PSYCHOMETRIKA—VOL. 61, NO. 1, 7–29
MARCH 1996

PROBABILITY MATRIX DECOMPOSITION MODELS

ERIC MARIS

UNIVERSITY OF NIJMEGEN

PAUL DE BOECK AND IVEN VAN MECHELEN

UNIVERSITY OF LEUVEN

In this paper, we consider a class of models for two-way matrices with binary entries of 0 and 1. First, we consider *Boolean matrix decomposition*, conceptualize it as a *latent response model* (LRM) and, by making use of this conceptualization, generalize it to a larger class of matrix decomposition models. Second, *probability matrix decomposition* (PMD) models are introduced as a probabilistic version of this larger class of deterministic matrix decomposition models. Third, an algorithm for the computation of the maximum likelihood (ML) and the maximum a posteriori (MAP) estimates of the parameters of PMD models is presented. This algorithm is an EM-algorithm, and is a special case of a more general algorithm that can be used for the whole class of LRMs. And fourth, as an example, a PMD model is applied to data on decision making in psychiatric diagnosis.

Key words: Boolean matrix decomposition, latent response model, clustering, two-way data, incomplete data, EM-algorithm, psychiatric diagnosis.

Within the domain of data analysis binary data have often taken a special place. This paper deals with a collection of models for binary data, which in the simplest case will be two-way two-mode (i.e., a binary matrix). Throughout this paper, the first mode will be referred to as *objects* and the second mode as *attributes*. Depending on the substantive context, the objects may be thought to denote persons, situations, stimuli, etc., and the attributes person characteristics, intelligence items, variables, etcetera. The data then indicate whether a person has a given characteristic, etcetera. This paper will also consider models for a slightly more complex type of data that arises when a binary two-way two-mode data set is extended with a replication mode (which can be conceived as a two-way matrix with multiple observations per cell).

Given such data, one may be interested in formal models that reveal the mechanisms according to which the data have come about. Quite a natural class of mechanisms that may be considered in this respect accounts for the data on the basis of the interplay of certain properties or events at the level of the objects, on the one hand, and properties or events at the level of the attributes, on the other hand. Existing models that make use of such mechanisms are the models of nonmetric factor analysis (Coombs, 1964), the Rasch model and various other models from Item Response Theory. The representation (i.e., the *properties or events* mentioned above) on which these models are based can be considered as *geometrical*: the objects and the attributes are represented as points in a space (possibly unidimensional as in the case of the Rasch model) whose geometrical relations determine (the probability of) the response. For nonmetric factor analysis, these geometrical relations are dominance relations between

This paper is based on a chapter of the first author's doctoral dissertation, written at the University of Leuven and supervised by Paul De Boeck.

Requests for reprints should be sent to Eric Maris, Nijmegen Institute for Cognition and Information, Department of Mathematical Psychology, University of Nijmegen, PO Box 9104, 6500 HE Nijmegen, THE NETHERLANDS (E-mail: U212776@VM.UCI.KUN.NL).

0033-3123/96/0300-94208$00.75/0

the coordinates of the objects and the attributes on each of the dimensions of this space. For the Rasch model, we also have a dominance relation, but this time it is with respect to points on a single line (i.e., the person and the item parameter) and it does not determine the response in an all-or-none fashion but probabilistically. And in the item factor analysis model (see Bock & Aitkin, 1981) the probability of a response depends on the position of the person and the item in the space through the scalar product of their coordinate vectors.

A distinctive characteristic of the models to be presented in this paper is that the properties or events they include (i.e., their representation) have the same nature as the data they intend to account for, that is, they are binary. The substantive relevance of such type of models can be exemplified by the case of successes of persons in intelligence items one wants to explain in terms of unobserved strategies that a person masters/does not master and via which an item can/cannot be solved. Another example are person by choice alternative (pick any out of n) data one wants to explain on the basis of latent requisites that a person poses/does not pose and that a choice alternative meets/does not meet.

Throughout this paper, the unobserved properties/events at the level of the objects and attributes will be referred to with the generic term of *latent responses*, and the modeling of data in terms of latent responses from the object side and from the attribute side will be referred to as *matrix decomposition*. Two cases of matrix decomposition will be considered: a first case in which the latent responses are constants, which will lead to deterministic models, and a second case in which the responses are random variables, which will lead to probabilistic models. The deterministic models are generalizations of the known models of Boolean factor analysis (Mickey, Mundle & Engelman, 1983) and hierarchical classes analysis (De Boeck & Rosenberg, 1988; Van Mechelen, De Boeck & Rosenberg, 1995). The probabilistic models are novel.

In the following, we begin by introducing the basic ideas of Boolean matrix decomposition by means of an example. We then show how these ideas can be generalized to a larger class of matrix decomposition models. Next, *probability matrix decomposition* (PMD) models are introduced as a probabilistic version of this larger class of deterministic matrix decomposition models. Then, an algorithm for the computation of the maximum likelihood (ML) and the maximum a posteriori (MAP) estimates of the parameters of PMD models is presented. Finally, as an example, a PMD model is applied to data on decision making in psychiatric diagnosis.

Basic Ideas

Example

We consider a hypothetical study on decision making in interpersonal relations. Fourteen subjects participated in this study. They were shown video recordings of 12 other persons presenting themselves in some standardized way (e.g., by giving a sketch of a typical week in their daily life, and telling about the most happy and most sad period in their lives). After each person's presentation, a subject had to indicate whether or not he or she would like this person as a member of a group with which he or she would spend a hypothetical one-week holiday. The data of this study can be presented in a 14-by-12 array, which is shown in Table 1(a).

The structure in this matrix can be understood in terms of some characteristics of the persons presenting themselves (further denoted as *candidates*), namely those that are relevant for the subject's judgments. In particular, it is assumed that each of the subjects has based his or her judgment on a subset of the following three person

TABLE 1

Data Matrix of a Hypothetical Study on Decision Making in Interpersonal Relations and its Numerical Representation

a. *Data Matrix*

Subject	*Candidate*											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Paul	1	1	0	0	0	0	1	1	0	0	0	0
Mary	1	1	0	0	0	0	1	1	0	0	0	0
Rebecca	1	1	0	0	0	0	1	1	0	0	0	0
Donald	0	0	1	1	1	1	1	1	0	0	1	1
Clyde	0	0	1	1	1	1	1	1	0	0	1	1
Fred	1	1	1	1	1	1	1	1	0	0	1	1
Susan	1	1	1	1	1	1	1	1	0	0	1	1
David	0	0	0	0	0	0	0	0	1	1	1	1
Jacky	0	0	0	0	0	0	0	0	1	1	1	1
Ellen	0	0	1	1	1	1	1	1	1	1	1	1
Jim	0	0	1	1	1	1	1	1	1	1	1	1
Lucy	0	0	1	1	1	1	1	1	1	1	1	1
Robert	1	1	1	1	1	1	1	1	1	1	1	1
George	0	0	0	0	0	0	0	0	0	0	0	0

b. *Numerical Representation*

Subject	*Bundle*		
	1	*2*	*3*
Paul	1	0	0
Mary	1	0	0
Rebecca	1	0	0
Donald	0	1	0
Clyde	0	1	0
Fred	1	1	0
Susan	1	1	0
David	0	0	1
Jacky	0	0	1
Ellen	0	1	1
Jim	0	1	1
Lucy	0	1	1
Robert	1	1	1
George	0	0	0

Candidate	*Bundle*		
	1	*2*	*3*
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	1	0
6	0	1	0
7	1	1	0
8	1	1	0
9	0	0	1
10	0	0	1
11	0	1	1
12	0	1	1

characteristics: (a) other-orientedness (i.e., being primarily oriented towards the needs, feelings, and thoughts of the other person, instead of ones own), (b) interestingness (i.e., being able to keep the other person interested in a conversation, and in the relation in general), and (c) physical attractiveness. Each of the subjects can be characterized in terms of the subset of these three person characteristics they want to see realized in a candidate in order to choose him or her as a member of his or her holiday group. Each of the candidates can be characterized in terms of the subset of the same three person characteristics that applies to him or her.

Apart from the characteristics on which the subjects base their judgments, there is also a *decision rule* to be specified. One rule is that a candidate must have *at least one* of the characteristics he or she considers important (belonging to his or her subset). Other rules are possible, and in the following we will consider some of them.

