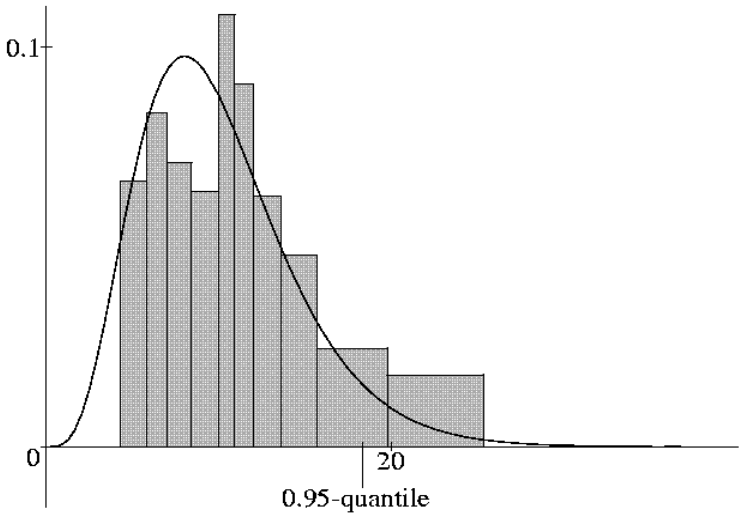
UNIVERSITÄT POTSDAM
Institut für Mathematik

Modeling and Scaling of Categorical Data

Henning Läuter

Ayad Ramadan



Mathematische Statistik und
Wahrscheinlichkeitsrechnung

Universität Potsdam – Institut für Mathematik

Mathematische Statistik und Wahrscheinlichkeitstheorie

Modeling and Scaling of Categorical Data

Henning Läuter and Ayad Ramadan

Institute of Mathematics, University of Potsdam
e-mail: laeuter@uni-potsdam.de
ayad_math@yahoo.com

Preprint 2010/03

März 2010

Impressum

Herausgeber: Mathematische Statistik und Wahrscheinlichkeitstheorie
am Institut für Mathematik der Universität Potsdam

Adresse: Am Neuen Palais 10
14469 Potsdam

Telefon: +49-331-977 1500
Fax: +49-331-977 1578
E-mail: neisse@math.uni-potsdam.de

ISSN 1613-3307

Modeling and Scaling of Categorical Data

Henning Läuter and Ayad Ramadan

University of Potsdam, Institute of Mathematics

laeuter@uni-potsdam.de, ayad_math@yahoo.com

Abstract

Estimation and testing of distributions in metric spaces are well known. R.A. Fisher, J. Neyman, W. Cochran and M. Bartlett achieved essential results on the statistical analysis of categorical data. In the last 40 years many other statisticians found important results in this field.
Often data sets contain categorical data, e.g. levels of factors or names. There does not exist any ordering or any distance between these categories. At each level there are measured some metric or categorical values. We introduce a new method of scaling based on statistical decisions. For this we define empirical probabilities for the original observations and find a class of distributions in a metric space where these empirical probabilities can be found as approximations for equivalently defined probabilities. With this method we identify probabilities connected with the categorical data and probabilities in metric spaces. Here we get a mapping from the levels of factors or names into points of a metric space. This mapping yields the scale for the categorical data.
From the statistical point of view we use multivariate statistical methods, we calculate maximum likelihood estimations and compare different approaches for scaling.

Key words: Multivariate scaling, discrimination, power of multivariate tests

1 Introduction

Estimation and testing for distributions of metric random variables are known since the end of the nineteenth century. R.A. Fisher and many other statisticians developed very efficient statistical methods for analyzing medical and biological data. These methods correspond to regression, multivariate analysis and in general to data analysis. Many procedures, e.g. the procedures of the analysis of variance belong to the basic methods in applied statistics.
Essential contributions about statistics of categorical data were developed first by R.A. Fisher, J. Neyman, W. Cochran and M. Bartlett. One finds very different strong results for analyzing categorical data since the 1960s. Mostly data structures from

