ELSEVIER

Chemometrics and Intelligent Laboratory Systems 59 (2001) 75–88

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

A robust version of the Tucker3 model

V. Pravdova, F. Estienne, B. Walczak [1], D.L. Massart [*]

ChemoAC, Fabi-Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussels, Belgium

Received 1 April 2001; accepted 9 July 2001

Abstract

A new procedure for identification of outliers in Tucker3 model is proposed. It is based on robust initialization of the Tucker3 algorithm using Multivariate trimming or Minimum covariance determinant. The performance of the algorithm is tested by a Monte Carlo study on simulated data sets and also on a real data set known to contain outliers.

Keywords: Robust Tucker3 model; Alternating least squares; *N*-way PCA

1. Introduction

N-way methods based on the alternating least squares (ALS) algorithm are least squares methods that are highly influenced by outlying data points. One outlying sample can strongly influence the resulting model. As for 2-way PCA and related methods, there are two possibilities to deal with outliers: statistical diagnostics can be used or a robust algorithm can be constructed. Statistical diagnostics tools can be applied to the already constructed models and are usually based on the detection of the 'leverage points', defined as points that are far away from the remaining data points in the model space. This approach does not always work for multiple outliers because of the so-called masking effect. Robust versions of modeling procedures aim at building models describing the majority of data without being influenced by the outlying objects. By data majority, we mean the data subset containing at least 51% of objects. Robust procedures are characterized by the so-called breakdown point, defined as a percentage of data objects that may be corrupted while the model still yields the proper estimates. A subset of data, containing no outliers is called a 'clean subset'.

In the arsenal of chemometrical methods, there are already many robust approaches, such as robust PCA, PCR, PLS [1–4]. The aim of our study was to construct a robust version of the Tucker3 approach, one of the most popular *N*-way methods.

2. Theory

2.1. N-way methods of data exploration

Several methods were proposed for *N*-way exploratory analysis, for instance CANDECOMP/PARAFAC [5,6] and the family of Tucker models

* Corresponding author. Tel.: +32-2-477-4737; fax: +32-2-477-4735.

E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

[1] On leave from Silesian University, Katowice, Poland.

0169-7439/01/$ - see front matter
PII: S0169-7439(01)00143-5

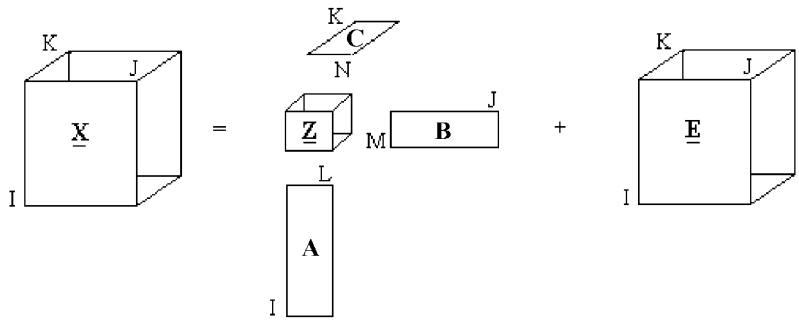

Fig. 1. The Tucker3 model.

[7,8]. In the present study, only the Tucker3 model is considered. Most of the *N*-way methods are based on ALS. The principle of ALS is to divide the parameters into several sets and for each set, the least squares solution is found conditionally on the remaining parameters. The estimation of parameters is repeated until a convergence criterion is satisfied. Fig. 1 shows the decomposition according to the Tucker3 model. The 3-way data matrix $\underline{\mathbf{X}}$ is decomposed into three orthogonal loading matrices $\mathbf{A}$ $(I \times L)$, $\mathbf{B}$ $(J \times M)$, $\mathbf{C}$ $(K \times N)$ and the core matrix $\underline{\mathbf{Z}}$ $(L \times M \times N)$ which describes the relationship among them. The largest squared elements of the core matrix $\underline{\mathbf{Z}}$ indicate the most important factors in the model of $\underline{\mathbf{X}}$. Mathematically, the Tucker3 model can be expressed as

$$x_{ijk} = \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} a_{il} b_{jm} c_{kn} z_{lmn} + e_{ijk} \qquad (1)$$

2.2. Data unfolding

For computational convenience, the Tucker3 algorithm used does not perform calculations directly on *N*-way arrays. The $\underline{\mathbf{X}}$ matrix is unfolded to standard 2-way matrices. This can be done in three different ways (see Fig. 2). Unfolded matrices are denoted as:

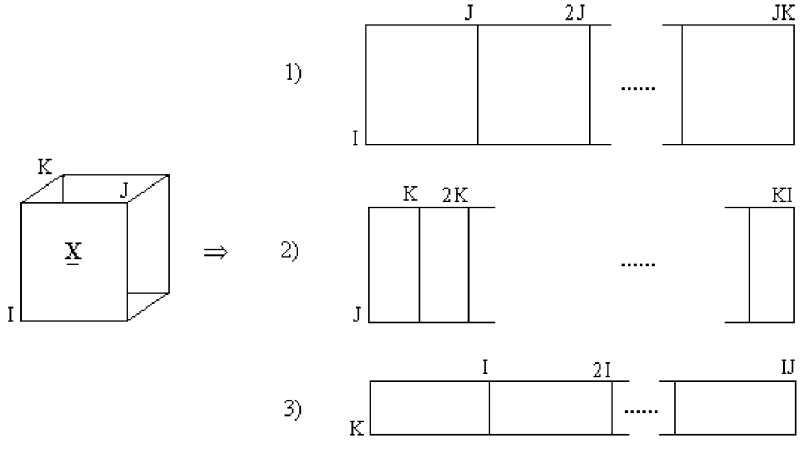

Fig. 2. Three different ways of unfolding of a 3-way data matrix.

$\mathbf{X}^{(I\times JK)}$, $\mathbf{X}^{(J\times IK)}$ and $\mathbf{X}^{(K\times IJ)}$. To calculate the loading matrices, several procedures can be used. Anderson and Bro [9] tested most of them with respect to speed and found NIPALS to be the fastest for large data arrays. In our algorithm, SVD is used for the estimation of **A**, **B** and **C** matrices.

2.3. Algorithm of Tucker3 model

(0) Initialize **B** and **C** (as random orthogonal matrices).

(1) $[\mathbf{A},v,d] = \mathrm{svd}(\mathbf{X}^{(I\times JK)}(\mathbf{C}\otimes\mathbf{B}),\ L)$.

(2) $[\mathbf{B},v,d] = \mathrm{svd}(\mathbf{X}^{(J\times IK)}(\mathbf{C}\otimes\mathbf{A}),\ M)$.

