

A STRATEGY FOR ANALYZING LINEAR MIXING DATA

Werner A. Stahel

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

June 2003

Abstract

The linear mixing model describes situations in which some basic components, like the emissions of some sources of air pollution, are mixed, and the mixture is observed on many variables. It resembles the model of factor analysis with factors that have a clear physical meaning.

This report focusses on describing some methods for fitting the linear mixing model to multivariate data, with an emphasis on graphical tools. Functions in S language available on the web implement these methods.

1 Introduction

The multivariate model of linear mixtures has wide applications in several fields of science.

Pollution data. An example appears in air pollution, when many chemical compounds can be measured. A data set to be used in this paper consists of 710 measurements of 16 volatile organic compounds (VOC). The concentrations of these compounds will come from a few sources of pollution, like exhaust and evaporation of gas from road traffic and solvents in industrial production. It is of interest to know how much of the pollution is due to which source.

A simple model for this situation is based on two assumptions:

- The sources emit compounds in constant proportions, i.e., if they are more active at one time than at another, the emission of all compounds is higher, but the proportions are still the same. These proportions define the sources “**profiles**” or “fingerprints.”
- Chemical reactions during transport of the polluted air from the site of emission to the site of the measurements are negligible.

The basic model of linear mixtures. With these assumptions, we get the model

$$X_i^{(j)} = Z_i^{(j)} + E_i^{(j)} = \sum_{k=1}^q S_i^{(k)} C_k^{(j)} + E_i^{(j)}, \quad i = 1, \dots, n, \quad j = 1, \dots, m. \quad (1)$$

Here, $X_i^{(j)}$ is the concentration of compound j for time i , $S_i^{(k)}$ reflects the k th source’s activity (as far as it is relevant at the site of measurement) for time i – called **scores** in this paper –, $C_k^{(j)}$ is the proportion of compound j in the k th source’s profile, and $E_i^{(j)}$ is a **random error** term comprizing the measurement error and possibly additional deviations from the ideal model.

The model states that the observed profiles are a linear mixture of the **source profiles**. Clearly, the values of the profiles (both $X_i^{(j)}$ and $C_k^{(j)}$) as well as the contributions $S_i^{(k)}$ cannot be negative.

Applications. This model applies to many situations in various sciences, some of which are

- Environment: Pollution in the air (see above) or in water.
- Geology: Formation of rock.
- Hydrochemistry: Soil layers are passed by water, which absorbes some minerals in relation to the composition and the thickness of the layers Akerjord and Christophersen (1996).
- Spectroscopy: Optical spectra of mixtures in chemical products or reactions.
- Chromatography: Intensity of chromatogram for wavelength j and distance (running time) i . Osten and Kowalski (1984) treat only 2 components, Vanderginste, Essers, Bosman, Reijnen and Kateman (1985) provide a basic reference.

- Remote Sensing: A spectrum can be obtained for each pixel of a satellite image. Different land uses have specific spectra and are mixed within a pixel to form a mixed spectrum.

Chemical Mass Balance and unmixing. In some applications, it is reasonable to assume that the source profiles are known. Then, the determination of the scores $S_i^{(k)}$ is quite straightforward. Often, there is not enough prior knowledge about the source profiles, and thus, they should be estimated from the data along with the scores. For some sources, a known profile may be a good approximation to the one that is active in the data, and it is desirable to use such knowledge.

Additional structure and external information. In our example, as in many other situations, the observations form a (multivariate) time series. It is reasonable to expect periodic – yearly and weakly – behavior and some smoothness of the “true values” $Z_i^{(j)}$. Similarly, a geographical location may be relevant in other studies.

There may be information about source activities, like traffic counts, or other processes influencing the measurements, like the weather. They may facilitate the interpretation of results, or they may even be used for finding meaningful solutions. Knowledge of this type have given rise to several application specific methodologies, often of an ad hoc nature.

Overview. The analysis of mixing data may proceed in several steps, as we will discuss in the next section. In the later sections, we will describe the steps in more detail and point to functions written in the S language and implemented for the software R or S-Plus. They are available from our website stat.ethz.ch

The most advanced method to achieve unmixing is, as far as we are aware, the Positive Matrix Factorization (PMF) program of Paatero and Tapper (1994). One of our functions is similar in nature, but less refined. Many of the graphical functions introduced here will be useful in combination with any algorithm for finding estimates for the source profiles $C_k^{(j)}$ and the scores $S_i^{(k)}$ in (1).