social, biological, medical and technical areas are analyzed. In biomedical applications categories as sex, race or social strata are considered, in technical problems one works with technical patterns or places. In social problems one uses verbal assessments or marks, in political or philosophical context one finds arrangements as "liberal", "moderate" or "conservative".
In this paper we introduce a method of scaling based on statistical decisions, especially classification methods are used. We will concentrate on methods and examples with categorical data. But it will be clear that the proposed procedures can be used as a pretreatment in other data structures for generating such transformed data which conform with assumptions in standard software.
Multidimensional scaling is considered by several authors. In most of the cases they use similarities or dissimilarities and then they find scales for the categories ([EvDu01]). Instead of such geometrical approaches we use here statistical decisions.

2 Basic model

We consider the q-way classification model which is used mostly in the analysis of variance. At least the structure is interesting for us. The basic assumptions can be explained in the 2-way classification. We are given data in the following structure.

	Factor B				
Factor A	level 1	level 2	level 3	$\cdots$	level b
level 1	z_{11j}	z_{12j}	z_{13j}	$\cdots$	z_{1bj}
	$j = 1, ..., m_{11}$	$j = 1, ..., m_{12}$	$j = 1, ..., m_{13}$	$\cdots$	$j = 1, ..., m_{1b}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
level a	z_{a1j}	z_{a2j}	z_{a3j}	$\cdots$	z_{abj}
	$j = 1, ..., m_{a1}$	$j = 1, ..., m_{a2}$	$j = 1, ..., m_{a3}$	$\cdots$	$j = 1, ..., m_{ab}$

At level s of the factor A and level t of factor B there are m_{st} observations and the total sample size is

$$n = \sum_{s=1}^{a} \sum_{t=1}^{b} m_{st}.$$

Here the categorical variables are the levels of the factors A and B. The $L := a \cdot b$ categories are described by

$$\{(s,t) : \quad s = 1, ..., a;\ t = 1, ..., b\}.$$

For finding a scale for the categories it is convenient to group the observations $\{z_{stj}\}$ in given classes. We assume that there are given k classes and h_{il} is the number of observations in the category $l = (s,t)$ falling in the class K_i. Then we have the

contingency table

	Factor B								
Factor A	level 1		level 2		level 3		$\cdots$	level b	
level 1	$h_{1(1,1)}$	$...h_{k(1,1)}$	$h_{1(1,2)}$	$...h_{k(1,2)}$	$h_{1(1,3)}$	$...h_{k(1,3)}$	$\cdots$	$h_{1(1,b)}$	$...h_{k(1,b)}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\ddots$	$\vdots$	
level a	$h_{1(a,1)}$	$...h_{k(a,1)}$	$h_{1(a,2)}$	$...h_{k(a,2)}$	$h_{1(a,3)}$	$...h_{k(a,3)}$	$\cdots$	$h_{1(a,b)}$	$...h_{k(a,b)}$

This table of frequencies is our starting data set. The interpretation of this table is the following. At the level $l = (s, t)$ we have observations from different classes and from the

$$h_{+l} := \sum_{i=1}^{k} h_{il} = \sum_{i=1}^{k} h_{i(s,t)}$$

observations the parts for the classes are given by

$$\frac{h_{1l}}{h_{+l}}, ..., \frac{h_{kl}}{h_{+l}}.$$

Such tables are obtained in a similar way if $q > 2$. The dimension of l depends on q in general.

2.1 Modeling of categorical data

We will find a model for data structures given in the last tables. For this we remember the discrimination of distributions or classes. There are given k distributions with densities $f_1, ..., f_k$ over a space $\mathbb{R}^p$ and for each point $y \in \mathbb{R}^p$ it is known that it can be a realization of one of the classes. Furthermore let $\pi_1, ..., \pi_k$ be prior probabilities for the classes. Then it is known that a given realization y of the random variable Y with the density f with

$$f(y) = \sum_{j=1}^{k} \pi_j f_{\vartheta_j}(y)$$

corresponds to the class K_i with the probability

$$\mathsf{P}(Y \in K_i \mid Y = y)) = \frac{\pi_i f_{\vartheta_i}(y)}{\sum_{j=1}^{k} \pi_j f_{\vartheta_j}(y)} =: \widetilde{p}_i(y). \tag{1}$$

Consequently under the assumption that a point y (or a point in a very near neighborhood of y) has the frequency $\tilde{m}$ in the data set then we expect $\tilde{m} \cdot \widetilde{p}_i(y)$ of these points corresponding to class K_i.