(3) $[\mathbf{C},v,d] = \mathrm{svd}(\mathbf{X}^{(K\times IJ)}(\mathbf{B}\otimes\mathbf{A}),\ N)$.

(4) Go to step 1 until the relative change in fit is small.

(5) $\mathbf{Z} = \mathbf{A}^T\mathbf{X}(\mathbf{C}\otimes\mathbf{B})$.

where symbols L, M, N denote numbers of factors in matrices **A**, **B** and **C**, respectively, and the symbol $\otimes$ denotes Kronecker multiplication: $\mathbf{A}\otimes\mathbf{B}$ yields the element-by-element multiplication of **B** with the elements from **A**, expressed as:

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

2.4. Robust PCA

One could think about robust initialization of the ALS algorithm, i.e. finding a clean subset for the matrix $\mathbf{X}^{(I\times JK)}$, but in reality, as the loading matrices **B** and **C** are only just initialized, the resulting matrix $(\mathbf{X}^{(I\times JK)}(\mathbf{C}\otimes\mathbf{B}))$ of dimensionality $I\times MN$ should be taken into account. The clean subset can be determined using such methods as for instance, multivariate trimming (MVT) [11] or minimum covariance determinant (MCD) [12]. Robust initialization of the Tucker3 algorithm seems to be the most important step to determine the final model and because this step is placed out of the main loop, the algorithm does not lead to oscillations. In the consecutive steps of ALS algorithm, the clean subset is constructed to decrease an objective function (see Eq. (4)), so that oscillations are avoided and convergence of the algorithm is achieved.

2.4.1. Multivariate trimming (MVT) [11]

The MVT procedure can be used for 'clean' subset selection when the input data matrix contains at least two times more objects than variables. The squared Mahalanobis distance (MD^2) is calculated according to the following equation:

$$\mathrm{MD}_i^2 = \left(\mathbf{t}_i - \bar{\mathbf{t}}\right)\mathbf{S}^{-1}\left(\mathbf{t}_i - \bar{\mathbf{t}}\right)^T \quad (2)$$

where $\mathbf{t}_i$ denotes the i-th object, $\bar{\mathbf{t}}$ denotes the vector containing means of data matrix columns and **S** is the covariance matrix.

A fixed percentage of objects (here 49%) with the highest MD^2 are removed and the remaining ones are used to calculate a mean and covariance matrix. MD^2 is calculated again for all objects using the new estimates of the mean and covariance matrix. Again, the 49% of objects with highest MD^2 are removed and the process is repeated until convergence of successive estimates of covariance matrix and mean. The subset of objects for which covariance and mean are stable is considered to be a clean subset of data.

2.4.2. Minimum covariance determinant (MCD) [12]

MCD aims at selecting a subset of h (out of m) objects, with the smallest determinant, i.e. the smallest volume in the p-dimensional space.

$$h = (m + p + 1)/2 \quad (3)$$

The MCD algorithm can be summarized as follows:

1. Randomly select 500 subsets of data containing $p + 1$ objects
2. For each subset:
   (a) Calculate its mean and covariance, **t** and **S**.
   (b) Calculate Mahalanobis distances for all objects using the estimates of data mean and covariance matrix calculated in step 2a.
   (c) Sort MD and take h objects with the smallest MD to calculate the next estimate of mean and covariance matrix.
   (d) Repeat steps b and c twice.
3. Take the 10 best solutions, i.e. the 10 subsets of h objects with the smallest determinants, and for each of them, repeat steps b and c until two subsequent determinants are equal.

4. Report the best solution, i.e. the subset with the smallest determinant.

The procedure starts with many very small data subsets (containing only $p+1$ objects) to increase the probability that these subsets do not contain outliers. Two iterations only are performed for all 500 subsets (steps 2b and 2c) to speed up the MCD procedure and, as demonstrated by Rousseeuw and Van Driessen [12], small number of iterations is sufficient to find good candidates of clean subsets. Only for the 10 best subsets are the calculations repeated till convergence of the algorithm.

2.5. Algorithm for robust Tucker3 model

To find possible multiple outliers in the first mode of the $\underline{\mathbf{X}}$, the following algorithm is proposed:

(0) Initialize loadings $\mathbf{B}$ and $\mathbf{C}$.

(1) Calculate $\mathbf{X}^{(I\times JK)}(\mathbf{C}\otimes\mathbf{B})$ and determine clean subset (using MVT or MCD).

(2) $[\mathbf{A}^*,\mathrm{v},\mathrm{d}] = \mathrm{svd}(\mathbf{X}^{(I^*\times JK)}(\mathbf{C}\otimes\mathbf{B}),\ L)$.

(3) $[\mathbf{B}^*,\mathrm{v},\mathrm{d}] = \mathrm{svd}(\mathbf{X}^{(J\times I^*K)}(\mathbf{C}\otimes\mathbf{A}^*),\ M)$.

(4) $[\mathbf{C}^*,\mathrm{v},\mathrm{d}] = \mathrm{svd}(\mathbf{X}^{(K\times I^*J)}(\mathbf{B}^*\otimes\mathbf{A}^*),\ N)$.

(5) $\mathbf{Z} = \mathbf{A}^{*T}\mathbf{X}^*(\mathbf{C}^*\otimes\mathbf{B}^*)$.

(6) Predict loadings A for all objects.

(7) Reconstruct $\mathbf{X}^{(I\times JK)}$: $\mathbf{X}^{(I\times JK)} = \mathbf{A}\mathbf{Z}^{(L\times MN)}(\mathbf{C}\otimes\mathbf{B})^T$.

(8) Calculate the sum of squared residuals for I objects in the first mode as the differences between the original data and the reconstructed one.

$$\text{residuals} = \mathrm{sum}\left(\left(\left(\mathbf{X}^{(I\times JK)} - \hat{\mathbf{X}}^{(I\times JK)}\right)^2\right)^T\right)$$

(9) Sort residuals along the first mode.

(10) Find h objects with the smallest residuals. They constitute the clean subset.

(11) Go to step 2 until the relative change in fit is small.

$\mathbf{A}^*$, $\mathbf{X}^*$, etc. are the matrices $\mathbf{A}$, $\mathbf{X}$, etc. limited to the clean subset of objects, and the notation $\mathbf{X}^{(I^*\times JK)}$ means that the unfolded data set contains objects reduced to the clean subset I^*. h is the number of objects in the clean subset.