2 The Linear Mixing Model

The Model. The general linear mixing model with n observations, m variables, q sources and a measurement error term $E_i^{(j)}$ is given by 1. It is conveniently written in matrix form,

$$\mathbf{X} = \mathbf{Z} + \mathbf{E} = \mathbf{S}\mathbf{C} + \mathbf{E} , \quad (2)$$

where \mathbf{X} is the data matrix showing the measurements \underline{X}_i for the i th observation as the i th row, and \mathbf{Z} , \mathbf{E} , and \mathbf{S} are defined accordingly. The source matrix \mathbf{C} collects the source profiles \underline{C}_k as its rows.

Since concentrations of compounds and contributions of sources cannot be negative, the following inequalities must hold:

$$(a) \quad X_i^{(j)} \geq 0 , \quad (b) \quad C_k^{(j)} \geq 0 , \quad (c) \quad S_i^{(k)} \geq 0 . \quad (3)$$

Figure 1 shows a simplified example with two source profiles \underline{C}_1 and \underline{C}_2 and three observed profiles \underline{X}_i without errors \underline{E}_i .

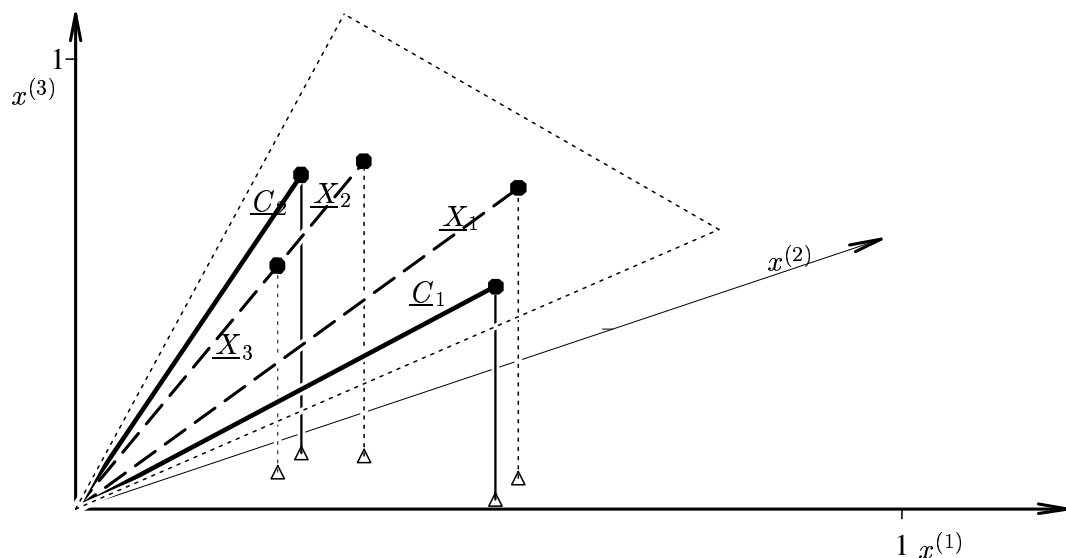


Figure 1: The Basic Model for three variables and two sources. Three possible data profiles without error terms are shown.

Assumptions about S . The source contributions S are often taken as fixed and unknown. They then play the role of incidental parameters. Alternatively, they can be seen as random variables following a parametric distribution. This leads to different methods for estimating parameters, source profiles, and scores.

Non-uniqueness of C . A basic difficulty of the model with fixed scores is that the parameters are not identifiable, since for any regular matrix T ,

$$SC = (ST^{-1})(TC) =: S^*C^*,$$

and if S^* and C^* fulfill the constraints, they lead to the same distribution of the observations.

A simple special case is a diagonal matrix T : each column of S can be multiplied by a constant and the corresponding row of C divided by the same constant without changing the observations. It is natural to avoid this ambiguity by setting a side condition on the rows of C . We will require the source profiles to be in compositional form (see below), that is,

$$\sum_j C_k^{(j)} = c \tag{4}$$

with a fixed c . We choose $c = m$ to make an “average” $C_k^{(j)} = 1$.

