MAJORIZATION ALGORITHMS FOR DISTANCE ASSOCIATION MODELS

JAN DE LEEUW

ABSTRACT. We construct majorization algorithms and programs for a class of multinomial maximum likelihood problems in contingency tables in which interactions are modeled using Euclidean distances between points representing the row and column objects.

1. INTRODUCTION

Suppose $F = \{f_{ij}\}$ is an $n \times m$ table with frequencies. The saturated log-linear model for the table is

$$\mathbf{E}(\underline{f}_{ij}) = \mu\alpha_i\beta_j \exp(\phi_{ij}), \tag{1a}$$

which decomposes the expected values into main effects and a first-order interaction[1]. Often saturated models are not what we want, because having as many $(n-1)(m-1)$ parameters to describe first-order interaction may not be desirable from a data reduction point of view.

For this reason, a large number of models have been proposed to reduce the number of parameters describing the interaction. Some of the most interesting ones are geometrical, in the sense that they

Date: July 27, 2006.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis, Ordination, Log-bilinear Models, Log-quadratic Models, Social Mobility, Confusion Matrices, Majorization.

[1]We follow the convention of underlining random variables [Hemelrijk, 1966].

represent the n row objects and the m column objects as points in low-dimensional space. The interaction term ϕ_{ij} is then modeled as a function of the form $\phi(x_i, y_j)$, where ϕ is often distance or at least distance-like. Thus we have a model for the expected values of the form

$$\mathbf{E}(\underline{f}_{ij}) = \mu\alpha_i\beta_j \exp \phi(x_i, y_j). \tag{1b}$$

We call such models *distance interaction models*, following De Rooij and Heiser [2005]. They are especially interesting if the interactions can be interpreted as some form of similarity for which a spatial representation is natural. Clearly distance interaction models are related to multidimensional scaling methods, which also relate measures of similarity to distance between points in low-dimensional space.

One particular type of table for which distance interaction models are often interesting is the square table, in which rows and columns refer to the same objects. In this case we can make both a spatial representation and capture the underlying symmetry of the table by a model of the more specific form

$$\mathbf{E}(\underline{f}_{ij}) = \mu\alpha_i\beta_j \exp \phi(x_i, x_j). \tag{1c}$$

Tables with different row and the column objects are often called *two-mode data*, tables in which both row and column objects are the same are *one-mode data*. Two-mode data are a mapping of $I \otimes J$ into the integers, while one-mode data are defined on $I \otimes I$.

The literature is on distance interaction models is difficult to summarize because the models have been discovered and discussed in many different applied contexts. We give a brief overview, which is mainly intended to introduce the type of models we have in mind, and to provide the reader with the major references. A similar overview has been published recently by De Rooij and Heiser [2005].

1.1. **Probability Model.** So far we have given models for the expected values only, which is clearly not enough to build a full likelihood. For frequency data there are at least three common alternatives [Bishop et al., 1975; Haberman, 1974].

1.1.1. *Poisson.* Suppose the $\underline{f}_{ij}$ are independent Poisson, with parameters λ_{ij}. Thus the Poisson log-likelihood, except for irrelevant constants, is

$$\mathcal{L}_{\mathcal{P}}(\Lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\log(\lambda_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_{ij}, \tag{2a}$$

where bullets are used to indicate summation over an index. Substituting (1a) gives

$$\mathcal{L}_{\mathcal{P}}(\mu,\alpha,\beta,\Phi) = f_{\bullet\bullet}\log\mu + \sum_{i=1}^{n} f_{i\bullet}\log\alpha_i + \sum_{j=1}^{m} f_{\bullet j}\log\beta_j + \\ + \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\phi_{ij} - \mu\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j\exp(\phi_{ij}), \tag{2b}$$

1.1.2. *Product Multinomial.* Suppose each row of $\underline{F}$ is an independent multinomial, with parameters $f_{i\bullet}$ and $\pi_{j|i}$. Thus the log-likelihood is

$$\mathcal{L}_{\mathcal{PM}}(\Pi) = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\log\pi_{j|i}. \tag{3a}$$