2.1.1 Modeling of categories

We denote by $\mathsf{Mult}(m, p_1, ..., p_k)$ the multinomial distribution where p_i is the probability for the class K_i and we repeat the experiment m times. Then we expect that under these $\tilde{m}$ observations approximately $\tilde{m} \cdot p_i$ correspond to K_i. As a consequence from the last two subsections we formulate the definition.

Definition 2.1 *The points* $\{s_l \in \mathbb{R}^p, l = 1, ..., L\}$ *are called scale points for the categories* $\{x_l, l = 1, ..., L\}$ *if*

$$(h_{1l}, ..., h_{kl})$$

are realizations of independent multinomial random variables W_l *with*

$$W_l \sim \mathsf{Mult}(h_{+l}, \widetilde{p}_1(s_l), ..., \widetilde{p}_k(s_l)),\ l = 1, ..., L$$

and

$$h_{+l} = \sum_{i=1}^{k} h_{il}.$$

This means that in the statistical model the same expected frequencies occur as in the table. In general the densities depend on some parameters. Then one has to estimate the distributional parameters and the scale parameters. This can be done using the likelihood principle. Here we use another criterion. We will find such scales that the classes will be discriminated as well as possible. Voinov and Nikulin considered in [VoNi93] multivariate multinomial distributions for identically distributed W_l, here we use a more general model.

2.2 Determination of observations

Scale points are to be constructed on the basis of the observations. The observations are those which are given by the categories and the frequencies. In our understanding the categories are identified with points $x_1, ..., x_L \in \mathbb{R}^p$ and these points are to be determined in an optimal way. The observations express the correspondence to some classes, denoted by $\{y_{11}, ..., y_{kn_k}\}$. Explicitly we have the observations

$$\{y_{11}, ..., y_{1n_1}\} = \{h_{11} \text{ times } x_1, h_{12} \text{ times } x_2, ..., h_{1L} \text{ times } x_L\},$$

hence we have $n_1 = h_{1+}$. Or we write

$$y_{1t} = x_1, t = 1, ..., h_{11}; y_{1t} = x_2, t = h_{11}+1, ..., h_{11}+h_{12}; ...; y_{1t} = x_L, t = h_{1+}-h_{1L}, ..., h_{1+}.$$

In an analogous way we have for the other classes $i = 2, ..., k$

$$y_{it} = x_1, t = 1, ..., h_{i1}; y_{it} = x_2, t = h_{i1}+1, ..., h_{i1}+h_{i2}; ...; y_{it} = x_L, t = h_{i+}-h_{iL}, ..., h_{i+}.$$

It holds $n_i = h_{i+}$. For statistical decisions one needs assumptions on the distributions.

2.3 Choice of distributions

In general one chooses $y_{ij} \in \mathbb{R}^q$ if one has data from a q-way classification model. This means $p = q$. But sometimes a lower-dimensional space is also possible especially if some factors have not a large influence on the results. The special case $p = 1$ is of interest if one likes to have an ordered scale for the categories.
Depending on the meaning of the observations we can choose the distributions. Quite often binomial, normal or Poisson distributions are useful, but especially in reliability or survival analysis exponential or Weibull distributions are to be chosen.
Assuming that we are given k distributions $\mathsf{P}_{\vartheta_1}, ..., \mathsf{P}_{\vartheta_k}$ and for each distribution P_{ϑ_i} with a density f_{ϑ_i} we have a random sample $Y_{i1}, ..., Y_{in_i}$. All random variables should be independent. For testing

$$\mathcal{H} : \mathsf{P}_{\vartheta_1} = ... = \mathsf{P}_{\vartheta_k}$$

against $\mathcal{K}$, that not all distributions are the same, we use the likelihood ratio test.
The joint density for $Y = (Y_{11}, ..., Y_{kn_k})$ is denoted by $f_{\vartheta_1,...,\vartheta_k}$. As usually the LRT is given by

$$\varphi(y) = 1 \quad \text{if} \quad R_n(y) := \frac{\max_{\vartheta_1,...,\vartheta_k} f_{\vartheta_1,...,\vartheta_k}(y)}{\max_{\vartheta} f_{\vartheta,...,\vartheta}(y)} \geq c,$$

where c ensures the significance level.