In order to visualize the remaining more general non-uniqueness problem, consider in figure 1 any vectors $\underline{C}_1^*, \underline{C}_2^*$ with $C_k^{(j)*} \geq 0$ which lie in the same plane as \underline{C}_1 and \underline{C}_2 outside the angle between them. Such a pair is also suitable for describing the given observations, as are \underline{C}_1 and \underline{C}_2 . The vectors which enclose the observations most tightly may be singled

out as the desired parametrization. On the other hand, the most extreme vectors which still satisfy the non-negativity constraints may also be a plausible solution. Note that this difficulty arises even if there are no errors!

A further assumption about the sources postulates that the source profiles be linearly independent, that is, that no source profile can be written as a linear combination of other source profiles.

Assumptions about the error term. Measurement errors may clearly be negative. The simplest distributional assumption is a normal law,

$$E_i^{(j)} \sim N(0, \sigma_{ij}^2), \quad (5)$$

Often, a constant standard deviation $\sigma_{ij} = \sigma$ is assumed (at least after suitable standardization). In reality, however, the precision of a measurement usually depends on its value, typically being proportional to it.

Compositional data and standardization In many applications, the data come as proportions or compositional data – that is, they add to 1, 100 percent or another suitable constant c .

For data without such restrictions, it may help to standardize the observations to obtain compositional form. Let

$$\tilde{X}_i^{(j)} = X_i^{(j)} / U_i, \quad U_i = \sum_j X_i^{(j)}$$

Then we may formally write

$$\tilde{X}_i^{(j)} = \tilde{Z}_i^{(j)} + \tilde{E}_i^{(j)}, \quad \tilde{Z}_i^{(j)} = \sum_k \tilde{S}_i^{(k)} C_k^{(j)}$$

with $\tilde{S}_i^{(k)} = S_i^{(k)} / U_i$, $\tilde{E}_i^{(j)} = E_i^{(j)} / U_i$, and the same source profiles $C_k^{(j)}$ as before. (These equations are useful for data but difficult to interpret as a model, since the $\tilde{S}_i^{(k)}$ depend on the random errors $E_i^{(j)}$.)

Steps. To fit the model to a data set, we proceed in these steps:

- A. Initial: Prepare and screen the data, give some auxiliary information,
- B. Subspace: Determine the subspace in which the data lies up to errors,
- C. Sources: Determine the source profiles,
- D. Model Checking: Assess the quality of the solution,
- E. Final: Analyze the results to give additional support for interpretation.

We will address these steps in Sections 4 to 8.

3 The script file

The steps which are typical for an analysis of mixing data have been collected into a script file in R or S-Plus that can be closely followed. It reflects some of the habits and conventions that we usually follow in any data analysis.

Graphical displays generated by our graphical functions will be “stamped” by the generation date and some minimal information about the project and the stage of the analysis. This is done by the function `g.stamp`, which looks for a project title and a string describing the current step. The stamp can be suppressed by writing `c.env$stamp <- 0`, i.e. for producing figures for a final report.

Objects generated in the course of the analysis begin by a letter and a dot, like `t.data`. The letters preceding the dot characterize the type of object. Functions provided by us begin by `f.` if their output is numeric and by `g.` if they produce a graphical display. Data to be saved and used throughout the analysis start by `d.`, and results of some longer lasting relevance, by `r.`. Temporary objects start by `t.`. They can be removed at the end of any step. The analysis also needs some constants which tune the analysis. They start by `c.`. One of them is the list `c.env` which provides the information about the project needed by `g.stamp` and determines if stamping should occur.

4 Step A: Data preparation and visualization

Step A 1: Data. The data to be analyzed is assumed to be stored as the data.frame `d.data`. Its precision can be given as a data.frame with the same dimension as `d.data`. Alternatively, a vector of relative errors and of absolute errors must be provided.

As a first step, the variables of `d.data` to be analyzed are selected and possibly rearranged. Since the **sequence of the variables** will be reflected in profile plots showing the components $X_i^{(j)}$ or $C_k^{(j)}$, the user should select a meaningful sequence.

These plots as well as other methods of analysis also require that the variables be expressed in **comparable units**. One way to achieve this goal is to scale the variables to equal medians. The result is stored in `d.dmat`.

The specifications of the error standard deviations σ_{ij} must be adjusted to these modifications.

