Correspondence Analysis of Identification Data.

Uwe Mortensen, Günter Meinhardt

Westfälische Wilhelms-Universität Münster
Fachbereich Psychologie und Sportwissenschaften, Institut III
D-48 149 Münster, Fliednerstr. 21

Abstract: Even when the main interest is in the sensory processes, the process of identifying a visual pattern under threshold conditions is determined by the way sensory and decision processes are interlocked. Important work concerned with these questions was done by e.g. Ashby and Townsend (1986) and Kadlec and Townsend (1992), who investigated the conditions of sampling independence and perceptual and decisional separability as tests for perceptual independence. These tests are computationally involved and the question is whether there exist simpler, more direct procedures allowing to test for sampling and perceptual independence. Here, methods of Dual Scaling and in particular Correspondence Analysis (CA), applied to confusion matrices, are suggested. CA of confusion matrices yields scale values for (i) the stimulus patterns, and (ii) for the responses. It will be argued that to the extend the sensory activity, generated by showing a stimulus pattern, can be represented by a random variable, the scale values of the stimuli will provide estimates of the mean values of these random variables, relative to their variances, an the scale values of the responses provide information about possible response biasses. The argument carries over to data from discrimination experiments; however, in this paper only data from identification experiments will be considered.

Contents

1 Introduction

Data from discrimination and identification experiments may provide insight into perceptual, in particular into coding processes. In order to explain errors in discrimination and identification judgments one usually assumes that the sensory representation of a stimulus pattern is, under the conditions of the experiment, noisy so that for nonidentical stimuli the overlap of the corresponding representations may turn out to be sufficiently large to generate wrong identity responses ("confusions"). The subject will, in some implicit way, set up boundaries allowing the classification of sensory representations. The responses will be biased if these boundaries are chosen sub-optimally.

Discrimination and identification experiments have a long history; Townsend and Landon (1983) provided an overview over available models in particular of the identification process. Ashby and Townsend (1986) discussed the notion of independent processing of features or stimulus dimensions, and Townsend and Ashby (1982) provided experimental tests of identification models. Nosofsky (1986) discussed the generalisation of identification processes to categorisation processes, a topic further investigated by Caelli et al. (1987), Rentschler et al. (1992), (1996) and Jüttner et al. (1997). Balakrishnan and Ratcliff(1996) and Balakrishnan (1998) suggested that comparisons are made by the subject with respect to a subjective likelihood ratio, and provided data that comply with this model (Balakrishnan, 1999); his results suggest that response bias has an effect on the encoding distributions, but no effect on the decision rule. This finding may be interpreted within the framework of models assuming processes of information integration (Busemeyer and Townsend, 1993; Diederich, 1997; Link and Heath, 1975). While these approaches are of high interest, a much simpler (but not necessarily better) ansatz will be made in this paper: it will be assumed that the subject evaluates the difference between the sensory activities, expressed as the difference between the corresponding random variables. It turns out that data from different experiments appear to be compatible with this approach.

One way to characterise the sensory activity is to specify a a random vector $\vec{X} = (x_1, \dots, x_s)'$, where the components x_i, $1 \leq i \leq s$ are random variables representing aspects of the neural activity that make up the sensory activity or representation. Given the x_i are normally distributed, they are specified once their means and variances are known. If $s = 1$ a measure for the sensitivity to a pattern is $d' = (\mu_{sn} - \mu_n)/\sigma$, where μ_{sn} is the mean when the pattern is presented, and μ_n is the mean when no pattern is presented, and σ is the standard deviation of X, assumed to be independent of the actual presentation of the pattern (Tanner and Swets, 1954). d' is independent of the boundaries and thus a measure of sensitivity that is free of bias (provided the equal-variance assumption is correct).

A violation of these assumptions may lead to misinterpretations of the data (Maloney and Thomas, 1991); however, an explicit test of the assumptions is usually rather time consuming. Thus it is useful to have some method that yield measures that are, on the one hand, equivalent to d' and that, on the other hand, are not based on particular assumptions concerning the distributions.

It will be argued that the application of Correspondence Analysis (CA) to discrimination and identification data provides such a measure, and additionally allows to evaluate possible response biases. CA belongs to the class of Dual-Scaling-methods. These methods allow to compute scale values for the row and column categories of a contingency table. The scale values refer to at least one (latent) dimension or attribute that is common to both sets of categories, which may therefore ”explain” the possibly existing dependencies among row and column categories.

The calculation of the scale values does not require a specific assumption about the conditional distributions of X. The only assumption that has to be made refers to the way the subject utilises the sensory information in order to make a judgment.

2 Models of the identification process

2.1 The structure of the experiment

Let $S = \{S_i | 1 \leq i \leq I\}$ and $T = \{T_j | 1 \leq j \leq J\}$ be two sets of stimulus patterns, with $S \subseteq T$, implying $I \leq J$. For the special case $S \subset T$ one has $I < J$, i.e. there exist $J - I$ patterns T_j that are not identical to any element S_i in S. All patterns T_j are defined such that, under given experimental conditions, they are confusable with at least some $S_i \in S$.

Let n_{ij} be the number of times S_i has been judged to be identical with T_j; the n_{ij} can be summarised in a table of the form The table will be referred to as a confusion

Table 1: Data: form of a confusion matrix

	T_1	T_2	$\cdots$	T_J	$\sum$
S_1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
S_2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$			$\cdots$		$\vdots$
S_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
$\sum$	n_{+1}	n_{+2}	$\cdots$	n_{+J}	N_+

matrix. In the following, $I = J$, so there is no classification of the stimuli.

2.2 Models, in particular the GRT

One may distinguish different types of models of pattern identification:

1. **Template matching as holistic identification:** The pattern is detected as a whole. Special models focus on the notion of a matched filter and the maximisation of cross correlations, e.g. Hauske, Wolf and Lupp (1976), Burgess (1985) and Meinhardt and Mortensen (1998),

2. **Component Identification:** Here the stimulus is identified according to the classification/identification of its components. A special case are stimuli defined by a

single parameter. The components may again be identified via template matching.

3. **The role of stimulus environment:** A general question is wether lateral, irrelevant stimulus components play a role. No explicit models will be discussed, although the effect of such components can directly be investigated.

Neural activities: Let $\mathbf{N}_i$ be the neural activity generated by the presentation of the stimulus S_i. The following characterisation of $\mathbf{N}_i$ follows the notation of Ashby and Townsend (1986). It will be assumed that in a given trial, $\mathbf{N}_i$ can be represented by a random vectors $\vec{x}_i$. The components of these vectors reflect aspects of the neural activities, in particular aspects representing the components of the stimuli that are varied among the stimuli: for instance, the stimuli may be defined by two components A and B, e.g. $S_i = A_j B_k$, where A_j and B_k are the j-th and k-th level of A and B, respectively. Then $\vec{x}_i = (x_{i1}, x_{i2})'$, and x_{i1} and x_{i2} represent the perceptual effects generated by A_j and B_k. $f_i(x_1, x_2)$ is the distribution of the perceptual effects.

It is conceivable that even if the stimuli are defined as in what Ashby and Townsend call a Complete Identification Experiment (see, however, Karlec and Townsend (19xx)), i.e. if the stimuli are defined as A_1B_1, A_1B_2, A_2B_1 and A_2B_2, suggesting a componentwise identification to the subject, the subject my process the stimuli in a holistic manner, so that decisions are made with respect to a single random variable, $\vec{x}_i = x_i$. Template matching would be an example for this, and X_i could be some measure of overall similarity. This may be seen in context with respect to the separability versus integrality dichotomy (see below).

General Recognition Theory (GRT) The GRT may be taken as a general framework to discuss the questions arising from identification experiments. General notions here, special notions for 2-dimensional stimuli later in the corresponding section.

1. **Physical dimensions:** X, Y physical measures of components. Single component stimuli: $X_1 \ldots, X_4$ values of the parameter with respect to which the stimuli differ (example: spatial frequency parameter).

2. **Perceptual dimensions:** x_1, x_2, as above.

3. **Decisional dimensions:** $r(x_1, x_2)$ response function for the selection of a particular response on the basis of x_1, x_2; $R_r(x_1, x_2)$ the global response function for performance over the course of many trials (d', or χ_i^2).

4. **Response** $a_j b_k$ to stimulus $A_j B_k$.

5. **Density functions:** In the GRT, the densities will be assumed to be Gaussians.

3 Representation of stimuli and responses by scale values:

3.1 Holistic decisions or single component stimuli

Suppose the stimulus is evaluated with respect to a single random variable, so that $\vec{x}_i = X_i$. It will be assumed that the subject decides for the response T_j if $X_i \in A_j$, where $A_j \subset \mathbb{R}$ is a set characteristic for S_j. Let A be the union of the ranges A_j:

$$A = \bigcup_{\forall j} A_j, \quad A_j \bigcap A_{j'} = \emptyset, \text{ if } j \neq j'. \tag{1}$$

The aim is to represent the S_i by scale values α_i such that the α_i reflect the average perceptual effect of the corresponding S_i. Similarly, each response may be represented by a scale value β_j. The general definition of the scale values will be introduced first; the interpretation in particular of the relation between the α_i and the corresponding β_i will be discussed later.

Definition 3.1 *Let X be the random variable representing the activity (the perceptual effect) generated by a stimulus. Then*

$$\alpha_i = E(X|S_i) \tag{2}$$

To prepare the definition of the scale values for the responses, the definition of α_i will be re-written. If S_i is presented, one may simply write X_i instead of X under the condition of S_i, or $f_i(X)$, where f_i is the density of X given S_i was presented. So

$$\alpha_i = \int_A X f(X|S_i) dX = \int_A X f_i(X) dX. \tag{3}$$

Because of (1), one may write

$$\alpha_i = \sum_j \int_{A_j} f_i(X) dX. \tag{4}$$

Writing X_i instead of X in order to indicate that X is considered given S_i was presented one may define

$$E(X_i \in A_j) = \int_{A_j} f_i(X) dX, \tag{5}$$

so that one has

$$\alpha_i = \sum_{j=1}^{J} E(X_i \in A_j). \tag{6}$$

One may then introduce the definition

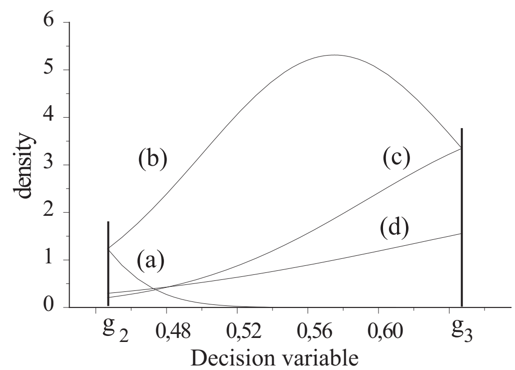
Definition 3.2 *The scale value β_j for the response T_j is given as the sum of expected values*

$$\beta_j = \sum_{i=1}^{I} E(X_i \in A_j) \tag{7}$$

Comments:

1. **Relation between definitions:** Formally, the definitions of α_i and β_j differ with respect to the summation: α_i is the sum of expected values $E(X_i \in A_j)$ over the responses T_j, while β_j is the sum of the expected values $E(X_i \in A_j)$ over the different stimuli S_i.

Figure 1: Densities over the interval $[g_2, g_2)$; Gaussian distributions (IV), see Fig. 5



2. **Relation between scale values:** Let us briefly consider the relation between the α_i and β_j. A sufficient condition for $\alpha_i \to E(X_i \in A_i)$ and consequently $\beta_i \to \alpha_i$ is $E(X_i \in A_j) \to 0$ for $i \neq j$. This condition may be relaxed; one will observe $\alpha_i \approx \beta_i$ if $n_{ii} > n_{ij}$, $i \neq j$ and if the n_{ij} decrease with decreasing similarity of S_i and S_j, i.e. with increasing difference of the parameter defining the difference between S_i and S_j. See also role of decision boundaries.

3. **Choice of variables:** So far, no further specification of the random variables X_i has been given, other than that they represent the "perceptual effects" or the "neural activities" generated by the stimuli. For certain stimuli, e.g. ones that are composed of line elements (Townsend and Nosofsky ?), one may simply assume that X represents the activity of a line detector. For other stimuli, there is no such straightforward interpretation, for instance when stimuli are defined by the superposition of two "discs" of different radii, of by the superposition of Gabor patches defined by different spatial frequency parameters. One may assume that the subject searches for aspects of the neural activity that allow to discriminate best among the stimuli. This assumption underlies the approach adopted to estimate the α_i and β_j, described in the following section. It turns out that to estimate these values, the A_j do not have to be estimated; however, since the values of β_j depend on their definition, one may refer to the A_j when it comes to interpretate the estiamted β_j-values.

3.2 The estimation of scale values

The scale values α_i are defined as expected values, and so one may try to estimate them by the arithmetic means of the X_i. Unfortunately, the X_i are not observable. However, given the assumption 3 above, one may arrive at estimates of these arithmetic means. To this end, the assumption has to be cast into a form that allows such an estimation. One possibility of achieving good discriminations is to choose the X_i in such a way that differences among the α_i, i.e. the variance of the α_i ("between" variance), is maximal relative to the variance of the X_i ("within" variance). As in discriminant analysis, one may therefore estimate the α_i by (i) decomposing the total sum of squares SS_{tot} of the X_i into components SS_b representing the "between", and SS_w, representing the "within" variances, and (ii) find estimates a_i of the α_i by maximising either the quotient SS_b/SS_w, or the quotient $\rho^2 = SS_b/SS_{tot}$. The latter quotient, ρ^2, is the correlation quotient introduced by Guttman (1941). The estimates of the scale values turn out to be equivalent to those arrived at by employing Correspondence Analysis; however, in contrast to the maximisation of ρ^2 the derivation of Correspondence Analysis does not refer explicitly to the X_i. Therefore, the maximisation of ρ^2 will be briefly indicated; the approach has been exposed, in a different context, by Nishisato (1980). Correspondence Analysis (CA) will be introduced more explicitly since it arrives at scale values for α_i and β_j and at the same time relates them to χ^2-components, thus facilitating the interpretation of the results, and will therefore actually be employed to analyse the data.

3.2.1 Estimation I: maximising the correlation ratio

To begin with, suppose that each of the stimulus patterns S_i is presented n times. Let I_{ij} be the set of integers indexing the trials when the response to S_i was T_j, so if $k \in I_{ij}$, then in the k-th trial the response T_j was given to S_i; the neural activity generated by the presentation of S_i can then be characterised by x_{ik}, and of course $k \leq n$. Let $n_{ij} = |I_{ij}|$, i.e. the response T_j was given altogether n_{ij} times when S_i was presented.

Let

$$\bar{x}_{ij} = \frac{1}{n_{ij}} \sum_{k \in I_{ij}} x_{ik}, \tag{8}$$

$$\bar{x}_{i+} = \frac{1}{n_{i+}} \sum_{j=1}^{J} n_{ij}\bar{x}_{ij}, \quad n_{i+} = \sum_j n_{ij} \tag{9}$$

$$\bar{x}_{+j} = \frac{1}{n_{+j}} \sum_{i=1}^{I} n_{ij}\bar{x}_{ij}, \quad n_{+j} = \sum_i n_{ij} \tag{10}$$

The means $\bar{x}_{i+}$ and $\bar{x}_{j+}$ could be taken as estimates of α_i and β_j if the x_{ik} were observable. Since the x_{ik} cannot directly be observed, some restrictions have to be introduced. Let

$$\bar{a}_i = \frac{1}{J} \sum_{j=1}^{J} \bar{x}_{ij}, \tag{11}$$

and suppose x_{ik} in (8) is replaced by $\bar{a}_i$, i.e. by their mean value as a least square approximation, taken with respect to the response alternatives T_j. Then $\bar{x}_{ij}$ is approximated by

$$\hat{\bar{x}}_{ij} = \frac{1}{n_{ij}} \sum_{k \in I_{ij}} \bar{a}_i = \frac{n_{ij}\bar{a}_i}{n_{ij}} = \bar{a}_i. \tag{12}$$

Replacing $\bar{x}_{ij}$ by the approximation $\hat{\bar{x}}_{ij}$ in (9) and (10) gives

$$\hat{\bar{x}}_{i+} = \frac{1}{n_{i+}} \sum_{j=1}^{J} n_{ij}\hat{\bar{x}}_{ij} = \frac{1}{n_{i+}} \sum_{j=1}^{J} n_{ij}\bar{a}_i = \bar{a}_i \tag{13}$$

$$\hat{\bar{x}}_{+j} = \frac{1}{n_{+j}} \sum_{i=1}^{I} n_{ij}\hat{\bar{x}}_{ij} = \frac{1}{n_{+j}} \sum_{i=1}^{I} n_{ij}\bar{a}_i = \bar{b}_j, \tag{14}$$

i.e. $\bar{a}_i$ is an estimate for $\bar{x}_{i+}$ and therefore for α_i, and $\bar{b}_j$ is an estimator for $\bar{x}_{+j}$ and therefore for β_j.

