FITTING UNREPLICATED CHOICE EXPERIMENTS

JAN DE LEEUW AND ELIZABETH BRUCH

1. Problem

In this paper we analyze choices of each of n individuals among a set of m alternatives. In fact, we use a slightly more general setup, in which only a subset of the choice set is available to a particular individual. Thus individual i chooses from m_i alternatives, with $1 \leq m_i \leq m$. We write $J_i \subseteq \{1, \cdots, m\}$ for the choice set of individual i.

We can code our data as a sequence $(y_1, \cdots, y_n)$ of binary vectors, where y_i has m_i elements and $y_{ij} = 1$ if individual i chooses $j \in J_i$. Thus the elements of each of the y_i add up to one.

In addition to the choices, for each individual i the alternatives $j \in J_i$ are characterized by p regressors. Thus for each individual there is a $m_i \times p$ matrix X_i with predictor values. The aim of our analysis is to study the relationship between the characteristics of the alternatives X_i and the choices y_i.

Observe that X_i has the characteristics of the choices as perceived by or as relevant for individual i. Some of the columns of the X_i may be characteristics of the individuals themselves, in which case they will be the same for each choice the

Date: June 12, 2005.

2000 *Mathematics Subject Classification.* 62J12,62J20,62P25,91B16.

Key words and phrases. Generalized Linear Models, Diagnostics, Applications to Social Science,Utility Theory.

individual makes. Some columns may be characteristics of the choices, in which case they will be the same for all individuals, i.e. they will be the same in each X_i.

2. Model

We suppose the y_i are realizations of binary random vectors[1] $\underline{y}_i$. Of course also $\sum_{j\in J_i}\underline{y}_{ij}\equiv 1$, so there is dependence between the $\underline{y}_{ij}$ *within* each individual i. We assume there is independence *between* different individuals.

We assume the *multinomial logit model*, which says there is a p element vector of regression coefficients β such that $\mathbf{E}(\underline{y}_i)=\pi_i(\beta)$, where

$$\pi_{ij}(\beta)=\frac{\exp(\beta'x_{ij})}{\sum_{\ell\in J_i}\exp(\beta'x_{i\ell})}. \tag{1}$$

Observe that x_{ij} is a row of X_i, i.e. a vector with p elements. Define the p-vectors

$$u_i\triangleq X_i'y_i,$$

$$\mu_i(\beta)\triangleq X_i'\pi_i(\beta)$$

Thus u_i can be thought of as a realization of the random p-vector $\underline{u}_i\triangleq X_i'\underline{y}_i$, with expectation $\mu_i(\beta)$.

For later reference we compute the derivatives of the π_{ij} with respect to the regression coefficients β. For the first derivatives

$$\mathcal{D}\log\pi_{ij}(\beta)=x_{ij}-\mu_i(\beta). \tag{2a}$$

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

To give a convenient expression for the second derivatives we define the positive semidefinite matrices

$$\Sigma_i(\beta) \stackrel{\Delta}{=} \sum_{j \in J_i} \pi_{ij}(\beta)\{x_{ij} - \mu_i(\beta)\}\{x_{ij} - \mu_i(\beta)\}'$$

and we find

$$\mathcal{D}^2 \log \pi_{ij}(\beta) = -\Sigma_i(\beta). \tag{2b}$$

Observe that this is the same for all j.

There is an alternative matrix expression for $\Sigma_i(\beta)$ which is sometimes useful. If

$$\Pi_i(\beta) \stackrel{\Delta}{=} \begin{bmatrix} \pi_{i1}(\beta) & 0 & \cdots & 0 \\ 0 & \pi_{i2}(\beta) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_{im(i)}(\beta) \end{bmatrix}$$

and

$$V_i(\beta) \stackrel{\Delta}{=} \Pi_i(\beta) - \pi_i(\beta)\pi_i(\beta)'$$

then $\Sigma_i(\beta) = X_i' V_i(\beta) X_i$.

3. Likelihood

The log-likelihood is

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \sum_{j \in J_i} y_{ij} \log \pi_{ij}(\beta). \tag{3a}$$

If u is the sum of the u_i we can write this as

$$\mathcal{L}(\beta) = u'\beta - \sum_{i=1}^{n} \log \sum_{j \in J_i} \exp(\beta' x_{ij}). \tag{3b}$$

This shows that u is a sufficient statistic for β, and that both the maximum likelihood estimate of β and the maximized likelihood are functions of u only.