Step A 2: Screening.

A scatterplot matrix reveals the relationships between the variables and helps to detect outliers at the same time. (Some statisticians may prefer to examine QQ-plots first.) A possible shift in the measurements may be seen if the distribution does not start at 0. Note that there is no assumption about the distribution of the X variables, and that transformations of variables are inadequate since they make the basic model meaningless.

Selected observation vectors \underline{X}_i are drawn as “data profiles” by the function `g.profile`. Figure 2 shows six series of eight consecutive observations in this way. All the figures are collected at the end of this report.

Plotting Variables. In many cases, including our example, the observations form a time series. It is therefore meaningful to plot each variable series in sequence. Since sources of pollution may be connected to weather as well as human activities, daily, weekly and yearly cycles are plausible. The function `g.ts.wrapped` shows such time series in an informative fashion. Figure 3 shows one of the variables with a period of 168 hours – a week.

Result of Step A. As a result of this step of analysis, the data should be plausible and meaningfully organized. For convenience, the data without observations containing missing values is stored in `d.wona`, and `d.sig` is adjusted accordingly.

5 Step B: Determination of the subspace and the number of sources

5.1 Methods for finding the subspace

Principal Component Analysis and Singular Value Decomposition. The most widespread technique for finding a linear subspace is principal components analysis. It provides a maximum likelihood solution for the model (1) if all the observations have the same precision, $\sigma_{ij} = \sigma$. The programs usually center the data first, and this does not conform to the model. Therefore, they can be used only if they include an option which bypasses centering. The estimated subspace is determined by

$$\mathbf{X} = \mathbf{Z} + \mathbf{R} = \mathbf{Y}\mathbf{Q} + \mathbf{R},$$

where \mathbf{Y} comprises the scores and \mathbf{Q} , the loadings of the first q principal components, and \mathbf{R} stands for the estimated error matrix or the residuals.

The Singular Value Decomposition is closely related to Principal Component Analysis. Programs will not center the data and are therefore suitable for finding the subspace.

Compositional data. Compositional data lies in an $m - 1$ -dimensional simplex. Its intersection with the space spanned by the q source profiles is $q - 1$ -dimensional. This subspace does not contain the origin of the coordinate system any more. Thus, the above methods are applied to centered data, resulting in

$$\tilde{\mathbf{X}} - \underline{\mathbf{1}}\tilde{\mathbf{X}}^T = \check{\mathbf{Y}}\check{\mathbf{Q}} + \mathbf{R}$$

(where $\tilde{\mathbf{X}}$ is the mean of the $\tilde{\mathbf{X}}_i$) and the desired decomposition of the non-centered data reads

$$\tilde{\mathbf{X}} = [\underline{\mathbf{1}} \quad \check{\mathbf{Y}}] \begin{bmatrix} \tilde{\mathbf{X}}^T \\ \check{\mathbf{Q}} \end{bmatrix} + \mathbf{R}. \quad (6)$$

Standardization of observations. If PCA (or SVD) are used, we recommend to standardize the data in such a way that the error standard deviations σ_{ij} may be assumed similar. To this end, the variables are first scaled to have equal mean variances $\sigma_j^2 = ave_i(\sigma_{ij})$, and then the observations are divided by the row sums $U_i = \sum_j X_i^{(j)}$.

Thus, a simple procedure to obtain an estimated subspace for a compositional or general data set is PCA applied to standardized data. `f.pcascaled` does this, applying first the scaling just mentioned.

Maximum likelihood for distinct σ_{ij} . Under the model with fixed scores and normal errors $E_i^{(j)}$ with general standard deviations σ_{ij} , the maximum likelihood estimator minimizes

$$Q(\mathbf{R}) = \sum_{i,j} \left(\frac{R_i^{(j)}}{\sigma_{ij}} \right)^2 \quad (7)$$

where \mathbf{R} is the matrix of residuals, $\mathbf{R} = \mathbf{X} - \mathbf{S}\mathbf{C}$.