Our model is of the form

$$\pi_{j|i} = \frac{\beta_j\exp(\phi_{ij})}{\sum_{\ell=1}^{m}\beta_\ell\exp(\phi_{ij\ell})}. \tag{3b}$$

and

$$\mathcal{L}_{\mathcal{PM}}(\beta,\Phi) = \sum_{j=1}^{m} f_{\bullet j}\log\beta_j + \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\phi_{ij} - \sum_{i=1}^{n} f_{i\bullet}\log\sum_{j=1}^{m}\beta_j\exp(\phi_{ij}). \tag{3c}$$

1.1.3. *Multinomial.* Suppose the whole table $\underline{F}$ is multinomial, with parameters $f_{\bullet\bullet}$ and π_{ij}. Thus the log-likelihood is

$$\mathcal{L}_{\mathcal{M}}(\Pi) = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\log\pi_{ij}. \tag{4a}$$

Our model is of the form

$$\pi_{ij} = \frac{\alpha_i\beta_j\exp(\phi_{ij})}{\sum_{k=1}^{n}\sum_{\ell=1}^{m}\alpha_k\beta_\ell\exp(\phi_{k\ell})}. \tag{4b}$$

and

$$\begin{aligned}\mathcal{L}_{\mathcal{M}}(\alpha,\beta,\Phi) = &\sum_{i=1}^{n} f_{i\bullet}\log\alpha_i + \sum_{j=1}^{m} f_{\bullet j}\log\beta_j + \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\phi_{ij} - f_{\bullet\bullet}\log\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j\exp(\phi_{ij}).\end{aligned} \tag{4c}$$

Theorem 1.1.

$$\max_{\mu}\mathcal{L}_{\mathcal{P}}(\mu,\alpha,\beta,\Phi) = \mathcal{L}_{\mathcal{M}}(\alpha,\beta,\Phi) + f_{\bullet\bullet}\log f_{\bullet\bullet} - f_{\bullet\bullet},$$

where the maximum is attained at

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

Proof. The Poisson log-likelihood is concave and differentiable so it is enough to solve the stationary equations. □

Theorem 1.2.

$$\max_{\mu,\alpha}\mathcal{L}_{\mathcal{P}}(\mu,\alpha,\beta,\Phi) = \mathcal{L}_{\mathcal{PM}}(\beta,\Phi) + \sum_{i=1}^{n} f_{i\bullet}\log f_{i\bullet} - f_{\bullet\bullet},$$

where the maximum is attained at

$$\hat{\mu}\hat{\alpha}_i = \frac{f_{i\bullet}}{\sum_{j=1}^{m}\beta_j\exp(\phi_{ij})}.$$

Proof. The Poisson log-likelihood is concave and differentiable so it is enough to solve the stationary equations. □

Corollary 1.3.

$$\max_{\alpha} \mathcal{L}_{\mathcal{M}}(\alpha, \beta, \Phi) = \mathcal{L}_{\mathcal{PM}}(\beta, \Phi) + \sum_{i=1}^{n} f_{i\bullet} \log f_{i\bullet} - f_{\bullet\bullet} \log f_{\bullet\bullet}.$$

Proof. We use the two previous theorems to arrive at the identities

$$\max_{\alpha} \max_{\mu} \mathcal{L}_{\mathcal{P}}(\mu, \alpha, \beta, \Phi) = \max_{\alpha} \mathcal{L}_{\mathcal{M}}(\alpha, \beta, \Phi) + f_{\bullet\bullet} \log f_{\bullet\bullet} - f_{\bullet\bullet} =$$

$$= \mathcal{L}_{\mathcal{PM}}(\beta, \Phi) + \sum_{i=1}^{n} f_{i\bullet} \log f_{i\bullet} - f_{\bullet\bullet}.$$

□

The theorems show that both the Poisson likelihood and the multinomial likelihood can both be maximized by maximizing the product multinomial likelihood. Moreover the additional parameters in the Poisson and multinomial models can be easily computed from the product multinomial parameter estimates. Thus it suffices to work with the product multinomial likelihood.

2. Application Areas

2.1. **One Mode Data.**

2.1.1. *Stimulus Recognition.* Consider the situation in which subjects are asked to identify stimuli. Some obvious examples are recognition of letters on an optimetricians card, or recognition of morse code signals in an army test. Replication over subjects leads to a square matrix, often called a *confusion matrix*, in which cell (i, j) indicates how many subjects responded j when presented with i.

Shepard [1957]Luce [1959]Luce [1963]

2.1.2. *Social Mobility.* Ever since Galton and Pearson, sociologists and statisticians have studied mobility tables, in which we count how many children of parents in profession or status group i wind up in profession or status group j.