If $\mu(\beta)$ is the sum of the $\mu_i(\beta)$ then we find for the derivatives

$$\mathcal{D}\mathcal{L}(\beta) = u - \mu(\beta), \tag{4a}$$

and if $\Sigma(\beta)$ is the sum of the $\Sigma_i(\beta)$ then the second derivatives are

$$\mathcal{D}^2\mathcal{L}(\beta) = -\Sigma(\beta). \tag{4b}$$

It is clear that, in fact, $\mathcal{L}(\beta)$ is infinitely many times continuously differentiable on $\mathbb{R}^p$.

In order to study existence and uniqueness of the maximum likelihood estimate let $\mathcal{L}_\infty \triangleq \sup_\beta \mathcal{L}(\beta)$ and $\mathbf{B} \triangleq \{\beta \mid \mathcal{L}(\beta) = \mathcal{L}_\infty\}$. Because $\mathcal{L}(\beta) < 0$ for all β we have $\mathcal{L}_\infty \leq 0$, i.e. the log likelihood is bounded above. Moreover $\lim_{\nu\to\infty} \mathcal{L}(\beta_\nu) = 0$ if and only if $\lim_{\nu\to\infty} \pi_{ij(i)}(\beta_\nu) = 1$ for all i. We can be more specific about the set of solutions.

Theorem 3.1. *If* $\mathbf{B}$ *is non-empty, then it is closed and convex.*

Proof. Because of Equation (4b) the log-likelihood is concave. □

Theorem 3.2. *If* $\mathbf{B}$ *is non-empty and*

$$\mathbf{rank}\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = p$$

then $\mathbf{B}$ *is a singleton.*

Proof. If the rank condition is true then $\Sigma(\beta)$ is positive definite for all β, and thus the log-likelihood is stricty concave. □

Theorem 3.3. *If there is a $\gamma > 0$ such that for all β and all z*

$$\frac{z'\Sigma(\beta)z}{z'z} \geq \gamma$$

then **B** *is a singleton.*

Proof. The condition, which says that the smallest eigenvalue of $\Sigma(\beta)$ is bounded away from zero, implies that the log-likelihood is uniformly concave. □

More generally, we know that the maximum likelihood estimate exists if and only if the likelihood function has no directions of recession [Bertsekas, 2003, Section 2.3].

4. The Likelihood equations

The log-likelihood, interpreted as a function both both u and β, is of the form $\mathcal{L}(\beta, u) = u'\beta - F(\beta)$,where

$$F(\beta) = \sum_{i=1}^{n} \log \sum_{j \in J_i} \exp(\beta' x_{ij}).$$

Here the function F is strongly convex, its first derivative is $\mu(\beta)$ and its second derivative is $\Sigma(\beta)$. F is, moreover, infinitely many times continuously differentiable.

The likelihood equations are of the form $\mu(\beta) = u$. Since μ, by assumption, has a uniformly positive definite derivative we can use Hadamard's global inverse function theorem [Krantz and Parks, 2002, Section 6.2]. This means that the inverse $\hat{\beta} \stackrel{\Delta}{=} \mu^{-1}$ exists, and is infinitely many times differentiable on all of $\mathbb{R}^p$.

For later reference we need the derivatives of the maximum likelihood estimates $\hat{\beta}(u)$, interpreted as a function of u. From the inverse function theorem

$$\mathcal{D}\hat{\beta}(u) = \Sigma^{-1}(\hat{\beta}(u)).$$

We also need derivatives of the maximized log-likelihood. If

$$F^\circ(u) \stackrel{\Delta}{=} \sup_\beta \mathcal{L}(\beta, u) = \mathcal{L}(\hat\beta(u), u) = F^\circ(u),$$

then

$$\mathcal{D}F^\circ(u) = \hat\beta(u),$$

and thus

$$\mathcal{D}^2F^\circ(u) = \Sigma^{-1}(\hat\beta(u)).$$

5. Asymptotics

5.1. **Sufficient Statistics.** In order to do large sample statistics, we have to imbed our experiment in an increasing sequence of experiments. Usually, in discrete generalized linear models, we assume replicated choices for each i, thus defining a multinomial distribution on the choices. We then let the number of choices for each i goes to infinity, and keep i fixed at n. We trivially get asymptotic normality from the CLT for iid bounded (binary) vectors.