Robust estimation. Since the errors $E_i^{(j)}$ should not be expected to follow a normal distribution, it is advisable to use a robust estimator, which minimizes a criterion of the form

$$Q(\mathbf{R}) = \sum_{i,j} \rho \left(\frac{R_i^{(j)}}{\sigma_{ij}} \right) \quad (8)$$

with a suitable function ρ instead of (7). If an additional source, which is not modelled by the subspace considered, were active for a few observations, then outliers would be generated. They would tend to show positive residuals $R_i^{(j)} > 0$, since the contributions of the additional source can only be positive. Therefore, and because positively skewed error distributions are common, outliers in the right tails of the distributions of $E_i^{(j)}$ should be treated more liberally than on the left side. This suggests an asymmetric ρ function.

The most simple, but not numerically optimal procedure to minimize (7) is called alternating least squares, and (8) is minimized analogously by an alternating robust regression. Both versions are available through `f.subspace`.

Least Trimmed Squares Subspace Estimator. ???

Step B 1: Finding a subspace for several numbers of dimensions. Principal component analysis (or SVA) yields an optimal subspace for all possible values of the number of dimensions, from 1 to $m - 1$. The other algorithms need to be applied for different numbers of dimensions separately in order to obtain the optimal subspace. A shortcut consists in first fitting a subspace with the largest plausible dimension and then performing a PCA on the fitted values. The subspaces found by this PCA can be used as approximately optimal subspaces of lower dimensions.

5.2 Step B 2: Choice of the number of sources

Scree Plot. A widespread informal method for choosing the dimension q in PCA consists of plotting the eigenvalues in decreasing order. More generally, the methods for fitting a subspace should provide a “merit” figure for each dimension, describing in some way the total variance explained by the subspace or something analogous to it. For PCA, the cumulative sum of the eigenvalues is the appropriate measure. `g.screeplot` is a version of the usual `screeplot` tailored to our application (Figure 4).

Residual variances. With each additional source, the residuals get smaller. More precisely, the empirical variance

$$v_j^{(q)} = \text{var}_i(R_i^{(j)(q)} / \sigma_{ij})$$

of the standardized residuals of each variable j decreases with increasing q . (This holds for PCA and SVD with $\sigma_{ij} = \sigma_j$ or $\sigma_{ij} = \sigma$, and it holds at least approximately for other methods of finding the subspace.) The $v_j^{(q)}$ are calculated by `f.resvar`.

A useful diagram compares the $v_j^{(q)}$ to the variances $v_j = \text{var}_i(X_i^{(j)})$, i.e., it shows the profiles $v_j^{(q)}/v_j$ (Figure 5). The profile of standard deviations σ_j/v_j may provide a useful reference. It may turn out that the residuals have less variability than the assumed measurement errors. This will often result from too pessimistic assumptions.

The profile of increments $(v_j^{(q)} - v_j^{(q+1)})/v_j$ can be drawn and examined for each increase in q separately (Figure 6). If only one of the m increments is large, it may well be that the additional potential source merely explains the measurement error of the respective variable.

Patterns in scores. If the observations come as a time series, are connected to geographical locations, or structured in any other meaningful way, it may help to study scores for specific dimensions accordingly. For time series, it is plausible that meaningful scores show seasonal smooth trends and stochastic dependencies, whereas pure measurement errors should exhibit less or no structure. Thus, plotting scores of a questionable additional dimension as a time series may help decide if it is needed (Figure 7). The function `g.ts.wrapped` helps.

Residuals. The distribution of raw residuals $R_i^{(j)}$ and standardized residuals $R_i^{(j)}/\sigma_{ij}$ is examined for each variable j (Figure 8). A long upper tail points to a missing source which may be active in a limited fraction of the observations. Note that the robust estimator (8) is designed to produce positively skewed residual distributions. Long tails in both directions point to unequal or misspecified variances.

The residuals of those variables for which the variance is larger than expected are plotted against time or other potentially informative explanatory variables. Autocorrelations may be expected. If the residuals show clear additional patterns, this will usually suggest that an additional source is active.

Result of Step B. The result of this step of the analysis is a subspace, expressed as a factorization either of the unstandardized or standardized data,

$$\mathbf{Z} = \mathbf{Y}\mathbf{Q} \quad \text{or} \quad (9)$$

$$\tilde{\mathbf{Z}} = \underline{\mathbf{1}}\tilde{\mathbf{X}}^T + \check{\mathbf{Y}}\check{\mathbf{Q}} \quad (10)$$

We write $\mathbf{Z} = \mathbf{Y}\mathbf{Q}$ instead of $\mathbf{Z} = \mathbf{S}\mathbf{C}$ since the decomposition provided by the methods mentioned here is usually not a sensible estimate of the scores and source profiles.