To arrive at an explicit expression for the estimation of the $\bar{a}_i$ and $\bar{b}_j$ the substitution of $\bar{a}_i$ for x_{ik} introduced above will be formulated formally in terms of a mapping s: $x_{ik} \mapsto \bar{a}_i$, i.e. s is a function of x_{ik} such that $s(x_{ik}) = \bar{a}_i$ for all j if $k \in I_{ij}$. Since x_{ik} is a random variable, $s(x_{ik})$ will be a random variable as well, so it makes sense to define means and variances for s. Writing s_{ik} instead of $s(x_{ik})$ for short, one has

$$\bar{s}_j = \frac{1}{n_{i+}} \sum_{i=1}^{I} \sum_{k \in I_{ij}} s_{ik} = \frac{1}{n_{i+}} \sum_{i=1}^{I} n_{ij}\bar{a}_i. \tag{15}$$

$\bar{s}_j$ is the mean of the s_{ik} over all i for a given T_j. Because of (14),

$$\bar{s}_j = \bar{b}_j. \tag{16}$$

The overall mean of the s_{ik} is given by $\bar{s} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k \in I_{ij}} s_{ik}/N$. The total variance of the s_{ik}

$$SS_{tot} = \sum_{j=1}^{J} \sum_{k=1}^{n_{+j}} (s_{ik} - \bar{s})^2. \tag{17}$$

As is well known from the analysis of variance, SS_{tot} can be decomposed into a "within"-component and a "between"-component:

$$SS_{tot} = \sum_{j=1}^{J} \sum_{k=1}^{n_{+j}} (s_{ik} - \bar{s}_j + \bar{s}_j - \bar{s})^2$$

$$= \sum_{j=1}^{J} \sum_{k=1}^{n_{+j}} (s_{ik} - \bar{s}_j)^2 + \sum_{j=1}^{J} n_{i+}(\bar{s}_j - \bar{s})^2. \tag{18}$$

$$= = SS_w + SS_b. \tag{19}$$

The first sum on the right of (18) is the "within"-component (SS_w), and the second is the "between"-component SS_b. SS_{tot} depends on the scale values $\bar{a}_i$, as is clear from the definition $s_{ik} = \bar{a}_i$ and from (15). The $\bar{a}_i$ may now be estimated by maximising SS_b relative to the value of SS_{tot}, which implies that SS_w will be minimised relative to SS_{tot}.

Without loss of generality one may put $\bar{s} = 0$, so that

$$SS_{tot} = \sum_{j=1}^{J} \sum_{k=1}^{n_{+j}} s_{ik}^2, \quad SS_{bt} = \sum_{j=1}^{J} n_{i+} \bar{s}_j^2. \tag{20}$$

Let $K = (n_{ij})$ be the matrix of confusion frequencies (as given in Table 1). Further, let

$$D_{rs} = \text{diag}(n_{1+}, n_{2+}, \ldots, n_{I+}), \quad D_{cs} = \text{diag}(n_{+1}, n_{+2}, \ldots, n_{+J}). \tag{21}$$

be the diagonal matrices of the row- and column sums, respectively, of K. Let $\bar{\mathbf{a}} = (\bar{a}_1, \ldots, \bar{a}_I)'$ and $\bar{\mathbf{b}} = (\bar{b}_1, \ldots, \bar{b}_J)'$. In terms of these vectors SS_{tot} can be expressed in the form

$$SS_{tot} = \sum_{i=1}^{I} \sum_{k \in I_{ij}} s_{ik}^2 = \sum_{i=1}^{I} n_{ij} \bar{a}_i^2 = \bar{\mathbf{a}}' D_{rs} \bar{\mathbf{a}}. \tag{22}$$

and making use of (15), one gets

$$SS_b = \sum_i \frac{1}{n_{i+}} (\sum_j n_{ij} \bar{a}_i)^2 = \bar{\mathbf{a}}' K' D_{rs}^{-1} K \bar{\mathbf{a}}. \tag{23}$$

Guttman's (1941) correlation ratio is then given by

$$\rho^2 = \frac{SS_b}{SS_{tot}} = \frac{\bar{\mathbf{a}}' K' D_{rs}^{-1} K \bar{\mathbf{a}}}{\bar{\mathbf{a}}' D_{rs} \bar{\mathbf{a}}}. \tag{24}$$

To illustrate the meaning of ρ^2, consider the following cases:

1. $\rho^2 \to 0$ if $SS_b \to 0$; so a small value of ρ^2 means that it is difficult for the subject to discriminate between patterns, since $SS_b \to 0$ means that the differences $\bar{s}_i - \bar{s}$ are small compared to the $s_{ik} - \bar{s}_i$, meaning that the S_i are perceived as being very similar. The similarity of the S_i will therefore be reflected by scale values for the S_i that are close together.

2. $\rho^2 \to 1$ for $SS_b \to SS_{tot}$, meaning that the $s_{ik} - \bar{s}_i$ are small compared to the $\bar{s}_i - \bar{s}$, so the S_i can well be discriminated, and the a_i are well separated.

The *estimated* values for the $\bar{a}_i$ and $\bar{b}_j$ may be conceived as components of the vectors $\mathbf{a} = (a_1, \ldots, a_I)'$ and $\mathbf{b} = (b_1, \ldots, b_J)'$. One finds then

Theorem 3.1 *ρ^2 is maximal relative to the value of SS_{tot} if $\bar{\mathbf{a}} = \mathbf{a}$, satisfying the equation*

$$D_{rs}^{-1} K D_{cs}^{-1} K' \mathbf{a} = \mu \mathbf{a}, \quad \mu = \rho_{max}^2. \tag{25}$$

i.e. $\bar{\mathbf{a}}$ is estimated by the eigenvector $\mathbf{a}$ of the matrix $D_{rs}^{-1} K D_{cs}^{-1} K'$, and the corresponding eigenvalue μ equals ρ^2. $\mathbf{b} = (b_1, \ldots, b_J)'$ is given by (16), i.e. by $b_j = \sum_{i=1}^{I} n_{ij} a_i / n_{i+}$, $j = 1, \ldots, J$.

Proof: see the Appendix, section 7.2. □

Remark: There may exist more than a single eigenvector **a** satisfying (25). The additional eigenvectors may reflect the fact that the subject makes use of more than a single aspect of the neural activities representing the patterns. This possibility will be discussed below in context with solutions arrived at by employing Correspondence Analysis in order to find the estimators **a** and **b**.

Interpretation of the scale values: The a_i were introduced first as estimates $\hat{\bar{x}}_{i+}$ for $\bar{x}_{ij}$, see (13), and thus represent mean values, without reference to the variances of the random variables representing the neural representation. The estimates (25), on the other hand, are defined relative to the variation as defined by SS_{tot}. The difference $a_i - a_{i'}$ will be "small" when the difference between the corresponding means is small relative to the variance of the distributions, and "large" when the difference between the means is large relative to their variances. It will be demonstrated in section **??** that unequal variances may lead to a lack of proportionality of the a_i to the $\bar{x}_{i+}$.

3.2.2 Estimation II: Correspondence Analysis

Let again $K = (n_{ij})$, $i, j = 1, \cdots, m$ be a confusion matrix; n_{ij} is the frequency with which the stimulus pattern S_i is confused with stimulus S_j (indicated by giving the response R_j). Let n_{i+} and n_{+j} be defined as before, and let $N = \sum_{i,j} n_{ij}$. $n_{i+}n_{+j}/N$ is the expected number of confusions of S_i with S_j, provided the subject judges randomly. Let

$$x_{ij} = \frac{n_{ij} - n_{i+}n_{+j}/N}{\sqrt{n_{i+}n_{+j}/N}}. \tag{26}$$

x_{ij} are called residuals, or or weighted residuals. Obviously, the residuals x_{ij} reflect the dependencies among stimuli and responses. Certainly,

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2, \tag{27}$$

Let $p_{ij} = x_{ij}/N$, $r_i = n_{i+}/N$, $c_j = n_{+j}/N$, and

$$t_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} = x_{ij}/\sqrt{N}, \tag{28}$$

then

$$\frac{\chi^2}{N} = \sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}^2. \tag{29}$$

χ^2/N is called the inertia of the confusion matrix, denoted by $In(K)$, i.e. $In(K) = \chi^2/N$. $In(K)$ (and, analogously, χ^2) can be written as the sum of components characterising either the patterns S_i or T_j.

$$In_i(K) = \chi_i^2/N = \sum_{j=1}^{J} t_{ij}^2, \quad In_j(K) = \chi_j^2/N = \sum_{i=1}^{I} t_{ij}^2. \tag{30}$$

Certainly, $In(K) = \sum_j \chi_j^2/N = \sum_i \chi_i^2/N$. The scaling provided by CA can be interpreted with respect to inertia (equivalently: χ^2-) components.

In the context of CA, the inertia instead of the χ^2 will be considered in the following, in particular since statistical packages refer to the inertia rather that to the χ^2 (e.g. STATISTICA).

Spatial representations and inertia decompositions: Let $r = (r_1, \ldots, r_I)'$ and $D_r = \text{diag}(r_1, \ldots, r_I)$ the diagonal matrix having r_i in its diagonal, with $r_i = n_{i+}/N$, $n_{i+} = \sum_j n_{ij}$. Analogously, $c = (c_1, \ldots, c_J)'$, $c_j = \sum_i n_{ij}/N$ and $D_c = \text{diag}(c_1, \ldots, c_J)$. Let $T = (t_{ij})$ be the matrix of t_{ij}-values; note that according to (28),

$$T = D_r^{-1/2}(P - rc')D_c^{-1/2}. \tag{31}$$

As is well known from linear algebra, the column or row vectors of T can be represented as linear combinations of some orthogonal basis vectors of the I- or J-dimensional vector space, respectively. The Singular Value Decomposition (SVD) provides such basis vectors, which turn out to be related to the vectors **a** and **b** of scale values as characerised in Theorem 3.1.

Theorem 3.2 *The SVD of T is given by*

$$T = U\Lambda^{1/2}V', \tag{32}$$

with U the matrix of normalised eigenvectors of TT', V the matrix of normalised eigenvectors of $T'T$ and $\Lambda^{1/2}$ the diagonal matrix diag$(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_s})$, λ_k, $1 \leq k \leq s \leq \min(I, J)$ the nonzero eigenvalues of TT' and $T'T$.

Proof: see Appendix, section 7.3.

Comments: Let u_{ik} be the i-th component of the k-th column vector of U, and let v_{jk} be the j-th component of the k-th column vector of V. The u_{ik} or $u_{ik}\sqrt{\lambda_k}$ may be taken as coordinates of a point representing the pattern S_i in an s-dimensional space $\mathbb{R}^s$ whose axes (possibly) represent some aspects of the neural representations $\mathbf{N}_i$ of S_i. Analogously, the v_{jk} or $v_{jk}\sqrt{\lambda_k}$ may be taken as coordinates of a point representing T_j in $\mathbb{R}^s$. The simultaneous presentation of S_i- and T_j-points is known as biplot, which reflects structural relations between the S_i and the T_j. The interpretation of the biplot will be based upon the distances between the S_i-points on the one hand and the T_j-points on the other. However, the choice of the coordinates determines a particular metric, and the interpretation will therefore be influenced by the particular metric; an appropriate re-scaling of the coordinates may thus lead to better interpretations. □

Definition of scale values: The coordinates of the S_i and T_j are defined as follows:

$$F = D_r^{-1/2}U\Lambda^{1/2} \tag{33}$$
$$G = D_c^{-1/2}V\Lambda^{1/2}, \tag{34}$$

(Greenacre, 1984, p. 89). The element f_{ik} in the i-th row and k-th column of F is the coordinate of a point representing S_i on the k-th axis. Analogously, the element g_{jk} of G

is the coordinate of the point representing T_j on the same axis. The coordinates F and G turn out to be equivalent to those found by maximising the correlation ratio ρ^2. Note that between the f_{ik} and the u_{ik} the following relations exist:

$$f_{ik} = u_{ik}\frac{\sqrt{\lambda_k}}{\sqrt{r_i}}, \qquad u_{ik} = f_{ik}\frac{\sqrt{r_i}}{\sqrt{\lambda_k}}, \tag{35}$$

$$g_{jk} = v_{jk}\frac{\sqrt{\lambda_k}}{\sqrt{cj}}, \qquad v_{jk} = g_{jk}\frac{\sqrt{c_j}}{\sqrt{\lambda_k}}. \tag{36}$$

Relation to the square ρ^2 of the correlation ratio: The following theorem shows that Correspondence Analysis is equivalent to maximising ρ^2, as discussed in section 3.2.1:

Theorem 3.3 *The F and G, as defined in (33) and (34), satisfy the following eigenvector equations:*

$$(D_r^{-1}PD_c^{-1}P')F = \Lambda F \tag{37}$$

$$(D_c^{-1}P'D_r^{-1}P)G = \Lambda G. \tag{38}$$

Proof: The proof may be found in section 7.4 of the Appendix. □

Remark: Starting from the SVD (32) makes it difficult to see how the scale values, given in the matrices F and G, relate to the mean values of the underlying distributions of the criterion variable η. The relations become obvious noting that the equations (37) and (38) correspond to (25) of Theorem 3.1. It follows that the scaling provided by CA is equivalent to the obtained when the correlation is maximised. Note that the equations (37) and (38) result from a rescaling of the eigenvectors in U and V, as given by the SVD (32), which is basically a principal component analysis (PCA) of the matrix T of residuals. The equivalence of these equations with (25) thus establishes a relation between PCA and Discriminant Analysis.

Relation between inertia components and spatial representation: To see how the spatial representation of the S_i and the T_j relates to χ^2 or inertia components let us first look at the t_{ij}: According to (32), the element t_{ij} of T is given by $t_{ij} = \sum_k u_{ik}v_{jk}\sqrt{\lambda_k}$. Then $t_{ij}^2 = \sum_k u_{ik}^2 v_{jk}^2 \lambda_k + 2\sum_{k\neq k'} u_{ik}v_{jk}u_{ik'}v_{jk'}\lambda_k$, and

$$\chi_i^2/N = \sum_{k=1}^{s} u_{ik}^2\lambda_k = r_i\sum_{k=1}^{s} f_{ik}^2, \tag{39}$$

because of (35) and because $\sum_j v_{jk}^2 = 1$ (the vectors in V are normalised), and $\sum_j v_{jk}v_{jk'} = 0$ (the vectors in V are orthogonal). Analogously, one finds

$$\chi_j^2/N = \sum_{k=1}^{s} v_{jk}^2\lambda_k = c_j\sum_{k=1}^{s} g_{jk}^2. \tag{40}$$

$In_i(K) = \chi_i^2/N$ is the inertia component due to the pattern S_i, and $In_j(K) = \chi_j^2/N$ is the inertia component due to the pattern T_j.

As an immediate consequence of (39) and (40) one has

$$\chi^2/N = \sum_{i=1}^{I} \chi_i^2/N = \sum_{k=1}^{s} \lambda_k \sum_{i=1}^{I} u_{ik}^2 = \sum_{k=1}^{s} \lambda_k, \tag{41}$$

since the eigenvectors are normalised, i.e. $\sum_i u_{ik}^2 = 1$, so that

$$\pi_k = \frac{\lambda_k}{\chi^2/N} \tag{42}$$

is the proportion of the total inertia due to the k-th latent dimension; this corresponds to the role of the λ_k in in an ordinary PCA, where the eigenvalues λ_k reflect the proportion of variance explained by the corresponding dimension.

To see how the spatial representations of the S_i and T_j in the biplot relate to the inertia components, let us consider the point of S_i, say. The coordinates of this point are given by $f_{i1}, \ldots, f_{is}$. The square of the Eucledian distance from the origin to this point is given by

$$d_i^2 = \sum_{k=1}^{s} f_{ik}^2, \tag{43}$$

so that, from (39),

$$\chi_i^2/N = r_i d_i^2, \tag{44}$$

and, analogously,

$$\chi_j^2/N = c_j d_j^2, \tag{45}$$

where d_j is the distance from the origin of the point representing T_j. The distance between two points, one for S_i and the other for $S_{i'}$, is given by

$$d_{ii'}^2 = \sum_{k=1}^{s} (f_{ik} - f_{i'k})^2; \tag{46}$$

there is, however, no simple relation to the inertia of K. The distance between two points representing T_j and $T_{j'}$ is defined analogously.

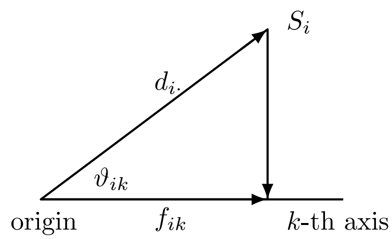
Scale values and inertia components: Consider the vector from the origin of the coordinate system to the point representing S_i. The vector has length d_i, and according to (44), $r_i d_i^2 = \chi_i^2/N$. The projection of d_i on the k-th axis equals, by definition, f_{ik}, see Fig. 2, and one has

$$d_i. \cos\vartheta_{ik} = f_{ik}. \tag{47}$$

It follows that then $\cos^2\vartheta_{ik} = f_{ik}^2/d_i^2$, and multiplying and dividing the right hand side by r_i yields

$$cos^2\vartheta_{ik} = \frac{r_i f_{ik}^2}{r_i d_i^2} = \frac{r_i f_{ik}^2}{In_i(K)} \tag{48}$$

Figure 2: Projection of S_i on k-th axis



But $r_i f_{ik}^2$ is the contribution of the k-th dimension to $In_i(K)$, the component of the inertia (χ^2) due to S_i. Therefore, $cos^2\vartheta_{ik}$ measures the contribution of the k-th dimension to $In_i(K)$, i.e. to the inertia or χ^2-component generated by S_i.

The interpretation of $\cos^2\vartheta_{jk} = c_j g_{jk}^2/In_j(K)$ is analogous: this value reflects the contribution of the k-th dimension to the inertia or χ^2-component due to T_j.