In our situation, however, there is only a single choice per individual, and thus the usual setup does not make sense. We have to let n go to infinity to get into asymptopia. So let us assume there is an infinite sequence of individuals i, each with a choice set J_i with m_i elements, and with a corresponding $m_i \times p$ matrix of predictors X_i.

Consider the independent random vectors $\underline{u}_i$, which have expected value $\mu_i(\beta)$ and variance-covariance matrix $\Sigma_i(\beta)$. Define the averages

$$\underline{u}_{(n)} \stackrel{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n}\underline{u}_i,$$

$$\mu_{(n)}(\beta) \stackrel{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n}\mu_i(\beta),$$

$$\Sigma_{(n)}(\beta) \stackrel{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n}\Sigma_i(\beta).$$

Thus, to make the notation explicit once again, the u_i are the observed values in our experiment, and the $\underline{u}_i$ are the random variables these observed values are supposed to be realizations of. For both observed values and random variables we compute the running averages. We have made an effort, throughout the paper, to keep the notation both explicit and consistent. In many statistical publications the distinction between random variables and their realizations is swept under the notational rug.

Clearly $\underline{u}_{(n)}$ (exactly) has expectation $\mu_{(n)}(\beta)$ and dispersion matrix $n^{-1}\Sigma_{(n)}(\beta)$. By using a suitable version of the central limit theorem [Hahn and Klass, 1981] we conclude that

$$n^{1/2}\Sigma_{(n)}^{-1/2}(\beta)(\underline{u}_{(n)} - \mu_{(n)}(\beta)) \stackrel{\mathcal{L}}{\Rightarrow} \mathcal{N}(0, I).$$

5.2. **Maximum Likelihood Estimates.** The sequence of maximum likelihood estimates is computed as $\hat{\underline{\beta}}_{(n)} = \mu_{(n)}^{-1}(\underline{u}_{(n)})$. By the Delta Method [Vaart, 1998, Chapter 3]

$$n^{1/2}\Sigma_{(n)}^{+1/2}(\beta)(\hat{\underline{\beta}}_{(n)} - \beta) \stackrel{\mathcal{L}}{\Rightarrow} \mathcal{N}(0, I)$$

which is sometimes (very imprecisely) paraphrased as $\hat{\underline{\beta}}_{(n)}$ is asymptotically multivariate normal with mean β and dispersion $n^{-1}\Sigma_{(n)}^{-1}$. In fact we even have the

"Studentized" version [Vu et al., 1996]

$$n^{1/2}\Sigma_{(n)}^{+1/2}(\hat{\underline{\beta}}_{(n)})(\hat{\underline{\beta}}_{(n)}-\beta)\overset{\mathcal{L}}{\Rightarrow}\mathcal{N}(0,I).$$

6. Diagnostics

6.1. **Standardized Residuals.** There are various ways to define residuals in unreplicated choice experiments. The most straightforward one is to use the studentized residuals

$$r_{ij}\overset{\Delta}{=}\frac{y_{ij}-\pi_{ij}(\hat{\beta})}{\sqrt{\pi_{ij}(\hat{\beta})(1-\pi(\hat{\beta}))}}.$$

The statistical properties of these residuals are not very satisfactory. In fact

$$r_{ij}=\begin{cases}-\sqrt{\frac{\pi_{ij}(\hat{\beta})}{1-\pi_{ij}(\hat{\beta})}} & \text{if } y_{ij}=0,\\ +\sqrt{\frac{1-\pi_{ij}(\hat{\beta})}{\pi_{ij}(\hat{\beta})}} & \text{if } y_{ij}=1.\end{cases}$$

This means that the asympotic distribution of the studentized residual is a mixture of two normal distributions.