2.1.3. *Interactions.* The basic data in a stimulus recognition experiment also occur in other situations. Consider, for example, the number of tourists from country i who vacation in country j. Or the number of container ships traveling from country i to country j. In both these examples there may be a problem with defining the diagonal of the matrix, and in that sense the examples are different from stimulus recognition.

Another obvious example in this class is sociometric data, in which we count the number of interactions between individual i and individual j in a social group. Again defining the diagonal is problematic. Or, more generally, we count the flows in any network from node i to node j.

2.2. **Two Mode Data.**

2.2.1. *Ordination.*

2.2.2. *Ideal Point Discriminant Analysis.* In Takane et al. [1987] a technique for discriminant analysis was proposed, and in Takane [1987] the technique was extended to the analysis of contingency tables. The basic model for the probability of choosing reponse j under condition i is

$$\pi_{j|i}(X,Y,\beta)=\frac{\beta_j\exp(-d_{ij}^2(X,Y))}{\sum_{\ell=1}^m\beta_\ell\exp(-d_{i\ell}^2(X,Y))}.$$

In discriminant analysis applications X is usually assumed to be in the span of a number of regressors, and we assume $X=ZB$. For various reasons, it is also often assumed in IPDA that $Y=D^{-1}F'X$, i.e. the column points are in the centroids of the row points.

2.2.3. *RC Association Models.*

$$\mathbf{E}(f_{ij}) = \mu + \alpha_i + \beta_j + \sum_{s=1}^{p} x_{is}y_{js}$$

3. Problem

3.1. **Loss Function.** The problem we study in this paper is minimization of the *deviance*, which is minus two times the log-likelihood,

$$\mathcal{D}(X, Y, \beta) = -2\sum_{i=1}^{n} f_{i\bullet}\sum_{j=1}^{m} p_{j|i}\log\pi_{j|i}(X, Y, \beta),$$

with

$$\pi_{j|i}(X, Y, \beta) = \frac{\beta_j \exp(\phi(x_i, y_j))}{\sum_{\ell=1}^{m}\beta_\ell \exp(\phi(x_i, y_\ell))}.$$

Also $f_{i\bullet}$ is the vector of row sums of F, and $p_{j|i} = f_{ij}/f_{i\bullet}$.

Since

$$\mathcal{D}(X, Y, \beta) \geq -2\sum_{i=1}^{n} f_{i\bullet}\sum_{j=1}^{m} p_{j|i}\log p_{j|i},$$

we actually compute the non-negative loss function

$$\mathcal{G}(X, Y, \beta) = 2\sum_{i=1}^{n} f_{i\bullet}\sum_{j=1}^{m} p_{j|i}\log\frac{p_{j|i}}{\pi_{j|i}(X, Y, \beta)},$$

which is asymptotically chi-squared in the case that the rows of F are independent multinomials.

3.2. **Combination Rules.** The basic algorithm we develop applies to arbitrary *combination rules* ϕ, but we shall be especially interested in

$$\begin{aligned}
\phi_S(x_i, y_j) &= -d_{ij}^2(X, Y) = -(x_i - y_j)'(x_i - y_j),\\
\phi_D(x_i, y_j) &= -d_{ij}(X, Y) = -\sqrt{(x_i - y_j)'(x_i - y_j)},\\
\phi_P(x_i, y_j) &= c_{ij}(X, Y) = x_i'y_j.
\end{aligned}$$

These are, repectively, the *negative squared distance rule*, the *negative distance rule*, and the *inner product rule*.

If the data have a single mode, then the symmetric versions of the rules are used, which we obtain by simply setting $X = Y$.

$$\begin{aligned}\phi_{\overline{S}}(x_i, x_j) &= -d_{ij}^2(X) = -(x_i - x_j)'(x_i - x_j),\\ \phi_{\overline{D}}(x_i, x_j) &= -d_{ij}(X) = -\sqrt{(x_i - x_j)'(x_i - x_j)},\\ \phi_{\overline{P}}(x_i, x_j) &= c_{ij}(X) = x_i'x_j.\end{aligned}$$