In each iteration of the ALS subroutine, the loadings $\mathbf{A}^*$, $\mathbf{B}^*$ and $\mathbf{C}^*$ are calculated for the clean subset of objects only. In step 6, the loadings A are predicted for all objects and the set $\mathbf{X}^{(I\times JK)}$ is reconstructed with the predefined number of factors. Residuals between the initial $\mathbf{X}^{(I\times JK)}$ and the reconstructed $\hat{\mathbf{X}}^{(I\times JK)}$ are calculated and sorted, and 51% of objects with the smallest residuals is selected to form the clean subset for the next ALS iteration. The objective function, F, to be minimized, is the sum of squared residuals for the h clean objects from the first mode.

$$F = \sum\sum\left(\mathbf{X}^* - \hat{\mathbf{X}}^*\right)^2 \tag{4}$$

There is no guarantee that the selected clean subset is optimal, but convergence of the ALS approach is secured.

In this algorithm, the outliers are identified in the first mode only, but as all modes are treated symmetrically, one can look for outliers in any mode. This can be done simply by inputting the $\underline{\mathbf{X}}$ matrix with dimension of interest in the first mode.

2.5.1. Outlier identification

Once the robust Tucker3 model is constructed, the standardized residuals from that model are calculated for all objects of the first mode according to the following equation [10].

$$\mathrm{rs}_i = \mathrm{res}_i \Big/ \left[3\times 1.48\sqrt{\mathrm{median}\left(\left(\mathrm{res}_i - \mathrm{median}(\mathrm{res}_i)\right)^2\right)}\right] \tag{5}$$

where

$$\mathrm{res}_i = \sqrt{\sum\nolimits_j\left[\left(\mathbf{X}_{ij} - \hat{\mathbf{X}}_{ij}\right)^2\right]} \tag{6}$$

for $i = 1\ldots, I$ and $j = 1\ldots, JK$. In Eq. (5), the residuals are divided by the robust version of standard deviation. Using $1.48\sqrt{\mathrm{median}\left(\left(\mathrm{res}_i - \mathrm{median}(\mathrm{res}_i)\right)^2\right)}$, the residuals for 51% of objects, which fit the model best, are calculated. This corresponds to the robust standard deviation of the data residuals. Objects with standardized residuals higher than three times the robust standard deviation are considered as outlying and are removed from the data set. This is equivalent with using the ratio presented in Eq. (5) and cut-off equals one. The final Tucker3 model is constructed as the least squares model for the data after outlier elimination.

3. Data

3.1. Simulated data set

A systematic Monte Carlo study was performed to evaluate performance of the algorithm. A data set of dimensionality $(50 \times 10 \times 10)$ was simulated with two factors in all modes. Two Tucker3 models (X1 and X2) were constructed to explain 60% and 90% of data variance. The initial data sets were then contaminated with different types (T1–T4) and different percentages (20% and 40%) of outliers.

The different types of outliers (T1–T4) can be characterized as follows:

T1 Data set constructed according to the same model as the initial data, but with a certain percentage of randomly permuted variables.
T2 Data set with the same dimensionality and the same level of noise, but constructed according to a different tri-linear model.
T3 Data set with the same level of noise but with a higher dimensionality than the initial data set.
T4 Data set with the same level of noise but with a lower dimensionality than the initial data set.

The simulation of tri-linear data structure was performed as follows: first, orthogonal loading matrices **A**, **B** and **C** with predefined dimensions were randomly initialized. For the selected structure and core matrix $\underline{\mathbf{Z}}$, the $\underline{\mathbf{X}}$ matrix was constructed as $\mathbf{X}^{I\times JK} = \mathbf{A}\mathbf{Z}^{L\times MN}(\mathbf{C}\otimes\mathbf{B})^T$. Then, the Tucker3 model was built, and new $\underline{\mathbf{X}}$ was reconstructed with chosen number of factors in each mode and used as initial data set with tri-linear structure. At the end, white Gaussian noise was added to $\underline{\mathbf{X}}$. In this way models, which differ in percentage of explained variance, data complexity and structure of core matrix, can be constructed.

The two following types of calculations were performed for two data models (X1 and X2), each with four types of outliers (T1–T2) and two percentages (20% and 40%):

1. One contaminated data set was constructed and the Tucker3 and robust Tucker3 models were built 100 times with random initialization of loadings **B**, **C**.
2. The construction of Tucker3 and robust Tucker3 models was repeated 100 times for the predefined type and percentage of outliers, but this time outliers were simulated randomly according to the chosen type in each run.

The performance of the algorithms is presented in the form of a percentage of unexplained variance for the constructed final models. In the case of robust Tucker3 approach, the final model is considered to be the Tucker3 model after outlier removal. The MVT procedure was applied in the Monte Carlo study to speed up calculations.

3.2. Real data set

An electroencephalographic (EEG) data set was used. The principle of electroencephalography is to give a representation of the electrical activity of the brain [13]. This activity is measured using metal electrodes placed on the scalp. The data was acquired during the testing phase of a new antidepressant drug. The effect of the drug was followed in time over a 2-day period (12 measurements). The EEGs were measured on 28 leads located on the patient's scalp. Each of the EEG was decomposed using the fast Fourier transform into seven energy bands commonly used in neurosciences [14]. Only the numerical values corresponding to the average energy of specific frequency bands are taken into account. This leads, for each patient, to a 3-way array with dimensions $(28 \times 7 \times 12)$. The study was performed on 12 patients. Only the results corresponding to two patients are shown here. Patient #6 shows a very typical behaviour, while patient #9 has aberrant results for electrode #12.

4. Results and discussion

4.1. Monte Carlo study

Let us consider the data set X1 contaminated with 20% outliers of type 1 (T2). The Tucker3 model for this data set is presented in Fig. 3. As one can notice there are 10 objects far away from the remaining ones, and the Tucker3 model is highly influenced by them.

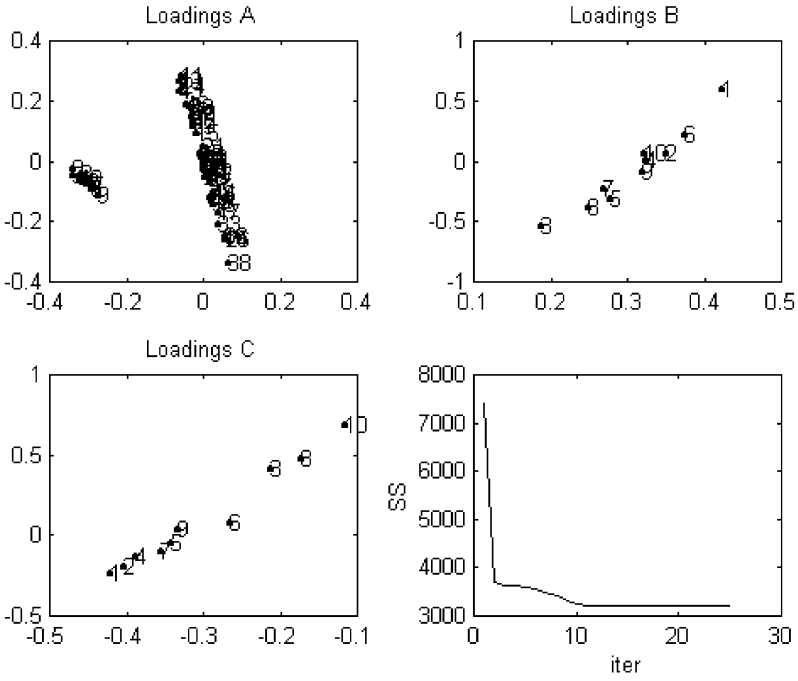

Fig. 3. Tucker3 model for data set X1 (90% of explained variance) with 20% of outliers (type T1).