It is more useful to look at the p-vectors $\rho_i\overset{\Delta}{=}u_i-\mu_i(\hat{\beta})=X_i'(y_i-\pi_i(\hat{\beta}))$. Observe that the ρ_i add up to zero. After some calculation we find, for $\underline{\rho}_{i(n)}\overset{\Delta}{=}\underline{u}_{(n)}-\mu_{(n)}(\hat{\underline{\beta}}_{(n)})$, that

$$n^{1/2}\{\Sigma_i(\beta)-\Sigma_i(\beta)\Sigma^{-1}(\beta)\Sigma_i(\beta)\}^{-1/2}\underline{\rho}_{i(n)}\overset{\mathcal{L}}{\Rightarrow}\mathcal{N}(0,I)$$

Or, to paraphrase again, its asymptotic distribution is normal with mean zero and variance $n^{-1}(\Sigma_i(\beta)-\Sigma_i(\beta)\Sigma^{-1}(\beta)\Sigma_i(\beta))$. Clearly this gives directly

$$n\underline{\rho}_{i(n)}'\{\Sigma_i(\beta)-\Sigma_i(\beta)\Sigma^{-1}(\beta)\Sigma_i(\beta)\}^{-1}\underline{\rho}_{i(n)}\overset{\mathcal{L}}{\Rightarrow}\chi^2(p).$$

The result also implies we can use

$$\tau_i\overset{\Delta}{=}(\Sigma_i-\Sigma_i\Sigma^{-1}\Sigma_i)^{-1/2}\rho_i$$

as more interesting vectors of studentized residuals.

6.2. **Hat Matrix.** In linear regression the *hat matrix* transform observations y into predicted values $\hat{y}$. This notion can be generalized in various ways in the multinomial logit model. The most useful generalization, perhaps, is to look at the transformation of u_i to $\mu_i(\hat{\beta}) = \mu_i(\mu^{-1}(u))$. Of course this transformation is nonlinear, and the hat matrix is defined as its derivative. Thus

$$\mathcal{H}_i(\hat{\beta}) = \Sigma_i(\hat{\beta})\Sigma^{-1}(\hat{\beta}).$$

Observe that these hat matrices add up to the identity. It is of some interest that the derivative of $\mu_i(\hat{\beta})$ with respect to u_k, where $k \neq i$, is also $\mathcal{H}_i(\hat{\beta})$.

We can summarize the size of the hat matrix by computing some overall size measure such as the determinant or the trace. For more detailed information about the size in various directions we can look at the diagonal or the eigenvalues.

6.3. **Likelihood Displacement.** We use data weights w, one w_i for each $i = 1, \cdots, n$ [Cook, 1986]. Define

$$\mathcal{L}(\beta, w) = u'\beta - \sum_{i=1}^{n} w_i \log \sum_{j \in J_i} \exp(\beta' x_{ij}),$$

where now

$$u = \sum_{i=1}^{n} w_i u_i.$$

The maximum likelihood estimate $\hat{\beta}$ and the maximum of the likelihood $\mathcal{L}(\hat{\beta}(w), w)$ can now be thought of as functions of w.and some calculation gives

$$\mathcal{D}_i \mathcal{L}(\hat{\beta}(w), w) = u_i' \hat{\beta}(w) - \log \sum_{j \in J_i} \exp(\hat{\beta}(w)' x_{ij}),$$

$$\mathcal{D}_i \hat{\beta}(w) = -\Sigma^{-1}(\hat{\beta}(w), w)\rho_i(\hat{\beta}(w), w).$$

Suppose $\overline{w}$ has elements n^{-1} and $\tilde{w}$ has elements $(n-1)^{-1}$ everywhere except for element i, where it has a zero. Thus $\beta(\overline{w}$ is our ususla maximum likelihood estimate. Using $\tilde{w}$ studies the effect of deleting observation i. We find

$$\hat{\beta}(\tilde{w}) = \hat{\beta} + \Sigma^{-1}(\hat{\beta})\rho_i(\hat{\beta}) + o((n-1)^{-1}).$$

Observe we get the same formula for $\tilde{\beta}$ if we take a single Newton-Raphson step from $\hat{\beta}$ towards the maximum of the likelihood of the remaining $n-1$ observations [Pregiborn, 1981].