This way of describing the structure in the data can be represented numerically. The numerical representation of the subjects is shown in the left-hand side of Table 1(b). It involves that each of the subjects is assigned a binary vector containing as many elements as there are relevant person characteristics. The elements correspond to other-orientedness, interestingness, and physical attractiveness, respectively. If a person characteristic is relevant for the subject's judgments, the corresponding element takes the value 1, 0 otherwise.

The numerical representation of the candidates is shown in the right-hand side of Table 1(b). The elements of the binary vectors that are assigned to the candidates take the value 1 if the corresponding person characteristic applies to the candidate, 0 otherwise (the order of the elements is the same as for the subjects).

At this point, it is convenient to introduce some terminology and notation. We will refer to the row entries of the data matrix (shown in Table 1(a)) as the *objects*, and to the column entries as the *attributes*. The number of objects and attributes will be denoted by, respectively, O and A. The data matrix will be denoted by $\mathbf{M}$ and its elements by M_{oa} ($o = 1, \ldots, O$, and $a = 1, \ldots, A$). The two sets of binary vectors constituting the numerical representation can be considered as the row vectors of two matrices (one for the objects, and one for the attributes). The columns of these matrices are called *bundles*, and the matrices themselves *bundle matrices*. The number of bundles will be denoted by B. The bundle matrices for the objects and the attributes will be denoted by, respectively, $\mathbf{S}$ and $\mathbf{P}$, and their elements by, respectively, S_{ob} and P_{ab} ($o = 1, \ldots, O$, $a = 1, \ldots, A$, and $b = 1, \ldots, B$). If $S_{ob}(P_{ab})$ equals 1, we say that the o-th object *belongs* to the b-th bundle. The o-th row of $\mathbf{S}$ will be denoted by $\mathbf{S}_o^t = (S_{o1}, \ldots, S_{oB})$, and the a-th row of $\mathbf{P}$ by $\mathbf{P}_a^t = (P_{a1}, \ldots, P_{aB})$.

Now, it is possible to describe how the bundle matrices are related to the data matrix. Every element M_{oa} in $\mathbf{M}$ is related to the column vector $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$ in a way that is determined by the subjects' decision rule. In particular, the rule that the candidate must have at least one of the characteristics the subject considers important, leads to the *Boolean scalar product* as the function relating M_{oa} and $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$. The Boolean scalar product will be denoted by v, and is defined as follows:

$$v(\mathbf{S}_o, \mathbf{P}_a) = \bigoplus_{b=1}^{B} (S_{ob} \times P_{ab}), \tag{1}$$

in which $\oplus$ denotes the Boolean sum, which has function value 0 if all terms are 0 and 1 otherwise. It is clear that $v(\mathbf{S}_o, \mathbf{P}_a)$ equals 1 *iff* there is at least one b for which $S_{ob} = P_{ab} = 1$, and 0 otherwise. Now, the relation between some M_{oa} and the corresponding vector $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$ is simply the following:

$$M_{oa} = v(\mathbf{S}_o, \mathbf{P}_a).$$

This relation can be expressed for the matrix **M** as a whole in the following way:

$$\mathbf{M} = \mathbf{S} \otimes \mathbf{P}^t, \tag{2}$$

in which $\otimes$ denotes the Boolean matrix product. The Boolean matrix product is defined the same way as the ordinary matrix product, except for the fact that the familiar scalar product is replaced by the Boolean scalar product. Equation (2) is the characteristic equation of *Boolean matrix decomposition*. Boolean matrix decomposition is the problem of finding two matrices **S** and **P** (with a minimal number of bundles) such that (2) holds.[1] This type of model has been presented by De Boeck and Rosenberg (1988); see also Van Mechelen and De Boeck (1990).

As an aside, it can be mentioned that the matrix **M** can be considered both as the data matrix itself or as a *model* for the data matrix. The advantage of considering it as a model is that it allows one to deal with (a) deviances between the data (which may be denoted by **D**, for example) and the model and (b) the fact that the observations in some cells may be missing.

Now, assume that the subjects use a different decision rule. In particular, assume that in order for a subject to choose a candidate as member of his or her group, the latter must have *all* the characteristics the subject considers important. This decision rule leads to a different function relating the M_{oa}'s and the $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$'s. In particular, it leads to the following:

$$v(\mathbf{S}_o, \mathbf{P}_a) = \prod_{b=1}^{B} \{1 - [S_{ob} \times (1 - P_{ab})]\}. \tag{3}$$

This function has function value 1 *iff* P_{ab} is larger than or equal to S_{ob} for all b, and 0 otherwise.

By introducing an alternative for the Boolean scalar product, we have in fact defined a new type of matrix decomposition. In particular, referring to (2), we only have to consider $\otimes$ as a symbol that denotes a new type of matrix product, namely one in which the Boolean scalar product is replaced by the function defined in (3). In this way, Boolean matrix decomposition can be generalized by simply defining alternatives for the Boolean scalar product.

Latent Response Models

The way the structure in **M** was formally described in the previous section can be looked upon from the distinction between latent and observed responses.[2] The S_{ob}'s and P_{ab}'s are the latent responses, and they *explain* the observed responses, the M_{oa}'s, as formally specified by the Boolean scalar product or its alternative defined in (3).

The basic idea involved in this conceptualization of Boolean matrix decomposition can be easily generalized to a much broader class of models, both with respect to the type of data to which they apply (categorical/continuous, scalar/vector-valued) and the type of structure they impose on the data. This class of models are the so-called *latent response models* (LRMs) (see also, Maris, 1992, 1995). In the following, we will give a formal definition of LRMs and will show how Boolean matrix decomposition and its generalizations are special cases.

[1] It should be noted here that in Kim's (1982) book on Boolean matrix theory, the term *decomposition* is used in a different sense.

[2] Were it not for possible confusion with *latent variable models*, the term *responses* could be replaced by *variables*.

Presenting these matrix decompositions within the framework of LRMs has the advantage that they can be considered as formalizations of the psychological structure or process that generates the observations (as was already shown implicitely in the example). Besides this, there are two more advantages of this way of presenting these matrix decompositions. First, it shows how a probabilistic version of these matrix decompositions can be formulated in a straightforward way, because deterministic and probabilistic LRMs are related in a simple and well-specified way. In this section, we will only consider deterministic LRMs. Probabilistic LRMs will be considered in the next section. A second advantage is that, for the probabilistic versions of the matrix decomposition models, we can make use of a general algorithm for the computation of the maximum likelihood (ML) and maximum a posteriori (MAP) estimates of the parameters of probabilistic LRMs.

A LRM is a model for a set of N response variables Y_n ($n = 1, \ldots, N$) in which each variable is explained on the basis of (a) a number of unobserved underlying variables that can be conceived of as latent responses, and (b) a rule to combine these latent responses. For the models that are considered here, N always equals $O \times A$, the number of objects times the number of attributes, and every n is associated with a particular object-attribute pair (o, a). In particular, for these models, every Y_n corresponds to one element M_{oa} in the matrix $\mathbf{M}$.

The definition of LRMs involves two key aspects: (a) the definition of *latent response variables*, and (b) the definition of a *condensation rule*. So, first, for every Y_n, a set of K latent variables X_{nk} ($k = 1, \ldots, K$) is defined. In vector notation, we write $\mathbf{X}_n = (X_{n1}, \ldots, X_{nK})^t$. In principle, not every X_{nk} has to be defined for a particular Y_n. However, in order to keep the notation simple, it is assumed that for every Y_n a complete vector $\mathbf{X}_n$ can be defined. For the matrix decomposition models that will be considered here, $\mathbf{X}_n$ is the $(2 \times B)$-dimensional vector $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$.

Second, a condensation rule is defined as a function that specifies the relationship between Y_n and $\mathbf{X}_n$. Using C as a generic symbol for a condensation rule, this relationship can be expressed as follows:

$$Y_n = C(\mathbf{X}_n). \tag{4}$$

This function may be different for different values of n, but in order to keep our notation simple, we will not index C. Instead, we let its argument (with n being some particular value) indicate its particular form. In the following, a class of matrix decomposition models will be presented by means of the condensation rules that are involved in them. In fact, these condensation rules are simply different modifications of the Boolean scalar product.

Loosely speaking, a LRM starts from a conceptualization of the unobserved (latent) process that underlies the observed response. This unobserved process is described in terms of latent variables whose values determine the observed response in an all-or-none fashion. By giving a particular psychological interpretation to these latent variables, LRMs are well suited for testing hypotheses about the psychological process involved in the coming about of the observed response.