6 Step C: Determination of source profiles

6.1 The problem

Functions which find a subspace fitting the data determine this subspace by a more or less arbitrary basis of vectors. Usually, these vectors are by no means suitable as source profiles. A next step is therefore to identify meaningful source profiles which span the same subspace as the basis available at this point of the analysis.

A basis suitable for graphical displays. For graphical analysis, it is useful to examine the data in compositional form. Then, the subspace found in the previous step has dimension $q - 1$ rather than q . More importantly, convex combination of source profiles then fill the interior of a polyhedron, which appears as a polygon in any projection of the $q - 1$ -dimensional space.

Therefore, if the algorithm for finding the subspace has not been the PCA of compositional data, we use this procedure on the fitted observations $\mathbf{Z} = \mathbf{Y}\mathbf{Q}$ now to obtain the decomposition (10) of the fitted, standardized data $\tilde{\mathbf{Z}}$.

It is instructive to examine a scatterplot matrix of the scores in this factorization (Figure 9). We will discuss below how to use such a display to find suitable source profiles.

Known source profiles. Some of the source profiles may be known a priori. The hypothesis that a source profile \underline{c}^* be relevant for the data under study can be tested. This problem is called target testing in the literature. It means that the profile must be contained in the subspace fitted to the data – up to random errors. We will not pursue this problem here.

Often, a known source profile will only be approximately adequate for the data at hand. It is therefore forced to lie in the subspace fitted to the data by projection, i.e., scores \underline{g}^* are found by fitting the elements c_j^* to the rows of \mathbf{Q} by weighted linear regression. The details of an optimal projection would again be driven by information about the precision of the elements c_j^* or of the fitted subspace. We propose to use the relative errors used for the data to define weights for the regression fit. This procedure can be used for any factorization of a subspace found in the previous step. If a basis for centered compositional data is used, \underline{c}^* must be standardized first, and then the center $\overline{\mathbf{X}}$ used for defining the factorization (10) must be subtracted.

6.2 Minimum volume polyhedron

As discussed above, the (standardized) true source profiles form a polyhedron that contains all the observations up to the errors $E_i^{(j)}$. This is the geometrical consequence of the requirement of positive scores $S_i^{(k)} \geq 0$. In order to select a plausible set of source profiles, one may search for the polyhedron with minimal volume subject to this restriction and the constraint of positive $C_k^{(j)}$.

An ad-hoc algorithm for doing this has been written by Marcel Wolbers Wolbers (2002) and is named `f.minpoly`.

6.3 Visual identification of extreme points

Step C 1: Eyeballing potential source profiles. It is often possible to identify observations which form corner points in the projections shown in a scatterplot matrix of the scores \check{Y} of the standardized observations. They may serve as tentative source profiles. They are usually extreme in at least one of the scores. Thus, they can be identified if the points which have an extreme value in any of the $q - 1$ scores axes directions are labelled. This is done by `f.markextreme` (Figure 9).

Graphical display with tentative sources. Once a complete set of tentative source profiles is available (which belong to the subspace or are projected onto it), it is useful to display the scores with respect to it (Figure 10). Note that the scores of standardized data are again more informative than the unstandardized ones, since convex combinations form a polyhedron, as mentioned above. This also means that the scores should be limited by 1 for a suitable set of source profiles. (Here we assume that the source profiles have been standardized.)

If such a set of tentative source profiles \mathbf{Z}_0 is expressed in the basis \mathbf{C} of the subspace, $\mathbf{Z}_0 = \mathbf{S}_0\mathbf{C}$, then the new factorization reads

$$\mathbf{Z} = \mathbf{S}^*\mathbf{C}^*, \quad \mathbf{C}^* = \mathbf{Z}_0, \quad \mathbf{S}^* = \mathbf{S}\mathbf{S}_0^{-1}$$

Alternatively, new scores can be obtained by projecting the fitted data \mathbf{Z} onto the basis $\mathbf{C}^* = \mathbf{Z}_0$ as discussed for “known source profiles” above. If \mathbf{C}^- is any matrix with $\mathbf{C}^*\mathbf{C}^- = \mathbf{I}_q$, then $\mathbf{S}^* = \mathbf{Z}\mathbf{C}^-$. The basis vectors \mathbf{Z}_0 should be standardized to sum to c (see (4)).