χ^2-distances and differences between scale values: The distance between two points respresenting two stimuli is meant to represent the subject's ability to discriminate between these two stimuli. It is plausible to assume that the discriminability does not only depend upon the difference between the expected values of the random variables representing the sensory activities resulting from the stimulus presentations, but also upon their variances. The variances determine the distribution of frequencies in the rows representing the stimuli. The more similar the frequency distributions of two rows, the more difficult it will be to discriminate between the corresponding stimuli. The relation between distances and variances turns out to be of importance for the interpretation of the results of a correspondence analysis. This leads to the following definition:

Definition 3.3 *Consider the rows corresponding to the stimuli S_i and $S_{i'}$. Let*

$$\delta_{ii'}^2 = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2 = \sum_{j=1}^{J} \frac{1}{n_{\cdot j}}\left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}}\right)^2. \tag{49}$$

$\delta_{ii'}^2$ *is called the* χ^2-distance *between the stimuli S_i und $S_{i'}$.*

$$\delta_{jj'}^2 = \sum_{i=1}^{I} \frac{1}{r_i}\left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}}\right)^2 = \sum_{i=1}^{I} \frac{1}{n_{i\cdot}}\left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}}\right)^2 \tag{50}$$

is called the χ^2-distance *between the responses (column categories) R_j und $R_{j'}$.*

$\delta_{ii'}^2$ will be small if $n_{ij}/n_{i\cdot} \approx n_{i'j}/n_{i'\cdot}$ for all j; since all stimuli occur with equal frequency, $n_{i\cdot} = n_{i'\cdot}$, so $\delta_{ii'}^2$ will be small when $n_{ij} \approx n_{i'j}$ for all j, i.e. when the frequency distributions of row i and row i' are sufficiently similar. In this case the differences between the two distributions do not contribute much to the χ^2-value of the confusion matrix. The following theorem establishes the relation between a χ^2-distance and the corresponding Eucledian distance:

Theorem 3.4 *The relation between the Euclidean distance $d_{ii'}$ and the corresponding χ^2-distance $\delta_{ii'}$ is given by*

$$d_{ii'} = \delta_{ii'} \tag{51}$$

Analogously one has for the responses

$$d_{jj'} = \delta_{jj'}. \tag{52}$$

The proof is given in section 7.5 of the Appendix. □

So the value of $d_{ii'}$ reflects the amount as to which the differences between the confusion frequencies for stimulus S_i and those for $S_{i'}$ contribute to the overall-χ^2 of the confusion matrix. $d_{ii'}$ is thus a measure for the ability of the subject to discriminate between S_i and $S_{i'}$. If only a single dimension is relevant to explain the confusion, i.e. if the subject refers to a single decision variable, then the differences between two scale values - one for S_i and the other for $S_{i'}$ - reflects the ability to discriminate between S_i and $S_{i'}$, and since the scale values are meant to be proportional to the expected values of the decision variable, given a particular stimulus was shown, $d_{ii'}$ is thus a measure of sensitivity equivalent to d'; see section **??** for a more explicit discussion of this point.

Relation between the S_i- and T_j-points. The relation between the points for the S_i and those for the T_j are, of course, also of interest. However, this relation cannot be discussed in terms of distances, since the distance between the point representing S_i and the point representing T_j is not defined. Instead, the relation between the coordinates F and G is given by the following theorem:

Theorem 3.5 *The relation between the coordinates F of the S_i and G of the T_j is given by the equations*

$$\begin{aligned} F &= D_r^{-1} P G \Lambda^{-1/2}, & (53) \\ G &= D_c^{-1} P' F \Lambda^{-1/2} & (54) \end{aligned}$$

Proof: The proof may be found in section 7.6 of the Appendix. □

To illustrate, let us consider the relation (54) between scale values of $T_1, \dots, T_J$ and and $S_1, \dots, S_I$ on the k-th dimension; one gets the equations

$$\begin{aligned} g_{1k} &= \frac{p_{11}}{c_1}\frac{f_{1k}}{\sqrt{\lambda_k}} + \frac{p_{21}}{c_1}\frac{f_{2k}}{\sqrt{\lambda_k}} + \cdots + \frac{p_{I1}}{c_1}\frac{f_{Ik}}{\sqrt{\lambda_k}} \\ g_{2k} &= \frac{p_{12}}{c_2}\frac{f_{1k}}{\sqrt{\lambda_k}} + \frac{p_{22}}{c_2}\frac{f_{2k}}{\sqrt{\lambda_k}} + \cdots + \frac{p_{I2}}{c_2}\frac{f_{Ik}}{\sqrt{\lambda_k}} \\ &\vdots \\ g_{Jk} &= \frac{p_{1J}}{c_J}\frac{f_{1k}}{\sqrt{\lambda_k}} + \frac{p_{2J}}{c_J}\frac{f_{k}}{\sqrt{\lambda_k}} + \cdots + \frac{p_{IJ}}{c_J}\frac{f_{k}}{\sqrt{\lambda_k}}. \end{aligned} \tag{55}$$

So the scale value g_{jk} of T_j on the k-th dimension may be seen as a sum of the "weighted" f_{ik}. One may ask when the point for T_j is near the point for S_i. This will be the case when $f_{ik} \approx g_{jk}$ for all k, that is if the f_{ik} and the g_{jk} have similar values for all k. A sufficient condition for f_{ik} and g_{jk} to assume similar values is that p_{ij} is large compared to the values of $p_{ij'}$, $j \neq j'$. So if S_i is in particular identified with T_j then the points for S_i and T_j will be close together.

4 The 1-dimensional case

4.1 Numerical evaluations

The definition of the scale values α_i and β_j was made without reference to a specific type of density for the random variables X_i, $i = 1, \ldots, I$. In order to find out to which extent the CA of a confusion matrix yields estimates of the α_i and β_j, specific assumptions have to made. With reference to the General Recognition Model of Ashby and Townsend (1986) it will now be assumed that the densities are Gaussian. However, the Gaussian assumption does not appear to be necessary; it will be demonstrated that densities differing considerably from the Gaussian in shape also yield, under certain mild side conditions, scale values that are just linear transformations of the expected values.

Gaussian densities: The mean and the variance of a distribution may depend upon the individual stimulus. So one may have equal and unequal spacing of expected values, and equal or unequal variances, so there are four different classes of density configurations to be considered. The following table contains a characterisation of the considered cased with respect to the expected values and the variances. The labels A to B refer to a

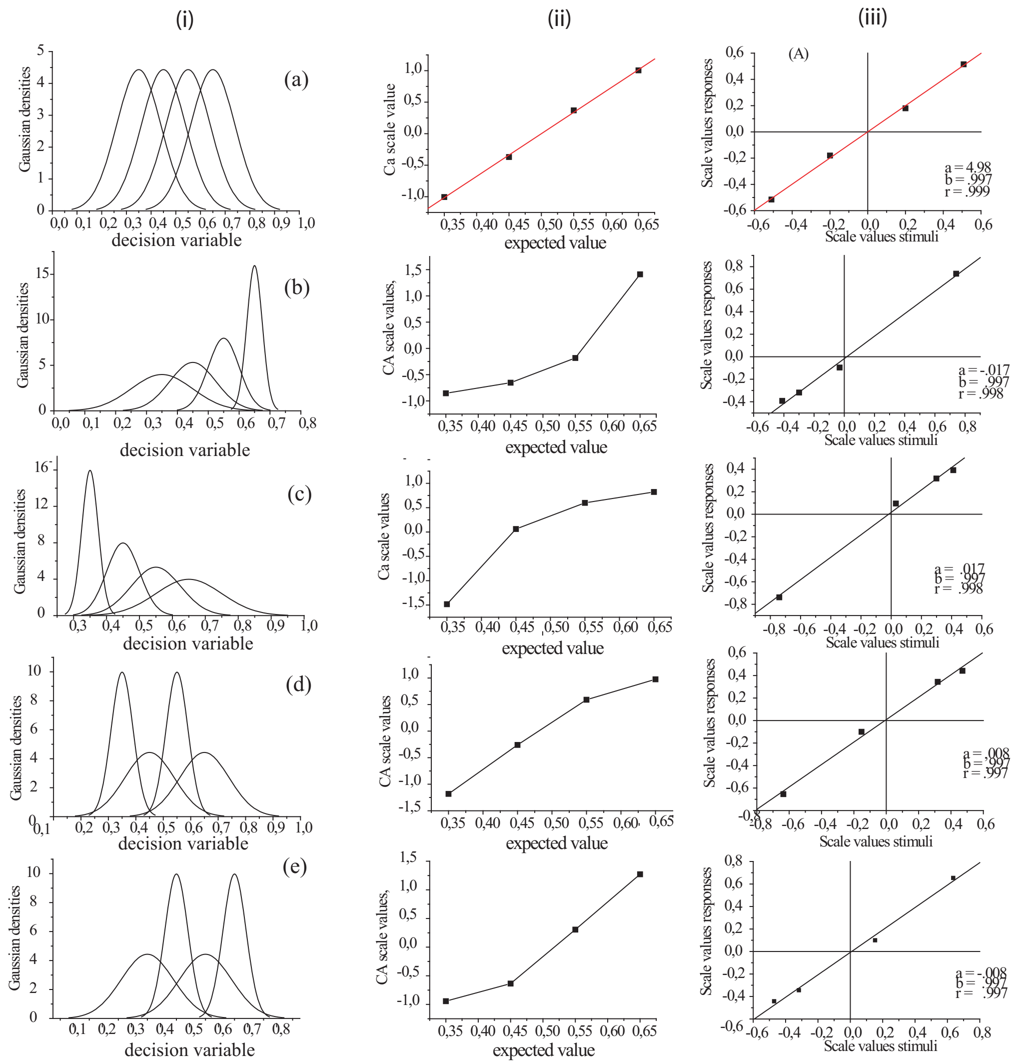
Table 2: Means and variances for density configurations

Label	μ_i (I)	standard deviations σ					μ_i (II)
A	.350	.090	.100	.025	.040	.090	.350
B	.450	.090	.075	.050	.090	.040	.375
C	.550	.090	.050	.075	.040	.090	.475
D	.650	.090	.025	.100	.090	.040	.750

configuration of densities as given in Fig. 3 for equally spaced mean values (μ_i, (I)), and Fig. 4 for not equally spaced mean values, (μ_i, (II)). The entries between μ_i, (I) and μ_i, (II) provide the standard deviations for the corresponding configuration of densities. Figures 3 and 4 show the configurations A to B together with the corresponding plot of scale values on the first dimension and expected values μ_i.

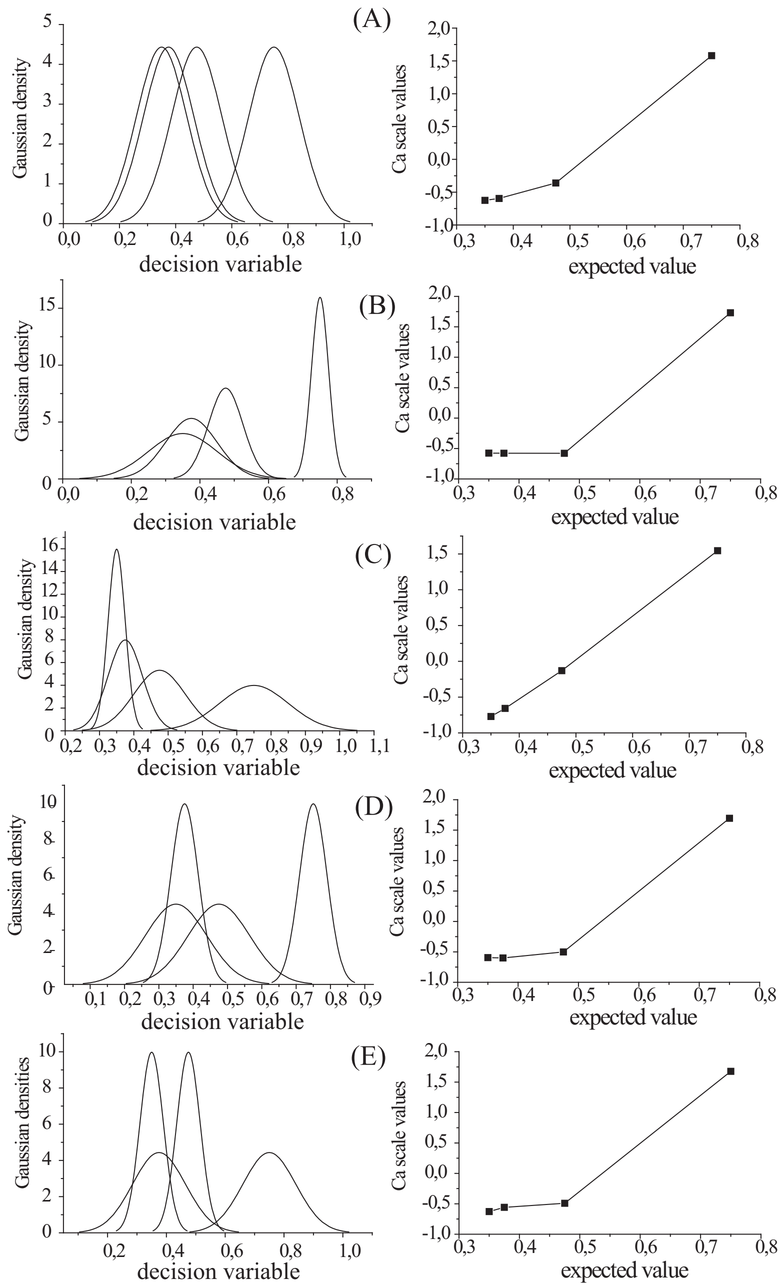
The cases (D) and (E) represent two configurations where the standard deviations do not increase or decrease with μ_i. The nonlinearity of the relation between scale values and expected values comes in only when a density with a (sufficiently) smaller standard deviation is followed by one with larger one, i.e. (D), or when a density with a larger standard deviation is followed by one with a smaller standard deviation, i.e. (E). If the differences between the standard deviations are not sufficiently large, the configurations (D) and (E) may well yield linear relations between the scale values and the μ_i.

Figure 3: Gaussian densities, equal spacing of expected values ((a1) to (e1)), and corresponding plots of scale values for stimuli versus expected values ((a2) to (e2)).



The bottom part of Table 2 gives the parameters for not equally spaced expected values. In Fig. 4. The case (a1)(unequally spaced expected values, equal standard deviations) is of interest because it shows that equal variances do not yet imply that the scale values are a linear function of the expected values. The difference between two scale values depends upon the difference of the corresponding expected values as well as upon the amount of overlap of the densities, i.e. upon their standard deviations. The relation between expected values and scale values is positively accelerated, meaning that the difference between two scale values is a nonlinear function of the corresponding

Figure 4: Gaussian densities, unequal spacing of expected values ((a1) to (e1)), and corresponding plots of scale values versus expected values ((a2) to (e2))



difference of the expected values. This comes about even more pronounced when the standard deviations decrease with increasing difference of the expected values, see the case (b2). Conversely, if the standard deviations increase with increasing difference of the expected values, the acceleration of the scale values may be reduced to a mere linearity, as demonstrated in (c2). The interpretation of (d1), (d2) and (e1), (e2) is obvious.

To summarise, a linear relationship between expected values and scale values does not necessarily imply that the variances of the underlying distributions are about equal; provided the variances do not co-vary with the expected values and do not differ too much from each other, the relation may turn out to be linear, although the values of r^2 as given in Fig. 6 may not turn out as high; still, the cases (d2) and (e2) in Fig. 4 yield a value of $r^2 = .971$ each. In general, lower values of r^2 in case of an overall linear appearing relationship will indicate variances that do not vary systematically with the expected values.

Apart from the means and variances of the underlying distributions the decision boundaries have to be chosen. It has been assumed here that the subject tries to decide optimally, meaning that the boundaries are chosen appropriately, i.e. such as to minimise the number of errors. The question is, how the subject proceeds to find them. This question is not explicitly dealt with in this paper. To compute the confusion frequencies, the boundaries were assumed to be given by the points x_{oi} at which two neighbouring densities have identical values, i.e. $f_i(x_{0i}) = f_{i+1}(x_{0i})$, $i = 1, \dots, I-1$, except for the densities in Fig. 5, where the boundaries were assumed to be given by $(\mu_i + \mu_{i+1})/2$. Generally, these two types of boundary definitions yield confusion frequencies differ only to a negligible amount, and the coordinates for the responses T_j are almost identical to those of the stimuli. In other words, there is no bias in the decisions. This corresponds to the findings for simple patterns, reported in section 4.2.1. For more complex patterns, a bias may exist.

The beta case etc In the Gaussian case one has

$$f_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right). \tag{56}$$

Since neural activity can vary only within finite limits it will be of interest to consider an alternative to this assumption. Here, the Beta-distribution will be investigated, normalised such that the random variables all vary on the interval $[0, 1]$,

$$f_i(x) = \frac{1}{B(a_i, b_i)} x^{a_i-1}(1-x)^{b_i-1}, \quad 0 \le x \le 1. \tag{57}$$

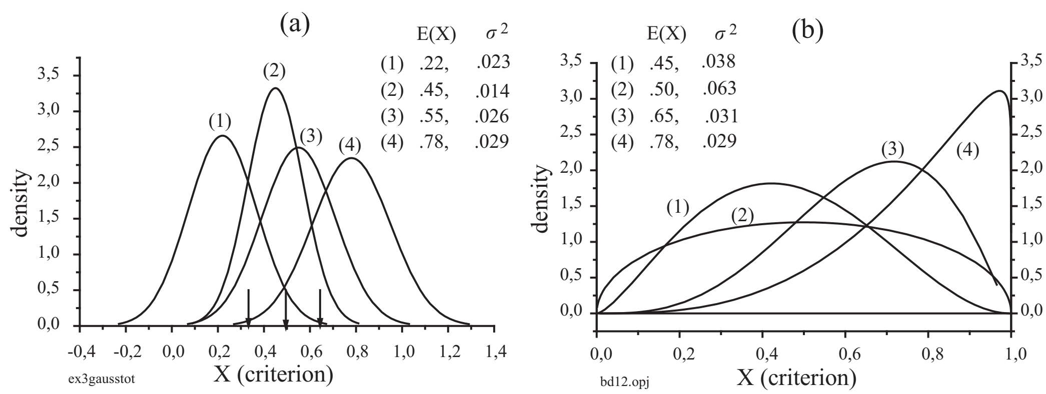
The shape of these densities depends upon the values of its parameters; β-densities thus allow to test the robustness of the estimates of α_i and β_j. The expected value and the variance of a β-density are given by the following expressions:

$$\mu_i = \frac{a_i}{a_i + b_i} \tag{58}$$

$$\sigma_i^2 = \frac{a_i b_i}{(a_i + b_i)^2(a_i + b_i + 1)}. \tag{59}$$

The parameters of the Gaussians were chosen from a range of values corresponding to that of the Beta-densities. The normalisation implies the need to choose the parameters

Figure 5: Distributions of criterion variable, (a) Gaussian distributions, (b) Beta-distributions (I)



from a certain range. The means μ_i were chosen not to increase in an equally spaced way, and the standard deviations σ_i were chosen such that they do not increase with the μ_i. The parameters a_i and b_i of the Beta-densities can be computed from (58) and (59),

The expected values μ_i increased and the variances varied moderately, but irregularly with μ_i, see Fig. 5. Confusion probabilities were computed as areas between decision boundaries under the corresponding densities; multiplied with some sufficiently large integer (the number of presentations of each stimulus) and rounded the resulting matrix of confusion probabilities is then turned into a matrix of confusion frequencies which was then analysed by a CA. Note that only *a single* latent dimension is employed, so one would expect that the CA yields a 1-dimensional solution.