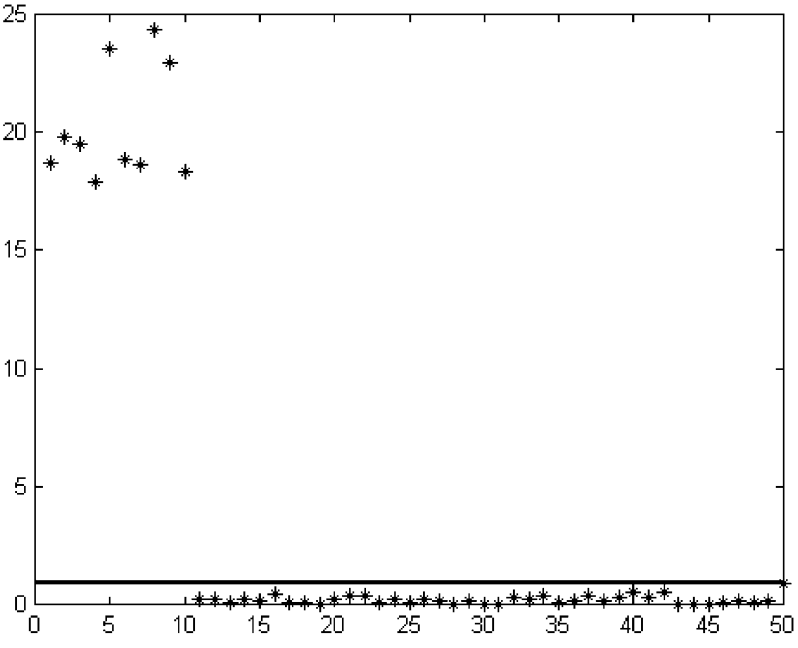

Fig. 4. Residuals from the robust Tucker3 model, data set X1, 20% contamination, type T1.

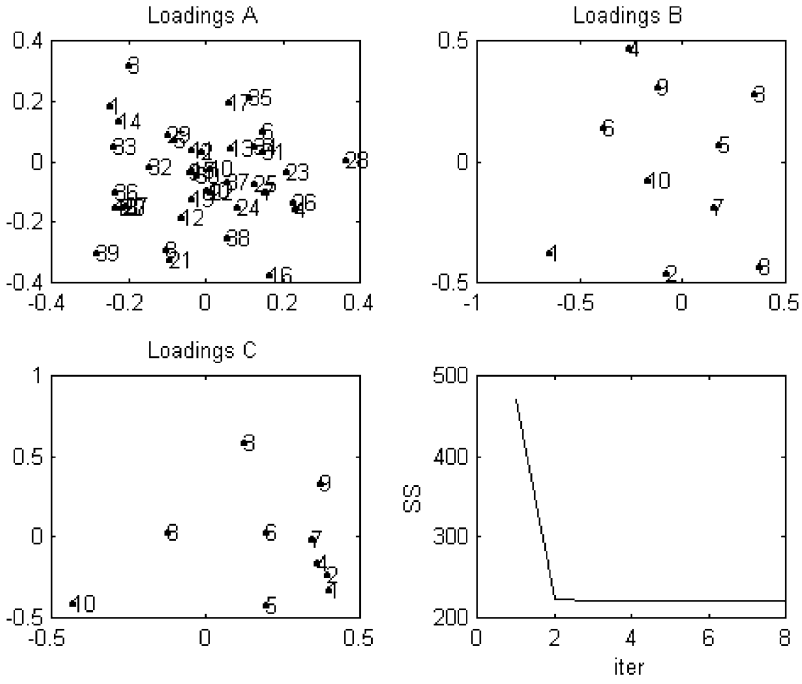

Fig. 5. Final Tucker3 model after elimination of identified outliers.

For the same data set, the robust Tucker3 model was constructed and the object residuals from that model are presented in Fig. 4. The first 10 outlying objects are correctly identified as the outlying ones. After their removal, the final Tucker3 model is constructed and its results are presented in Fig. 5.

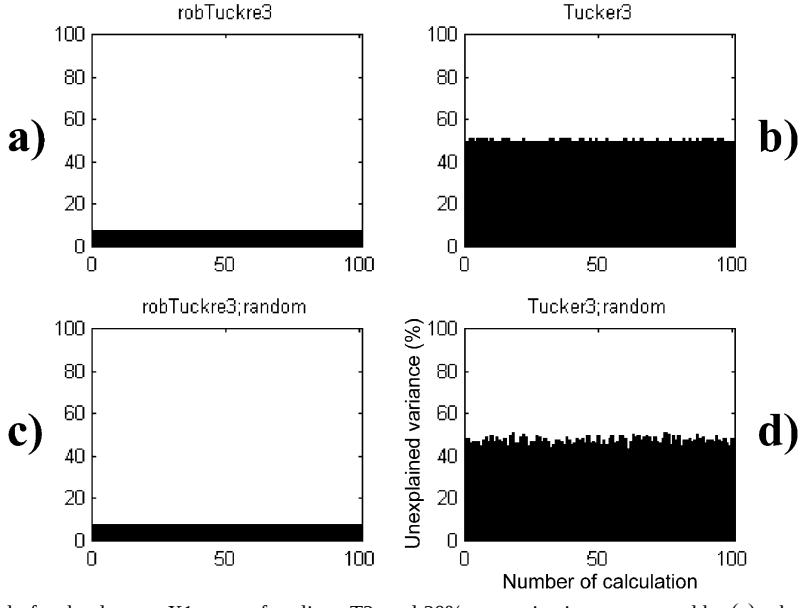

Fig. 6. Monte Carlo study for the data set X1, type of outliers, T2, and 20% contamination constructed by (a) robust Tucker3, (b) Tucker3 model with random initialization and (c) robust Tucker3, (d) Tucker3 model with each time randomly generated outliers.