2.3.1 Normal distributions

We assume that

$$\begin{matrix} Y_{11} & \dots & Y_{1n_1} \\ \vdots & \ddots & \vdots \\ Y_{k1} & \dots & Y_{kn_k} \end{matrix}$$

are independent and normally distributed p-dimensional random variables, $Y_{ij} \sim \mathsf{N}_p(\mu_i, \Sigma)$. Then we consider the test problem

$$\mathcal{H} : \quad \mu_1 = ... = \mu_k \quad \text{against} \quad \mathcal{K} : \quad \text{not}\, \mathcal{H}. \tag{2}$$

We denote the sample mean for the ith distribution by $y_{i\cdot}, \quad i = 1, ..., k$, the total mean by

$$y_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{j=1}^{k} n_j y_{j\cdot} \quad .$$

The unbiased estimator for the variance is

$$S = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\cdot})(Y_{ij} - Y_{i\cdot})^t.$$

Then

$$T_0^2(Y) = \frac{n-k-p+1}{(k-1)(n-k)p} \sum_{i=1}^{k} n_i (Y_{i\cdot} - Y_{\cdot\cdot})^t S^{-1} (Y_{i\cdot} - Y_{\cdot\cdot})$$

is approximately F-distributed. H. Ahrens and J. Läuter proposed in [AhLa81] the approximation $T_0^2(Y) \approx \mathsf{F}_{g_1,g_2}$ for

$$g_1 = \begin{cases} \frac{(k-1)(n-k-p)p}{n-(k-1)p-2} & \text{if} \quad n-(k-1)p-2 > 0 \\ \infty & \text{otherwise,} \end{cases}$$

$$g_2 = n-k-p+1.$$

Then an admissible test is given by

$$\varphi(y) = \begin{cases} 1 & \text{if} \quad T_0^2(y) > \mathsf{F}_{g_1,g_2;\alpha} \\ 0 & \text{otherwise,} \end{cases}$$

for the α-fractile of the F_{g_1,g_2}-distribution.

2.3.2 Exponential distributions

We choose $p = 1$. We consider independent exponentially distributed variables

$$\begin{array}{ccc} Y_{11} & \dots & Y_{1n_1} \\ \vdots & \ddots & \vdots \\ Y_{k1} & \dots & Y_{kn_k}. \end{array}$$

With the densities

$$f_\mu(x) \;=\; \mu \exp(-\mu x)\, 1_{(0,\infty)}(x)$$

we assume that Y_{ij} has the density f_{μ_i}. The likeliood ratio test statistic has the form (up to a factor)

$$R(y) \;=\; \frac{y_{\cdot\cdot}^n}{y_{1\cdot}^{n_1} \cdot \ldots \cdot y_{k\cdot}^{n_k}}.$$

If $R(y)$ is large enough then we reject the hypothesis.

3 Criteria for scaling

M.G. Kendall and A. Stuart [KeSt67] and lateron H. Ahrens and J. Läuter in [AhLa81] introduced a method for scaling which bases on a test statistic. This will be generalized

for higher dimensional q-way classification tables. This was considered by H. Läuter in [La07] too. At first we denote the levels of the q factors in an arbitrary way by real numbers. The factor i has ν_i levels. Then we put τ_{ij} for the level j of the factor i, all levels are described by

$$\tau \;=\; (\tau_{11}, ..., \tau_{1\nu_1}, ..., \tau_{q\nu_q})^t$$

and altogether we have $\nu \;=\; \sum_i \nu_i$ levels. In section 2.2 the categories were identified by $x_1, ..., x_L$ and we introduced the y_{ij}. For any x_l we find a $p \times \nu$ matrix C_l with $x_l = C_l\tau$. Every y_{ts} is one of the values $C_1\tau, ..., C_L\tau$. We use