Table 3: Confusion frequencies n_{ij}, and scale values, Gauss distributions (I)

	T_1	T_2	T_3	T_4	n_{i+}	μ	σ^2	f_{i1}
S_1	778	191	29	2	1000	.22	.023	-1.098
S_2	169	493	302	37	1001	.45	.014	-.221
S_3	90	288	387	236	1001	.55	.026	.251
S_4	4	45	200	751	1000	.78	.029	1.069
n_{+j}	1041	1017	918	1026	4002			
g_{j1}	-1.059	-.0249	.293	1.059		$\chi^2 = 3548.69$		

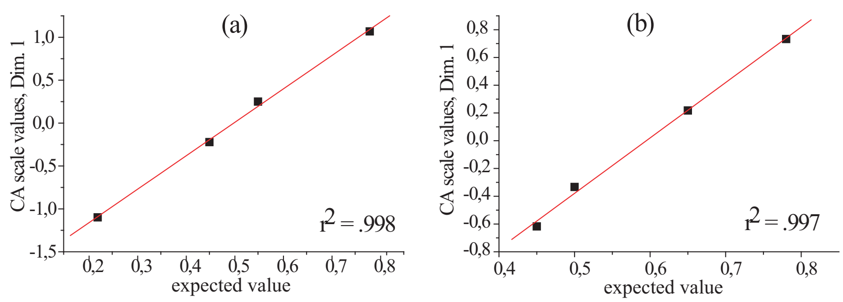
Further, the scale values f_{i1} for the S_i and g_{j1} for the T_j on the first dimension, resulting fro a CA of the n_{ij}. The χ^2- or inertia component due to the first dimension is only 69.35 %; this result will be further commented upon below. Note that the r^2-values for the regression of the CA-scale values on the μ_i amount to .99.

Table 4 shows the results when the distributions are Beta-distributions; the notation is as in Table 3. The plots of the f_{i1} versus the expected values are presented in Fig. 6, (b). The values $r^2 = .998$ for the Gaussian case and $r^2 = .997$ for the beta-case indicate an excellent approximation of the expected values by the scale values in terms of a linear relationship. This is remarkable insofar as the parameters of the beta-distributions were *not* chosen to generate distributions that look "nice", i.e. look like Gaussians.

Table 4: Confusion frequencies n_{ij} and scale values, beta-distributions (I)

	T_1	T_2	T_3	T_4	n_{i+}	μ	σ^2	f_{i1}
S_1	556	168	175	101	1000	.45	.038	-.617
S_2	468	127	170	235	1000	.50	.063	-.333
S_3	174	150	278	398	1000	.65	.031	.218
S_4	65	68	169	697	999	.78	.029	.733
n_{+j}	1263	513	792	1431	3999			
g_{j1}	-.631	-.238	.048	.616		$\chi^2 = 1159.39$		

Figure 6: Regression of scale values on expected values: (a) Gaussian distributions (I), (b) Beta-distributions (I); see Fig. 5.

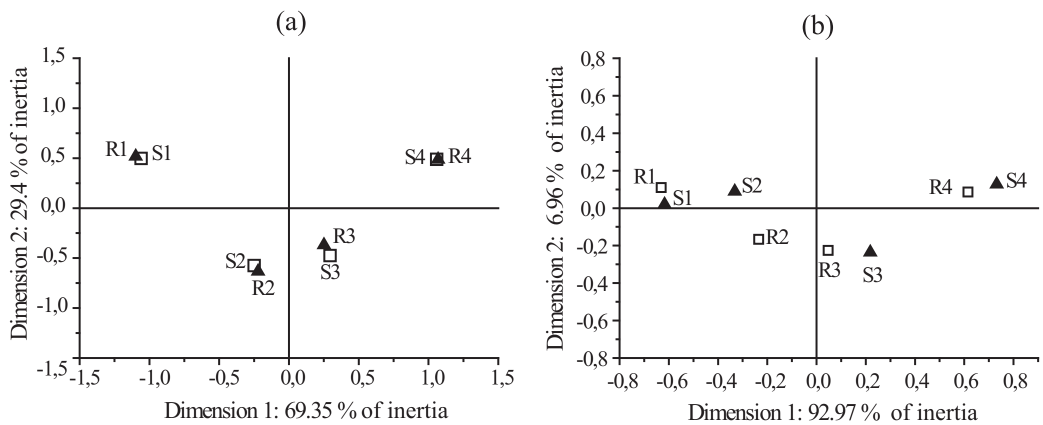


Biplots and the number of latent dimensions

Fig. 7, (a) shows the biplot for the confusion frequencies as generated by the Gaussian distributions. The main feature of the configuration is that S_i- and corresponding T_i-points assume very similar positions, i.e. there is a high degree of correspondence between stimulus pattern S_i- and corresponding T_i-patterns. The projections of the points on the first dimension, "explaining" 69.35 % of the total inertia (or χ^2), are the scale values f_{i1} and g_{j1} given in Table 3. There are, however, two points to be discussed with respect to the biplot:

1. Although only a single dimension (random variable) was used to generate the confusion frequencies, the first dimension accounts for only 69.35 % of the inertia of

Figure 7: Biplots for the confusion frequencies of Table 3 and 4, as generated by Gauss-densities (a), and beta-densities (b)



the data, 29.40 % are due to a second dimension. Indeed, the eigenvalues are $\lambda_1 = .615$, $\lambda_2 = .261$, $\lambda_3 = .011$, and (42) yields the corresponding percentages of χ^2 "explained" by the corresponding dimensions.

2. The plot shows the typical horseshoe effect (Greenacre (1984)): the points representing S_i and the corresponding T_i seem to lie on a U-shaped line.

The two points are related to each other. Inspection of the confusion frequencies in Table 3 shows that the number of correct responses (778, 493, 387 and 751) are always the largest numbers in the i-th row and the i-th column. The number of confusions levels off the farther apart $S_{i'}$ from a given S_i and $T_{i'}$ from the corresponding T_i. This causes the Matrix $T = (t_{ij})$ (c.f. (28)), to be of full rank and the SVD of T will yield $r = I = J$ eigenvalues $\lambda_k \neq 0$. The lesser the number of errors made the more the confusion matrix will resemble a diagonal matrix and the more will the eigenvalues λ_k, $k > 1$, differ from zero, and the more will the biplot based on the first two "dimensions" resemble a U-shaped or horseshoe like configuration. The eigenvalues λ_k, $k > 1$, therefore do not reflect a genuine second feature of the patterns, but have to be considered an artifact resulting from the linear decomposition (32). A more detailed discussion of the horseshoe-effect in Correspondence Analysis may be found in Greenacre (1984), p. 226. For the present purposes it is sufficient to note that the horseshoe-effect is indicative of the second dimension being a negligible artifact.

Fig. 7 (b) shows the biplot for the confusion frequencies generated by the beta-distributions. Note that here the CA-solution is basically 1-dimensional: the first dimension explains about 93% of the χ^2 (or the inertia χ^2/N) of the Table. Inspection of Table 4 shows that here the main diagonal does not always contain the largest number in a given row or column. This fact appears to suppress the horseshoe-effect. Corresponding to PCA-approximations of factor analyses of measurement, the second dimension, accounting for about 7 % of the inertia, here just reflects random error.

4.2 The analysis of empirical data

4.2.1 Gabor patches

An identification experiment was performed employing Gabor patches as stimuli, where a Gabor patch is defined as

$$s(x) = \exp(-x^2/\sigma^2)\cos(2\pi f x); \tag{60}$$

f is the spatial frequency of s, and σ^2 defines the width of s. Four stimulus patterns were defined by the spatial frequencies $f_1 = 3.25$, $f_2 = 3.75$, $f_3 = 4.25$ and $f_4 = 4.75$ c/deg. The patterns were presented in random order and the subject had to decide which of the four patterns was presented. There were 11 sessions with 100 trials each, i.e. each pattern was presented 25 times in a session. Individual CAs were computed[1] for each of the 11 confusion matrices, giving essentially the same picture, i.e. biplot for each session. So the data were lumped, i.e. a single CA was computed for the $4 \times 4 \times 11$ matrix. the results are shown in Fig. 4.2.1, together with the results of the matrix generated by averaging the confusion matrices over the 11 sessions. Table 5 shows the confusion matrix resulting from adding the data from the 11 sessions into one matrix, together with the scale values f_{i1} for the stimuli, and g_{j1} for the responses. The scale values with respect to the second dimension have been omitted since the second dimension does not appear to reflect a second attribute with respect to which the patterns were judged; as the biplots show, the configuration shows a horseshoe effect (Greenacre (1984), p. 226). Since the confusion matrix (as all individual matrices) have maximum frequencies in the diagonal cells the (generalised) SVD of $P - rc'$ will indicate at least two, if not three dimensions underlying the data. The less confusions occur, the more the matrix will become similar to a diagonal matrix requiring 3 dimensions to represent the data, even if a single latent criterion was employed by the subject to identify the stimuli. The regression of the scale

Table 5: Confusion frequencies, summed over the 11 sessions; $\chi^2 = 922.85$ equals sum of either χ_i^2 or χ_j^2; eigenvalues: $\lambda_1 = .600$, $\lambda_2 = .208$, $\lambda_3 = .03$.

	T_1	T_2	T_3	T_4	Σ	χ_i^2	f_{i1}
S_1	193	74	8	0	275	337.48	-.999
S_2	60	161	51	3	275	115.34	-.395
S_3	14	92	134	31	271	98.85	.307
S_4	4	15	102	154	275	371.18	1.091
Σ	271	342	295	188	1096		
χ_j^2	332.91	126.59	127.73	335.62		$\chi^2 = 922.85$	
g_{j1}	-.979	-.348	.538	1.201			

values f_{i1} on the spatial frequency parameters φ_i of the patterns is perfectly linear, see Fig. 9, suggesting (i) that the subject identified the patterns indeed with respect to a single latent dimension $\mathbf{x}$ representing the apparent spatial frequency characteristic of a pattern, (ii) that we may assume that the f_{i1}-values are indeed proportional to the the

[1] Correspondence Analysis of the Statistica-package

Figure 8: Biplots for Gabor patches; (a) different sessions, (b) averaged over sessions

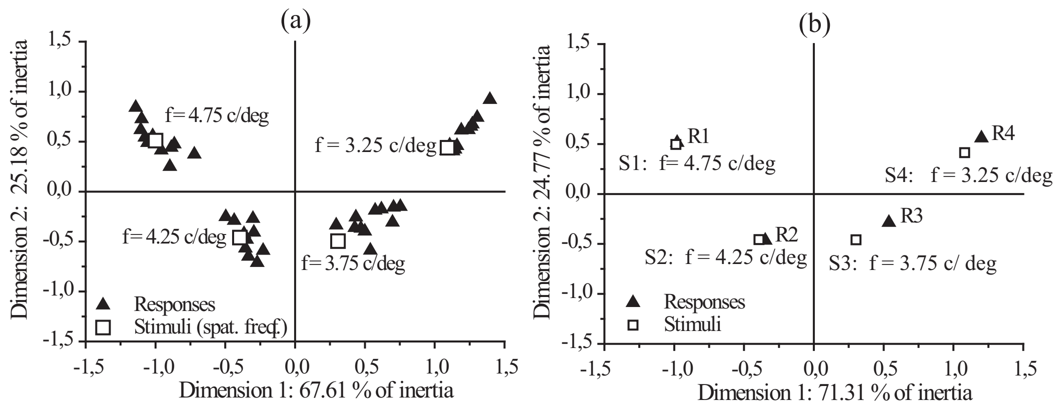
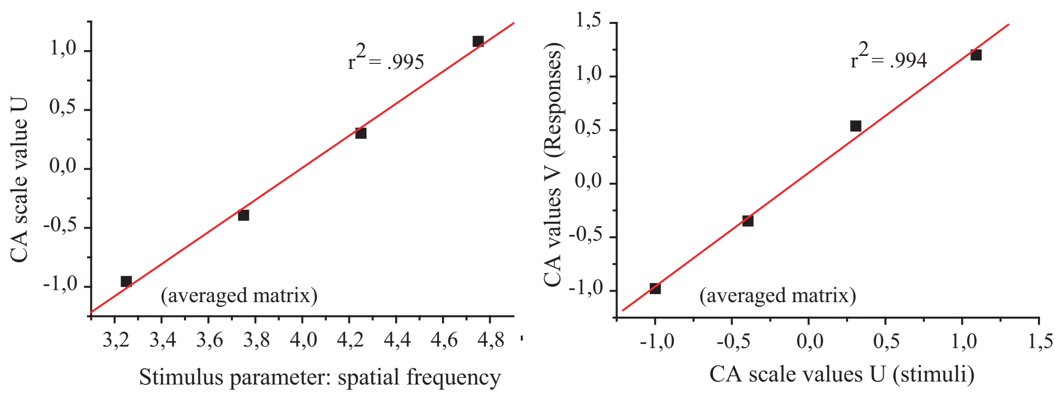


Figure 9: Regression of scale values of stimuli versus corresponding conditional expectations



conditional expectations of the random variables representing the activity generated by the stimuli, and (iii), with respect to Fig. 3, (A), that the variances of these random variabes are about equal, meaning that the patterns are about equally difficult to identify. The data point to the possibility that the spatial frequency parameter appears to be the relevant parameter in the identification process. One may also think of the subject concentrating on other features of the pattern that allow an identification of a pattern, e.g. the position of the luminance maxima; if variables like that are the relevant variables they have to be such that they are linearly related to the parameter values of the stimuli, i.e. the spatial frequencies φ_i.

Note that the χ_i^2-values are largest for S_1 and S_2; correspondingly, χ_j^2-values are largest for T_1 and T_4. So the pattern defined by the smallest value of φ and that defined by the largest value of φ generate the largest contribution to the total χ^2 of the table. Moreover, if one would plot the χ_i^2- or the χ_j^2-values with versus the subscript of either the S_i or the T_j one would get practically symmetrical, U-shaped curves. This means that the χ^2-contributions are symmetrical functions of the distance from the virtual mean pattern, defined by a $\bar{\varphi}$ being the arithmetic mean of the φ_i, $i = 1, \ldots, 4$. The row of confusion frequencies one would observe if such a pattern had indeed been presented would correspond to the mean row of frequencies, and the χ_i^2-values increase with increasing distance from this mean. A similar statement holds for the T_j and their χ_j^2-values. This is characteristic for the results of a CA: the origin of the biplot corresponds to the mean row and the mean column of the contingency table, and the more a row or a column deviates from the mean row or column, the larger the corresponding χ^2-component.

4.2.2 Superimposed Gabor patches with and without flankers

The stimuli are now given by superpositions of two Gabor-functions,

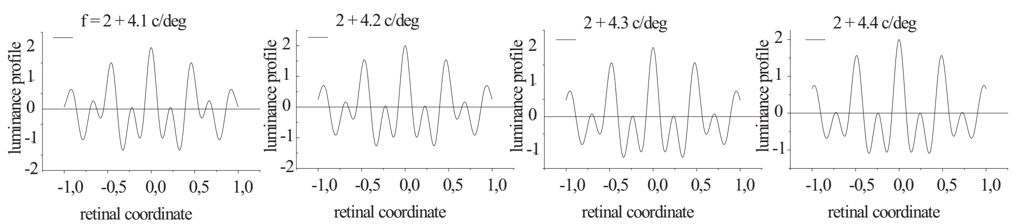
$$s_1(x; f_1) = \cos(2\pi f_1 x)\exp(-x^2/2\sigma^2),$$

with $f_1 = 2$ c/deg, and

$$s_i = \cos(2\pi f_i x)\exp(-x^2/2\sigma^2)$$

and $f_i = 4 + i \times .1$, $i = 1, \ldots, 4$. The stimulus patterns are then $S_i = s_1 + s_i$. The flanking patches were defined by either one of three possible spatial frequencies f_{0k}: $f_{01} = 3.8$, and $f_{01} = 4.5$, and $f_{03} = 5.0$. Figure 10 shows the luminance distributions for the four stimulus patterns; Figure 11 shows these distributions superimposed; the overall shape

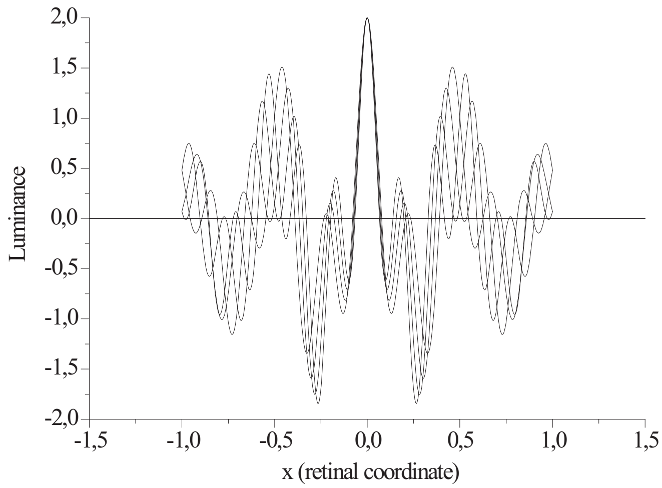
Figure 10: Luminance distributions of the stimuli



is obviously very similar, so the subject has to respond to relatively fine differences in the patterns.

To identify a stimulus the subject just has to identify the partial pattern s_i; in this case the subject can separate the two components s_1 and s_i. However, the pattern S_i may be perceived as a conjunction of the two components, or identify to particular partial aspects of the patterns, due to specific aspects of the superposition. In the first case one may predict that the relevant random variable with respects to whose value the patterns are predicted reflect just the spatial frequency f_i of s_i, and corresponding to the results of the experiment reported in section 4.2.1, the scale values for the stimuli, and also for the responses, should be linear functions of the spatial frequency parameter f_i. In the second case such a linear relationship may not occur. So a linear relationship is compatible in

Figure 11: Luminance distributions of the stimuli, superimposed



particular with the first hypothesis, although such a relationship is at least in principle conceivable with some sort of Gestalt-identification.