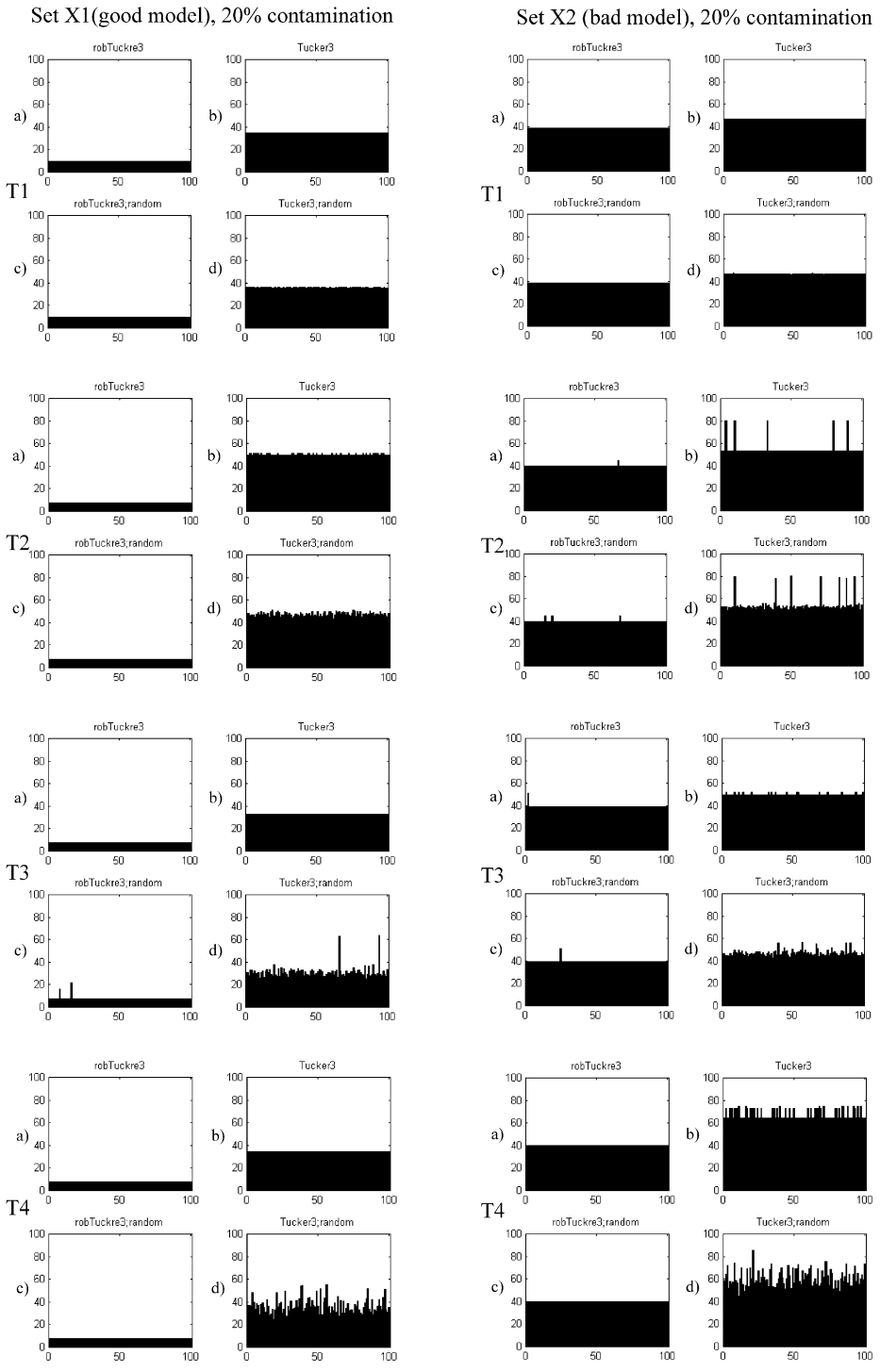

Fig. 7. Final results for Monte Carlo study for contamination 20% (data sets X1 and X2, type of outliers T1–T2).