3.3. **Distances or Inner Products.** Consider

$$\pi_{j|i}(X, Y, \beta) = \frac{\beta_j \exp(\phi_{ij})}{\sum_{\ell=1}^m \beta_\ell \exp(\phi_{i\ell})}$$

with $\phi_{ij} = -d_{ij}^2(X, Y)$. Then

$$\pi_{j|i} = \frac{\tilde\beta_j \exp(x_i'y_j)}{\sum_{\ell=1}^m \tilde\beta_\ell \exp(x_i'y_\ell)},$$

with

$$\tilde\beta_j = \beta_j \exp(-\frac{1}{2}y_j'y_j).$$

Thus fitting a negative squared distance model with bias parameters is equivalent to fitting an inner product model with bias parameters. Fitting a negative squared distance model *without* bias parameters can be done by fitting an inner product model with bias parameters, but the two models are equivalent only if we impose nonlinear constraints on the bias parameters.

4. Algorithm

The algorithm we propose alternates minimization over β for fixed X, Y and minimization over X, Y for fixed β. For single mode data we require $X = Y$. In both steps we use majorization [].

4.1. **Optimum β.** We have

$$-\log \pi_{j|i}(X, Y, \beta) = -\log \beta_j - \phi(x_i, y_j) + \log \sum_{\ell=1}^m \beta_\ell \exp(\phi(x_i, y_\ell)).$$

By concavity of the logarithm, for all pairs β and $\tilde{\beta}$,

$$\log \sum_{\ell=1}^{m} \beta_\ell \exp(\phi(x_i, y_\ell)) \leq \log \sum_{\ell=1}^{m} \tilde{\beta}_\ell \exp(\phi(x_i, y_\ell)) + \sum_{j=1}^{n} \frac{\pi_{j|i}(X, Y, \tilde{\beta})}{\tilde{\beta}_j}(\beta_j - \tilde{\beta}_j),$$

and thus

$$\kappa_j(\beta) \leq -\log \beta_j - x_j + \log \sum_{\ell=1}^{m} \tilde{\beta}_\ell \exp(\phi(x_i, y_\ell)) + \sum_{j=1}^{n} \frac{\pi_{j|i}(X, Y, \tilde{\beta})}{\tilde{\beta}_j}(\beta_j - \tilde{\beta}_j.$$

4.2. **Optimum X, Y.** Define

$$\pi_j(x) = \frac{\beta_j \exp(x_j)}{\sum_{\ell=1}^{n} \beta_\ell \exp(x_\ell)},$$

and

$$\kappa_j(x) = -\log \pi_j(x) = -\log \beta_j - x_j + \log \sum_{\ell=1}^{n} \beta_\ell \exp(x_\ell).$$

For the partials of κ_j we find

$$\frac{\partial \kappa_j}{\partial x_\ell} = -(\delta^{j\ell} - \pi_\ell(x)),$$

and for the second partials

$$\frac{\partial^2 \kappa_j}{\partial x_\ell \partial x_\nu} = \frac{\partial \pi_\ell}{\partial x_\nu} = \pi_\ell(x)(\delta^{\ell\nu} - \pi_\nu(x)),$$

which is the same for all j. This shows, by the way, that κ_j is a convex function of x.

Lemma 4.1. *Suppose p is any vector in the simplex $\mathbb{S}^{n-1}$, i.e. p has n non-negative elements that add up to one, Let $V(p) = P - pp'$, where P is the diagonal matrix with p on the diagonal, and let $\lambda_+(p)$ be the largest eigenvalue (or the spectral radius) of $V(p)$. Then*

$$\max_{p \in \mathbb{S}^{n-1}} \lambda_+(p) = \frac{1}{2}$$

Proof. The spectral radius of a matrix is less than or equal to any matrix norm. In particular

$$\lambda_+(p) \leq \max_{i=1}^{n} \sum_{j=1}^{n} |v_{ij}(p)|.$$

Now

$$\sum_{j=1}^{n} |v_{ij}(p)| = p_i(1-p_i) + \sum_{j\neq i}^{n} p_i p_j = 2p_i(1-p_i).$$

Since $p_i(1-p_i) \leq \frac{1}{4}$ for $0 \leq p_i \leq 1$ we see that

$$\lambda_+(p) \leq \frac{1}{2}.$$

If p has two nonzero elements, both equal to $\frac{1}{2}$, then we have $\lambda_+(p) = \frac{1}{2}$, so the maximum is attained. □