Biplots (I): Fig. 12 shows the biplots corresponding to the three spatial frequency parameters of the flanking patterns. Although there are only three such parameters, six biplots are shown: the data from trials with a flanking pattern and those without a flanking pattern have been analysed separately. One may expect that the data from trials without flanking patterns are very similar; on the other hand, the spatial frequency f_{ok} of the flanking pattern was kept constant within a session, and this may have an effect also on the identification process when no flankers were actually presented. The dominant feature of the biplots is that the patterns appear to be discriminated with respect to a single dimension: the inertia component due to the first dimension is always between $\approx 86\%$ and $\approx 90\%$. The second dimension seems to reflect a horse-shoe effect and represents most likely a numerical artifact and not a perceptually relevant feature of the patterns. The stimuli S_1, S_2, S_3 and S_4 are always well-ordered along the first dimension, corresponding to the spatial frequency parameter f_i of the second pattern component. The responses $R_1, \ldots, R_4$ are also well-ordered along the first dimension; however, with increasing value of f_i the precision of the responses seems to deteriorate when a flanking pattern was shown together with the stimulus pattern. In particular R_3 appears midway between S_3 and S_4, indicating that S_3 and S_4 are more often confused with each other than the remaining stimuli. It is of interest to have a look at the χ^2-values for the individual confusion matrices, which are given in Figure 12. For all values of f_{0k}, the χ^2-value for the stimulus presentations with flanking patterns is higher than for the stimulus presentations without flanking pattern; so the flanking patterns seem to facilitate the identification of the stimulus patterns.

Biplots (II): To investigate further the influence of the flanking patterns, the results of

Figure 12: Biplots, for the different spatial frequency parameters of flanking patterns

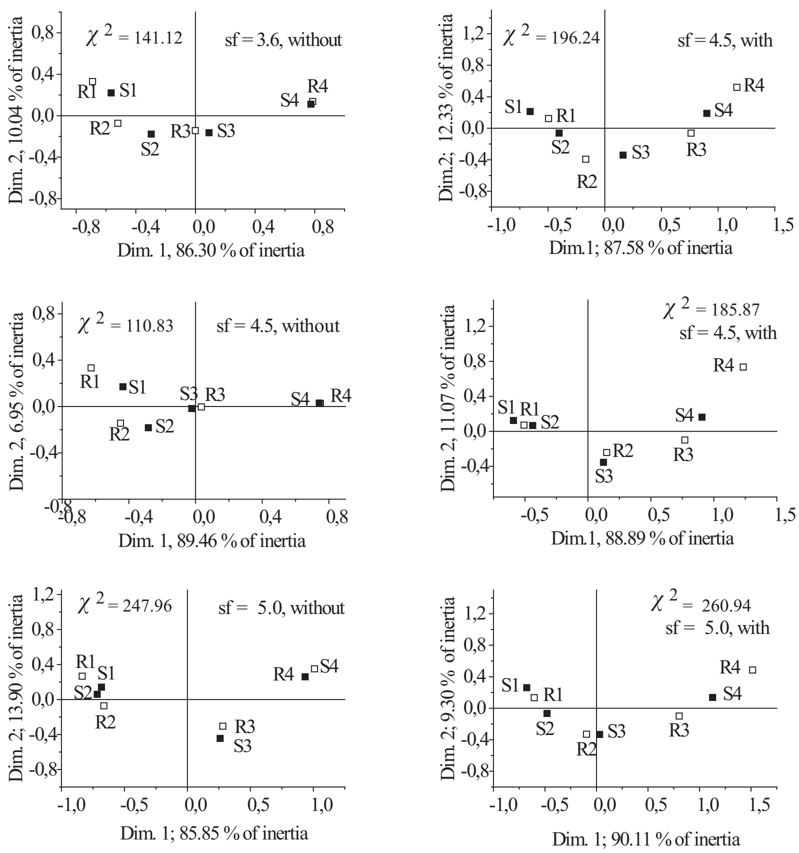
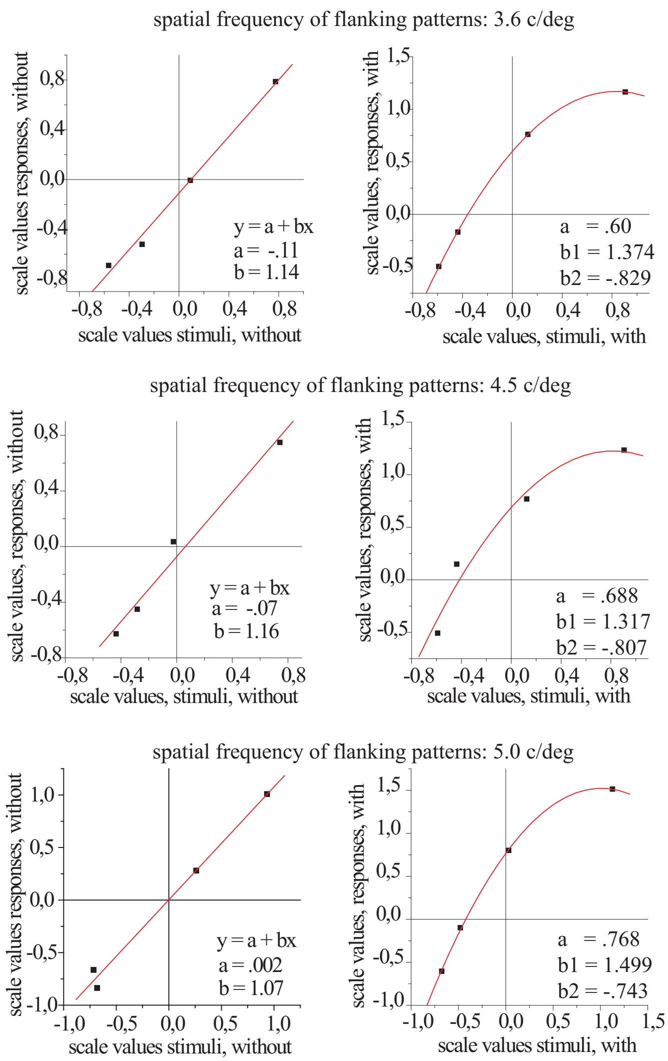


Figure 13: Stimulus versus response scale values (biplots), depending upon spatial frequency parameter of flanking pattern



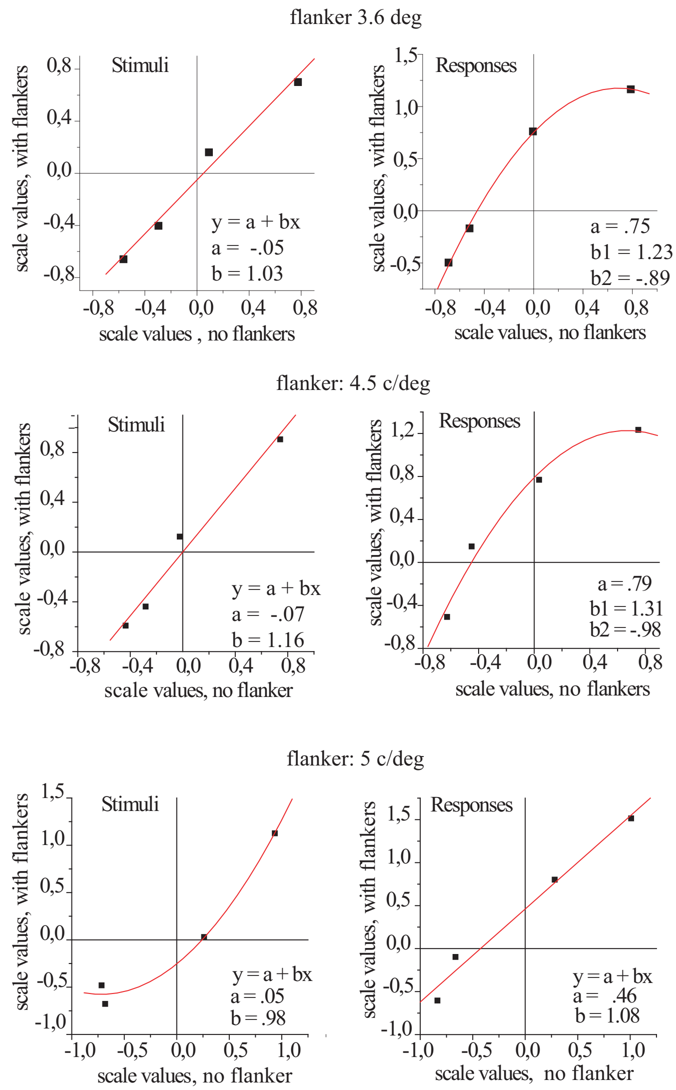
the Correspondence Analysis (CA) are presented in a different form in Fig. 13: the scale values of the stimuli when presented without flanking patterns are plotted against the scale values of the stimuli when presented with flanking patterns, and similarly the scale values for the responses from trials without flanking pattern are plotted against the scale values of the corresponding responses when the flanking patterns were shown. Only the scale values on the first dimension are presented.

The main result to be taken from Figure 13 is that the relation between the scale values for the stimuli and the responses is linear when no flanking pattern was presented with the stimuli, whereas this relation appears to nonlinear when the stimuli were presented with flanking patterns; again, from a statistical point of view the null-hypothesis, that the data are compatible with a linear relationship, may still be maintained. It is the trend over the different spatial frequencies that suggests the existence of nonlinearities. This trend would mean that in particular the difference between the responses R_3 and R_4 is reduced under the flanking condition. From the meaning of the scale values for the responses this indicates that the flanking patterns induce an adjustment of the decision boundaries.

Another aspect of the data is revealed when the scale values of the stimuli for the no-flanker condition are plotted against the scale values for the stimuli for the flanker condition, and similarly for the scale values of the responses, see Figure 14. The plots of the scale values for the stimuli are definitely linear for the flanking parameters $f_{01} = 3.6$ and $f_{02} = 4.5$, while the plots of the scale values of the responses may contain a nonlinear component. From a purely statistical point of view, a linear function will also account for the response data, but there is a definite improvement of the fit when a nonlinear function is fitted (a polynomial of 2nd order). This finding contrasts with that for the flanking parameter $f_{03} = 5.0$: here the plot for the stimuli seems to reflect a positively accelerated relation between the scale values of the stimuli, and while the fit of the 2nd order polynomial to the response data does not imply any improvement of the fit. If one inspects the parameters of the fitted curves, one finds that the slope parameter b for the stimulus plots increases with increasing value of f_{0k}, from $b = 1.03$ for $f_{01} = 3.6$ to $b = 1.16$ for $f_{02} = 4.5$, and for $f_{03} = 5.0$ the positively accelerated relation fits even better than the linear one. For the response plots the parameter b_1, representing the linear component of the polynomial, also increases from $b_1 = 1.23$ for f_{01} to $b_1 = 1.31$ for f_{02}, and at the same time the parameter b_2, reflecting the negative acceleration, decreases from -.89 for f_{01} to -.98 for f_{02}; for f_{03} the relation has turned linear. So there are systematic changes in the parameters describing the relation between scale values, depending on the value of f_{0k}. The flanking patterns appear to pull the scale values for stimuli and responses apart. With respect to Figs. **??** and **??** one may conclude that the flanking patterns exert an influence either on the expected values of the underlying random variables, or, more likely, on their variances. The neural representations appear to be more separable with increasing frequency of the flanking patterns, meaning that the variances become smaller with an increasing value of f_{ok}. This corresponds to the observed increase of χ^2-values when flanking patterns are presented.

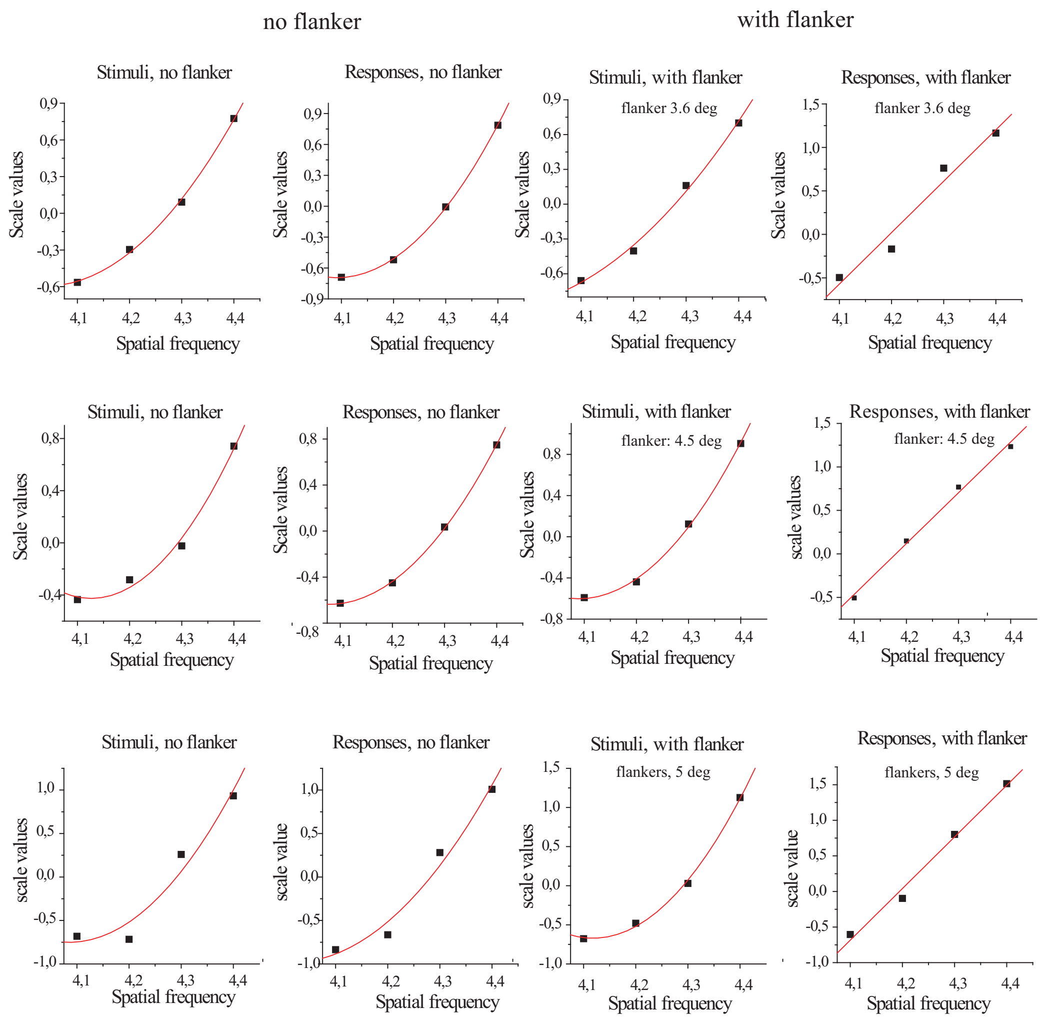
Further aspects of the data are revealed when the scale values of stimuli and responses are plotted against the spatial frequency parameter f_i of the second stimulus component, see Figure 15. The relations between the scale values and the f_i appear to be rather systematic. Except for the relation between the scale values of the responses and the f_i

Figure 14: Scale values of stimuli and responses, without versus with flankers



under the flanker condition, the relation between the scale values and the f_i appear to be nonlinear; in particular, they are positively accelerated. Only for $f_{0k} = 5.0$ a linear relation fits equally well when no flanking patterns were presented. The nonlinearity of the relations between scale values and f_i-values is remarkable since the differences $f_{i+1} - f_i$, $i \leq 3$ are identical for all i and the experiment reported in section 4.2.1 a linear relation between the parameter f_i, the spatial frequency defining a Gabor patch, and the scale values was observed, see Figures 8 and 9. Moreover, the difference $f_{i+1} - f_i$ equalled .25 in this experiment, while in the experiment reported in this section the difference

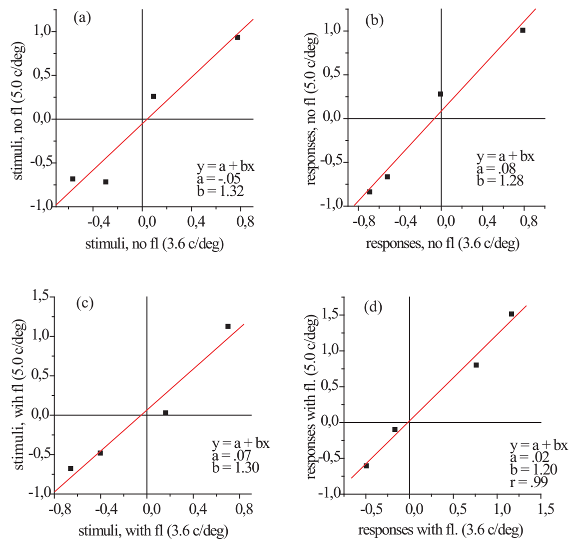
Figure 15: Scale values versus f_i-values



equalled only .1. Possibly, the fact that here superpositions of Gabor patches defined the stimuli plays a role. Alternatively, a change of the variances of the underlying random variables with the value of the spatial frequency-parameter of the flanking patterns may cause the observed nonlinearities; further data are required to clarify these points. In any case, the fact that linear relationships are observed for the responses, provided a flanking pattern was shown, is of interest. The flanking patterns appear to have a adjusting effect on the sets A_j, which determine the range of values of the random variable, representing the sensory effects, implying the response R_j. These adjustment effects may again be due to corresponding changes of the variances of the decision variable.

“[ht!]