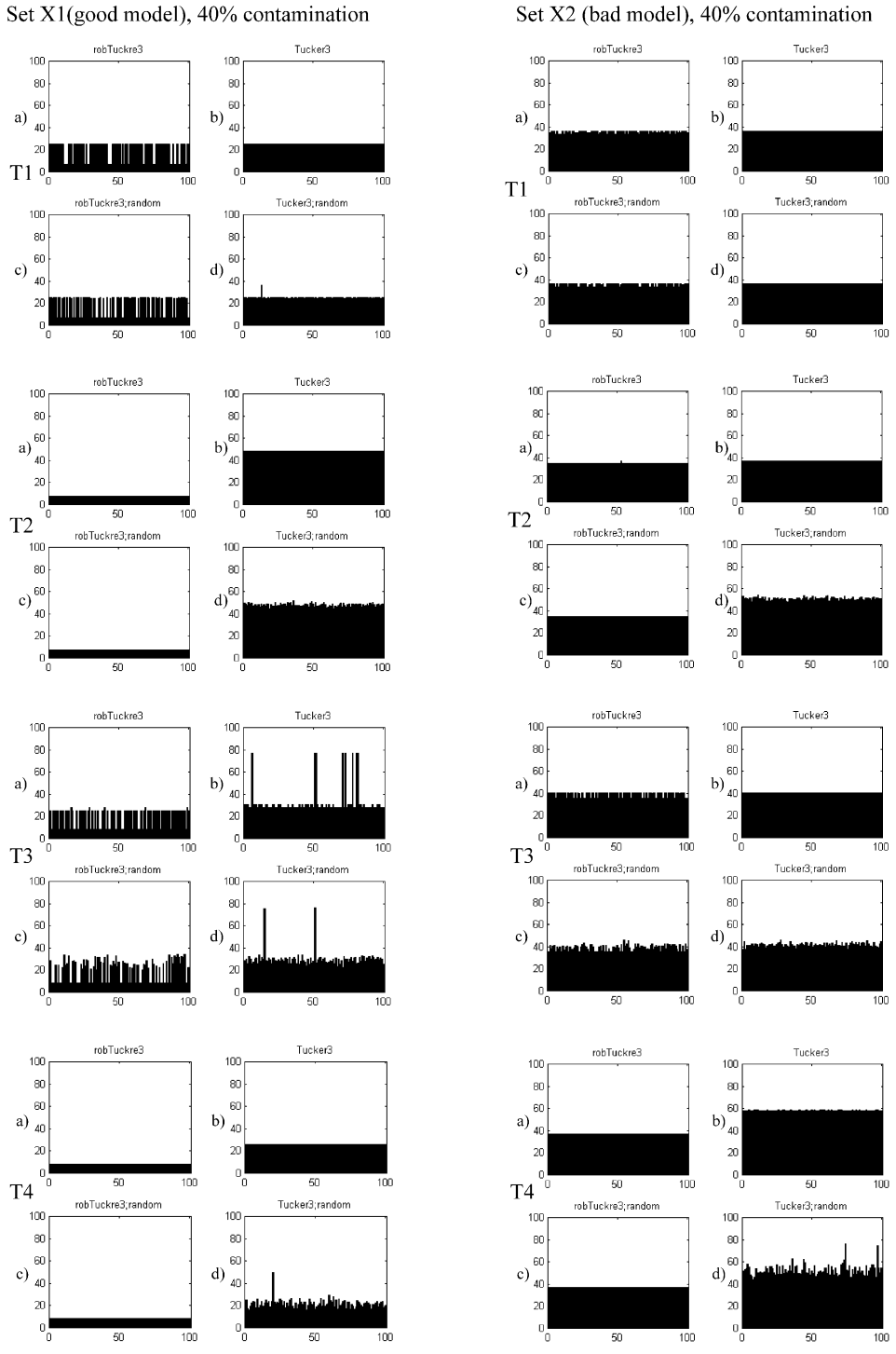

Fig. 8. Final results for Monte Carlo study for contamination 40% (data sets X1 and X2, type of outliers T1–T2).

For each studied data set, the Tucker3 and robust Tucker3 algorithms were run 100 times with random initialization of loadings. The results for the discussed data set, expressed as the percentage of the explained variance, are presented in bar form in Fig. 6a.

The observed results show that the robust Tucker3 algorithm always converges to the proper solution, and that the outlying objects do not influence the final Tucker3 model.

Analogous results for the (non-robust) Tucker3 model are presented in Fig. 6b. They indicate that the

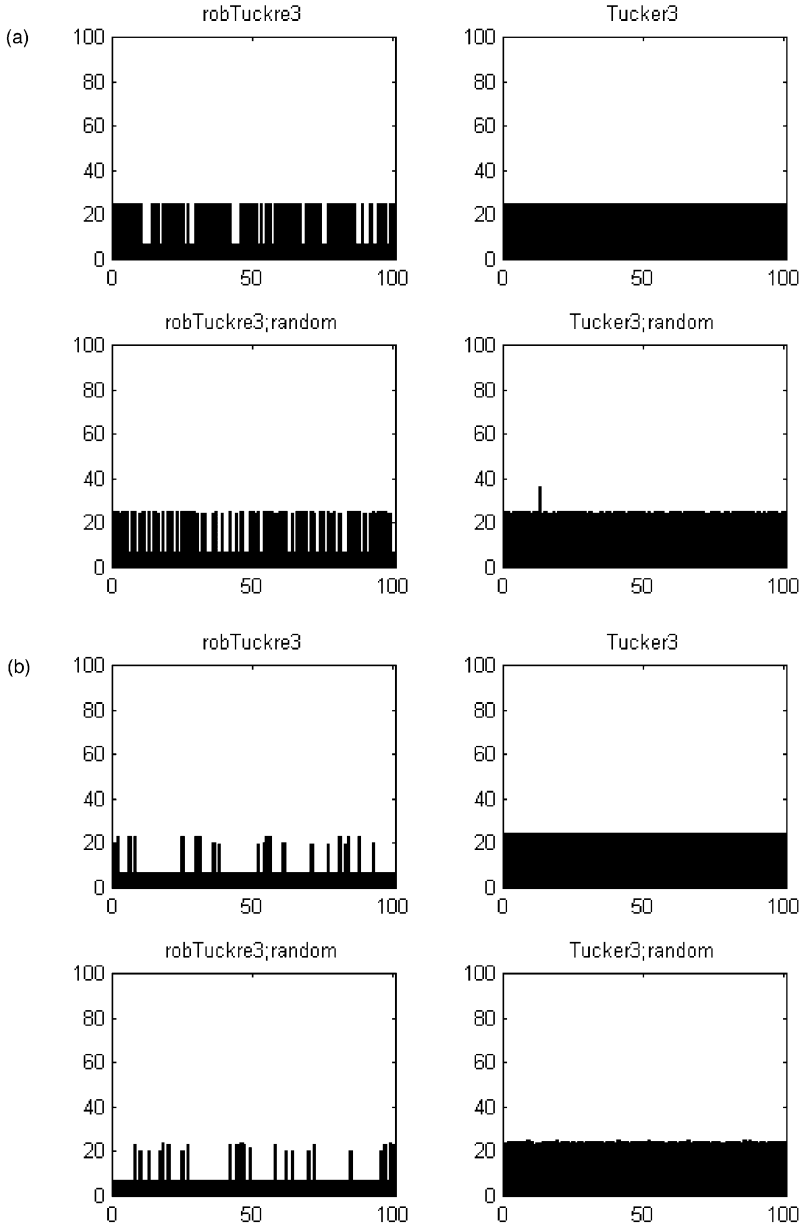

Fig. 9. Comparison of two algorithms for finding a clean subset. (a) Multivariate trimming (MVT). (b) Multivariate covariance determinant (MCD).

