CORRESPONDENCE ANALYSIS USING DISTANCE ASSOCIATION MODELS

JAN DE LEEUW

1. Distance

Suppose F and G are two tables with positive frequencies. Define the Poisson distance between F and G as

$$\mathcal{D}(F,G)=2\sum_{i=1}^n\sum_{j=1}^m f_{ij}\log\frac{f_{ij}}{g_{ij}}-(f_{ij}-g_{ij}).$$

Observe that $\mathcal{D}(F,G)\geq 0$, with equality if and only if $F=G$. Moreover $\mathcal{D}(F,G)\neq\mathcal{D}(G,F)$. If F and G are close then Poisson distances are close to chi-square distances. More precisely

$$\mathcal{D}(F,G)=\sum_{i=1}^n\sum_{j=1}^m\frac{(f_{ij}-g_{ij})^2}{f_{ij}}+o(\|G-F\|^2),$$

as well as

$$\mathcal{D}(F,G)=\sum_{i=1}^n\sum_{j=1}^m\frac{(f_{ij}-g_{ij})^2}{g_{ij}}+o(\|G-F\|^2).$$

We can easily generalize the Poisson distance to cases in which some elements of F are zero, by using $\lim_{x\to 0}x\log x=0$. We can also generalize to incomplete tables, by summing only over the set of index pairs for which we do have information.

Now consider the case in which F is an observed table of frequencies, which may have zeroes and which may be incomplete. The elements of G are given by a parametric family of tables, depending on parameters θ. We write $\lambda_{ij}(\theta)$ for the elements of $G(\theta)$. If *theta* varies over an open subset Θ

Date: March 8, 2006.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

of $\mathbb{R}^p$, then $G(\theta)$ varies over a manifold $\mathcal{G}$ in the positive orthant of $\mathbb{R}^{n\times m}$. The problem we study in this paper is to project F on this manifold $\mathcal{G}$, i.e. to find

$$\inf_{G\in\mathcal{G}}\mathcal{D}(F,G)=\inf_{\theta\in\Theta}\mathcal{D}(F,G(\theta)),$$

and to find

$$\operatorname*{\mathbf{argmin}}_{G\in\mathcal{G}}\mathcal{D}(F,G)=\operatorname*{\mathbf{argmin}}_{\theta\in\Theta}\mathcal{D}(F,G(\theta)),$$

if the minimum is attained. Observe

$$\mathcal{D}(F,G(\theta))=2\sum_{i=1}^n\sum_{j=1}^m f_{ij}\log\frac{f_{ij}}{\lambda_{ij}(\theta)}-(f_{ij}-\lambda_{ij}(\theta)),$$

and in order to solve the minimization problem it suffices minimize the simpler function

$$\mathcal{L}_P(F,G(\theta))=-\sum_{i=1}^n\sum_{j=1}^m\{f_{ij}\log\lambda_{ij}(\theta)-\lambda_{ij}(\theta)\},$$

which is the negative Poisson log-likelihood. If the observed frequencies have a Poisson distribution, then our estimates are maximum likelihood estimates.

2. Models for Row and Column Effects

We now specialize the parametric models we fit to be of the form

$$\lambda_{ij}(\theta)=\mu\alpha_i\beta_j\gamma_{ij}(\theta).$$

Thus the model includes a main effect μ, row effects α_i, column effects β_j, and a parametrized interaction $\gamma_{ij}(\theta)$. Using this specification we can derive the following useful results. We have

$$\mathcal{D}(F,G(\mu,\alpha,\beta,\theta))=$$
$$\mu\sum_{i=1}^n\sum_{j=1}^m\alpha_i\beta_j\gamma_{ij}(\theta)-f_{\bullet\bullet}\log\mu-\sum_{i=1}^n\sum_{j=1}^m f_{ij}\log\alpha_i\beta_j\gamma_{ij}(\theta),$$

where replacing an index with a bullet means summation over the index. This is minimized over μ at

$$\hat{\mu} = \frac{f_{\bullet\bullet}}{\sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \gamma_{ij}(\theta)}.$$

This implies that minimizing $\mathcal{D}(F, G(\mu, \alpha, \beta, \theta))$ can be done by minimizing

$$\mathcal{L}_M(F, G(\alpha, \beta, \theta)) = -\sum_{i=1}^n \sum_{j=1}^m p_{ij} \log \frac{\alpha_i \beta_j \gamma_{ij}(\theta)}{\sum_{k=1}^n \sum_{\ell=1}^m \alpha_k \beta_\ell \gamma_{k\ell}(\theta)}$$

which is the negative log-likelihood of a multinomial model for the table and we use $p_{ij} = f_{ij}/f_{\bullet\bullet}$.