Figure 16: Scale values for $f_{01} = 3.6$ versus $f_{02} = 5.0$



Finally, one may consider the plots of the scale values for stimuli for a certain value f_{0k} versus the scale values of the stimuli for another value of f_{0k}; similarly for the responses. Figure 14 shows these plots. We consider in particular the data for $f_{0k} = 3.6$ and $f_{0k} = 5.0$. Note first that the plots are definitely linear, and note further that for the stimuli the slope parameters assume the values $b = 1.32$ (no flanking patterns) and $b = 1.30$ (with flanking patterns). The difference between these two b-values is certainly negligible, and an equivalent statement holds for the additive constants a ($a = .05$ and $a = .07$), i.e. the difference has only a negligible effect on the prediction of the scale values for the stimuli under the $f_{0k} = 5.0$-condition on the basis of the scale values from the $f_{0k} = 3.6$-condition, whether the flanking patterns were actually presented or not. This is remarkable since the finding means that the flanking patterns employed in a particular session appear to determine the perception of the stimuli independent of their actual presentation. The fact that the values of b are larger than 1 imply that the scale values for the $f_{0k} = 5.0$-condition are more spread out than for the $f_{0k} = 3.6$-condition, corresponding to the larger χ^2-values for data from trials with flanking patterns. For the responses, an analogous statement holds, ($b = 1.28$ and $b = 1.20$). If only the scale values for the responses were affected by the flanking parameter one could easily argue that this parameter affects the definition of the sets A_j, but the fact that also the scale values of the stimuli are affected would mean that also the sensory representations change, provided the model considered in this paper is correct: an increase of the spatial frequency of the flanking patterns seems to reduce the variance of the underlying random variables.

5 The 2-dimensional case

The stimuli may vary with respect to more than a single aspect. Suppose the stimuli vary with respect to the aspects A and B. The subject may combine information on the aspects into a single percept that can be represented by a single random variable; this may be the case if, for instance, the stimuli are represented by templates in the subject's memory and the subject chooses the response according to the maximally activated template. If Y_j, $j = 1, \ldots, I$, represents the the acitivity of the j-th template, then the decision variable would be $Y = \max(Y_1, \ldots, Y_I)$. Alternatively, the subject may try to evaluate the aspects A and B individually in order to arrive at a decision about the presented stimulus. A sufficient, but not necessecary condition for this is that the subject is able to perceive the different levels A_i of A and B_j of B independenty. If the A_i and B_j are not perceived independently, in the sense that the neural activities corresponding toA_i and B_j are independent of each other, the subject may still be able to evaluate these activities independently. Ashby and Townsend (1986) and Karlec and Townsend (19??) have carefully disentangled these possibilities. In the following, the definitions and theorems presented by A & T will be briefly presented. It will then be shown that employing Correspondence Analysis to the confusion matrix allows, together with some tests proposed by A & T, some quick decisions about the type of process underlying the subject's performance.

5.1 Ashby and Townsend (1986)

x, y random variables representing perceptual effects of the two components defining the stimulus, defined as A_iB_j, i.e. i-th level of feature A and j-th level of feature B.

Definition 5.1 *Let $f_{A_iB_j}(x,y)$ the common distribution of x and y when the stimulus is defined by A_iB_j, and let $g_{A_iB_j}$ denote marginal distributions. The components A_i and B_j are perceptually independent, if*

$$f_{A_iB_j}(x,y) = g_{A_iB_j}(x)g_{A_iB_j}(y) \tag{61}$$

holds.

Remark: If f is Gaussian, independence holds if and only if the effects of A_i and B_j are uncorrelated over trials.

Definition 5.2 *Let (a,b) denote the event that a stimulus (A,B) is reported. Sampling independence holds if*

$$p(a_2b_2|A_2B_2) = p(A \textit{ sampled } |A_2B_2)p(B \textit{ sampled } |A_2B_2) \tag{62}$$

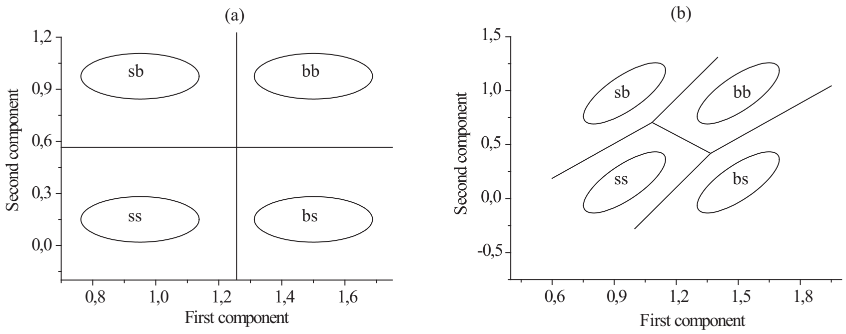
holds. For the complete identification experiment:

$$p(a_2b_2|A_2B_2) = [p(a_2b_1|A_2B_2) + p(a_2b_2|A_2B_2)][p(a_1b_2|A_2B_2) + p(a_2b_2|A_2B_2)]. \tag{63}$$

Remark:

$$p(a_2b_2|A_2B_2) = \frac{p(a_2b_2 \cap A_2B_2)}{p(A_2B_2)} \overset{general}{=} \frac{n_{ij}}{n_{i+}}. \tag{64}$$

Figure 17: Configurations of stimuli; (a) the separable case, (b) optimal boundaries for the perceptually dependent case



Theorem 5.1 *Consider the complete identification experiment. Then*

1. *Sampling independence holds for the stimulus A_iB_j, if perceptual independence holds and the decision bounds are parallel to the coordinate axes.*
2. *Perceptual independence holds for A_iB_j, if sampling independence holds for A, B for different decision criteria and if decision bounds are parallel to the coordinate axes.*
3. *If the decision bounds are not parallel to the axes then sampling independence and perceptual independence are logically unrelated, i.e. .*

Separability versus Integrality: According to Garner & Morton (1969), stimulus components are separable if they "act separately in the organism and thus can go independently of each other". Integrality if components "join one another such that it is extremely difficult for the subject to take note of one without at the same time taking note of the other". Separable stimuli can be separately attended to whereas components of integral stimuli cannot (this may happen to various degrees).

Definition 5.3 *Operational definition A: If components are separable, performance on a task demands a response based on a single component is unaffected by the level of other irrelevant components. With integral components, varying the level of irrelevant components degrades performance ("filtering task").*

Separability can be defined either at the perceptual or the decisional level (p. 164). Let

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy. \tag{65}$$

Components A and B are perceptually separable if the marginal distribution of A does not depend on the level of B, so that

$$g_{A_iB_1}(x) = g_{A_iB_2}(x), \quad i = 1,2 \tag{66}$$

Definition 5.4 *Consider a complete identification experiment, ie one in which all combinations A_iB_j occur. A and B are perceptually separable if the perceptual effect of one component does not depend on the level of the other, that is if*

$$g_{A_iB_1}(x) = g_{A_iB_2}(x), \quad i = 1, 2 \tag{67}$$

$$g_{A_1B_j}(y) = g_{A_2B_j}(y), \quad j = 1, 2 \tag{68}$$

Definition 5.5 *Consider a complete identification experiment with stimuli A_1B_1, A_1B_2, A_2B_1 and A_2B_2. The components A and B are decisionally separable if the decision about one component does not depend upon the level of the other, that is, if the decision bounds in the general recognition theory are parallel to the coordinate axes.*

This relates to Theorem 5.1: one may now say that if decisional separability holds, then perceptual independence is equivalent to sampling independence (when the latter holds for different decision criteria). If decisional separability is not found, then Theorem 5.1 indicates that sampling independence is logically unrelated to perceptual independence.

Theorem 5.2 *Complete identification experiment; Suppose the following three conditions hold:*

1. *All perceptual presentations are normally distributed (the general Gaussian recognition model holds),*
2. *No two $\vec{\mu}_{ij}$ are equal, where $\vec{\mu}_{ij}$ is the mean vector for stimulus A_iB_j,*
3. *the subjects responds optimally, maximising probability correct.*

Then perceptual separability and decisional separability together imply the perceptual independence of components A and B within each stimulus configuration.

Theorem 5.3 *Complete identification experiment: If perceptual and decisional separability hold for components A and B, then marginal response invariance holds, ie the probability of correctly recognising one component does not depend upon the level of the other:*

$$p(a_ib_1|A_iB_1) + p(a_ib_2|A_iB_1) = p(a_ib_1|A_iB_2) + p(a_ib_2|A_iB_2) \tag{69}$$

$$p(a_1b_j|A_1B_j) + p(a_2b_j|A_1B_j) = p(a_1b_j|A_2B_j) + p(a_2b_j|A_2B_j) \tag{70}$$

Relevant conclusions: Key question is: does perceptual independence hold?

1. Suppose response invariance does not hold, ie equations (69), (70) do not hold. Then the decision boundaries are not parallel to the coordinate axes, meaning that decisional *and* perceptual separability do not hold, (see Theor. 5.2), i.e. either perceptual or decisional separability fail. According to Definition 5.5, if decision bounds are parallel to coordinate axes, then the features are perceptually separable. So if perceptual separability does not hold, then the axes are not parallel to the axes.

2. According to Theor. 5.1, if perceptual independence and decision bounds parallel to the axes, then sampling independence, and sampling independence and decision bounds parallel to the axes, then perceptual independence.

   Sampling independence holds, if the conditions of eq. (63) are met. If the equations are not satisfied, sampling independence can be refuted. Then, according to Theor. 5.1, the conjunction sampling independence and parallel decision bounds does not hold. So, if sampling independence and, from **??**, parallelism of decision bounds do not hold, we cannot conclude that perceptual independence holds; the findings support the hypothesis that perceptual independence does not hold.

Relation to Correspondence Analysis

CA provides an orthogonal system of coordinates, with each axis representing a certain proportion of the total inertia χ^2/N, and therefore of the total χ^2. Suppose (x, y) represents the perceptual effect of a stimulus, and suppose further that f is 2-dimensional Gaussian with zero covariance and equal variances. Then perceptual independence holds. Suppose further that decisional and sampling independence holds and the subject decides optimally. The stimulus configuration (in the biplot) is then rectangular, where the scale differences of the first axis represent the differences between the two means of the feature with the larger parameter difference, and the difference of the scale values of the second dimension represent the feature with the smaller parameter difference. If the decision bounds are optimal, i.e. are defined by the corresponding mean of the scale values, the points for the responses will be identical to the stimulus point (error free case).

The scale differences will be proportional to the corresponding d'-values. This is intuitively clear: the scale values will be proportional to the corresponding mean values, and since the variances are supposed to be identical, the weighting will be identical as well. This will be illustrated numerically. The effect of the following distortions has to be illustrated numerically:

1. Different variances, optimal decision bounds, zero covariances,

2. Equal variances, equal correlation (positive and negative) for all stimuli,

3. The combination of unequal variances and equal correlation,

4. Shifts of the decision bounds; still parallel to the axes, but deviating from the optimal values (defined by the equal a priori probabilities and equal costs of wrong decisions).

5. Decision bounds that are not parallel to the axes, but linear.

The effects should be discussed with respect to the Ashby & Townsend and Karlec & Townsend papers: The discussion should aim at an identification of the perceptual independence case, because that's what is most interesting. In particular:

1. Definition of Sampling independence, see Def. 63,
2. Relation to Theorem 5.1,
3. and the other notions and corresponding Theorems.

The stimulus configuration in the disc-hat-experiment: So the labelling is counter-clockwise.

Table 6: Stimulus configuration in the disc-hat-experiment

Stim. 1	A_1B_1
Stim. 2	A_2B_1
Stim. 3	A_2B_2
Stim. 4	A_1B_2

The equivalent to the d'_{AB_j} and d'_{A_iB} is then

$$d'_{AB_1} \circeq \Delta'_{AB_1} = f_{31} - f_{21}, \qquad d'_{AB_2} \circeq \Delta'_{AB_2} = f_{41} - f_{11} \tag{71}$$

$$d'_{A_1B} \circeq \Delta'_{A_1B} = f_{12} - f_{22}, \qquad d'_{A_2B} \circeq \Delta'_{A_2B} = f_{42} - f_{32} \tag{72}$$

The theorems relating separability and independence on the one hand and d'_{AB_j} and d'_{A_iB} on the other should carry over to the equivalent statements relating independence and separability and Δ'_{AB_j} and Δ'_{A_iB}.

Alternative approach: suppose the stimulus is represented by some variable $y = \sum_k \alpha_k\phi_k + \xi$, ϕ_k a representation of a feature, α_k a weight, ξ some error. y is evaluated as in Thurstone's model, or Fisher's discriminant analysis.

5.2 Numerical evaluations

5.2.1 The separable case; $r = 0$ versus $r \neq 0$

The case $r = 0$: Consider the case that A_1 and A_2 differ more than B_1 and B_2 (radii of superimposed circular discs).

1. **The case $\sigma_1 = \sigma_2$:** The mean values determine the configuration. The first axis separates the stimuli with respect to the greater feature difference.

   Special case: all features have the same value. Then the two dimensions represent the same proportion of inertia. The stimuli form a square, they are positioned on the axes.

   If the features have different values, the configuration will also be rectangular, but the sides of the rectangular are parallel to the axes.

2. **The case $\sigma_1 \neq \sigma_2$:** If all features have the same value, the case $\sigma_1 < \sigma_2$ will separate the stimuli S_1, S_2 on the one hand versus S_3 and S_4 on the other. If $\sigma_1 > \sigma_2$, the stimuli S_1, S_4 on the one hand and S_2 and S_3 on the other.

The case $r \neq 0$: For equal variances:

1. **The case $r > 0$:** The configuration is turned clockwise.
2. **The case $r < 0$:** The configuration is turned counterclockwise.

For unequal variances, the reverse may hold (examples). In any case: the separability test always complies with the data, while for $r \neq 0$ the independence test does not.

5.2.2 The case $r \neq 0$, Gaussian classification

Let us assume that the subject is able to somehow construct the optimal decision bounds. For positive and negative correlations: the configurations are turned clock- or anticlockwise. In any case, neither the independence test nor the separability test hold. This is of importance since some configurations look like the separable case with $r \neq 0$: slightly turned configuration.

5.2.3 Conclusions

If the stimuli are identified with respect to the components, the stimulus configuration as well as the response configuration will be rectangular, regardless of the way the decision boundaries are defined. The Ashby-Townsend tests will reveal whether the boundaries are parallel to the axes defined by the components or not, i.e. whether separability holds or not. If the correlation between the feature representations are zero, the configuration will be parallel to the axes, otherwise the configuration appears rotated. It is the rectangular form of the configuration that indicates that decisions are made with respect to the individual components.

One may therefore say that if rectangularity is not observed, then the subject identifies the patterns not according to an evaluation of the component representation, but according to some function $\phi(A, B)$. ϕ may reflect template matching, or that indeed some function of the components is taken as a basis for the decisions, - or the variance-covariance matrices are different. A look at the inertias may reveal that indeed the decisions are based on some function ϕ; this will be discussed with respect to some data (the disc-hat-data). Indeed, if the var-covar-matrices are unequal, this would mean that the particular values of the features determine the interaction between the representation of the components is configuration specific, pointing to a holistic representation of the stimuli, which is a special case of a representation by some function ϕ.

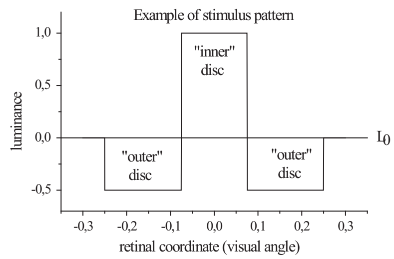
5.3 Empirical data: stimuli composed of circular "discs"

Stimulus patterns were defined according to

$$l(r) = L_0(1 + m\, s(r)), \quad r = \sqrt{x^2 + y^2}, \tag{73}$$

L_0 defines the space average luminance of the screen, and $m = (L_{max} - L_{min})/2L_0$ is the (maxwell contrast). $s(r)$ is defined as the sum of two concentrically superimposed

Figure 18: Example of stimulus: luminance profile



circular "discs" d_1 and d_2 of different radius and luminance, specified by

$$d_k(r) = \begin{cases} 1, & r \leq c_k \\ 0, & r > c_k \end{cases}, \quad k = 1, 2 \tag{74}$$

i.e. $s(r)$ is defined as

$$s(r) = (1 + \alpha)d_1(r) - \alpha d_2(r), \quad \alpha = 1/2. \tag{75}$$

Four patterns $s(r)$ were defined: r_1 could assume either the value r_{11} or r_{12}, and r_2 could assume the values r_{21} or r_{22}. The total width (diameter) of the stimulus pattern is defined by the "outer" disc and was either $2c_1 = .5^0$ or $2c_1 = .46^0$. The "inner" disc had either the diameter $2c_2 = .15^0$ or $2c_2 = .14^0$. Stimuli are characterised by a combination of two letters, b for "big", s for "small". Responses are characterised the same way, only with capital letters, a particular pattern s_i is defined by a combination (r_{1i}, r_{2i}), so there are four possible patterns. The actual values of the r_{ij} are summarized in Table 7. A more

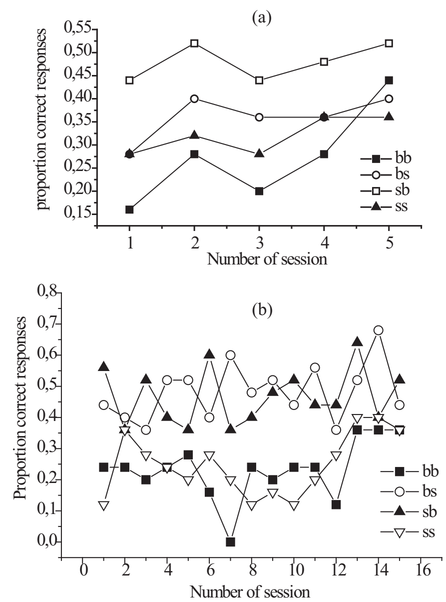
Table 7: Characterisation of stimuli and resonses; The first specifies the "outer", the second the "inner" disc", s and S for "small", b and B for "big". c_1 radius for inner, c_2 for outer disc.