$$h_{t\cdot} = \frac{1}{L}\sum_{l=1}^{L} h_{tl}, \quad h_{\cdot l} = \frac{1}{k}\sum_{t=1}^{k} h_{tl}, \quad h_{\cdot\cdot} = \frac{1}{kL}\sum_{t=1}^{k}\sum_{l=1}^{L} h_{tl},$$

$$h_{t\cdot}\cdot L = \sum_{l=1}^{L} h_{tl} = n_t, \quad h_{\cdot\cdot}\cdot kL = n.$$

Then we calculate

$$y_{t\cdot} = \frac{1}{n_t}\sum_{s=1}^{n_t} y_{ts} = \frac{1}{n_t}\Big(h_{t1}C_1 + ... + h_{tL}C_L\Big)\tau, \quad y_{\cdot\cdot} = \frac{k}{n}\Big(h_{\cdot 1}C_1 + ... + h_{\cdot L}C_L\Big)\tau$$

$$y_{t\cdot} - y_{\cdot\cdot} = \Big((\frac{h_{t1}}{n_t} - \frac{kh_{\cdot 1}}{n})C_1 + ... + (\frac{h_{tL}}{n_t} - \frac{kh_{\cdot L}}{n})C_L\Big)\,\tau =: D_t\,\tau.$$

These values are to be substituted in the test statistics. In the normal case we have T_0^2 as the test statistic. For calculating this statistic we use

$$H := \sum_{i=1}^{k} n_i\Big(y_{i\cdot} - y_{\cdot\cdot}\Big)\Big(y_{i\cdot} - y_{\cdot\cdot}\Big)^t = \sum_{i=1}^{k} n_i D_i\,\tau\,\tau^t D_i^t,$$

$$S := \frac{1}{n-k}\sum_{i=1}^{k}\sum_{s=1}^{n_i}\Big(y_{is} - y_{i\cdot}\Big)\Big(y_{is} - y_{i\cdot}\Big)^t = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{l=1}^{L} h_{il}F_{il}\,\tau\,\tau^t F_{il}^t$$

for

$$F_{il} = C_l - \frac{1}{n_i}\Big(h_{i1}C_1 + ... + h_{iL}C_L\Big)$$

and

$$T_0^2 = \frac{n-k-p+1}{(k-1)(n-k)p}\;\sum_{i=1}^{k} n_i(y_{i\cdot} - y_{\cdot\cdot})^t S^{-1}(y_{i\cdot} - y_{\cdot\cdot}) =$$

$$= \frac{n-k-p+1}{(k-1)(n-k)p}\;tr\Big(HS^{-1}\Big),$$

$$tr\Big(HS^{-1}\Big) \;=\; \tau^t\Big[\sum_{i=1}^{k} n_i D_i^t S^{-1} D_i\Big]\tau$$

with

$$S = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{l=1}^{L} m_{il} F_{il}\,\tau\,\tau^t F_{il}^t.$$

In the case of exponential distributions we considered the LRT. There we had $p = 1$ and so we have $L = \nu$ and the statistic was

$$R(y) \;=\; \frac{y_{\cdot\cdot}^{n}}{y_{1\cdot}^{n_1}\cdot\ldots\cdot y_{k\cdot}^{n_k}}. \tag{3}$$

Substituting here the values with τ we obtain

$$R \;=\; \frac{\Big[\frac{k}{n}\Big(h_{\cdot 1}C_1 + \ldots + h_{\cdot L}C_L\Big)\tau\Big]^n}{\Big[\frac{1}{n_1}\Big(h_{11}C_1 + \ldots + h_{1L}C_L\Big)\tau\Big]^{n_1}\cdot\ldots\cdot\Big[\frac{1}{n_k}\Big(h_{k1}C_1 + \ldots + h_{kL}C_L\Big)\tau\Big]^{n_k}}. \tag{4}$$