Theorem 4.2. *For all x, y*

$$\kappa_j(x) \leq \kappa_j(y) + \frac{1}{4}\sum_{\ell=1}^{n}[x_\ell - \tau_{j\ell}(y)]^2 - \frac{1}{4}\sum_{\ell=1}^{n}[y_\ell - \tau_{j\ell}(y)]^2$$

where

$$\tau_{j\ell}(y) = y_\ell + 2(\delta^{j\ell} - \pi_\ell(y)).$$

Proof. By the mean value theorem and Lemma 4.1

$$\kappa_j(x) \leq \kappa_j(y) - \sum_{\ell=1}^{n}(\delta^{j\ell} - \pi_\ell(y))(x_\ell - y_\ell) + \frac{1}{4}\sum_{\ell=1}^{n}(x_\ell - y_\ell)^2.$$

Completing the square gives the required result. □

Now think of

$$\kappa_j(\beta) = -\log\beta_j - x_j + \log\sum_{\ell=1}^{m}\beta_\ell \exp(x_\ell)$$

as a function of β for fixed x. By concavity of the logaritm

$$\log\sum_{\ell=1}^{m}\beta_\ell\exp(x_\ell) \leq \log\sum_{\ell=1}^{m}\tilde{\beta}_\ell\exp(x_\ell) + \sum_{j=1}^{n}\frac{\pi_j(\tilde{\beta})}{\tilde{\beta}_j}(\beta_j - \tilde{\beta}_j,$$

and thus

$$\kappa_j(\beta) \leq -\log \beta_j - x_j + \log \sum_{\ell=1}^{m} \tilde{\beta}_\ell \exp(x_\ell) + \sum_{j=1}^{n} \frac{\pi_j(\tilde{\beta})}{\tilde{\beta}_j}(\beta_j - \tilde{\beta}_j.$$

We first rewrite our optimization problem as minimization of the deviance, which is minus two times the log-likelihood. Now the problem is to minimize

$$\mathcal{D}(X, Y, \theta) = -2 \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} \log \pi_{j|i}(X, Y, \beta),$$

with

$$\pi_{j|i}(X, Y, \beta) = \frac{\exp(\psi_{ij}(X, Y, \theta))}{\sum_{\ell=1}^{m} \exp(\psi_{i\ell}(X, Y, \theta))}.$$

We use the majorization result from Appendix **??**. After some computation we find

(5)

$$\mathcal{D}(X, Y, \theta) \leq \mathcal{D}(\tilde{X}, \tilde{Y}, \tilde{\theta}) + \frac{1}{2} \sum_{i=1}^{n} f_{i\bullet} \sum_{\ell=1}^{m} [\psi_{i\ell}(X, Y, \theta) - \tau_{i\ell}(\tilde{X}, \tilde{Y}, \tilde{\theta})]^2 +$$

$$- \frac{1}{2} \sum_{i=1}^{n} f_{i\bullet} \sum_{\ell=1}^{m} [\psi_{i\ell}(\tilde{X}, \tilde{Y}, \tilde{\theta}) - \tau_{i\ell}(\tilde{X}, \tilde{Y}, \tilde{\theta})]^2, \tag{6}$$

where

$$\tau_{i\ell}(\tilde{X}, \tilde{Y}, \tilde{\theta}) = \psi_{i\ell}(\tilde{X}, \tilde{Y}, \tilde{\theta}) + 2(p_{\ell|i} - \pi_{\ell|i}(\tilde{X}, \tilde{Y}, \tilde{\theta})).$$

Here $f_{i\bullet}$ are the row sums of F, and $p_{j|i} = f_{ij}/f_{i\bullet}$.

For the majorization algorithms we have to minimize (or at least decrease) the single term on the right in (5) depending on X, Y, and θ. This means minimizing the least squares loss function

$$\sigma(X, Y, \theta) = \sum_{i=1}^{n} f_{i\bullet} \sum_{\ell=1}^{m} [\psi_{i\ell}(X, Y, \theta) - \tau_{i\ell}(\tilde{X}, \tilde{Y}, \tilde{\theta})]^2. \tag{7}$$