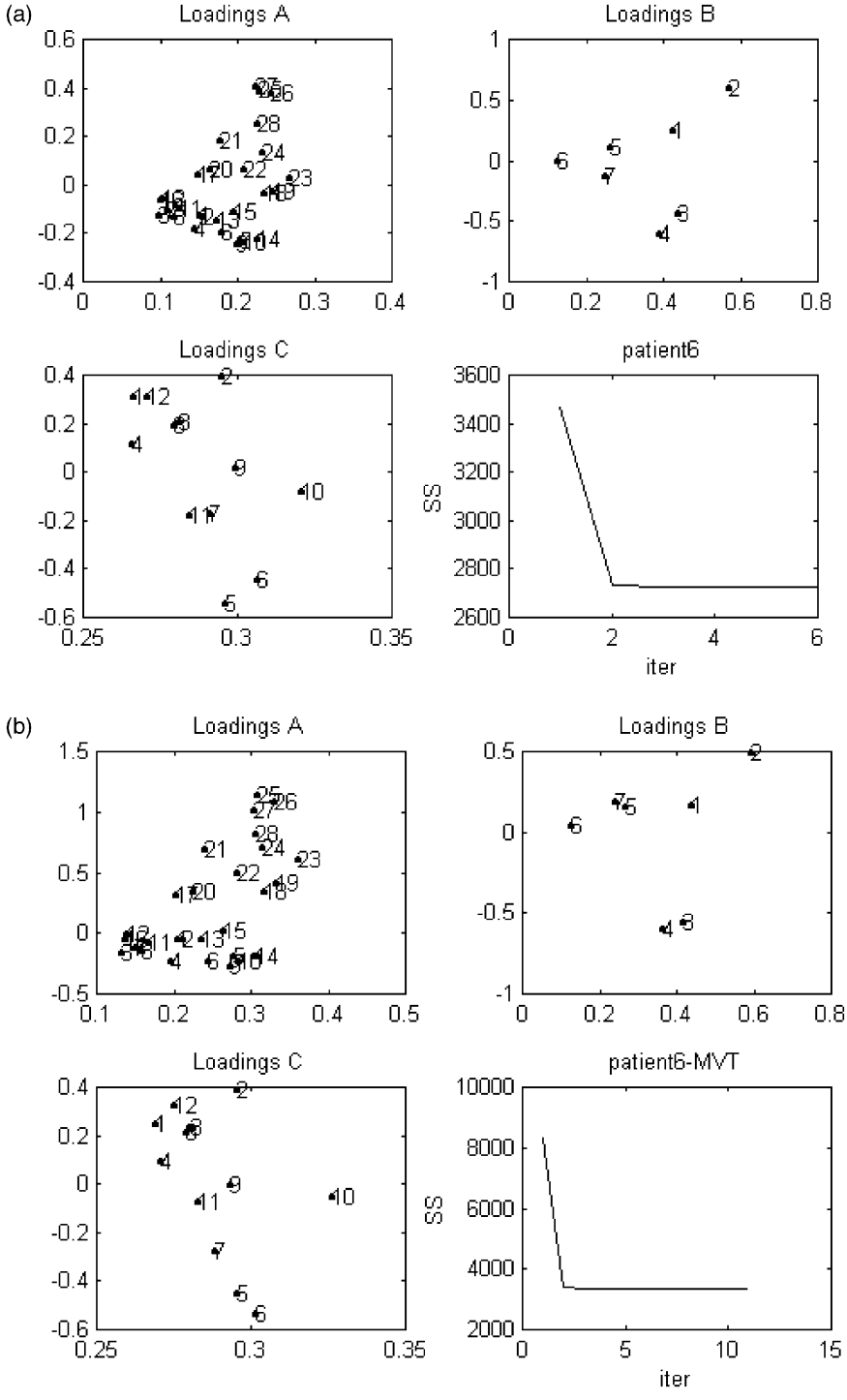

Fig. 10. **A**, **B** and **C** loading matrices and convergence times for patient #6: (a) Tucker3 model. (b) Robust Tucker3 model.

Tucker3 algorithm is highly influenced by outliers and, depending on the initialization of the loadings, the algorithm converges to different solutions.

In the next step of our study, both algorithms, i.e. Tucker3 and robust Tucker3, were run 100 times, each time once for a different data set contaminated

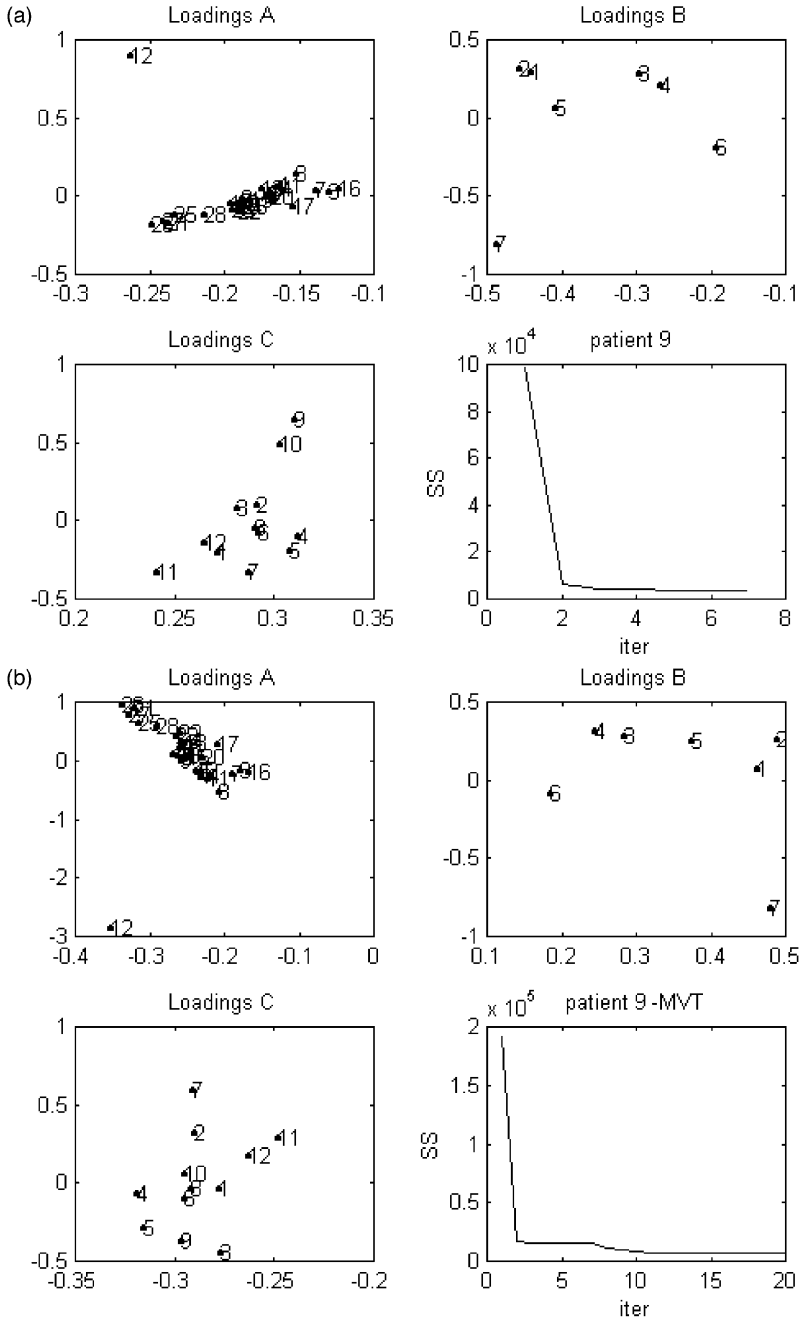

Fig. 11. **A**, **B** and **C** loading matrices and convergence times for patient #9 : (a) Tucker3 model. (b) Robust Tucker3 model.

randomly with 20% of outliers constructed according to the chosen model (type T2). The results are presented in Fig. 6c and d. The robust Tucker3 algorithm always leads to the proper model not influenced by outlying objects, whereas the Tucker3 models are highly influenced by them.