The formula for $\tilde{\beta}$ suggests the Cook's distance approximation

$$(\tilde{\beta} - \hat{\beta})'\Sigma(\hat{\beta})(\tilde{\beta} - \hat{\beta}) \approx \rho_i(\hat{\beta})'\Sigma^{-1}(\hat{\beta})\rho_i(\hat{\beta})$$

7. Example: Neighborhood Choice

The data used in this example are from the 1992-1994 Multi-City Study of Urban Inequality (hereafter MCSUI)[2]. Each respondent is given a card showing a stylized neighborhood containing 15 empty houses. The respondent's house is in the center of the neighborhood. The respondent is then asked:

> Now I'd like you to imagine an ideal neighborhood that has the ethnic and racial mix you personally would feel most comfortable

[2] Bobo, Lawrence, James Johnson, Melvin Oliver, Reynolds Farley, Barry Bluestone, Irene Browne, Sheldon Danziger, Garry Green, Harry Holtzer, Maria Krysan, Michael Massagli, and Camille Zubrinsky Charles. 1998 .*Multi-City Study of Urban Inequality, 1992-1994*: [Atlanta, Boston, Detroit, and Los Angeles][Computer file]. ICPSR version. Atlanta, GA: Mathematica/-Boston, MA: University of Massachusetts, Survey Research Laboratory/Ann Arbor, MI: University of Michigan, Detroit Area Study and Institute for Social Research, Survey Research Center/Los Angeles, CA: University of California, Survey Research Program [producer]. Inter-university Consortium for Political and Social Research [distributor].

> in. Here is a blank neighborhood card like those we have been using. Using the letters A for Asian, B for Black, H for Hispanic, and W for White, please put a letter in each of these houses to represent your ideal neighborhood where you would most like to live. Please be sure to fill in all the houses.

For example, Figure 7 shows a card filled out by one black respondent. The respondent's own house is shown as an X in the middle of the neighborhood. This respondent indicated that his "ideal" neighborhood consisted of 3 Asian households, 5 white households, 4 black households, and 2 Hispanic households. If we ignore the position of each neighbor, each respondent chooses his or her ideal neighborhood from 680 possible configurations of the four race/ethnic groups.

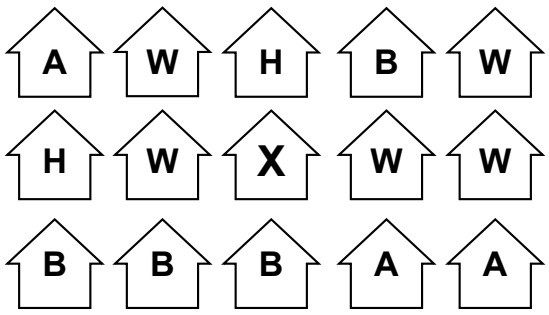

FIGURE 1. Card Filled out by a Black Respondent

We use the information given in the card to construct the x_{ij} vectors for the p regressors. The x_{ij} vectors include information on the race-ethnic composition of each possible ideal neighborhood, which may affect its attractiveness to survey respondents. Our models allow for the possibility that this effect is nonlinear. For

example, neighborhoods that have almost no black residents may be very unattractive to blacks, neighborhoods in which blacks have significant representation may be very attractive, and neighborhoods that are almost 100 percent black may also be unattractive. Thus, we include the linear and quadratic terms for the proportions in each of four race-ethnic groups (whites, blacks, Asians, and Hispanics) as separate variables in the discrete choice models. We treat neighborhood proportion white as the omitted category. For the neighborhood shown in Figure 7, the x_{ij} vector is shown below:

$$x_{ij} = [0.214, 0.286, 0.143, 0.046, 0.082, 0.020]$$

where, for example, $0.214 = \frac{3}{14}$ = the proportion of Asian households shown in this neighborhood, and 0.046 is the proportion of Asian households squared. The scalar y_{ij} in this case would be 1, indicating that this is the chosen neighborhood. In this case, we have six covariates so $p = 6$.

We use the conditional logit model to examine the relationship between the race of the respondent and the probability that he or she selects a neighborhood of a given race composition. Since there are 14 empty houses, and four possible races for each neighbor, there are $(14 + 4 - 1)!/(14!(4 - 1)!) = 680$ possible neighborhoods that respondents can choose from. In the present application, each respondent has the potential of moving to every possible neighborhood. Thus, $m_i = m$ for all i.