Stimulus	Response	c_1	c_2	c_1/c_2	c_2/c_1
bb	BB	.25	.075	3.333	.300
bs	BS	.25	.07	.3571	.28
sb	SB	.23	.075	3.067	.320
bb	BB	.23	.07	3.285	.304

complete description of the experiment is given in Mortensen and Meinhardt (1999).

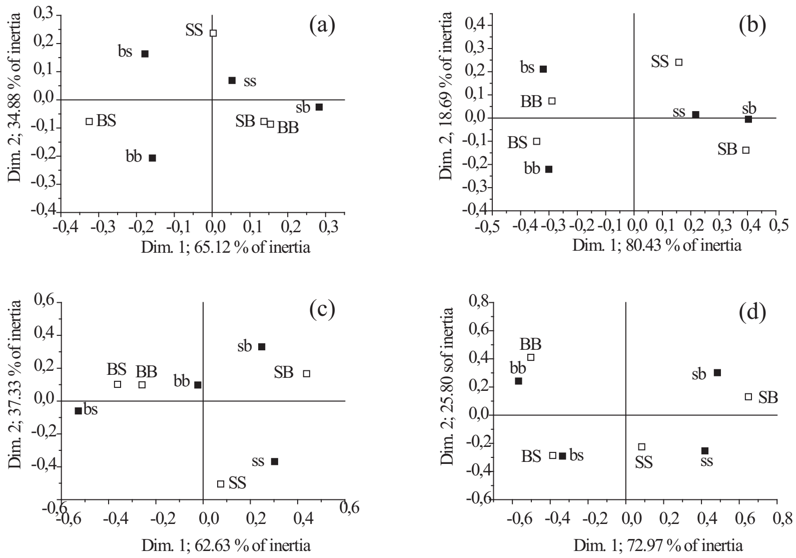
The conditions of stimulus presentation were such that it was difficult for a subject to arrive at a correct response, so there is some learning involved. One may expect that the subject will try different strategies. To minimise the effect of averaging over different strategies during a session, the number of stimulus presentations per session was kept relatively small: stimuli were each presented 25 times per session. The first question one may ask is whether the effect of learning may be seen from a plot of correct responses for the four stimulus patterns, seed Fig. 19, (a) for subject GM and (b) for subject KF. Obviously, the course of the proportions of correct responses does not follow classical

Figure 19: Proportion of correct responses over sessions.



learning curves, i.e. here is no more or less monotonic increase of the proportions of correct responses. All one can say is that the proportions for the patterns *sb* and *bs* are above those for the patterns *bb* and *ss*for both subjects; it seems that the patterns *sb* and *bs* can be discriminated more easily than the patterns *ss* and *bb*. To find out more about any learning effects one may thus investigate the biplots for each session. Figure 20 shows the biplots for subject GM, who participated in only 5 sessions. In the first session (a), the responses show a strong random component; the χ^2 for the data from this session is not significant; a similar result holds for the second session, which is not presented here for lack of space. Still, a first systematic element shows already up in this plot, which remains invariant in the plots for all the following session: the patterns *bb* and *bs* show up in the left quadrants of the plot, and the patterns *ss* and *sb* appear in the right quadrants. So the patterns with a ”big” outer disc seem to form a class, and those with a small outer disc another. This means that the outer disc defines the meaning of the first dimension. Further, the patterns with a ”small” inner disc appear either in the upper two quadrants ((a) and (b)) or in the lower two quadrants ((c) and (d)), so the sizes of the discs for a basis for the decisions from the first session on. The responses do not show a clear pattern, though. The points for the responses SB and BB are almost identical in the first session, and their position deviates considerably from the positions of the respective stimulus patterns. For the third session - (c) - , the positions of the responses

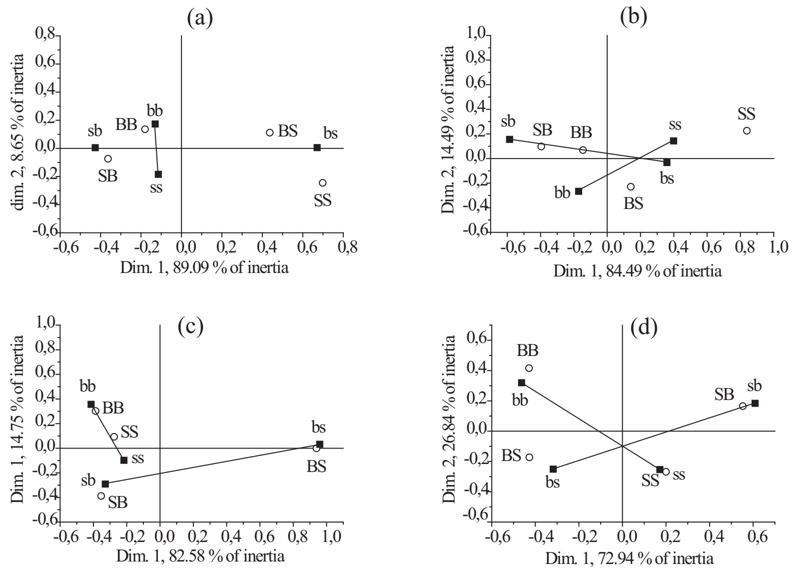
Figure 20: Biplots, Subject GM. (a): session 1, $\chi^2 = 5.83$, $p = .80$; (b): session 3, $\chi^2 = 15.52$, $p = .19$; (c): session 4, $\chi^2 = 17.27$, $p = .044$; (d): session 5, $\chi^2 = 28.93$, $p = .001$. See text for further explanation.



correspond to those of the stimulus patterns at least with respect to the size of the outer disc. The fourth session provides an improved picture, and in the final, fifth session - (d) - the subject's decisions have improved such that each stimulus pattern corresponds to one quadrant, and the corresponding response point is situated in the same quadrant. The subject knows that the stimulus patterns are defined by an orthogonal variation of the sizes of the inner and the outer disc and tries to evaluate the stimulus components, i.e. the discs, independent of each other. In other words, the subject evaluates the patterns with respect to two, not with respect to a single decision variables. Now the scale values are meant to reflect conditional means; each stimulus pattern is thus represented by two such means, one for the distribution of the representation of the inner, the other for that of the outer disc. While the representation of the stimuli is relatively stable over the sessions, that of the responses is not, it stabilises over the sessions, though. This may mean that the bounds of the areas for each stimulus have to be learned. In this sense, learning amounts to a reduction of response bias.

A second subject (KF) provided data from altogether 15 sessions; the number of stimulus presentations was again 25 in each session. Fig. 21 the biplots for a selection of the sessions. The subject KF agrees with the subject GM insofar as he eventually arrived at a classification scheme that is practically identical with that of GM: the two features, inner and outer disc, are evaluated independent of each other, with the outer disc being the feature that is easier to discriminate. However, in all sessions preceding the fifteenth KF's decisions appear to be guided by the relation among the disc sizes, and less so by an independent evaluation of their sizes. In the first session - (a) - the patterns that seem

Figure 21: Biplots, Subject KF. (a): session 1, $\chi^2 = 18.17$, $p = .029$; (b): session 6, $\chi^2 = 19.17$, $p = .021$; (c): session 14, $\chi^2 = 37.71$, $p = .000$; (d): session 15, $\chi^2 = 24.52$, $p = .000$.



to define the first dimension are *sb* and *bs*, explaining about 89 % of the total inertia, i.e. of the total χ^2; the patterns *ss* and *bb* appear to cluster near *sb* and generate a second dimension accounting, however, for only 8.65 % of the inertia. Note that the responses BB and SB are close to the points representing the corresponding stimuli, while the responses BS and in particular SS are not so, indicating that the decision bounds for the inner disc are inappropriately chosen. This may result from the fact that the feature are not evaluated independent of each other.

In the sixth session - (b) - the position of the points representing the stimuli has changed, the first dimension seems to be defined to some extent by the inner disc. The response points are not too far from the stimulus points, i.e. the subject begins to succeed in finding decision bounds relative to the dimensions that serve to distinguish among the patterns. In the 14th session the subject's decisions appear again strongly determined by the relation among the features, the pattern *bs* versus the remaining patterns seems to define the 1st dimension, accounting for about 83 % of the inertia. The outer disc appears to define the second dimension, accounting, however, only for maximally 15 % of the inertia (a part of the proportions reflects just noise!). Only in the last, 15th session the subject appears to abandon the strategy to evaluate the pattern in a Gestalt-like manner and concentrate on the features independent of each other.

That the subject KF judged the stimuli according to the relation among the features may be further corroborated by looking at the data from another view point. For instance, Fig. 22 shows the biplots for the sessions 9 - (a) - and 10, (b). In session 9, the first

Figure 22: Biplots for the sessions 9 - (a) - and 10 - (b), and ratio of disc diameters as criterion variable, (c) session 9, (d) session 10

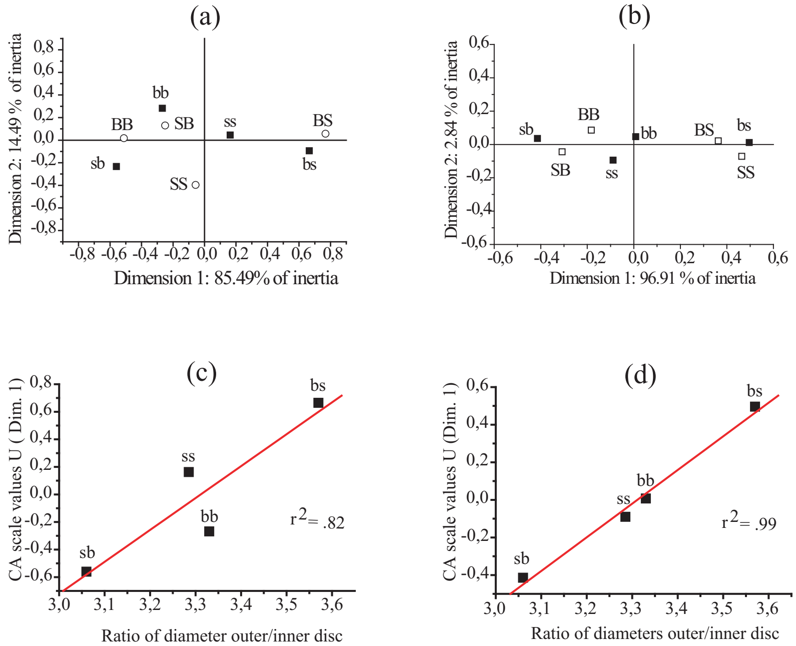
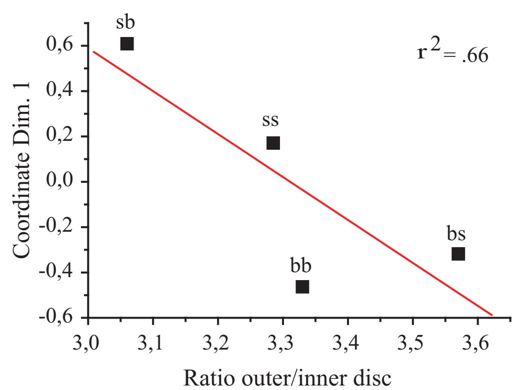


Figure 23: Scale values for the first dimension versus quotients of disc ratios for session 15, subject KF



dimension, accounting for the larger part of the inertia (85.5 %), is defined by the patterns *bs* and *sb*; only the points for *bb* contributes a little to the second dimension. This holds to an even larger extent for the 10th session. Figure 22 (c) and (d) shows plots of the scale values of the stimuli versus the quotient of the radii outer to inner disc, corresponding to the biplots (a) and (b), respectively. For session 10 the fit is practically perfect. So one may argue that subject KF tries to base his decisions on the value of a single variable, amounting to a gestalt-like evaluation of the patterns. Only in session 15 the decisions appear to be based on an independent evaluation of the features; Fig. 23 shows the plot of the scale values for the first dimension versus the ratios of the radii, corresponding to the biplot (d) in Fig. 20, and now the value of r^2 has decreased to .66.

6 Summary and discussion

It was argued that the scale values for the stimuli (row categories) and responses (column categories) provided by a Correspondence Analysis are linear functions of conditional expectations of random variables representing the sensory activity generated by the stimuli. These random variables may actually be random vectors, the components of which represent different aspects or features of the stimuli with respect to which they may be distinguished. It is assumed that the subject selects the stimulus features such that the number of errors is minimised, relative to the experimental conditions. Stimuli and responses can simultaneously be represented by points in a coordinate system whose axes represent, analogous to the axes in principal component analysis, independent stimulus dimensions; these latent axes correspond to those found in a discriminant analysis of the data. However, care has to be taken with respect to numerical artifacts; for instance, a second dimension may result from the fact that correct responses outnumber the number of confusions, in which case the horseshoe effect will appear, as in Fig. 7, where it is clear from the start that the data, i.e. the confusion matrix, is generated by a single "latent" dimension. If one has reason to assume that the second dimension does not just reflect numerical artifacts, the Eucledian distance $d_{ii'}$ between point representing stimuli may be taken as a parameter free sensitivity measure equivalent to the d' in SDT; if the data suggest that only a single dimension represents a perceptually relevant stimulus dimension, the differences of the scale values on this dimension may be interpreted as indices of discriminability or sensitivity, analogous to the d'-measure. $d_{ii'}$ has the advantage of not requiring the assumption that the underlying random variables are Gaussians with equal variance; the nonlinear relations between stimulus parameters and scale values may actually indicate deviations from the equal-variance case.

The position of points representing the responses provides information concerning possibly biased confusions of the correct stimulus with other stimuli. The Eucledian distance between stimulus and response points is not explained though; the relation between stimulus and response points is given by the systems of equations (53) and (53), or, in explicit form, by (55). For instance, for a certain response, say R_1, one has

$$\begin{aligned} g_{11} &= \frac{p_{11}}{c_1}\frac{f_{11}}{\sqrt{\lambda_1}} + \frac{p_{21}}{c_1}\frac{f_{21}}{\sqrt{\lambda_1}} + \cdots + \frac{p_{I1}}{c_1}\frac{f_{I1}}{\sqrt{\lambda_1}} \\ g_{12} &= \frac{p_{11}}{c_1}\frac{f_{12}}{\sqrt{\lambda_2}} + \frac{p_{21}}{c_1}\frac{f_{22}}{\sqrt{\lambda_2}} + \cdots + \frac{p_{I1}}{c_1}\frac{f_{I2}}{\sqrt{\lambda_2}} \end{aligned}$$

The points for S_1 and R_1 will be close to each other when $g_{11} \approx f_{11}$ and $g_{12} \approx f_{12}$, and a sufficient condition for this is $p_{i1} \approx 0$ for all $i \neq 1$, and $\lambda_1 \approx \lambda_2$. However, as the inspection e.g. of Table 3 and the corresponding Figure 7 (a) shows, this condition is by no means necessary.

The Correspondence Analysis of some experimental data yields results that correspond to the above interpretation of the scale values for stimuli and responses. When the stimuli are "simple" Gabor -patches, as described in section 4.2.1, the scale values for the stimuli are practically perfect linear functions of the spatial frequency parameter f_i defining the Gabor patch. This suggests that the expected values of the random variables representing the sensory activity are again linear functions of the f_i.

The situation may become more complicated when the stimuli are two-dimensional, as in section 5.3. The stimuli were constructed by an orthogonal variation of two components, an "inner" and an "outer" disc. During the first sessions, the subjects seem to try to evaluate the patterns with respect to a single dimension, representing the combination of the two components; in this case, most likely the ratio between the radii of inner and outer disc. After some sessions, the subjects seem to restructure their decisions, trying to evaluate each dimension separately. The usual learning curves - percent correct responses versus number of trial or session - do not reveal such restructuring processes; the biplots may, however, provide some insight into these processes.

The effects of additional, i.e. flanking patterns on the identification of stimuli may also be explored investigating the relations between the scale values and the parameters of the stimulus pattern, as in section 4.2.2. Flanking patterns seem to pull apart the scale values of stimuli; since the scale values are related to the χ^2-values of a confusion matrix this means that patterns are better discriminated when they are presented together with flanking patterns; for the patterns considered here, discrimination is improved the higher the value of the spatial frequency parameter of the flanking Gabor patches, see in particular Figure 14. Most likely the variances of the random variables representing the neural responses are reduced with increasing flanker parameter; more data are required before further interpretations concerning the underlying neuronal processes are suggested.

So far, only data from identification experiments have been considered; the collection of data from discrimination experiments is under way.

7 Appendix

7.1 Decomposition of variance

$$
\begin{aligned}
SS_{tot} &= \sum_{i=1}^{I}\sum_{k=1}^{n_{i+}}(s_{ik}-\overline{s})^2 = \sum_{i=1}^{I}\sum_{k=1}^{n_{i+}}(s_{ik}-\overline{s}_i+\overline{s}_i-\overline{s})^2 \\
&= \sum_{i=1}^{I}\sum_{k=1}^{n_{i+}}(s_{ik}-\overline{s}_i)^2 + \sum_{i=1}^{I} n_{i+}(\overline{s}_i-\overline{s})^2 - 2\sum_{i=1}^{I}\sum_{k=1}^{n_{i+}}(s_{ik}-\overline{s}_i)(\overline{s}_i-\overline{s}) \\
&= \sum_{i=1}^{I}\sum_{k=1}^{n_{i+}}(s_{ik}-\overline{s}_i)^2 + \sum_{i=1}^{I} n_{i+}(\overline{s}_i-\overline{s})^2 \\
&= SS_{wt} + SS_{bt},
\end{aligned}
$$

and $\sum_i\sum_k(\xi_{ik}-\overline{s}_i)(\overline{s}_i-\overline{s}) = \sum_i(\overline{s}_i-\overline{s})\sum_k(\xi_{ik}-\overline{s}_i) = 0$ since $\sum_k(\xi_{ik}-\overline{s}_i) = 0$. □

7.2 Proof of Theorem 3.1

Let μ be a Lagrange-multiplicator, and

$$Q(\bar{\mathbf{a}}) = \bar{\mathbf{a}}' K' D_{rs}^{-1} K \bar{\mathbf{a}} - \mu(\bar{\mathbf{a}}' D_{cs} \bar{\mathbf{a}} - S). \tag{76}$$

Let $\epsilon_j = (0, \cdots, 0, 1, 0, \cdots, 0)'$ be a vector whose components equal zero except for the j-th component, which is equal to 1. Then the vector $\mathbf{a}$ yielding a maximum value of ρ^2 relative to the value of SS_{tot} is the solution of

$$\begin{aligned}\frac{\partial Q}{\partial a_i} &= \epsilon_i' K' D_{rs}^{-1} K \mathbf{a} + \mathbf{a} k' D_{rs}^{-1} K \epsilon_j - \mu(\epsilon_j' D_{cs} \mathbf{a} + \mathbf{a} D_{cs} \epsilon_j) \\ &= 2K' D_{rs}^{-1} K \mathbf{a} - 2\,\mu D_{cs} \mathbf{a} = K' D_{rs}^{-1} K \mathbf{a} - \mu D_{cs} \mathbf{a} = 0\end{aligned}$$

i.e.

$$K' D_{rs}^{-1} K \mathbf{a} = \mu D_{rs} \mathbf{a} \tag{77}$$

If this equation is multiplied from the left by D_{rs}^{-1} one gets (25).