At this point, we do not yet have a model for the Y_n's. In the deterministic case, the model for the Y_n's assumes that the values of these variables are constants. This model is a set of functions (one for each Y_n, but possibly identical) expressing the relationship between the Y_n's and some unknown parameters. In the probabilistic case, which will be considered in the next section, the model assumes that the Y_n's are *random* variables whose values are *not* constant (which is expressed by saying that a random variable has multiple *realizations*). And their model is a set of *probability*

distribution functions (PDFs) (one for each Y_n, but possibly identical) that depend on some unknown parameters.

The parameters will be denoted by θ_t ($t = 1, \ldots, T$). In vector notation, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)^t$. No distinction is made between different types of parameters, as for example, parameters that are associated with the objects, and parameters that are associated with the attributes. Now, the model for the Y_n's in the deterministic case can be expressed as follows:

$$Y_n = f_n(\boldsymbol{\theta}), \tag{5}$$

in which f_n is some function. At first sight, this equation may be confusing because in (4) Y_n is expressed as a function of $\boldsymbol{\theta}$. However, the relationship between these two equations is established by the following essential feature of LRMs: given a particular condensation rule, the model for the Y_n's is completely specified by the model for the $\mathbf{X}_n$'s. The model for the $\mathbf{X}_n$'s (in the deterministic case) will be expressed as follows:

$$\mathbf{X}_n = g_n(\boldsymbol{\theta}), \tag{6}$$

in which g_n is some function. If $K > 1$, g_n is in fact a set of K more elementary functions, which can be denoted by g_{nk} ($k = 1, \ldots, K$). Now, equations (4)–(6) can be combined as follows:

$$Y_n = f_n(\boldsymbol{\theta}) = C(\mathbf{X}_n) = C[g_n(\boldsymbol{\theta})].$$

Boolean Matrix Decomposition and its Generalizations as LRMs

We will now show how Boolean matrix decomposition and its generalizations can be considered as LRM's. For the matrix decomposition models, it holds that Y_n and $\mathbf{X}_n$ can be replaced by M_{oa} and $\mathbf{X}_{oa}$, respectively. Moreover, M_{oa} is a binary scalar, and $\mathbf{X}_{oa}$ is a $(2 \times B)$-dimensional binary vector.

The model for the $\mathbf{X}_{oa}$'s is the same for all matrix decomposition models; the difference between them is determined by a difference in condensation rules. The parameters of the model for the $\mathbf{X}_{oa}$'s are the two matrices $\mathbf{S}$ and $\mathbf{P}$ ($\boldsymbol{\theta} = (\mathbf{S}, \mathbf{P})$). The function values of the g_{oa}'s are $(2 \times B)$-dimensional, and are defined as follows:

$$g_{oa}(\mathbf{S}, \mathbf{P}) = (S_{o1}, \ldots, S_{oB}, P_{a1}, \ldots, P_{aB})^t.$$

$$= (\mathbf{S}_o^t, \mathbf{P}_a^t)^t.$$

Through the relation $\mathbf{X}_{oa} = g_{oa}(\mathbf{S}, \mathbf{P})$ the $O \times A \times (2 \times B)$ latent responses are expressed as a function of $(O + A) \times B$ parameters. It is instructive to stress the fact that when we write $\mathbf{X}_{oa} = (\mathbf{S}_o^t, \mathbf{P}_a^t)^t$, as we have done previously, we have in fact already adopted a model for the $\mathbf{X}_{oa}$'s.

We will now present a class of matrix decomposition models by specifying modifications of the Boolean scalar product, which, in the framework of LRMs, are simply alternative condensation rules. They all involve a mapping from a $(2 \times B)$-dimensional Boolean space into a 1-dimensional one. In presenting these condensation rules, we will use a notation that is specific for the model chosen for the $\mathbf{X}_{oa}$'s. In particular, we replace $\mathbf{X}_{oa}$ by $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$.

The first condensation rule is the Boolean scalar product. Here, we define it in a way that can be more easily interpreted in psychological terms than in the previous section. In particular,

$$v(\mathbf{S}_o, \mathbf{P}_a) = 1 \quad \text{if } \exists\ b\text{: } (S_{ob} = 1) \wedge (P_{ab} = 1)$$
$$= 0 \quad \text{otherwise,} \tag{7}$$

in which $\wedge$ denotes the logical *and*. Both for psychological interpretation and estimation it is useful to consider (7) as a condensation rule consisting of two steps. In the first step, for each of the B bundles, it is determined whether the condition $(S_{ob} = 1) \wedge (P_{ab} = 1)$ holds. Whether or not this condition holds, is indicated by the variable U_{oab}, which has the value 1 if the condition holds, and 0 otherwise. U_{oab} can be considered as a variable that indicates whether or not object and attribute have something *in common*. In the second step, it is determined whether there is a bundle for which U_{oab} equals 1. The rule involved here, is the *disjunctive* one (one U_{oab} being 1 is enough for the vector product to be 1 also). The complete condensation rule is denoted as *disjunctive communality*.

The second condensation rule is the one we proposed as an alternative for disjunctive communality in the previous section. It is defined as follows:

$$v(\mathbf{S}_o, \mathbf{P}_a) = 1 \quad \text{if } \forall\ b\text{: } P_{ab} \geq S_{ob}$$
$$= 0 \quad \text{otherwise.}$$

Again, this condensation rule can be considered as consisting of two steps. In the first step, for each of the bundles, it is determined whether the condition $P_{ab} \geq S_{ob}$ holds. We will use U_{oab} to denote whether this condition holds ($U_{oab} = 1$) or not ($U_{oab} = 0$). U_{oab} can be considered as a variable that indicates whether or not the attribute *dominates* the object. It is clear that another version of this condensation rule is obtained by replacing $P_{ab} \geq S_{ob}$ with $P_{ab} \leq S_{ob}$, in which case U_{oab} denotes whether the object dominates the attribute. In the second step, it is determined whether U_{oab} equals 1 for every bundle. The rule involved here is the *conjunctive* one (all U_{oab}'s have to be 1 for the vector product to be 1 also). The complete condensation rule is denoted as *conjunctive dominance*.

Two other condensation rules can be simply defined as modifications of, respectively, disjunctive communality and conjunctive dominance. As such, they are illustrations of the generality of the framework. As their first step, both condensation rules determine, for each of the bundles, whether the condition $S_{ob} = P_{ab}$ (the *equality* condition) holds. And as their second step, they involve, respectively, the disjunctive and the conjunctive rule. Therefore, the resulting complete condensation rules are denoted as, respectively, *disjunctive* and *conjunctive equality*.

Probability Matrix Decomposition Models

Probability matrix decomposition (PMD) models are the probabilistic version of the class of deterministic matrix decomposition models presented in the previous paragraph. Extending these models to the probabilistic case makes sense for two reasons.

First, it may be that the process that generates the observations is of a probabilistic nature. This probabilistic nature may be introduced by random *between-subject* variability, or random *within-subject* variability. An example of the former are the judgments of the members of some homogeneous population of subjects with respect to whether or not a particular politician (the object) has a particular personality characteristic (the attribute), like for example conscientiousness. And an example of the latter are the responses (recognized-not recognized) of a particular subject (the object) to tachistoscopic presentations of a particular character (the attribute). In both examples,

it makes sense to assume that multiple responses on the same object-attribute combination are realizations of the same random variable.

The second reason for considering this probabilistic version is of a practical nature. In particular, in most applications it will turn out that the number of bundles (B) that is required for the deterministic matrix decomposition is far too large to be useful. The solution for this problem in the deterministic case, is to select an *approximate* decomposition that has a relatively small number of bundles, but that is nevertheless good enough as evaluated by means of some goodness-of-fit measure. Such a measure depends on the discrepancies between the observations and the values that are predicted by the model (decomposition). This approach is followed by De Boeck & Rosenberg (1988). Now, using probabilistic models, we also drop the goal of obtaining a perfect decomposition, but we do not do this by introducing some error *afterwards*, in the form of discrepancies between observations and predicted values. Instead, we consider the observations as being *generated* by some well-specified stochastic process.