The estimated utility functions are shown in Figure 7, below.

References

D.R. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, Belmont, Massachusetts, 2003.

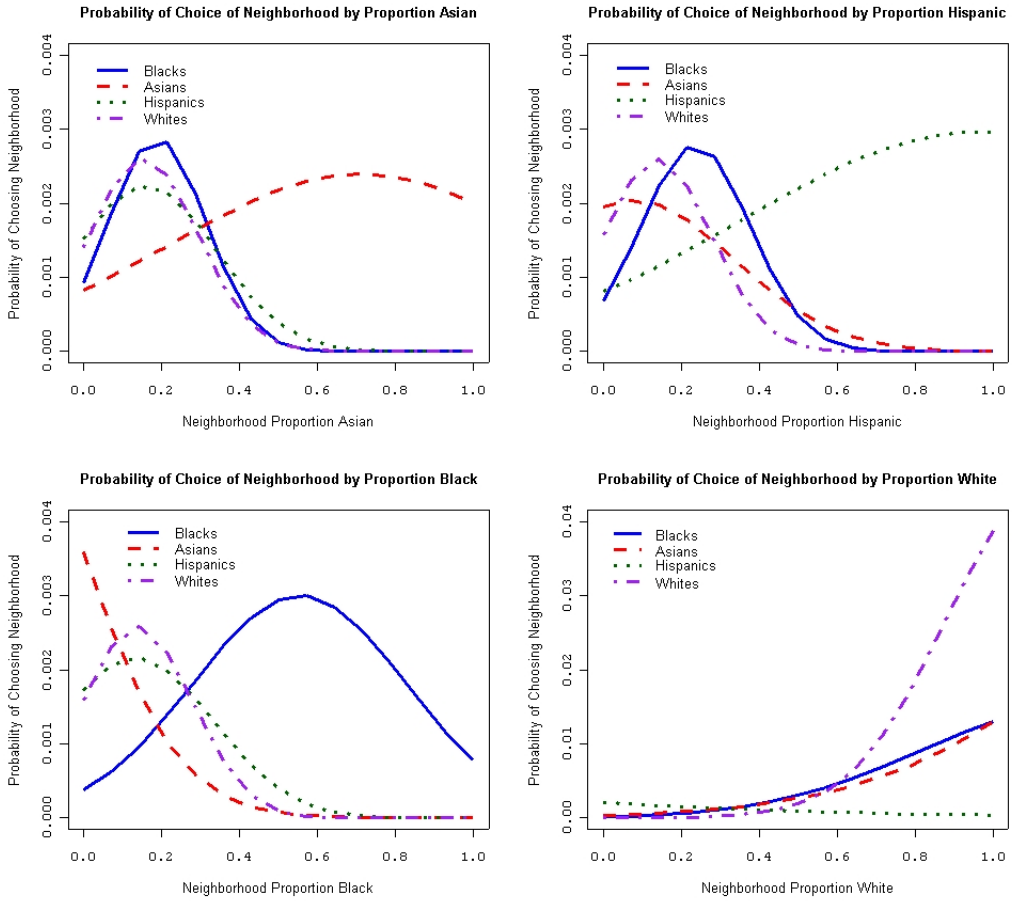

FIGURE 2. Estimated Utility Functions

R.D. Cook. Assessment of Local Influence (with Discussion). *Journal of the Royal Statistical Society*, B 48:133–169, 1986.

M.G. Hahn and M.J. Klass. The Multidimensional Central Limit Theorem for Arrays Normed by Affine Transformations. *Annals of Probability*, 9(4):611–623, 1981.

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

S.G. Krantz and H.R. Parks. *The Implicit Function Theorem. History, Theory, and Applications*. Birkhäuser, 2002.

D. Pregiborn. Logistic Regression Diagnostics. *Annals of Statistics*, 9:705–724, 1981.

A.W. Van Der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

H.T.V Vu, R.A. Maller, and M.J. Klass. On the Studentization of Random Vectors. *Journal of Multivariate Analysis*, 57:142–155, 1996.

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

E-mail address, Elisabeth Bruch: bruch@stat.ucla.edu

URL, Elisabeth Bruch: http://www.stat.ucla.edu/~bruch