Step C 2: Improvement of the visual identification. A scatterplot matrix of the new scores is now examined. If it exhibits high correlations, one of the involved trial sources should be replaced by a point that shows up as an extreme point in this new display. Then, the revised set of trial source profiles (corner points) is used and checked in the same way. Figure 11 shows the scores for an improved selection of “corner points”, which was obtained by inspecting the previous scatterplot matrix: Tentative source B was replaced by g because of the correlation between B and D in Figure 10. Again, C was replaced by k, because of its correlation with E.

Step C 3: Improving on non-negativity. If extreme points are selected to form a preliminary guess of the source profile matrix \mathbf{C} , some or many scores will usually be negative, since the extreme points represent still a mixture of source profiles.

The situation may be improved by “pushing the corners out” as far as possible. A simple way to do this is to determine the maximal γ_k such that

$$\tilde{\mathbf{C}}_k = \tilde{\mathbf{X}} + \gamma_k(\mathbf{C}_k^* - \tilde{\mathbf{X}}) \tag{11}$$

still fulfills the non-negativity constraint $\mathbf{C}_k^{(j)} \geq 0$, and then using the $\tilde{\mathbf{C}}_k$ as the new basis. Figure 12 shows the resulting source profiles as well as those obtained in the previous step.

6.4 Step C 4: Enforcing non-negativity.

An alternating non-negative least squares algorithm can be applied to enforce non-negativity. It is implemented as `f.altnnls` and uses the R library `quadprog`.

Approximately known source profiles. If \mathbf{C} was known up to some measurement error, the natural approach would be to use a penalized least squares method, that is, to optimize the criterion

$$Q(\mathbf{R}) + \gamma \sum_j \left[\sum_{k,j} \left(\frac{(C_k^{(j)} - C_{kj}^*)}{\eta_{kj}} \right)^2 \right] \quad (12)$$

Here, \mathbf{C}^* is the “known” source matrix, η_{kj} is the precision of $C_k^{(j)}$, γ is a penalization constant, and $Q(\mathbf{R})$ is given by (7) or (8).

The idea of penalizing deviations from a matrix \mathbf{C}^* can be applied to ensure that the alternating non-negative least squares solution is as close as possible to the initial solution found before. `f.altnnls` allows for such a penalty. Figure 13 shows the scatterplot matrix of the scores resulting from applying this procedure.

6.5 Additional criteria

Simple structure of profiles. Knowledge about the meaning of source profiles may suggest that they should have a kind of “simple structure.” It may be desirable to have as many zero contributions $C_k^{(j)}$ as possible, for example. In factor analysis, a loadings matrix is often chosen according to a criterion quantifying its “simplicity,” the most popular being the varimax criterion. We do not yet have an algorithm for maximizing the number of zeros or any other suitable criterion in our context.

Patterns in scores. There may be knowledge about patterns expected in the scores. In chemical reactions, one usually knows that some sources, i.e. some chemical compounds, are absent at the beginning of a reaction, and others may be absent at the end. In this and other applications, we may expect a smooth time series of scores.

If external information is available on the activities of some source, its scores should show a high multiple correlation with the respective explanatory variables.

All of these additional informations can be built into the criterion function for alternating non-negative least squares by adding another penalty term (in addition to (12)).

7 Step D: Assessing the Quality of the Solution

When a solution has been derived by the previous steps, the following graphical displays may help to check its quality:

- The source profiles are plotted and assessed by subject matter knowledge (Figure 14).
- The scores are plotted against time or other useful variables (Figure 15). If explanatory variables for the activities of some sources are available, the respective scatter plots should be examined.

Note that residuals have already be examined after finding the subspace in Section 5.2.

8 Step E: Exploiting the Results

Contributions of the Sources. In many applications in pollution assessment, it is important to quantify the contributions of the various sources to the total immissions. Let

$$Q_k^{(j)} = \text{ave}_i S_i^{(k)} C_k^{(j)} \quad , \quad \tilde{Q}_k^{(j)} = Q_k^{(j)} / \sum_j Q_k^{(j)} \quad (13)$$

Then, $\tilde{Q}_k^{(j)}$ is such a measure. The $\tilde{Q}_k^{(j)}$ profiles for the sources k are shown either individually or in cumulative form (Figure 16).