In the case $q = 1$ it holds $C_l\tau = \tau_{1l} =: \tau_l$ and therefore we get

$$R \;=\; \frac{k^n n_1^{n_1}\cdot\ldots\cdot n_k^{n_k}}{n^n}\,\frac{\Big(h_{\cdot 1}\tau_1 + \ldots + h_{\cdot L}\tau_L\Big)^n}{\Big(h_{11}\tau_1 + \ldots + h_{1L}\tau_L\Big)^{n_1}\cdot\ldots\cdot\Big(h_{k1}\tau_1 + \ldots + h_{kL}\tau_L\Big)^{n_k}}. \tag{5}$$

The aim is to find such a scale that the distributions or here classes can be discriminated as well as possible. Therefore we have to determine such a vector τ^* that maximizes the corresponding test statistic. In the normal case the test bases on T_0^2, for the exponential distributions the likelihood ratio test statistic R was proposed.

Definition 3.1 *If LR denotes the test statistic where large values of LR lead to the rejection of the hypothesis then τ^* with*

$$LR(\tau^*) = \max_{\tau} LR(\tau) \tag{6}$$

is called a most separating scale.

4 Calculation of most separating scales

In general one has to use some optimization software for finding a maximal τ^*. We will consider in some detail the special case of normal distributions. In section 2.3.1 we considered the statistic T_0^2 as the statistic to be maximized. Up to a factor this coincides with

$$tr(HS^{-1}) = \tau^t\Big[\sum_{i=1}^{k} n_i D_i^t S^{-1} D_i\Big]\tau \tag{7}$$

with

$$S = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{l=1}^{L} h_{il} F_{il}\, \tau\, \tau^t F_{il}^t.$$

4.1 One-dimensional normal distributions

We consider $p = 1$. Then we have the $1 \times L$ matrices C_l, D_i, F_{il} and get

$$H = \tau^t A \tau, \quad S = \tau^t B \tau \tag{8}$$

for the matrices

$$A = \sum_{i=1}^{k} n_i D_i^t D_i, \quad B = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{l=1}^{L} h_{il} F_{il}^t F_{il}. \tag{9}$$

Therefore the $\tau^* = (\tau_1^*, ..., \tau_L^*)^t$ maximizing $tr(HS^{-1})$ is determined by the eigenvector to the maximal eigenvalue of

$$A\tau = \lambda B \tau. \tag{10}$$

Therefore in the case $p = 1$ the optimal scale vector τ^* can be determined exactly. The level l corresponds to the number τ_l^*.

4.2 Higher-dimensional case

Now we consider q-way classification models and $p \leq q$. Then we have the $p \times \nu$ matrices C_l, D_i, F_{il} and with $S_\tau := S$ we have

$$tr(HS^{-1}) = tr(HS_\tau^{-1}) = \tau^t \Big[\sum_{i=1}^{k} n_i D_i^t S_\tau^{-1} D_i \Big] \tau \tag{11}$$

for

$$S_\tau = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{l=1}^{L} h_{il} F_{il} \tau \tau^t F_{il}^t. \tag{12}$$

Defining

$$\psi(\tau, a) := a^t \Big[\sum_{i=1}^{k} n_i D_i^t S_\tau^{-1} D_i \Big] a \tag{13}$$

and then τ^* fulfills

$$\psi(\tau^*, \tau^*) = \max_{\tau} \psi(\tau, \tau). \tag{14}$$

We see that ψ does not change if τ is substituted by $\mu\tau$ for any real μ.

Definition 4.1 *$\widetilde{\tau}$ is called a local extremum if*

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi\Big((1-\lambda)\widetilde{\tau}+\lambda v,(1-\lambda)\widetilde{\tau}+\lambda v\Big)|_{\lambda=0}\leq 0 \quad \forall v\in\mathbb{R}^p.$$

We are interested in characterizing such a local extremum. This gives us the next theorem.