The precise algorithm used to minimize (7) depends, of course, on the combination rule. For our three differrent rules we have,

respectively,

$$\sigma_S(X,Y,\theta) = \sum_{i=1}^n f_{i\bullet} \sum_{\ell=1}^m [-\frac{1}{2}d_{i\ell}^2(X,Y) - \tau_{i\ell}(\tilde{X},\tilde{Y},\tilde{\theta})]^2, \tag{8a}$$

$$\sigma_D(X,Y,\theta) = \sum_{i=1}^n f_{i\bullet} \sum_{\ell=1}^m [[-d_{i\ell}(X,Y) - \tau_{i\ell}(\tilde{X},\tilde{Y},\tilde{\theta})]^2, \tag{8b}$$

$$\sigma_P(X,Y,\theta) = \sum_{i=1}^n f_{i\bullet} \sum_{\ell=1}^m [x_i'y_\ell - \tau_{i\ell}(\tilde{X},\tilde{Y},\tilde{\theta})]^2. \tag{8c}$$

These least squares appriximation subproblems can all be tackled by standard multidimensional scaling, multidimensional unfolding, singular value or eigen value algorithms. Because generally σ_P is somewhat easier to deal with than σ_S, we generally use the equivalence pointed out in Appendix 3.3 to reduced the negative squared distance rule with bias to the inner product rule with bias.

Appendix A. Alternative Majorization

To show the versatily of the majorization method, we also give a direct majorization of the negative Poisson likelihood (**??**) for the negative distance and the negative squared distance rules. The part that depends on ϕ_{ij} is

$$\mathcal{F}(\Phi) = -\sum_{i=1}^n \sum_{j=1}^m f_{ij}\phi_{ij} + \mu \sum_{i=1}^n \sum_{j=1}^m \alpha_i\beta_j \exp(\phi_{ij})$$

Now if both $\phi_{ij} \leq 0$ and $\tilde{\phi}_{ij} \leq 0$ we have

$$\exp(\phi_{ij}) = \exp(\tilde{\phi}_{ij}) + \exp(\tilde{\phi}_{ij})(\phi_{ij} - \tilde{\phi}_{ij}) + \frac{1}{2}\exp(\xi_{ij})(\phi_{ij} - \tilde{\phi}_{ij})^2$$

for some ξ_{ij} between ϕ_{ij} and $\tilde{\phi}_{ij}$. Thus $\xi_{ij} \leq 0$, and

$$\exp(\phi_{ij}) \leq \exp(\tilde{\phi}_{ij}) + \exp(\tilde{\phi}_{ij})(\phi_{ij} - \tilde{\phi}_{ij}) + \frac{1}{2}(\phi_{ij} - \tilde{\phi}_{ij})^2$$

As a consequence

$$\mathcal{F}(\Phi) \leq \mathcal{F}(\tilde{\Phi}) - \sum_{i=1}^{n}\sum_{j=1}^{m}(f_{ij} - \mu\alpha_i\beta_j\exp(\tilde{\phi}_{ij}))(\phi_{ij} - \tilde{\phi}_{ij}) + \\ + \frac{1}{2}\mu\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j(\phi_{ij} - \tilde{\phi}_{ij})^2$$

and it suffices to minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j(\phi_{ij} - z_{ij})^2,$$

where

$$z_{ij} = \tilde{\phi}_{ij} + \frac{f_{ij} - \mu\alpha_i\beta_j\exp(\tilde{\phi}_{ij})}{\mu\alpha_i\beta_j}.$$

This is (again) a weighted multidimensional scaling problem, which can be alternated with updating α and β by iterative proportional fitting. Observe the derivation depends on the fact that negative distances are squared distances are non-positive, and thus it does not work for the inner product rule.

References

Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, 1975.

M De Rooij and W.J. Heiser. Graphical Representations and Odds Ratios in a Distance Association Model for the Analysis of Cross-Classified Data. *Psychometrika*, 70:99–122, 2005.

S.J. Haberman. *The Analysis of Frequency Data*. University of Chicago Press, 1974.

J. Hemelrijk. Underlining Random Variables. *Statistica Neerlandica*, 20:1–7, 1966.

R.D. Luce. Detection and Recognition. In R.D. Luce, R.R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume 1, chapter 3, pages 103–189. Wiley, 1963.

R.D. Luce. *Individual Choice Behavior*. Wiley, 1959.

R.N. Shepard. Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space. *Psychometrika*, 22:325–345, 1957.

Y. Takane. Analysis of Contingency Tables by Ideal Point Discriminant Analysis. *Psychometrika*, 52:493–513, 1987.

Y. Takane, H. Bozdogan, and T. Shibayama. Ideal Point Discriminant Analysis. *Psychometrika*, 52:371–392, 1987.

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