The calculations described above were performed for the data sets contaminated with different percentages of outliers of different types. The final results, presented in Fig. 7, reveal that the proposed robust version of the Tucker3 model works properly for data sets containing no more than 20% of outlying samples. The robust models constructed for data sets X1 and X2 with 20% of outliers, i.e. data sets with a different percentage of explained variance, are not influenced by outliers.

The final results for data sets X1 and X2 with 40% of outliers are presented in Fig. 8. The robust model performed properly only for two types of outliers (T2 and T4). The results for the types T1 and T3 were strongly influenced by the procedure for the selection of the clean subset. Here, MVT results are presented; those with MCD are somewhat better.

Analogous calculations were performed for the data sets with clustering tendency. The results of the Monte Carlo study for these data sets lead to the same conclusions.

While working with the highly contaminated data sets (40%), it was noticed that there is an essential difference depending on the methods used to select a clean subset. In Fig. 9, the results for X1 (40% of outliers T1; simulation type 2) achieved with MVT and MCD are presented for illustrative purposes.

The observed differences in MVT and MCD performance for highly contaminated data (40%) are associated with different breakdown points of those methods. MCD with breakdown point 50% performs better, but due to the relatively long computation time required, it was not used in the Monte Carlo study.

4.2. Real data set

The classical and robust Tucker3 algorithms were applied on the real data set. The results obtained for patient #6 (the one without outlying object) show (Fig. 10a–b) that the classical and the robust Tucker3 models are equivalent on this normal patient.

Moreover, convergence is as fast in both cases. The results obtained for patient #9 with the classical Tucker3 model (Fig. 11a) already spots object #12 as

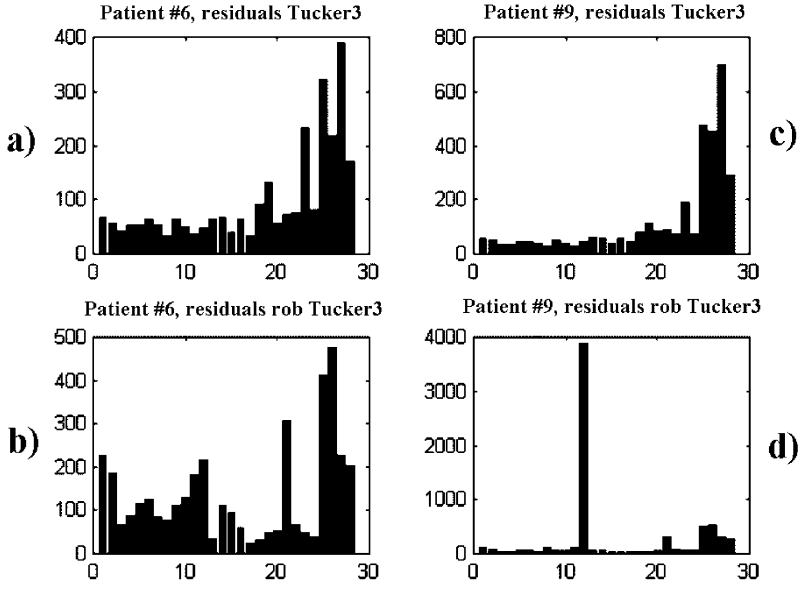

Fig. 12. Residuals obtained for the reconstruction of the objects on the 1st mode (12 electrodes): (a) patient #6, Tucker3 model; (b) patient #6, robust Tucker3 model; (c) patient #9, Tucker3 model; (d) patient #9, robust Tucker3 model.

an outlier on the A loading plot (corresponding to the electrodes dimension). This is even more obvious when using the robust version of the algorithm (Fig. 11b) as scale is different.

In the case of the robust Tucker3, the loadings on B and C are not influenced anymore by electrode #12 as the corresponding slice of the matrix is not used in the model construction. For patient #6, the residuals obtained for the 1st mode (electrodes dimension) with the classical method (Fig. 12a) and the robust method (Fig. 12b) show the same pattern. The situation is very different for patient #9. For the classical Tucker3 model, the residuals for electrode #12 (Fig. 12c) are not higher than the residuals of other points corresponding to good electrodes. The outlying electrode is therefore not found by the model residuals. For the robust Tucker3 model the residuals for electrode #12 (Fig. 12d) are extremely high and the outlier can be found and eliminated. In the robust Tucker3 approach, the loadings on A, B, and C are really robust. The reconstruction is good for all of the points except electrode #12.

5. Conclusion

The performed study shows that the robust version of the Tucker3 model always converges to a good solution when the data are contaminated by 20% outliers. For 40% contamination, the algorithm converges to a good solution only for two types of outliers (T2 and T4). It can be concluded that MCD is better algorithm for finding the clean subset than MVT. The robust Tucker3 algorithm gives good results also for the real data set.

Acknowledgements

DLM thanks the FWO project (G.0171.98) and EU project NWAYQUAL (G6RD-CT-1999-00039) for funding this research.

References

[1] Y.L. Xie, J.H. Wang, Y.Z. Liang, L.X. Sun, X.H. Song, R.Q. Yu, J. Chemom. 7 (1993) 527–541.
[2] B. Walczak, D.L. Massart, Chemom. Intell. Lab. Syst. 27 (1995) 354–362.
[3] I.N. Wakeling, H.J.H. Macfie, J. Chemom. 6 (1992) 189–198.
[4] J.D. Carroll, J.J. Chang, Psychometrica 35 (1970) 283–319.
[5] R.A. Harshman, UCLA Work. Pap. Phonetic 16 (1970) 1–84.
[6] L.R. Tucker, Problems in Measuring Change, The University of Wisconsin Press, Madison, 1963, pp. 122–137.
[7] L.R. Tucker, Psychometrica 31 (1966) 279–311.
[8] C.A. Andersson, R. Bro, Chemom. Intell. Lab. Syst. 42 (1998) 93.
[9] L.P. Ammann, J. Am. Stat. Assoc. 88 (1994) 505–514.
[10] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[11] R. Gnanadesikan, J.R. Kettenring, Biometrics 28 (1972) 81–124.
[12] P.J. Rousseeuw, K. Van Driessen, Technometrics 41 (1999) 212.
[13] M.J. Aminoff, Electrodiagnosis in Clinical Neurology, 2nd edn., Churchill Livingstone, NY, 1986.
[14] H.H. Jasper, Electroencephalogr. Clin. Neurophysiol. 10 (1958) 370.