Theorem 1 *$\widetilde{\tau}$ is a local extremum if and only if $\alpha(\widetilde{\tau})=0$ with*

$$\alpha(\tau):=\sum_{i=1}^{k}n_iD_i^tS_\tau^{-1}D_i\tau-\frac{1}{n-k}\sum_{i=1}^{k}n_i\sum_{j=1}^{k}\sum_{l=1}^{L}h_{jl}F_{jl}^tS_\tau^{-1}D_i\tau\tau^tF_{jl}^tS_\tau^{-1}D_i\tau.$$

BEWEIS: We put $\tau_\lambda=(1-\lambda)\widetilde{\tau}+\lambda v$ and obtain

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\tau_\lambda=v-\tau_\lambda,\quad \frac{\mathrm{d}}{\mathrm{d}\lambda}\tau_\lambda\tau_\lambda^t|_{\lambda=0}=(v-\widetilde{\tau})\widetilde{\tau}^t+\widetilde{\tau}(v-\widetilde{\tau})^t,$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}S_{\tau_\lambda}^{-1}=-S_{\tau_\lambda}^{-1}(\frac{\mathrm{d}}{\mathrm{d}\lambda}S_{\tau_\lambda})S_{\tau_\lambda}^{-1}$$

and consequently

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}S_{\tau_\lambda}^{-1}|_{\lambda=0}=-\frac{1}{n-k}S_{\widetilde{\tau}}^{-1}\sum_{j=1}^{k}\sum_{l=1}^{L}h_{jl}F_{jl}(v\widetilde{\tau}^t+\widetilde{\tau}v^t-2\widetilde{\tau}\widetilde{\tau}^t)F_{jl}^tS_{\widetilde{\tau}}^{-1}.$$

Now we calculate in a direct way

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi(\tau_\lambda,\tau_\lambda)|_{\lambda=0}=2v^t\alpha(\widetilde{\tau})$$

and so the theorem is proven.

□

This theorem gives us a proposal for the calculation of a local extremum.
Step 1: Choice of an initial point τ_0.
Step 2: Set $w:=\frac{1}{|\alpha(\tau_0)|}\alpha(\tau_0)$ and $\widetilde{\tau}_\lambda=(1-\lambda)\tau_0+\lambda w$ for euclidian norm $|\alpha(\tau_0)|$ of $\alpha(\tau_0)$.
Step 3: Determine such λ_1 that

$$\psi(\widetilde{\tau}_{\lambda_1},\widetilde{\tau}_{\lambda_1})=\max_\lambda\psi(\widetilde{\tau}_\lambda,\widetilde{\tau}_\lambda).$$

Step 4: Set $\tau_1:=\widetilde{\tau}_{\lambda_1}$ and calculate $\alpha(\tau_1)$. Now we set $w:=\frac{1}{|\alpha(\tau_1)|}\alpha(\tau_1)$ and $\widetilde{\tau}_\lambda=(1-\lambda)\tau_1+\lambda w$ and so on.
In this way we get a sequence of q-vectors $\tau_0,\tau_1,\tau_2,...$ and have

$$\psi(\tau_0,\tau_0)\leq\psi(\tau_1,\tau_1)\leq\psi(\tau_2,\tau_2)\leq...$$

In each step one can check $\alpha(\tau_j)$ and decide to proceed in the sequential calculation or to break up. Under $\alpha(\tau_j)\approx 0$ one reaches the optimum.

Acknowledgement The authors are very grateful to Prof. H. Liero and Dr. Martin Läuter for their helpful comments and suggestions.

References

[AhLa81] Ahrens, H. and Läuter, J.: Mehrdimensionale Varianzanalyse. Wiley, Akademie-Verlag, Berlin (1981)

[EvDu01] Everitt, B.S. and Dunn, G.: Applied Multivariate Data Analysis. Hodder Education London (2001)

[KeSt67] Kendall, M.G. and Stuart, A. : The Advanced Theory of Statistics. Wiley, Griffin&Comp. London. (1967)

[La07] Läuter, H.: Modeling and scaling of categorical data. Preprint, University of Linz (2007)

[VoNi93] Voinov, V.G. and Nikulin, M.S.: Unbiased Estimators and Their Applications. Kluwer Academic Publishers. (1993)