In a probabilistic model, the Y_n's are considered as random variables for which a set of PDFs is specified. These PDFs will be denoted by $f_n(Y_n; \boldsymbol{\theta})$, in which $\boldsymbol{\theta}$ is the set of parameters on which the PDFs depend.

The M_{oa}'s are considered as Bernoulli random variables whose PDF is specified by the probability $P(M_{oa} = 1)$. Loosely speaking, we want to impose some structure on the $P(M_{oa} = 1)$'s such that the characteristics of the deterministic matrix decomposition are reflected in it. This can be obtained in a simple way by considering the S_{ob}'s and the P_{ab}'s as independent Bernoulli random variables. The PDFs of S_{ob} and P_{ab} are specified by the probabilities $P(S_{ob} = 1)$ and $P(P_{ab} = 1)$, which will be denoted by ρ_{ob} and τ_{ab}, respectively. Thus, in PMD models, the fact of belonging to a particular bundle is considered as a random variable.

Using the symbols for the general case of LRMs, our transition from a deterministic to a probabilistic model involves considering $\mathbf{X}_n$ as a (vector-valued) random variable. In the following, the PDF of $\mathbf{X}_n$ will be denoted by $g_n(\mathbf{X}_n; \boldsymbol{\theta})$.

Now, the $P(M_{oa} = 1)$'s that characterize the PDFs of the M_{oa}'s can be expressed as functions of the ρ_{ob}'s and τ_{ab}'s, which are the basic parameters of the model. Thus, for this LRM, $\boldsymbol{\theta}$ consists of an $(O \times B)$-matrix of ρ_{ob}'s, which will be denoted by $\boldsymbol{\rho}$, and an $(A \times B)$-matrix of τ_{ab}'s, which will be denoted by $\boldsymbol{\tau}$. Denoting $P(M_{oa} = 1)$ by π_{oa}, it is easy to show that, for disjunctive communality the following holds:

$$\pi_{oa} = 1 - \prod_{b=1}^{B} (1 - \psi_{oab}), \tag{8}$$

in which ψ_{oab} denotes $P(U_{oab} = 1)$, which itself is defined as follows:

$$\psi_{oab} = \rho_{ob}\tau_{ab}.$$

For conjunctive dominance, the following holds:

$$\pi_{oa} = \prod_{b=1}^{B} \psi_{oab}, \tag{9}$$

in which ψ_{oab} again denotes $P(U_{oab} = 1)$, which is defined as follows for this condensation rule (in the version of attributes dominating objects):

$$\psi_{oab} = 1 - [\rho_{ob}(1 - \tau_{ab})].$$

Finally, for disjunctive and conjunctive equality, the same structure as in, respectively, (8) and (9) holds. The only difference with the previous two condensation rules is the definition of ψ_{oab}. For both disjunctive and conjunctive equality, this definition is the following:

$$\psi_{oab} = [\rho_{ob}\tau_{ab}] + [(1 - \rho_{ob})(1 - \tau_{ab})].$$

This is an appropriate place to mention the close connection between PMD models and the so-called *structural statistical reliability theory* (see Gertsbakh, 1989), which was pointed out by one of the reviewers. In this theory, one considers the probabilistic behavior of systems (machines, water supply circuits, logistic networks, ...) that are composed of components that may or may not operate appropriately. It is assumed that appropriate behavior of the *system* (the machine works, water is supplied to every town, every refugee camp gets the necessary medicine, ...) is some function of the behavior of the *components*, in the same way that M_{oa} is some function of $(\mathbf{S}_o^t, \mathbf{P}_a^t)^t$. Also, the *components*' behavior is assumed to be probabilistic, similar to our assumption of independent Bernoulli PDFs for the S_{ob}'s and P_{ab}'s. The problems this theory deals with are methods for obtaining accurate upper and lower bounds to system reliability (the counterpart of $P(M_{oa} = 1)$) and for accurate approximations to it.

Estimation

We will consider a general algorithm for the computation of the ML and the MAP estimates of the parameters of probabilistic LRMs, and will show how this algorithm can be used to compute the estimates of the parameters of PMD models. The algorithm will be presented first in the context of ML estimation, and later it will be shown how it can be adapted for MAP estimation.

ML Estimation

We start by introducing some notation. The PDF of $\mathbf{Y} = (Y_1, \ldots, Y_N)^t$ is denoted by $f(\mathbf{Y}; \boldsymbol{\theta})$, and the PDF of $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^t$ by $g(\mathbf{X}; \boldsymbol{\theta})$. Making the assumption of *local stochastic independence* (LSI), these PDFs are defined as follows:

$$f(\mathbf{Y}; \boldsymbol{\theta}) = \prod_{n=1}^{N} f_n(Y_n; \boldsymbol{\theta}),$$

and

$$g(\mathbf{X}; \boldsymbol{\theta}) = \prod_{n=1}^{N} g_n(\mathbf{X}_n; \boldsymbol{\theta}).$$

Now, the conditional PDF $h(\mathbf{X}|\mathbf{Y}; \boldsymbol{\theta})$ can be written as follows:

$$h(\mathbf{X}|\mathbf{Y}; \boldsymbol{\theta}) = \frac{g(\mathbf{X}; \boldsymbol{\theta})}{f(\mathbf{Y}; \boldsymbol{\theta})}. \tag{10}$$

The essential point in (10) is the fact that the joint PDF of $\mathbf{X}$ and $\mathbf{Y}$ (which appears in the numerator of the formula of the conditional PDF) is the same as the marginal PDF of $\mathbf{X}$. This fact allows us to use the EM-algorithm (Dempster, Laird, & Rubin, 1977) to maximize $\ln f(\mathbf{Y}; \boldsymbol{\theta})$ (and therefore also $f(\mathbf{Y}; \boldsymbol{\theta})$). In the terminology of the EM-algo-

rithm, $g(\mathbf{X}; \boldsymbol{\theta})$ is denoted as the *complete data likelihood*, and $f(\mathbf{Y}; \boldsymbol{\theta})$ as the *observed data likelihood.*

We will now show how the EM-algorithm can be used to compute the ML estimates of the parameters of PMD models. We first introduce some notation. The number of 1- and 0-responses on the (observed) random variable M_{oa} will be denoted by n_{oa1} and n_{oa0}, respectively. Their sum will be denoted by n_{oa+}. The total number of responses to the o-th object will be denoted by n_{o++}, and the total number of responses to the a-th attribute by n_{+a+}. The number of 1- and 0-responses on the (latent) random variable S_{ob} that occur in the process that generates the responses on M_{oa} will be denoted by r_{oab1} and r_{oab0}, respectively. Their sum will be denoted by r_{oab+}. Similar numbers can be defined for the responses on the (latent) random variable P_{ab}. They will be denoted by t_{oab1}, t_{oab0}, and t_{oab+}, respectively. It is clear that both r_{oab+} and t_{oab+} are equal to n_{oa+}. In the formula's below, r_{oab0} and t_{oab0} will be replaced by $(n_{oa+} - r_{oab1})$ and $(n_{oa+} - t_{oab1})$, respectively.

Now, $g(\mathbf{X}; \boldsymbol{\theta})$ is defined as follows:

$$g(\mathbf{X}; \boldsymbol{\theta}) = \prod_{o=1}^{O} \prod_{a=1}^{A} \prod_{b=1}^{B} [\rho_{ob}^{r_{oab1}}(1 - \rho_{ob})^{(n_{oa+} - r_{oab1})}]$$

$$\times \prod_{o=1}^{O} \prod_{a=1}^{A} \prod_{b=1}^{B} [\tau_{ab}^{t_{oab1}}(1 - \tau_{ab})^{(n_{oa+} - t_{oab1})}],$$

which is a product of $O \times A \times (2 \times B)$ Bernoulli random variables, each having n_{oa+} realizations. Thus, in this case, $\boldsymbol{\theta}$ consists of an $(O \times B)$-matrix of object parameters, denoted by $\boldsymbol{\rho}$, and an $(A \times B)$-matrix of attribute parameters, denoted by $\boldsymbol{\tau}$. The rows of $\boldsymbol{\rho}$ will be denoted by $\boldsymbol{\rho}_o^t = (\rho_{o1}, \ldots, \rho_{oB})$. The rows of τ will be denoted by $\boldsymbol{\tau}_a^t = (\tau_{a1}, \ldots, \tau_{aB})$.

In this definition of $g(\mathbf{X}; \boldsymbol{\theta})$ we have implicitely made use of the assumption of *local statistical independence* (LSI) between the latent responses. This assumption involves (a) that the realizations of *different* S_{ob}'s en P_{ab}'s are LSI and (b) that also the *multiple* realizations of each S_{ob} (n_{o++} realizations) and each P_{ab} (n_{+a+} realizations) are LSI. These multiple realizations can be both within a single subject responding to multiple (object, attribute)-pairs and over multiple subjects responding to a single (object, attribute)-pair. This latter type of LSI (multiple realizations coming from different subjects) is the classical assumption of *experimental independence.*

It is easy to show the following:

$$\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau}) = \sum_{o=1}^{O} \sum_{b=1}^{B} \{[r_{o+b1} \ln \rho_{ob}] + [(n_{o++} - r_{o+b1}) \ln (1 - \rho_{ob})]\}$$

$$+ \sum_{a=1}^{A} \sum_{b=1}^{B} \{[t_{+ab1} \ln \tau_{ab}] + [(n_{+a+} - t_{+ab1}) \ln (1 - \tau_{ab})]\}. \quad (11)$$

In (11), r_{o+b1} and t_{+ab1} denote the sum of r_{oab1} and t_{oab1} over all attributes and objects, respectively.

The maximization of (11) is simple. In particular, the ML estimates, denoted by $\hat{\rho}_{ob}$ and $\hat{\tau}_{ab}$, are given by the following equations:

$$\hat{\rho}_{ob} = \frac{r_{o+b1}}{n_{o++}}, \tag{12}$$

$$\hat{\tau}_{ab} = \frac{t_{+ab1}}{n_{+a+}}. \tag{13}$$

The same equations that can be used for maximizing $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$ can also be used for the maximization of $\ln f(\mathbf{Y}; \boldsymbol{\rho}, \boldsymbol{\tau})$, the loglikelihood of the PMD model. This is because using the EM-algorithm involves that in the $(p + 1)$-th EM-cycle we maximize $E[\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})|\mathbf{Y}; \boldsymbol{\rho}^{(p)}, \boldsymbol{\tau}^{(p)}]$, the conditional expected value of the complete data loglikelihood given the observed data and the parameter values of the p-th EM-cycle (denoted by $\boldsymbol{\rho}^{(p)}$ and $\boldsymbol{\tau}^{(p)}$). Since $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$ is linear in the statistics r_{o+b1} and t_{+ab1}, $E[\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})|\mathbf{Y}; \boldsymbol{\rho}^{(p)}, \boldsymbol{\tau}^{(p)}]$ differs from $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$ only in the fact that these statistics are replaced by their conditional expected values, which can be expressed as follows:

$$E(r_{o+b1}|\mathbf{Y}, \boldsymbol{\rho}^{(p)}, \boldsymbol{\tau}^{(p)}) = \sum_{a=1}^{A} \{[n_{oa1} E(S_{ob}|M_{oa} = 1, \boldsymbol{\rho}_o^{(p)}, \boldsymbol{\tau}_a^{(p)})] + [n_{oa0} E(S_{ob}|M_{oa} = 0, \boldsymbol{\rho}_o^{(p)}, \boldsymbol{\tau}_a^{(p)})]\}, \tag{14}$$

and

$$E(t_{+ab1}|\mathbf{Y}, \boldsymbol{\rho}^{(p)}, \boldsymbol{\tau}^{(p)}) = \sum_{o=1}^{O} \{[n_{oa1} E(P_{ab}|M_{oa} = 1, \boldsymbol{\rho}_o^{(p)}, \boldsymbol{\tau}_a^{(p)})] + [n_{oa0} E(P_{ab}|M_{oa} = 0, \boldsymbol{\rho}_o^{(p)}, \boldsymbol{\tau}_a^{(p)})]\}. \tag{15}$$

The right-hand sides of (14) and (15) follow from their corresponding left-hand sides because of the assumption of LSI between the M_{oa}'s.

Now, for the maximization of $E[\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})|\mathbf{Y}; \boldsymbol{\rho}^{(p)}, \boldsymbol{\tau}^{(p)}]$, we can make use of equations (12) and (13). In particular, since $E[\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})|\mathbf{Y}; \boldsymbol{\rho}^{(p)}, \boldsymbol{\tau}^{(p)}]$ differs from $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$ only in that the statistics r_{o+b1} and t_{+ab1} are replaced by their conditional expected values, it follows that the maximization of this function is possible by making use of (12) and (13) with the numerators being replaced by these conditional expected values.

The computation of the conditional expected values of the statistics r_{o+b1} and t_{+ab1} is the *E-step* of the EM-algorithm. What is needed for this computation are the expected values on the right-hand side of (14) and (15). They are conditional probabilities defined by the PDFs of the $\mathbf{S}_o$'s and $\mathbf{P}_a$'s, and the condensation rule. The derivation of the formula's for these conditional probabilities is straightforward but tedious. As an example, we will derive the formula's for disjunctive communality. Because the derivations for the S_{ob}'s and P_{ab}'s are completely analogous, we will only consider the latter.

For all condensation rules, the formulas for the conditional probabilities are all based on the following basic formula:

$$P(P_{ab} = 1|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)$$
$$= [P(P_{ab} = 1|U_{oab} = 1; \rho_{ob}, \tau_{ab}) \times P(U_{oab} = 1|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)]$$
$$+ [P(P_{ab} = 1|U_{oab} = 0; \rho_{ob}, \tau_{ab}) \times P(U_{oab} = 0|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)] \quad (16)$$

Because $P(U_{oab} = 0|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)$ equals $[1 - P(U_{oab} = 1|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)]$, we can rewrite (19) as follows:

$$P(P_{ab} = 1|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)$$
$$= \{[P(P_{ab} = 1|U_{oab} = 1; \rho_{ob}, \tau_{ab}) - P(P_{ab} = 1|U_{oab} = 0; \rho_{ob}, \tau_{ab})]$$
$$\times P(U_{oab} = 1|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)\} + P(P_{ab} = 1|U_{oab} = 0; \rho_{ob}, \tau_{ab}). \quad (17)$$

Three different conditional probabilities appear on the right-hand side of (17). The two that involve conditioning on U_{oab} will be considered first. Obviously, they depend on how U_{oab} is defined (the first step of the condensation rule). Then we will consider the conditional probability that involves conditioning on M_{oa}. This probability only depends on the second step of the condensation rules (i.e., disjunctive or conjunctive).

According to the definition of U_{oab} in disjunctive communality, the formula's for the two conditional probabilities that involve conditioning on U_{oab} can be shown to be the following:

$$P(P_{ab} = 1|U_{oab} = 1; \rho_{ob}, \tau_{ab}) = 1,$$

and

$$P(P_{ab} = 1|U_{oab} = 0; \rho_{ob}, \tau_{ab}) = \frac{(1 - \rho_{ob})\tau_{ab}}{1 - (\rho_{ob}\tau_{ab})}.$$

The formulas for $P(U_{oab} = 1|M_{oa}; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a)$ are most easily expressed in terms of the ψ_{oab}'s (the probabilities of the U_{oab}'s being equal to 1). For the disjunctive major condensation rule, they can be shown to be the following:

$$P(U_{oab} = 1|M_{oa} = 1; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a) = \frac{\psi_{oab}}{\pi_{oa}},$$

and

$$P(U_{oab} = 1|M_{oa} = 0; \boldsymbol{\rho}_o, \boldsymbol{\tau}_a) = 0,$$

in which π_{oa} is determined according to disjunctive communality.

Next, in the *M-step*, the conditional expected values at the right-hand sides of (14) and (15) replace r_{o+b1} and t_{+ab1} in (12) and (13). In this way, new values for the parameters are obtained, that replace the given values $\boldsymbol{\rho}^{(p)}$ and $\boldsymbol{\tau}^{(p)}$ for the E-step of the next cycle.

Summarizing, we can say that an EM-algorithm has been specified whose E-step involves the computation of the conditional expected values of the r_{o+b1}'s and the t_{+ab1}'s, and whose M-step involves the maximization of a function that has the same structure as $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$.

We still have to deal with the question whether the EM-algorithm does what it is being used for, namely maximizing $\ln f(\mathbf{Y}; \boldsymbol{\theta})$. In this respect, the EM-algorithm is better nor worse than the existing algorithms (e.g., steepest ascent, Newton-Raphson, Davidon-Fletcher-Powell). In particular, under certain regularity conditions (see Wu, 1983),

which are fulfilled for all PMD models, it can be proved that the EM-algorithm converges to a stationary point (i.e., a solution of the likelihood equations) of $\ln f(\mathbf{Y}; \boldsymbol{\theta})$. Whether or not this stationary point is also a local or global maximum, depends on the particular form of $\ln f(\mathbf{Y}; \boldsymbol{\theta})$ and the starting values (i.e., $\boldsymbol{\theta}^{(0)}$).

MAP Estimation

Considering MAP estimation was motivated by the fact that, depending on the particular set of observations, ML estimates that are in the interior of the parameter space may not exist. In the case of multinomial logistic regression, this problem has been dealt with by Albert and Anderson (1984). For the PMD models, this means that ML estimates of the ρ_{ob}'s and τ_{ab}'s in the interior of [0, 1] may not exist. This fact is problematic because it results in over/underflow during computation.

In the Bayesian framework, this problem does not exist. For reasons that will become clear in the following, we will consider MAP estimation. With respect to the arbitrariness of the prior PDF, it has to be noted that, except for a constant, the likelihood function and the posterior PDF are asymptotically equivalent. Therefore, MAP and ML estimates are asymptotically equivalent.

Although ML and MAP estimates are defined in a different statistical framework, their actual computation may be very similar. In particular, the choice of a particular prior PDF in some cases is formally equivalent to adding a *prior sample* within the ML framework (see, e.g., Jannarone, Yu, & Laughlin, 1990; Novick & Jackson, 1974).

A prior PDF that (for certain values of its parameters) is formally equivalent to a prior sample for the case of PMD models, is the *beta distribution* (see Mood, Graybill, & Boes, 1974, p. 115). This PDF is defined on the domain]0, 1[only, as it should be for probabilities. Assuming that the two parameters of this PDF are both equal to 2, it follows that

$$f(W; 2, 2) = c \times W(1 - W),$$

in which c denotes a constant value. This PDF has expected value and variance equal to 0.5 and 0.05, respectively. Now, considering W to be any parameter of the PMD model (ρ_{ob} or τ_{ab}), and disregarding the constant c (which has no effect on parameter estimation), it follows that $f(W; 2, 2)$ has the same functional form as the joint probability of a 1- and a 0-response on the latent Bernoulli random variable (S_{ob} or P_{ab}) whose PDF is specified by this parameter.

Making use of the prior sample interpretation of the prior PDF, it is clear that computing MAP estimates is the same as computing ML estimates using an extended sample. This extended sample involves both $\mathbf{Y}$ and the prior sample, which will be denoted by $\mathbf{Z}$. The array $\mathbf{Z}$ is of order $(T \times 2)$, and contains one pair of observations, Z_{t1} and Z_{t2}, for every value of t. For the PMD models, T equals $B \times (O + A)$.

The MAP estimates can be computed by means of the EM-algorithm. The complete data are $\mathbf{X}$ and $\mathbf{Z}$. The Z_{tj}'s are considered as a special type of latent random variables because they are mapped in observed random variables by means of a function, which is the identity function in this case. The function to be maximized in the M-step is the conditional expected value of $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$, as defined for the ML estimates, plus the loglikelihood of $\mathbf{Z}$. This latter log likelihood has the same structure as $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$. It follows that, disregarding the constant c, their sum differs from the conditional expected value of $\ln g(\mathbf{X}; \boldsymbol{\rho}, \boldsymbol{\tau})$ only in that one has to add 1 to each of the conditional expected values of the r_{o+b1}'s and t_{+ab1}'s, and 2 to each of the n_{o++}'s and n_{+a+}'s. In the M-step, these quantities (with the 1's and the 2's added) appear in, respectively, the numerator and the denominator of (12) and (13). Since the conditional

expected values of the r_{o+b1}'s and t_{+ab1}'s are bounded above by, respectively, n_{o++} and n_{+a+}, and since one observation in the prior sample is a 0 and the other is a 1, it is clear that this algorithm cannot result in estimates on the boundary of the parameter space (0 or 1).

Uniqueness

Many estimation methods involving the optimization (either maximization or minimization) of some function, have to deal with the following two problems: (i) the optimization algorithm may not always find the (a) solution for this optimization problem (i.e., the problem of local maxima and minima), and (ii) the solution may not be unique. It is instructive to note that both problems can be considered as uniqueness-problems: the first concerns the uniqueness of the results of the optimization algorithm, and the second the uniqueness of the solution of the optimization problem itself. Now, for the estimation of the parameters of PMD models, either ML or MAP, no analytical results have been obtained with respect to either the uniqueness of the results of the optimization *algorithm*, or the uniqueness of the solution of the optimization *problem*. Therefore, at this point, in applications, the only way in which we can get evidence with respect to these two problems, is by running the optimization algorithm several times, each time using different random starting values.

Application: Decision Making in Psychiatric Diagnosis

In this section, we will present the results of an analysis of data that were collected by Van Mechelen and De Boeck (1990). In their study, 15 psychiatrists were asked to judge 30 patients (objects) with respect to 23 symptoms and four psychiatric diagnoses (thus, 27 attributes). The four psychiatric diagnoses are four major categories of the first axis of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) of the American Psychiatric Association (1980), namely substance use disorders, schizophrenic disorders, affective disorders (with exclusion of the manic), and anxiety disorders. The psychiatrists were not instructed to give a single diagnosis to every patient; they were allowed to give more than one. The data can be organized in a (30 × 27)-matrix, in which each cell contains 15 observations.

Notice that the four diagnostic categories are considered attributes just like the symptoms. Using a PMD model to analyze the responses to both symptoms and diagnostic categories involves that we implicitely assume these responses to be LSI. This assumption is violated if the psychiatrists use some decision rule in which they combine their responses to the symptoms to determine their responses to the diagnostic categories. However, this assumption does *not* imply *global* statistical independence: the responses to the symptoms and the diagnostic categories may very well be statistically dependent, as long as this dependence disappears after conditioning on the parameters (making the independence *local*).

These data were analyzed according to the PMD model involving disjunctive communality. The analysis involved the computation of the ML and MAP estimates for models with from one to four bundles. However, except for the model with one bundle, no ML estimates in the interior of the parameter space could be found. Therefore, in the following we will only consider the MAP estimates.

Uniqueness

Because no formal proof of the uniqueness (in the two senses that were described previously) could be given, we took the approach to apply the algorithm several times

to the same optimization problem, each time using different random starting values. Now, for the models with from one to three bundles, using 10 different random starting value configurations for each of these three models, the algorithm converged to the same solution. However, for the model with four bundles, we found four different solutions of which the corresponding likelihoods were different but not substantially. Thus, these different four bundle solutions did not differ much from each other with respect to the likelihood criterion.[3]

As will be argued in the following, the fact of having found four different solutions for the four bundle model does not have to be considered problematic. In particular, there is good evidence that the fourth bundle is only very weakly identified by the data. This evidence was found by examining the pair-wise correspondences between the bundles of the different four bundle solutions. For every pair of different solutions, $4^2 = 16$ correspondences between pairs of bundles were examined. We computed the average absolute difference between the elements of these pairs, which will be denoted by V_{bc} (in which b indexes the first bundle of the pair, and c the second). The differences between the elements of the object bundles and those between the elements of the attribute bundles were added in order to obtain an overall statistic per bundle. Denoting the estimated probabilities of the two solutions by $\hat{\rho}_{ob}(\hat{\tau}_{ab})$ and $\tilde{\rho}_{ob}(\tilde{\tau}_{ab})$, respectively, V_{bc} can be defined formally as follows:

$$V_{bc} = (O + A)^{-1}\left[\left(\sum_{o=1}^{O} |\hat{\rho}_{ob} - \tilde{\rho}_{oc}|\right) + \left(\sum_{a=1}^{A} |\hat{\tau}_{ab} - \tilde{\tau}_{ac}|\right)\right].$$

Now, the V_{bc}-values for the comparisons between different pairs of solutions all show the same pattern. It was decided not to compute the average V_{bc}-values over the $6 = (4 \times 3)/2$ tables that were obtained, because this would involve a rearrangement of the rows and columns of these tables according to some expected pattern (see further). Such a rearrangement can result in a spurious pattern in the table of averages. Therefore, in Table 2, the V_{bc}-values for only one of the six comparisons are shown. For three of the four rows (columns) in Table 2 (and in the tables for the other comparisons) there is one element that is substantially smaller than the others. Moreover, the bundle of a given solution that is different from all the bundles in some other solution, is always the same (i.e., for all other solutions). Thus, it appears that, by way of speaking, the data provide information that is *strong enough* to always let three particular bundles show up in the solution, but not to let a particular fourth one show up.

The explanation given above is consistent with the fact that the three bundles that always appear in the four bundle solutions, correspond very highly with the three bundle solution. This can be seen in Table 3, where the V_{bc}-values are shown of the comparison between the three bundle solution and one of the four bundle solutions. The V_{bc}-values for the other four bundle solutions are similar.

The Solution

The three bundle solution is presented in Tables 4 and 5 for, respectively, the attributes (symptoms and syndromes) and the objects (the patients). In order to get a clearer picture of the main structure in these matrices of probability estimates, we transformed (trichotomized) them into the ordered categories 0, *mid*, and 1, depending

[3] In an analysis of another data set (see Candel & Maris, 1994), we also found convergence to the same solution for different random starting values for models with a *low* number of bundles (up to 4 for these data), whereas convergence to different solutions was found for models with a high number of bundles. Also with respect to the likelihood criterion, these different solutions did not differ much from each other.

TABLE 2

Average Absolute Differences Between the Elements of Bundles of Different Solutions for the Four Bundle Model

Bundles of Solution 2	*Bundles of Solution 1*			
	1	*2*	*3*	*4*
1	0.016713	0.376705	0.245678	0.368399
2	0.364948	0.061612	0.323485	0.413740
3	0.287653	0.271789	0.275048	0.235633
4	0.368113	0.434569	0.214719	0.103347

TABLE 3

Average Absolute Differences Between the Elements of the Bundles of the Solution for the Three Bundle Model and one of the Solutions for the Four Bundle Model

Bundles of the Solution for the Three Bundle Model	*Bundles of a Solution for the Four Bundle Model*			
	1	*2*	*3*	*4*
1	0.035308	0.379390	0.226820	0.365584
2	0.376718	0.014500	0.353927	0.392403
3	0.379243	0.414455	0.262051	0.048235

TABLE 4

Three Bundle Solution for the Attributes

	MAP Estimates			*Trichotomized Values*		
	Bundle			*Bundle*		
Attribute	*1*	*2*	*3*	*1*	*2*	*3*
•*speech disorganisation*	0.799	0.09	0.083	1	0	0
•*inappropriate affect/ behavior*	0.932	0.081	0.282	1	0	0
•**schizophrenic disorder*	0.853	0.008	0.015	1	0	0
•hallucinations	0.346	0.007	0.013	*mid*	0	0
•denial of illness	0.419	0.019	0.184	*mid*	0	0
•intellectual impairment	0.433	0.28	0.246	*mid*	0	0
•*depression*	0.088	0.98	0.111	0	1	0
•*anxiety*	0.14	0.772	0.247	0	1	0
•**affective disorder*	0.107	0.967	0.202	0	1	0
•suicide/self–mutilation	0.011	0.487	0.017	0	*mid*	0
•excessive somatic concerns	0.231	0.344	0.054	0	*mid*	0
•*narcotics/drugs*	0.009	0.009	0.855	0	0	1
•**substance use disorder*	0.007	0.007	0.941	0	0	1
•alcohol abuse	0.008	0.006	0.457	0	0	*mid*
•*social isolation*	0.793	0.702	0.125	1	1	0
•social dullness	0.807	0.368	0.165	1	*mid*	0
•retardation/ lack of emotion	0.528	0.338	0.091	*mid*	*mid*	0
•*anxiety disorder	0.019	0.621	0.339	0	*mid*	*mid*
•*role impairment*	0.929	0.835	0.685	1	1	1
•disturbance in daily routine/leisure time	0.805	0.543	0.401	1	*mid*	*mid*
•*agitation/excitement*	0.281	0.137	0.156	0	0	0
•*disorientation/ memory impairment*	0.218	0.268	0.143	0	0	0
•*antisocial*	0.056	0.024	0.327	0	0	0
•*belligerence/negativism*	0.212	0.21	0.028	0	0	0
•*thoughts of grandeur*	0.051	0.006	0.176	0	0	0
•*suspicion/persecution*	0.269	0.071	0.031	0	0	0
•*impulse control impairment*	0.21	0.076	0.197	0	0	0

TABLE 5

Three Bundle Solution for the Objects

Trichotomized Values			
Bundle			
1	*2*	*3*	*Number of Patients*
1	0	0	*6*
mid	0	0	3
0	1	0	*9*
0	*mid*	0	1
0	0	1	*2*
mid	1	0	2
mid	*mid*	0	1
mid	0	1	1
mid	0	*mid*	1
0	*mid*	*mid*	1
mid	*mid*	*mid*	1
0	0	0	*2*

on whether their value was in the range]0.0–0.33 . . .],]0.33 . . . –0.66 . . . [, or [0.66 . . . –1.0[, respectively. For interpretation, we will make the simplification that the trichotomized values 1 and 0 indicate that the object (attribute) *always* belongs to (value 1) or *never* belongs to (value 0) the bundle. The value *mid* will be considered as indicating that the object (attribute) either belongs to or does not belong to the bundle with about equal probability. In Table 4, the solution for the attributes is given, together with the trichotomized values. The names of the attributes with a clear pattern (i.e., a

pattern without a *mid*) is printed in italic. In Table 5, the solution for the objects is given by indicating the number of patients having each of the possible patterns of trichotomized values. The numbers of patients with a clear pattern are printed in italic. This information is sufficient for an interpretation.

Interpretation

We first consider the interpretation of the solution for the attributes, to which we will give the most attention, and then the one for the objects. The interpretation of the solution for the attributes is straightforward. In particular, ignoring for the moment the position of the anxiety disorder, each of the three bundles specifies the symptoms of one major disorder. This is the case because, ignoring the anxiety disorder, each of the remaining major disorders belongs to a *single different* bundle. Thus, the basic structure in the solution for the attributes is such that a patient that is diagnosed as having a particular major disorder, has all the symptoms of the bundle that corresponds to this major disorder. In the following, this inference will be weakened somewhat, but its essence will remain.

In uncovering the implicit rules that govern the psychiatrist's diagnoses, it is useful to look for the symptoms that are *specific* for a particular diagnosis. Obviously, these specific symptoms may only belong to the bundle that corresponds to this diagnosis. However, since an object (attribute) may have both the value 1 and the value *mid* for a particular bundle, we have two kinds of diagnosis-specific symptoms. The difference between these two kinds of symptoms will be explained in the following.

The first bundle contains the symptoms of the schizophrenic disorder. The specific symptoms speech disorganisation (which refers to disorganisation in the *content* of what is being said as, e.g., lack of coherence) and inappropriate affect and behavior, do have a 1 for the first bundle. Ignoring the probabilistic nature of bundle-membership, one can say that these symptoms are *necessary conditions* for the schizophrenic disorder diagnosis (however, they are not the *only* necessary conditions). In other words, every patient that is diagnosed as schizophrenic has these two symptoms. This is not the case for three other specific symptoms with a *mid*-value on the first bundle, namely hallucinations, denial of illness, and intellectual impairment. For these symptoms, it is also possible that they do *not* belong to the first bundle. This reflects the fact that not every patient that is diagnosed as schizophrenic exhibits hallucinations, denial of illness, and intellectual impairment. But, because these symptoms are nevertheless specific for the schizophrenic disorder diagnosis, they must possess high *cue validity*. Notice, however, that the corresponding probability of bundle membership (τ_{a1}) may not be interpreted as the probability that a patient who shows such a symptom, is considered a schizophrenic by a psychiatrist. The correct interpretation is that, if the probabilities for all the other bundles are zero, this probability of bundle membership on the attribute side is the probability of having such a symptom *if* the patient is being considered a schizophrenic by a psychiatrist ($S_{o1} = 1$).

These findings illustrate an advantage of PMD models over their deterministic equivalents. In particular, by allowing an attribute to have a *moderate* probability on a particular bundle, one can represent the fact that it sometimes but not always applies to the objects of its bundle. A specially interesting case are attributes that do not apply to objects of other bundles; they are interesting because of their cue validity. It is obvious that, in an analogous way, an *object* may have some but not all attributes of a particular bundle.

The second bundle contains the symptoms of the affective disorder. For this disorder, there are two specific symptoms that are also necessary conditions, namely depression and anxiety. And again, we have two symptoms that are not necessary, but

that nevertheless have a high cue validity, namely thoughts about and attempts at suicide and/or automutilation, and excessive somatic concerns.

The third bundle contains the symptoms of the substance use disorder. The diagnosis-specific symptom that is also a necessary condition (excessive use of narcotics/drugs) is rather obvious. And again, since alcohol abuse is only one particular case of the substance use disorder, we also have a non-necessary symptom with high cue validity for this disorder.

We will now consider the position of the anxiety disorder in the solution. The reason why there is not a bundle that is uniquely associated to the anxiety disorder, is that this diagnosis is mainly given together with either the affective (mostly) or the substance use disorder. It probably is so that the anxiety disorder is not considered as a diagnosis like the others, but more as a way to indicate that anxiety is an important aspect of the patient's mental illness. The fact that the *symptom* anxiety (contrary to the anxiety disorder diagnosis) does not belong to the bundle of the substance use disorder, possibly reflects the fact that anxiety (as an emotion) is assumed to occur in some of these patients, but that it does not show itself as an observable symptom because of the influence of narcotics and/or alcohol. With respect to anxiety, it is also interesting to note that neither anxiety as a symptom nor the anxiety disorder ever applies to patients that are diagnosed as schizophrenic (disregarding the patients with a multiple diagnosis).

The symptoms that belong to both the first *and* the second bundle apply to both the patients that are diagnosed as having the schizophrenic, and those that are diagnosed as having the affective disorder. In particular, both types of patients are judged to have a disturbed social life, whereas this is not the case for patients that were given the substance use disorder diagnosis.

And from the kind of attributes that belong to all three the bundles it can be inferred that what all patients had in common is the fact that they were considered as not being able to fulfill some of their roles in daily life (e.g., parent, partner, employee, student).

Finally, we will briefly consider the solution for the objects (patients). The first point to be made with respect to this structure is that the majority of the patients were given a simple diagnosis in one of three categories: schizophrenic, affective, or substance use disorder. These patients are the ones that belong to a single bundle. It also has to be noted here that the affective and the substance use disorders may be complemented by the anxiety disorder as a secondary diagnosis.

The second point to be made is that none of the patients that belong to two or more bundles has a clear pattern (i.e., one without a *mid*). This reflects the fact that there is no agreement among the psychiatrists with respect to which combined diagnosis should be given. This disagreement resulted in moderate observed proportions in the cells that correspond to these patient-diagnosis combinations.

Related Models and Conclusion

The PMD models presented here can be characterized with three important features: (a) they are models for binary two-way two-mode data (possibly with replications); (b) they are decomposition models in that the data are considered a function of properties of the elements from each mode separately; (c) they are probabilistic. Other models exist with one, two, or even all three of these features, and therefore it is of value to situate the PMD models in the broader context of those other models and to indicate where exactly their specificity is to be found.

Many models exist for two-mode binary data, such as item response models, item factor analysis, latent class analysis, lattice models, hierarchical classes models, etc.

Natural ways to analyze binary data are to use set-theory and Boolean algebra, like in lattices and hierarchical classes models, or to model the probabilities of the data, like in item response and latent class models. In the PMD models, a combination of both approaches is used. First, the data are considered as resulting from different partitions of the element sets of the two modes (each partition defines a bundle) and a Boolean rule applied on multiple partition class membership (the condensation rule), and second, the partitions are random partitions (bundle membership is probabilistic).

Like item response models (e.g., Fischer, 1974) and models for item factor analysis (Bock & Aitkin, 1981), PMD models are decomposition models, in that they model the probabilities of the data from separate parameter sets for the two modes. Typical for the PMD models is that the decomposition can be thought of in terms of separate *latent response* vectors for the two modes, with the probability *parameters* applying to these covert data. In other models, either there are no intermediate covert data behind the observed data, like in item response models, or these covert data are not random variables but constants instead, like in the hierarchical classes model (which makes the model deterministic). What is decomposed in the PMD models are the data, like in a singular value decomposition of the raw data (Greenacre, 1984; Nishisato, 1980), but unlike in traditional singular value decomposition, the components are binary and have a probabilistic nature. One can of course also present the PMD models as models that decompose the probabilities of the data, namely into single mode probabilities that are combined in nonlinear ways.

PMD models are probabilistic models, in that they take the data as results from a random process and in that they consequently contain parameters describing this random process. A different approach would be to consider the random process unknown and consequently not to try to model the random process. The corresponding strategy is to construct a theoretical data set from which the observed one deviates as little as possible while the theoretical data set that is used for this approximation is still the result of a rather simple but now deterministic process. The random element is then captured in the deviations of the observed data from the hypothetical data. This is a possible rationale behind what Arabie and Hubert (1992) have called "combinatorial data analysis." It is also the approach followed with hierarchical classes analysis. Important drawbacks are that one cannot make use of ML estimation and statistical theory.

The PMD models offer a new way of analyzing two-mode binary data, which is especially interesting if the data can be considered to result from binary random variables for the elements of both modes. An important advantage is that one can select a PMD model with a condensation rule that corresponds to one's theory about how the two modes interact to yield the data. Furthermore, an algorithm exists that allows for estimating the parameters of the model, that is, the parameters of the Bernoulli PDF for each random variable. Although the uniqueness problem has not been solved in a satisfactory way, it was shown with an application how one can gain information about uniqueness from using different randomly chosen starting values. The results of the application are also quite encouraging as far as their meaningfulness is concerned.

Besides uniqueness, another topic for future research is the development of methods for assessing accuracy of estimation and goodness-of-fit. With respect to accuracy of estimation, the obvious thing to do is to compute the inverse of the information matrix at the MAP (or ML) estimates. And for goodness-of-fit testing, a statistic whose usefulness should be examined is Pearson's chi-square. Finally, PMD models in their present version require more than one observation in each cell (i.e., for each combination of elements from the two modes), since otherwise the consistency of the ML and MAP estimates cannot be proved. However, when the parameters would be considered

random variables themselves, a kind of *random effects* or *marginal* version of the PMD models can be formulated, which allows for a corresponding estimation method that can be applied to data with a single observation in each cell.

References

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*, 1–10.

American Psychiatric Association. (1980). *The diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

Arabie, P., & Hubert, L. J. (1992). Combinatorial data analysis. *Annual Review of Psychology, 43*, 169–203.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Candel, M., & Maris, E. (1994). *Probability matrix decomposition: Perceptual analysis of two-way two-mode binary data*. Manuscript submitted for publication.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

De Boeck, P. (1986). *HICLAS computer program: version 1.0*. Leuven, Belgium: University of Leuven.

De Boeck, P., & Rosenberg, S. (1988). Hierarchical classes: Model and data analysis. *Psychometrika, 53*, 361–381.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Fischer, G. H. (1974). *Einfuhrung in die Theorie psychologischer Tests* [Introduction to the theory of psychological tests]. Bern: Huber.

Gertsbakh, I. B. (1989). *Statistical reliability theory*. New York: Marcel Dekker.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Jannarone, R. J., Yu, K. F., & Laughlin, J. E. (1990). Easy Bayes estimates for Rasch-type models. *Psychometrika, 55*, 449–460.

Kim, K. H. (1982). *Boolean matrix theory*. New York: Marcel Dekker.

Maris, E. (1992). *Psychometric models for psychological processes and structures*. Unpublished doctoral dissertation, University of Leuven, Belgium.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523–547.

Mickey, M. R., Mundle, P., & Engelman, L. (1983). Boolean factor analysis. In W. J. Dixon (Ed.), *BMDP statistical software* (pp. 538–546, 692). Berkeley: University of California Press.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics*. London: McGraw-Hill.

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

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Van Mechelen, I., & De Boeck, P. (1990). Projection of binary criterion into a model of hierarchical classes. *Psychometrika, 55*, 677–694.

Van Mechelen, I., De Boeck, P., & Rosenberg, S. (1995). The conjunctive hierarchical classes model. *Psychometrika, 60*, 505–521.

Wu, C. F. J. (1983). On the convergence properties of the EM-algorithm. *The Annals of Statistics, 11*, 95–103.

Manuscript received 5/16/94
Final version received 1/18/95