

where $P_0 = 0$. The inverse transformation can be rewritten in the following way:

$$y_i = \log(\frac{P_i - P_{i-1}}{1 - P_i}) = \text{logit}\left(\frac{P_i - P_{i-1}}{1 - P_{i-1}}\right) = \text{logit}\left(\text{P}(u \in C_i \middle| u \in C_i \text{ or greater})\right), \tag{6}$$

for $i = 1, \ldots, d$, as it can be easily verified. Now, the right hand side of equation (6) is precisely a quantity called *Continuation-Ratio Logit* and indicated by Agresti (1990) as one of the useful parameterizations in an ordered categorical data response model.

The point is that, by two different lines of reasoning, the literature on compositional data and the literature on categorical data have come to propose the same transformation in two separate but related fields, as it is the purpose of this brief note to show.

A cross fertilization of the two areas suggests for example the use of parameterization (6) together with the assumption of a multivariate normal distribution on the y vector to obtain a suitable Bayesian prior for ordinal categorical data. The Bayesian analysis of categorical data has taken a different direction in the past few years due to the recent developemnts of computational methods like MCMC, which, to a certain degree, allow the researcher to be less interested in exact distributions and to divert attention to piecemeal model building instead. An example is the book by Johnson and Albert (1999). But rather than embarking in the impossible task of reviewing the whole of Bayesian literature on categorical data, in the next section a categorical data problem arising in phase I clinical trials and its connection to compositional ideas is illustrated. In such problem, covariates appear as different doses in a toxicity model under the control of the experimenter, rather than random quantities.

4. Bayesian dose finding

Dose finding trials are designed to estimate, out of a set of prespecified doses, the highest dose with a probability of toxicity closest to a preassigned target and called for convenience *maximum tolerated dose.*

Bayesian methods for modelling the dose escalation scheme have been proposed in the last dozen years, notably the continual reassessment method, or CRM (O'Quigley and others 1990).

The CRM is made up of two distinct components:

(1) an allocation rule to assign sequentially the incoming patients to one of d possible doses, with the intent of assigning doses ever closer to, and eventually recommending, the MTD;
(2) a statistical procedure based on Bayes theorem which updates the information on the probabilities of toxicity in light of the results obtained for the patients already observed.

The CRM has been criticised for being too aggressive in recommending escalation, and for treating therefore too many patients above target. The problem is that the CRM is based on a rigid low-dimensional parametric model and a few observations are enough to create a tight posterior that leads the researcher to trust a highly informative inference on the unknown model.

To countermeasure this tendency of the CRM, alternatives like the modified CRM (Goodman and others 1995) and a curve-free method (Gasparini and Eisele 2000) have been proposed. In particular, the curve-free method is based on a parameterization formally similar to the partial sum parameterization (3) for ordinal categorical data. Here the parameter is vector

$$(\pi_1, \dots \pi_d), \quad 0 \leq \pi_1 \leq \pi_2 \leq \dots \leq \pi_d \leq 1 \tag{7}$$

where

$$\pi_i = \mathrm{P}(\text{toxicity if } i\text{−th dose is applied}) \tag{8}$$

and assigning a prior to it is an equivalent problem to assigning a prior to $(P_1, \dots P_d)$ in the ordered categorical Bayesian analysis.

In Gasparini and Eisele (2000) the following reparameterization is proposed:

$$\theta_i = \frac{1-\pi_i}{1-\pi_{i-1}}, \quad i = 1, \dots, d \tag{9}$$

where $\pi_0 = 0$. Now notice that

$$\mathrm{logit}\,(\theta_i) = \mathrm{logit}\left(\frac{1-\pi_i}{1-\pi_{i-1}}\right) = \log\big(\frac{1-\pi_i}{\pi_i - \pi_{i-1}}\big) = -\log\big(\frac{\pi_i - \pi_{i-1}}{1-\pi_i}\big) \tag{10}$$

which is formally the opposite of the continuation-ratio logit from formula (6) applied to the π's.

In Gasparini and Eisele (2000), $\theta_1, \dots, \theta_d$ are taken to be independent and beta distributed, but that has created several problems for simulation and, more fundamentally, for lack of flexibility in the prior. Some of the problems have been identified in Cheung (2002). It would have probably been better to follow the compositional recommendation of transforming the ordered simplex parameterization in (7) not to an alternative reparameterization on the simplex such as (9) but, rather, to a full space reparameterization such as (10) and possibly assign a d-dimensional normal distribution to the vector of $\mathrm{logit}\,(\theta_i)$.

The recommendation will be taken into consideration in future work extending the toxicity data from simple binary data (presence/absence of toxicity) to more complex ordinal data (no/mild/serious toxicity).

5. References

Agresti A., 1990, Categorical data analysis: Wiley, New York.

Aitchison J., 1986, The statistical analysis of compositional data: Chapman and Hall.

Cheung Y., 2002, On the use of nonparametric curves in phase I trials with low toxicity tolerance: Biometrics, v. 58, no. 1, p. 237-240.

Ferguson T., 1973, A bayesian analysis of some nonparametric problems: The Annals of Statistics, v. 1, p. 209–230.

Gasparini M. and Eisele J., 2000, A curve-free method for phase I clinical trials: Biometrics, v. 56, no. 2, p. 609–615, Correction in Biometrics, v. 57, no. 2, p. 659-660.

Goodman S., Zahurak M. and Piantadosi S., 1995, Some practical improvements in the continual reassessment method for phase I studies: Statistics in Medicine, v. 14, p. 1149–1161.

Johnson V. and Albert J., 1999, Ordinal data modeling: Springer.

Lenk P.J., 1988, The logistic normal distribution for Bayesian, nonparametric, predictive densities: Journal of the American Statistical Association, v. 83, p. 509–516.

Leonard T., 1972, Bayesian methods for binomial data: Biometrika, v. 59, p. 581–589.

O'Quigley J., Pepe M. and Fisher L., 1990, Continual reassessment method: a practical design for phase I clinical trials in cancer: Biometrics, v. 46, p. 33–48.