7.3 Proof of Theorem 3.2

Let $X = (x_{ij})$ an arbitrary $m \times n$-matrix, i.e. $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, $x_{ij} \in \mathbb{R}$, with rank $r \leq \min(m, n)$. Let $X_1, X_2, \ldots, X_n$ denote the column vectors of X. The X_j can be represented as linear combinations of r linear independent, in particular orthogonal, n-dimensional basis vectors. Equivalently, the row vectors can be represented as lineqar combinations of r lineqar independent, in particular orthogonal, m-dimensional basis vectors. Let $L_1, \ldots, L_r$ be r orthogonal, n-dimensional vectors. There exist r coefficients $v_{j1}, \ldots, v_{jr}$ such that

$$X_j = v_{j1} L_1 + \cdots + v_{jr} L_r, \quad j = 1, 2, \ldots, n. \tag{78}$$

This is equivalent to writing

$$X = UV', \tag{79}$$

where $U = [U_1, \ldots, U_r]$, $V = (v_{kj})$, with $k = 1, 2, \ldots, r$ and of course $j = 1, 2, \ldots, n$. From the postulated orthogonality of L one finds immediately

$$X'X = VU'UV' = V\Lambda V, \quad U'U = \Lambda = \begin{pmatrix} \lambda_1, & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \vdots & 0 \\ 0 & 0 & \cdots & \lambda_r \end{pmatrix}, \tag{80}$$

and λ_k is the square of the length of the basis vector L_k. the equation $X'X = V\Lambda V'$ suggests that the column vectors of A are the eigenvectors of $X'X$, and the λ_k are the corresponding eigenvalues. The roots $\sqrt{\lambda_k}$, i.e. the lengths of the L_k, are also known as singular values of $X'X$.

Multiplying the U_k with $1/\sqrt{\lambda_k}$ means to norm the L_k, i.e. $Q = U\Lambda^{-1/2}$ contains the orthonormal basis vectors, and certainly $U = Q\Lambda^{-1/2}$. The eigenvectors of $X'X$

are also known to be orthogonal, and without loss of generality can be postulated to be normalised. Consequently one may re-write (79) as

$$X = U\Lambda^{1/2}V'. \tag{81}$$

Further, since $UU' = V\Lambda^{1/2}\Lambda^{1/2}V' = U\Lambda U'$, U must contain the orthogoonal eigenvectors of XX', corresponding to the nonzero eigenvalues of XX' (equal to those of $X'X$).

7.4 Proof of Theorem 3.3

According to (102), $(D_c^{-1}P')F = G\Lambda^{1/2}$. Multiplication from the left with $D_r^{-1}P$ yields

$$D_r^{-1}P(D_c^{-1}P')F = D_r^{-1}PG\Lambda^{1/2}.$$

But according to (101) one has $D_r^{-1/2}PG = F\Lambda^{1/2}$, so that

$$(D_r^{-1}PD_c^{-1}P')F = F\Lambda \tag{82}$$

follows. Therefore the coordinates F are given as the eigenvectors of the matrix $D_r^{-1}P(D_c^{-1}P'$. The validity of

$$(D_c^{-1}P'D_r^{-1}P)G = G\Lambda \tag{83}$$

is shown analogously. □

7.5 Proof of Theorem 3.4

Proof: $F = D_r^{-1/2}U\Lambda^{1/2}$ implies

$$f_{is} = u_{is}\sqrt{\lambda_s}/\sqrt{r_i},$$

where u_{is} is the element in the i-th row and the s-th column of U. It follows that

$$FV' = D_r^{-1/2}U\Lambda^{1/2}V' = D_r^{-1/2}X.$$

Then

$$\frac{x_{ij}}{\sqrt{r_i}} = \sum_{s=1}^{r} f_{is}v_{js}$$

and

$$\frac{x_{kj}}{\sqrt{c_k}} = \sum_{s=1}^{r} f_{ks}v_{js}$$

so that

$$\frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{c_k}} = \sum_{s=1}^{r}(f_{is} - f_{ks})v_{js}.$$

But

$$\begin{aligned}
\frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{c_k}} &= \frac{1}{\sqrt{r_i}} \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} - \frac{1}{\sqrt{c_k}} \frac{p_{kj} - c_k c_j}{\sqrt{c_k c_j}} \\
&= \frac{1}{\sqrt{c_j}} \left(\frac{p_{ij}}{r_i} - c_j \right) - \frac{1}{\sqrt{c_j}} \left(\frac{p_{kj}}{c_k} - c_j \right) \\
&= \frac{1}{\sqrt{c_j}} \left(\frac{p_{ij}}{r_i} - \frac{p_{ik}}{c_k} \right), \qquad (84)
\end{aligned}$$

so that

$$\sum_{j=1}^{J} \left(\frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{c_k}} \right)^2 = \sum_{j=1}^{J} \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - \frac{p_{kj}}{c_k} \right)^2 = \delta_{ik}^2 \qquad (85)$$

and therefore

$$\begin{aligned}
\delta_{ik}^2 &= \sum_{j=1}^{J} \left(\sum_{s=1}^{r} (f_{is} - f_{ks}) v_{js} \right)^2 \\
&= \sum_{j=1}^{J} v_{js}^2 \sum_{s=1}^{r} (f_{is} - f_{ks})^2 \\
&+ 2 \sum_{j=1}^{J} \sum_{s<s'} (f_{is} - f_{ks})(f_{is'} - f_{ks'}) v_{js} v_{js'}. \qquad (86)
\end{aligned}$$

But

$$\begin{aligned}
& 2 \sum_{j=1}^{J} \sum_{s<s'} (f_{is} - f_{ks})(f_{is'} - f_{ks'}) v_{js} v_{js'} \\
= \; & 2 \sum_{s<s'} [(f_{is} - f_{ks})(f_{is'} - f_{ks'}) \sum_{j=1}^{J} v_{js} v_{js'}] = 0
\end{aligned}$$

and $\sum_j v_{js} v_{js'} = 0$ because of the orthogonality of the eigenvectors B_s and $B_{s'}$. Further one has $\sum_j v_{js}^2 = 1$, because the eigenvectors are normalised, i.e. have the lengths 1. Therefore one has on the left of (86) the χ^2-distance between the i-th and the k-th row, and on the right one has the corresponding Eucledian distance defined by the coordinates in F. □

7.6 Proof of Theorem 3.5

We need the following relations:

$$\begin{aligned}
r'F &= 0 \qquad (87) \\
c'G &= 0. \qquad (88)
\end{aligned}$$

Proof: Let $\mathbf{1} = (1, 1, \ldots, 1)'$. the number of components of $\mathbf{1}$ will depend upon the equation with respect to which $\mathbf{1}$ is used.

Certainly, one has

$$D_r^{-1} r = \mathbf{1}, \quad D_c^{-1} c = \mathbf{1}, \tag{89}$$

which is easily illustrated for the example $J = 2$:

$$\begin{pmatrix} 1/r_1, & 0 \\ 0, & 1/r_2 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Further,

$$r'\mathbf{1} = \sum_i r_i = 1, \quad c'\mathbf{1} = \sum_j c_j = 1. \tag{90}$$

These equations follow immediately from the definition of the r_i and the c_j.

The SVD of T is given by $T = D_r^{-1/2}(P - rc')D_c^{-1/2} = U\Lambda^{1/2}V'$. Therefore

$$P - rc' = D_r^{1/2} U \Lambda^{1/2} V' D_c^{1/2}. \tag{91}$$

Let

$$A = D_r^{1/2} U, \quad B = D_c^{1/2} V, \tag{92}$$

so that (91) can be written in the form

$$P - rc' = A\Lambda^{1/2}B'; \tag{93}$$

this expression is also known as the generalised SVD of the matrix $P - rc'$.

Since U and V are orthonormal it follows that $U'U = I$, $V'V = I$, I the unity matrix. But $U = D_r^{-1/2}A$, $V = D_c^{-1/2}B$. Then

$$U'U = A'D_r^{-1}A = I, \quad V'V = B'D_c^{-1}B = I, \tag{94}$$

To prove (87) und (88) it should be noted that F may be written in the form $F = D_r^{-1}(P - rc')D_c^{-1}B$: if $A\Lambda^{1/2}B'$ substituted for $P - rc'$ one gets

$$F = D_r^{-1}A\Lambda^{1/2}B'D_c^{-1}B = D_r^{-1}A\Lambda^{1/2}.$$

On the other hand one has

$$D_r^{-1}(P - rc')D_c^{-1}B = (D_r^{-1}P - D_r^{-1}rc')D_c^{-1}B = (D_r^{-1}P - \mathbf{1}c')D_c^{-1}B.$$

this yields

$$F = (D_r^{-1}P - \mathbf{1}c')D_c^{-1}B \tag{95}$$

$$G = (D_c^{-1}P' - \mathbf{1}r')D_r^{-1}A, \tag{96}$$

where the expression for G was derived in an analogous way. Then it follows that

$$r'F = r'(D_r^{-1}P - \mathbf{1}c')D_c^{-1}B.$$

But it is $r'(D_r^{-1}P - \mathbf{1}c') = \mathbf{1}'P - r'\mathbf{1}c'$, and further one has $\mathbf{1}'P = c'$, $r'\mathbf{1} = 1$, so that $\mathbf{1}'P - r'\mathbf{1}c' = 0$, according to (90). Therefore it follows that $r'F = 0$. $c'G = 0$ is shown in an analogous way.

□

The relation between coordinates

Proof: $P - rc' = A\Lambda^{1/2}B'$ implies

$$D_r^{-1}(P - rc')D_c^{-1} = D_r^{-1}A\Lambda^{1/2}B'D_c^{-1}, \tag{97}$$

or, in transposed form,

$$D_c^{1/2}(P' - cr')D_r^{1/2} = D_c^{-1}B\Lambda^{1/2}A'D_r^{-1}. \tag{98}$$

Because of $F = D_r^{-1}A\Lambda^{1/2}$, $G = D_c^{-1}B\Lambda^{1/2}$ this yields

$$D_r^{-1}PD_c^{-1} - D_r^{-1}rc'D_c^{-1} = FB'D_c^{-1}, \tag{99}$$

and on the other hand

$$D_c^{-1}P'D_r^{-1} - D_c^{-1}cr'D_r^{-1} = GA'D_r^{-1}. \tag{100}$$

If (99) is multiplied from the right with $B\Lambda^{1/2}$ and (100) with $A\Lambda^{1/2}$, one gets because of (94)

$$\begin{aligned} D_r^{-1}PD_c^{-1}B\Lambda^{1/2} - D_r^{-1}rc'D_c^{-1}B\Lambda^{1/2} &= F\Lambda^{1/2} \\ D_c^{-1}P'D_r^{-1}A\Lambda^{1/2} - D_c^{-1}cr'D_r^{-1}A\Lambda^{1/2} &= G\Lambda^{1/2}. \end{aligned}$$

But $D_c^{-1}B\Lambda^{1/2} = G$, and $D_r^{-1}A\Lambda^{1/2} = F$, and $c'G = 0$, $r'F = 0$, so that

$$D_r^{-1}PG = F\Lambda^{1/2} \tag{101}$$

$$D_c^{-1}P'F = G\Lambda^{1/2} \tag{102}$$

follows. Multiplication from the right with $\Lambda^{-1/2}$ yields the equations (54) and (53). □

References

[1] Ashby, G.F. & Townsend, J.T. Varieties of perceptual independence. *Psychological Review*, 1986, 93, 154-179

[2] Ashby, F.G. & Gott, R.E. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1988, 14, 33-53

[3] Balakrishnan, J.D. (1998) Some more sensitive measures of sensitivity and resonse bias. *Psychological Methods, 3*, 68-90

[4] Balakrishnan, J.D., Ratcliff, R. (1996) Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 615-633

[5] Balakrishnan, J.D. (1999) Decision Processes in discrimination: Fundamental misrepresentations of Signal Detection Theory. *Journal of Experimental Psychology, 25*, 1189-1206

[6] Bishop, C.M., Neural networks for pattern recognition. Clarendon Press, Oxford 1995

[7] Bradley, A., Skottun, B., Ohzawa, I., Sclar, G., Freeman, A.D. (1987) Visual orientation and spatial frequency discrimination: A comparison of single neurons and behavior. J. Neurophysiol. 57: 755-772

[8] Britten, K.H., Shadlen, M.N., Newsome, W.T., Movshon, J.A. (1993) Responses of neurons in macaque MT to stochastic motion signals. Vis. Neuroscience, 10:1157-1169

[9] Burgess, A.E., Ghandeharian, H.: Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation. *J. Opt. Soc. Am. A 2*, 1985, 1498 – 1507

[10] Tanner, Jr., W. P., Swets, J.A. (1954) A decision-making theory of visual attention. *Psychological Review, 61*, 410-409

[11] Busemeyer, J.R., Townsend, J.T. (1993). Decision Field Theory: A dynamic cognition approach to decision making. *Psychological Review, 100*, 432-459

[12] Caelli, T., Rentschler, I., Scheidler, W. (1987) Visual pattern recognition in humans. I. Evidence for adaptive filtering. *Biological Cybernetics*, 57, 233-240

[13] Dean, A. F. (1981) The variability of discharge of simple cells of cat striate cortex. Exp. Brain Res. 44: 437-440

[14] Diederich, A: (1997) Dynmic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology, 41*, 260-274

[15] Gershon, E.D., Wiener, M.C., Latham, P.E., Richmond, B.J. (1998) Coding strategies in monkey V1 and inferior temporal cortices. J. Neurophysiol. 79, 1135-1144

[16] Guttman, L. (1941) The quantification of of class attributes: A theory and method of scale construction. In Horst, P. (ed.) The prediction of personal adjustment. Social Science Research Council, New York, 319-348,

[17] Greenacre, M.J. Theory and applications of correspondence analysis. *Academic Press*, London, 1983

[18] Hauske, G., Wolf, Lupp, U. (1976) Matched filters in human vision. *Biolog. Cybern.*, 22, 181 – 188

[19] Jüttner, M., Caelli, T. Rentschler, I. (1997) Evidence-based pattern classification: a strucutal approach to human perceptual learning and generalization. *Journal of Mathematical Psychology*, 41, 144-259

[20] Kadlec, H., Townsend, J.T. (1992) Implications of marginal and conditional detection parameters for the separabilities and independence of perceptual dimensions. *Journal of Mathematical Psychology*, 36, 325-374

[21] Lancaster, H.O. (1957) Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44, 289

[22] Link, S.W., Heat, R.A. (1975) A sequential theory of psychological discrimination. *Psychometrica, 40*, 77-105

[23] Luce, R.D., Individual choice behavior. A theoretical analysis. John Wiley & sons, New York 1959

[24] Luce, R.D., Detection and recognition. In: Luce, R.D., Bush, R.R. and Galanter, E. (eds.) Handbook of Mathematical Psychology, Vol. 1, Wiley, New York 1963

[25] Maloney, L.T., Thomas, E.A.C. (1991) Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology, 35*, 443-470

[26] Meinhardt, G., Mortensen, U. (1998) Detection of aperiodic test patterns by pattern specific detectors revealed by subthreshold summation. *Biol. Cybern.*, 79, 413 – 425

[27] Nishisato, S. (1980) Analysis of categorical data: Dual scaling and its applications. Toronto University Press, Toronto

[28] Nosofsky, R.M. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 1986, 115, 39-57

[29] Rentschler, I., Jüttner, M., Caelli, T. (1993) Probabilistic analysis of human supervised learning and classification. *Vision Research*, ??, 669-687

[30] Rentschler, I., Barth, E., Caelli, T., Zetsche, C., Jüttner, M.. (1996) Generalization of form in visual pattern classification. *Spatial Vision*, 10, 59-85

[31] Ripley, B. D., Pattern recognition and neural networks. *Cambridge University Press*, 1996

[32] Snowden, R.J., Treue, S., Andersen, R.A. (1992) The response of neurons in areas V1 and MT of the rhesus monkey to moving random dot patterns. Exp. Brain Res. 88:3898-400

[33] Swets, J.A. (1986a) Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99*, 110-117

[34] Swets, J.A. (1986b) From empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin, 99*, 191-198

[35] Townsend, J.T., Theoretical analysis of an alphabetic confusion matrix. *Perception and Psychophysics*, 1971 (a), 9, 40 - 50

[36] Townsend, J.T., Alphabetic confusion: A test of models for individuals. *Perception and Psychophysics*, 1971 (b), 9, 449-454

[37] Townsend, J.T., & Ashby, G.F. Experimental test of contemporary mathematical models of visual letter recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 1982, 8, 834-864

[38] Townsend, J.T. & Landon, D.E. Mathematical models of recognition and confusion in psychology. *Mathematical Social Sciences*,1983, 4, 25-71

[39] Townsend, J.T., Hu, G.G., & Ashby, F.G. A test of visual feature sampling independence with orthogonal straight lines. *Bulletin of the Psychonomic Society*, 1980, 15, 163-166