We can take this one step further by writing

$$\mathcal{D}(F, G(\alpha, \beta, \theta)) = \sum_{i=1}^n \alpha_i \sum_{j=1}^m \beta_j \gamma_{ij}(\theta) - \sum_{i=1}^n f_{i\bullet} \log \alpha_i - \sum_{i=1}^n \sum_{j=1}^m f_{ij} \log \beta_j \gamma_{ij}(\theta),$$

where we have absorbed the μ into the α_i. This is minimized at

$$\hat{\alpha}_i = \frac{f_{i\bullet}}{\sum_{j=1}^m \alpha_i \beta_j \gamma_{ij}(\theta)},$$

which implies that we can minimize $\mathcal{D}(F, G(\mu, \alpha, \beta, \theta))$ by minimizing

$$\mathcal{L}_{PM}(F, G(\beta, \theta)) = -\sum_{i=1}^n f_{i\bullet} \sum_{j=1}^m p_{j|i} \log \frac{\beta_j \gamma_{ij}(\theta)}{\sum_{\ell=1}^m \beta_\ell \gamma_{i\ell}(\theta)}$$

which is the negative log-likelihood of a product multinomial model for the rows of the table and we use $p_{j|i} = f_{ij}/f_{i\bullet}$. A similar derivation could be applied, of course, to finding a product multinomial model for the columns by eliminating β.

So we have seen in this section that allowing row and column effects shows that we are not just computing maximum likelihood estimates of the parameters in the Poisson case, but also in the multinomial and product multinomial cases (for suitably normalized versions of the same parametric model). Or, to put it differently, although we start by using Poisson distances between tables, we can show that our treatment covers multinomial and product multinomial distances between tables as well.

We concluded the section with a slightly more general result. We need this for the extension of our procedures to include versions of multiple correspondence analysis. Suppose we have more than one table of observed frequencies. In particular, we will be interested in the case in which we have tables F_r, with $r = 1, \cdots s$, which all have n rows, but they may have a different number of columns m_r. Consider the row and column effect model

$$\lambda_{ijr}(\theta) = \alpha_{ir}\beta_{jr}\gamma_{ijr}(\theta).$$

The Poisson distance between the observed and parametrized table can be written as

$$\mathcal{D}(F, G(\alpha, \beta, \theta)) = \sum_{i=1}^{n}\sum_{r=1}^{s}\alpha_{ir}\sum_{j=1}^{m_r}\beta_{jr}\gamma_{ijr}(\theta) - \sum_{i=1}^{n}\sum_{r=1}^{s} f_{i\bullet r}\log\alpha_{ir} - \sum_{i=1}^{n}\sum_{j=1}^{m_r}\sum_{r=1}^{s} f_{ijr}\log\beta_j\gamma_{ijr}(\theta),$$

and minimizing out α gives the product multinomial negative log-likelihood

$$\mathcal{L}_{PM}(F, G(\beta, \theta)) = -\sum_{i=1}^{n}\sum_{r=1}^{s} f_{i\bullet r}\sum_{j=1}^{m_r} p_{j|ir}\log\frac{\beta_{jr}\gamma_{ijr}(\theta)}{\sum_{\ell=1}^{m_r}\beta_{\ell r}\gamma_{i\ell r}(\theta)}.$$

3. Binary Tables

Observe that the Poisson distance can also be used to measure distance between binary tables, which only consists of zeroes and ones. We simply treat them as a special case of frequency tables.

This is especially interesting in the case in which we have a set of indicator matrices(or dummies), i.e. a number of binary $n \times m_r$ tables whose rows add up to one. The product multinomial negative log-likelihood from the previous section becomes

$$\mathcal{L}_{PM}(F, G(\beta, \theta)) = -\sum_{i=1}^{n}\sum_{r=1}^{s}\log\frac{\beta_{j_{ir}r}\gamma_{ij_{ir}r}(\theta)}{\sum_{\ell=1}^{m_r}\beta_{\ell r}\gamma_{i\ell r}(\theta)},$$

where j_{ir} is the unique index for which $f_{ijr} = 1$. Now suppose

$$\beta_{j_{ir}r}\gamma_{ij_{ir}r}(\theta) = \max_{\ell=1}^{m_r}\beta_{\ell r}\gamma_{i\ell r}(\theta), \tag{1}$$

and suppose γ is positively homogeneous, in the sense that for each $\tau > 0$ and for some $u > 0$ we have $\gamma(\tau\theta) = \tau^u\gamma(\theta)$. Then, as $\tau \to \infty$,

$$\inf_\tau \mathcal{L}_{PM}(F, G(\beta, \tau\theta)) = \lim_{\tau\to\infty} \mathcal{L}_{PM}(F, G(\beta, \tau\theta)) = -ns,$$

and the minimum is not attained. Clearly in that case the infimum of $\mathcal{D}(F, G(\alpha, \beta, \theta))$ is zero, and we have perfect fit. Thus, for homogeneous models, we can interpret our Poisson distance methods as ways to fit the system of inequalities (1). In most cases the system will not actually be solvable, and we compute an approximate solution. If the system is solvable, then our method will find the infimum by letting parameters go to infinity.

In the special case in which the indicator matrices have only two columns, the system becomes

$$\beta_{j_{ir}r}\gamma_{ij_{ir}r}(\theta) > \beta_{\ell_{ir}r}\gamma_{i\ell_{ir}r}, \tag{2}$$

where j_{ir} is the unique index for which $f_{ijr} = 1$ and ℓ_{ir} is the unique index for which $f_{ijr} = 0$.

4. Distance Association Models

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu