Paper presented in Versailles. Octobre 1985, "Data analysis and informatics V". Will appear in E Diday (Ed.) Data Analysis and Informatics II, Amsterdam: North Holland

THE ANALYSIS OF TIME-BUDGETS WITH A LATENT TIME-BUDGET MODEL

JAN DE LEEUW AND PETER G.M. VAN DER HEIJDEN

Department of Data Theory and Department of Psychometrics, University of Leiden, Hooigracht 15, 2312 TW Leiden, The Netherlands.

1. INTRODUCTION

Event history data are data specifying the sequence and times in which objects or individuals are in specific states (categories) during some period of time, where the number of states is finite. These data are collected in for example sociology, anthropology, ethology and psychology in studies concerning the use of time by humans and non-humans. Aggregated forms of event history data are also known as *time-budget data*, or *time-allocation data*. In *sociology* data are often collected in the form of *diaries*. A very valuable review of contemporary model-based data analysis techniques for the analysis of diaries and other 'social dynamics' data is Tuma and Hannan [16]. These techniques require a great deal of prior knowledge, and can be applied only in a fairly restrictive class of situations (very many observations, and only very few states). In *ethology* attention for time allocation seems to be growing. It will be clear that, unlike in sociology, it is impossible to work with diaries in ethology; the behavior is observed by the researcher. In *anthropology* time allocation studies also seem to become more popular. In a recent paper Gross [6] reviews the use of time allocation methods in anthropology. Gross compares these methods in terms of their usefulness for anthropologists, who obviously have to deal with practical constraints that are quite different from those of ethologists. It appears from Gross' evaluation that the *random spot check method* is the preferable one. The studies of Erasmus [4] and Johnson [12] are given as major examples of this approach.

In spot check methods investigators take 'snapshot-like' recordings of behavior. The idea is that, if a person is baking bread 7 times out of 100 times we have made a spot-check of his home, and we assume that he spends approximately 7% of his time baking bread. Clearly there are various practical and theoretical limitations to this method, which almost immediately come to mind. They are discussed in detail by Gross ([6], p. 537-546), but they do not alter the final conclusion about the usefulness of the method.

2. WAYS TO ANALYZE TIME-BUDGET DATA

If we look at the ways time allocation data are usually analyzed, we find mostly purely descriptive and tabulatory techniques. In his review paper Gross ([6], p. 546-548) also has no suggestions about how one should proceed beyond the purely descriptive point. There are more specific proposals for the analysis of time budgets in the publications of the group involved in the Multinational Time Budget Project [14]. Converse [1] pioneered the use of multivariate analysis techniques on such data, and this was subsequently taken up in [15] and [8]. The multivariate analysis techniques that are used, however, are standard packaged techniques that do not take special properties of time allocation data into account. This makes their use tentative,

at best. It seems that correspondence analysis (CA) is better adapted to analyze time-budget data. The number of applications of CA is growing in the context of time-budget data (see, for example [3], [11], [13], and [9]).

Although the chi square distance of CA seems to provide us with an appropriate measure for differences between time budgets, one important problem is not solved. Clark et al. (in [15], p. 67) formulate this as follows: "Multivariate techniques are very ill-suited to serve as communications devices between the community of scientific researchers and the larger society. Indeed, the techniques are not even particular appropriate for communicating ideas within many parts of the scientific community." In this paper we shall discuss a special purpose multivariate technique for the analysis of spot-check and other time-budget data. It is our opinion that the type of representation derived by this technique can be communicated very well, both inside and outside the scientific community. The representations are much more economical than long lists of tables or descriptive diagrams, and they emphasize the most important variation in the data. We will compare this technique with CA, both theoretically as well as empirically.

3. A LATENT TIME-BUDGET MODEL

Suppose an individual or group is engaged, during some period, in any one of m different activities. The basic data in this paper are measurements of the distribution of time over the different activities for the different individuals or groups, without taking the time of day into account. If the groups are men and women in a particular culture, for instance, we can see which activities belong to the task of the typical woman, and which to the task of the typical man, but we cannot see how the available time is used to plan and execute activities. Another piece of information that we loose if we go from event history to time budget data is the sequence of activities. We aggregate over time, and the sequence, or the count of the transitions, simply gets lost. Thus event history data are inherently richer, but this is also their weakness. As always rich data structures can easily lead to many empty cells, and to overparametrization. For this reason the time budgets, which are as it were a marginal of the event history data matrix (see [13], [9]), are more robust data.

The data can be collected in an n x m matrix, where n is the number of groups or individuals and m the number of activities. The entries of the matrix are integers n_{ij}, which is the number of individuals in group i engaged in activity j during random spot-checks. The total number of spot-checks for row i is n_{i+}, which is supposed to be a fixed number, determined by the design of the experiment. We have decided, beforehand, that we are going to check in on the families of some tribe, say, 75 times. Because the family sizes are fixed during the experiment, this means, for example, that we have 18 x 75 observations on adult males, and 94 x 75 observations on juveniles. It is possible that various things go wrong during the experiment (people may be out hunting, persons may die, and so on), but this does not change the fact that essentially the n_{i+} are fixed. But the n_{ij} are a different matter. They are the outcomes of the experiment, and it is best to conceptualize them as random variables. If we repeat the experiment with the same n_{i+}, we shall undoubtedly find somewhat different n_{ij}. Now $p_{ij} = n_{ij}/n_{i+}$ can be used as an estimate of π_{ij}, the proportion of time spent by a typical member of group i on activity j. If we assume independence between observations, the n_{ij} in row i are multinomially distributed with means $E(n_{ij}) = n_{i+}\pi_{ij}$. A first result that interests us, as a sort of baseline, is whether the π_{ij} are different for the different groups. We expect that they will be very different indeed, otherwise our categories of behavior or our groups of individuals must have been defined in a rather uninteresting way. The usual test for equi-distribution in a rectangular table is the chi-square test.

Here we discuss a model for the analysis of time budgets, which can be considered to be a specially adapted form of factor analysis. We assume that the equidistribution model is untenable, and we analyze the difference in the distributions for the various i. The theoretical time budgets are given by π_{ij}, where i = 1, ... , n, j = 1, ... , m, and where $\Sigma_{j=1}^{m} \pi_{ij} = 1$ for all i. The model is

$$\pi_{ij} = \Sigma_{k=1}^{p} \alpha_{ik}\beta_{jk}, \tag{1}$$

with restrictions $\alpha_{ik} \geq 0$, $\beta_{jk} \geq 0$, and $\Sigma_{k=1}^{p} \alpha_{ik} = 1 = \Sigma_{j=1}^{m} \beta_{jk}$. The number of degrees of freedom equals the number of independent cells minus the number of independent parameters, being n(m - 1) - {p(m - 1) + (p - 1)n}. This model can be interpreted as a model describing how a theoretical time budget in row i is the result of p latent time budgets, given by β_{jk}, on which i 'loads'. Each theoretical as well as each latent time budget sums to 1. Values α_{ik} show for which proportions the theoretical time budget of row i is made up from latent time budget k. The number of latent time budgets has to be specified by the researcher. In case of p = 1, model (1) is equal to the usual equidistribution model, having (n - 1)(m - 1) degrees of freedom. In case we postulate that our time budgets are sampled under a product multinomial distribution, maximum likelihood estimation can be accomplished using the EM algorithm [2].

4. MAXIMUM LIKELIHOOD ESTIMATION

The logarithm of the likelihood is, except for an irrelevant constant,

$$\mathfrak{L}(\alpha,\beta) = \Sigma_{i=1}^{n}\Sigma_{j=1}^{m} n_{ij} \ln \Sigma_{k=1}^{p} \alpha_{ik}\beta_{jk}. \tag{2}$$

This must be maximized over the unknowns, with the restrictions given above. We present an elementary derivation of an EM-type algorithm for this problem. Suppose $\underline{\alpha}_{ik}$ and $\underline{\beta}_{jk}$ are the current best estimates at some point during the iterations of the algorithm, giving theoretical values $\underline{\pi}_{ij}$ according to (1). Also define $n_{ijk} = n_{ij}\underline{\alpha}_{ik}\underline{\beta}_{jk}/\underline{\pi}_{ij}$. We use the notational convention of replacing an index over which we have summed by a plus. Observe that $n_{ij+} = n_{ij}$, and the n_{ijk} can be seen as distributing the observed budgets over the p dimensions.

Theorem 1. Consider the algorithm which computes updates by the rules $\beta_{jk}^{+} = n_{+jk}/n_{++k}$ and $\alpha_{ik}^{+} = n_{i+k}/n_{i++}$. Then $\mathfrak{L}(\alpha^{+},\beta^{+}) \geq \mathfrak{L}(\underline{\alpha},\underline{\beta})$.

Proof. From the concavity of the logarithm,

$$\ln \pi_{ij}/\underline{\pi}_{ij} = \ln \{\Sigma_{k=1}^{p} \underline{\alpha}_{ik}\underline{\beta}_{jk}(\alpha_{ik}\beta_{jk}/\underline{\alpha}_{ik}\underline{\beta}_{jk})/\underline{\pi}_{ij}\} \geq$$

$$\geq \{\Sigma_{k=1}^{p} \underline{\alpha}_{ik}\underline{\beta}_{jk} \ln(\alpha_{ik}\beta_{jk}/\underline{\alpha}_{ik}\underline{\beta}_{jk})\}/\underline{\pi}_{ij}. \tag{3}$$

Substitution in (2), using (3), gives

$$\mathfrak{L}(\alpha,\beta) \geq \mathfrak{L}(\underline{\alpha},\underline{\beta}) + \mathfrak{D}(\alpha,\beta,\underline{\alpha},\underline{\beta}), \tag{4}$$

where

$$\mathfrak{D}(\alpha,\beta,\underline{\alpha},\underline{\beta}) =$$

$$= \Sigma_{i=1}^{n}\Sigma_{j=1}^{m}\Sigma_{k=1}^{p} n_{ijk} \ln \alpha_{ik}\beta_{jk} - \Sigma_{i=1}^{n}\Sigma_{j=1}^{m}\Sigma_{k=1}^{p} n_{ijk} \ln \underline{\alpha}_{ik}\underline{\beta}_{jk}. \qquad (5)$$

Moreover (4) is an equality if $(\alpha,\beta) = (\underline{\alpha},\underline{\beta})$. It is easy to see that we find (α^+,β^+), the successor of $(\underline{\alpha},\underline{\beta})$, by maximizing $\mathfrak{D}(\alpha,\beta,\underline{\alpha},\underline{\beta})$ over (α,β). Thus

$$\mathfrak{L}(\alpha^+,\beta^+) \geq \mathfrak{L}(\underline{\alpha},\underline{\beta}) + \mathfrak{D}(\alpha^+,\beta^+,\underline{\alpha},\underline{\beta}) \geq$$

$$\geq \mathfrak{L}(\underline{\alpha},\underline{\beta}) + \mathfrak{D}(\underline{\alpha},\underline{\beta},\underline{\alpha},\underline{\beta}) = \mathfrak{L}(\underline{\alpha},\underline{\beta}). \qquad (6)$$

QED.

So, starting with arbitrary $\underline{\alpha}_{ik}$ and $\underline{\beta}_{jk}$ the n_{ijk} are derived (Expectation-step). Then we compute the new α_{ik} and β_{jk} from the n_{ijk} (Maximization-step). From these new estimates we compute new n_{ijk}, etcetera. Theorem 1 shows that the algorithm increases the likelihood in each step. In fact if we build in the rule that the algorithm stops if $(\underline{\alpha},\underline{\beta})$ already maximizes $\mathfrak{D}(\alpha,\beta,\underline{\alpha},\underline{\beta})$, then either the algorithm stops at a stationary point where the likelihood equations are satisfied, or it generates an infinite sequence for which the increase in the likelihood is strict. The sequence of solutions that is computed has accumulation points, because the restrictions define a compact set, and each accumulation point has the same value of the likelihood and satisfies the likelihood equations. This means that we can say for all practical purposes that the algorithm converges.The stationary points of the algorithm have an interesting property, which we prove next. It turns out that the matrices of observed and expected values have the same marginals. This is obvious for the row marginals, but for the columns the situation is a bit more complex.

Theorem 2. At a stationary point $\Sigma_{i=1}^{n} n_{i+}\pi_{ij} = n_{+j}$.

Proof. At a stationary point of the algorithm we have

$$\Sigma_{j=1}^{m} (n_{ij}/\pi_{ij})\beta_{jk} = n_{i+}, \qquad (7a)$$

$$\Sigma_{i=1}^{n} (n_{ij}/\pi_{ij})\alpha_{ik} = \mu_k, \qquad (7b)$$

with μ_k Lagrange multipliers. If we multiply both sides of (7b) by β_{jk}, and sum over k, we find

$$n_{+j} = \Sigma_{k=1}^{p} \mu_k\beta_{jk}. \qquad (8)$$

If we multiply both sides of (7b) by β_{jk}, sum over j, and use (7a), we find

$$\Sigma_{i=1}^{n} n_{i+}\alpha_{ik} = \mu_k. \qquad (9)$$

Now multiply both sides of (9) by β_{jk}, and sum over k. This gives, using (8),

$$\Sigma_{i=1}^{n} n_{i+}\pi_{ij} = \Sigma_{k=1}^{p} \mu_k\beta_{jk} = n_{+j}. \qquad (10)$$

QED.

After the maximization of the likelihood is carried out, the difference between the observed and expected time budgets can be tested using chi-square statistics. However, this test could be problematic, since the asymptotic distribution only holds if in a specific time budget each observation has the same theoretical distribution, and subsequent observations are independent. The latter assumption will often be violated because activities of different objects will not be independent: for example, a mother is cooking, the child helps her. This dependence can be taken care of, for instance by noting only the behavior of one of the objects, for example, the mother. We find another type of violation in case the n_{ij} are derived from event histories, and a frequency corresponds to for example a minute spent on some activity. Clearly activities often take longer than a minute, and in this case the dependence of subsequent observations is considerable. So in this situation model (1) should only be used as a descriptive tool. However, violation of this assumption does not have to be severe in case of data sampling using the random spot check method discussed in the introduction. Here, by making the intervals between subsequent spot checks large, the dependence between the observations can probably be made small. So in this situation, model (1) can be used for inferential purposes.

The objective of the latent budget model will be clear: it aims at a sparse description of the data in terms of typical time budgets, provided by β_{jk}. Values α_{ik} show how (groups of) objects load on these typical time budgets. Model (1) seems to be very well suited for the analysis of time budgets due to its restrictions that loadings α_{ik} and estimated proportions β_{jk} should be larger than zero and add up to one over time budgets and categories, respectively.

5. RELATIONS WITH CA

The resemblance between the latent budget model and CA is large, in the sense that the approximation of π_{ij} provided by a t-dimensional CA solution will be often about the same as the approximation provided by the model in case of (t + 1) latent time budgets. Intuitively this can be made clear by realizing that in a two-dimensional CA plot the location of each profile point (observed time budget) can be expressed as a weighted average of three typical time budgets placed at the periphery of the cloud of profile points. Another way to make this clear is by comparing model (1) with the model fitted by CA, which is

$$\pi_{ij} = \pi_{i.}\pi_{.j}(1 + \Sigma_{s=1}^{t} r_{is}c_{js}\lambda_s), \qquad (11)$$

where s is an index for the dimension.

In case t = 0, and p = 1, the approximation of both model (1) and (11) gives the equiprobability model. In case of t = 1 and p = 2, (1) as well as (11) approximate π_{ij} using the sum of two products of a row term and a column term. In general, in case of t = q and p = (q + 1), both in (1) and (11) π_{ij} is approximated using the sum of q + 1 products of a row and a column term. Although the approximations given by (1) and (11) are often about the same, this is not necessarily the case. First of all this is due to the fact that the restrictions on parameters α_{ik} and β_{jk} are different from those on the scores r_{is} and c_{js}. Secondly, in case of model (1), maximizing the likelihood is asymptotically equivalent to minimizing the value of the chi-square statistic

$$S = n_{++} \Sigma_{i=1}^{n}\Sigma_{j=1}^{m}(p_{ij} - \pi_{ij})^2/\pi_{ij} \qquad (12a)$$

Thus the estimates computed by the EM-algorithm are efficient if the model is true. This is not the case for CA, where

$$S = n_{++} \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} (p_{ij} - \pi_{ij})^2/(p_{i+}p_{+j}) \qquad (12b)$$

is minimized.

The fact that approximations (1) and (11) are nearly the same for this example (although scores r_{is} and c_{js} are quite different from α_{ik} and β_{jk} for all k and s) can further be illustrated when we use the generalization of CA proposed by Escofier [5], see also Van der Heijden & De Leeuw [10], Van der Heijden [9]. This generalization of CA can be written in model form as

$$\pi_{ij} = \pi\!\!\!\pi_{ij} + \pi_{i.}\pi_{.j}\Sigma_{s=1}^{t}(r_{is}c_{js}\lambda_s), \qquad (13)$$

where $\pi\!\!\!\pi_{ij}$ is the expected frequency for cell (i,j) under some model.

6. AN EXAMPLE

Gross et al. [7] recently presented an analysis of random spot check data (compare the introduction) of males, females and kids in four tribes of Amazone Indians. They showed 12 figures (one for each of the three types of persons in each of the four tribes). In these figures for 7 two-hour periods 7 vertical bars display the proportion of time spent in 6 behavior categories, by subdividing each bar into six parts (one part for each category), the length of which represents the proportion of time spent into some category. By measuring the proportions in these bars we could derive a data block with elements n_{ij} representing the proportion of time that group i spends in category j. The tribes are the Mekranoti, the Kanela, the Bororo and the Xavante. The six behavior states are 'idle', 'sleep', 'care', 'nonsubst', 'domestic' and 'wild'. For a description of these tribes and a definition of behaviors we refer to Gross et al. [7]. We will analyze the data matrix, in which the tribes are coded interactively with the types (males, females, kids). So the matrix has order 12 x 6.

In Table 1 we find estimates for the independence model (row and column margins) in the first column. Table 1 also shows we find the estimates of α_{ik} and β_{jk} for the model with two latent budgets (columns 2 and 3), and for the model with three latent budgets (columns 4, 5, and 6). Considering the results for p = 2, we see that the first latent budget is the budget for the adults (for the females predominantly), being less idle and asleep than the marginal time budget, but performing more 'care', 'nonsubst' and 'wild' behavior; the second latent budget is that for the kids, being more idle and asleep, and doing the other behaviors less. Columns 4, 5, and 6 show us results for model (1) with p = 3 latent budgets. These latent budgets correspond roughly with the three types of persons, namely the males, the kids and the females respectively. Because the raw data were not available, we could only compute chi-square measures over the proportions. For the proportions these measures are 2.52 for p = 1 (df is 55), .96 for p = 2 (df is 46) and .37 for p = 3 (df is 18). Consider the case that for each row $n_{i+} = 50$, then the chi-squares are 124, 46 and 18 respectively. Comparing these chi-squares with the number of degrees of freedom shows that we cannot conclude for these data that more than three latent time-budgets are necessary to 'explain' the data.

The singular values from CA are .355, .224, .155, .068, and .039, which explain 61%, 24%, 12%, 2%, and 1% of the total inertia. The first two dimensions of the solution are shown in Figure 1. Not surprisingly, we see again the three clusters of males, females, and kids. We can now also use (13) to decompose residuals from model (1) for different numbers of latent time-budgets p. In case of p = 2, we find singular values .230, .155, .070, .047 and .016; in case of

p = 3 these values become .155, .069, .044, .019, .016. This illustrates that for this example approximation (1) in case of p, is about equal to approximation (11) in case of t = p - 1.

7. CONCLUSION AND DISCUSSION

We have illustrated the latent time budget model, and compared it with CA. We think we can conclude that the model might be useful as a descriptive tool in case it makes sense to think that the observed time budgets are generated by some typical latent ones. This was the case in the example shown. In principle the model can also be used for inferential purposes, whereas CA does not (directly) provide this possibility. On the other hand, CA provides us with plots that allow a fast interpretation.

Another aspect in comparing CA and these models is the following. The latter are perhaps easier to explain to lay men, as a method for finding "typical" time-budgets. On the other hand, it will sometimes be somewhat artificial to think of the observed time budgets as stemming from typical ones. However it was certainly useful here, because the first two dimensions of CA showed three clusters of time-budget points. The EM-algorithm for the time budget model can, in many cases, be considered as a form of CA in which the dimensions are rotated in such a way that an interpretation in terms of latent budgets becomes possible.

8. REFERENCES

[1] Converse, P.E. (1972). Country differences in time use. In A. Szalai[14].

[2] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1-38.

[3] Deville, J.-C. & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. Journal of Econometrics, 22, 169-189.

[4] Erasmus, C.J. (1955). Work patterns in a Mayo Village. American Anthropology, 57, 322-333.

[5] Escofier, B. (1983). Analyse de la difference entre deux mesures sur le produit de deux ensembles. Cahiers de l'analyse des données, 8, 325-329.

[6] Gross, D.R. (1984). Time allocation: A tool for the study of cultural behavior. Annual Review of Anthropology, 13, 519-558.

[7] Gross, D.R., Rubin, J., & Flowers, N.M. (1985). All in a day's time. Presented at the annual meeting of the AAAS, Los Angeles, 1985.

[8] Harvey, A.S., Szalai, A., Elliott, D.H., Stone, P.J., & Clarke, S.M. (1984). Time budget research. New York, Campus Verlag.

[9] Heijden, P.G.M. van der (1987). Correspondence analysis of longitudinal categorical data. Leiden: DSWO-press.

[10] Heijden, P.G.M. van der & Leeuw, J. de (1985). Correspondence analysis used complementary to loglinear analysis. Psychometrika, 50, 429-447.

[11] Jambu, M. & Lebeaux, M.-O. (1983). Cluster analysis and data analysis. Amsterdam: North Holland.

[12] Johnson, A. (1975). Time allocation in a Machiguenga community. Ethnology, 14, 310-321.

[13] Leeuw, J. de, Heijden, P.G.M. van der, & Kreft, I. (1985). Homogeneity analysis of event history data. Methods of Operations Research, 50, 299-316.

[14] Szalai, A. et al. (1972). The uses of time. The Hague: Mouton.

[15] Staikov, Z. (ed.). (1982). It's about time. Sofia, Bulgarian Academy of Science.

[16] Tuma, N.B. & Hannan, M.T. (1984). Social dynamics. Models and methods. New York: Academic Press.

Table 1: Parameter estimates for the time-budget model, for different values of p.

Table 1a: Budget weights, rowwise.

		p=1 k=1	p=2 k=1	k=2	p=3 k=1	k=2	k=3
Mekranoti	M	1.000	.763	.237	.107	.832	.061
	F	1.000	.725	.275	.253	.109	.638
	K	1.000	.054	.946	.876	.084	.040
Kanela	M	1.000	.558	.442	.346	.448	.206
	F	1.000	.782	.218	.220	.019	.761
	K	1.000	.038	.962	.929	.009	.063
Bororo	M	1.000	.458	.542	.363	.625	.012
	F	1.000	.828	.172	.146	.207	.647
	K	1.000	.108	.892	.783	.200	.017
Xavente	M	1.000	.331	.669	.520	.457	.024
	F	1.000	.982	.018	.002	.216	.783
	K	1.000	.134	.866	.799	.110	.091

Table 1b: latent budgets, columnwise.

Idle	.594	.391	.781	.817	.437	.391
Sleep	.060	.031	.087	.095	.032	.034
Care	.032	.068	.000	.000	.000	.116
Nonsubst	.174	.338	.023	.005	.271	.348
Domestic	.093	.105	.081	.080	.096	.110
Wild	.047	.067	.028	.003	.163	.000
Fit:	2.519	.963		.370		
Df:	55	38		21		

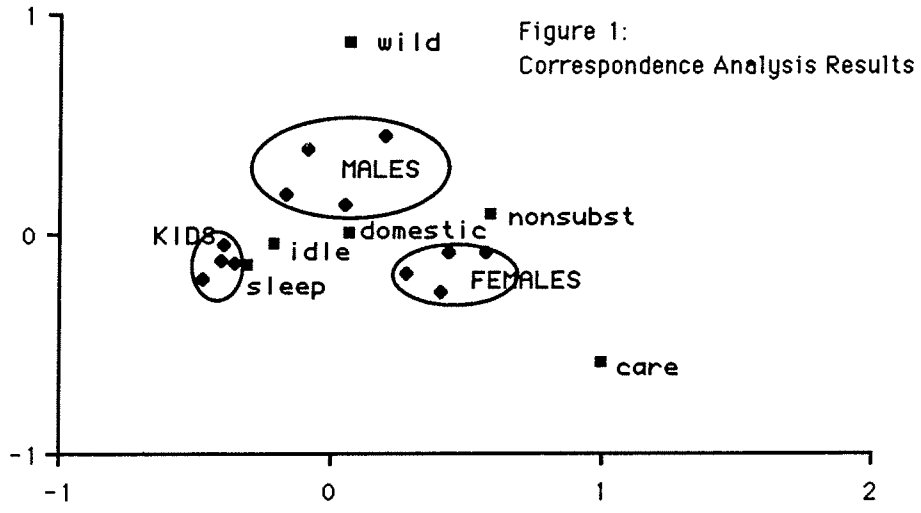

Figure 1:
Correspondence Analysis Results