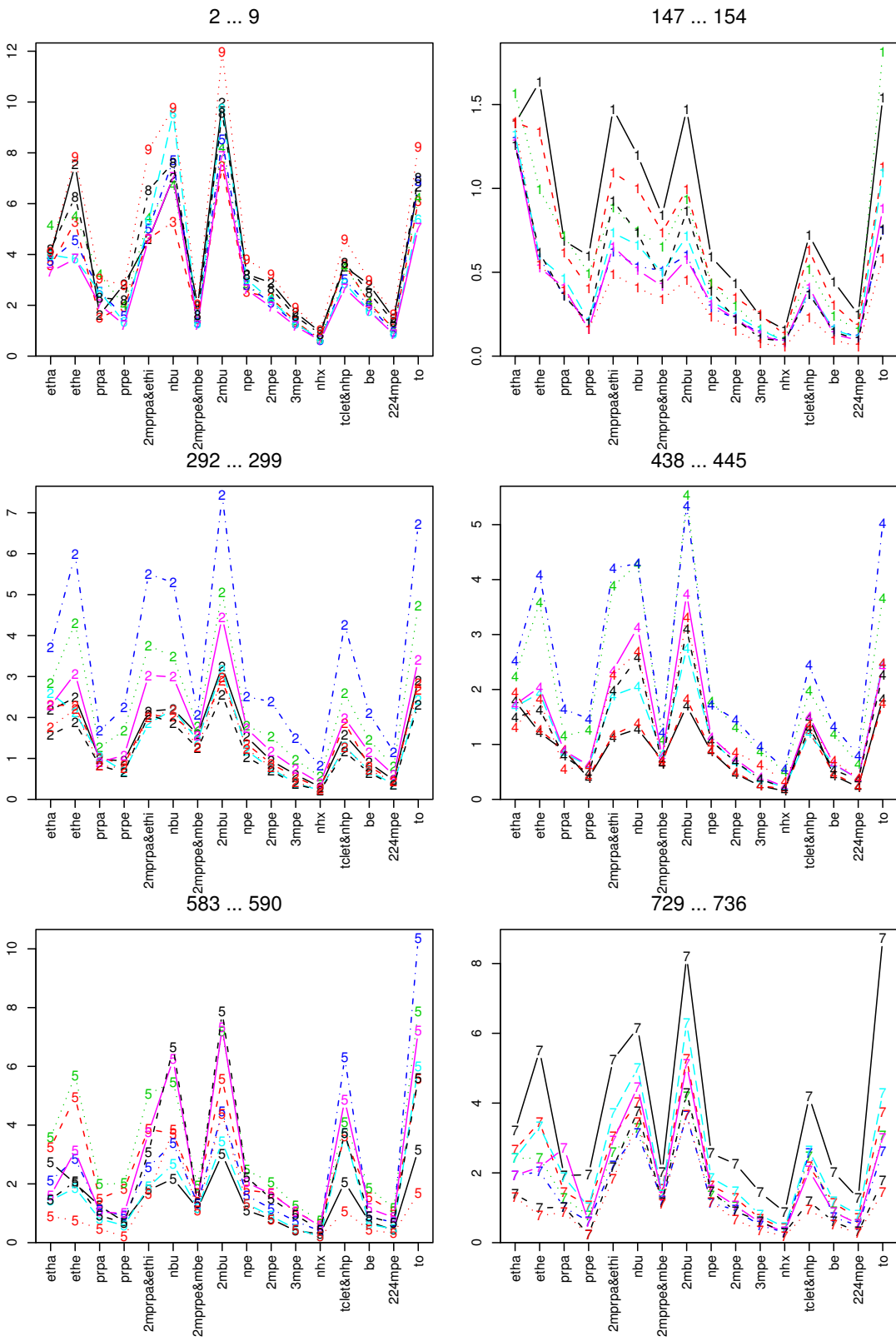
It is, of course, possible to show the observation specific contributions $S_i^{(k)} C_k^{(j)} / \sum_j S_i^{(k)} C_k^{(j)}$ for each pair of k and j as a time series.

Summation over some or all variables may also make sense.

Regression of scores on external variables. If external information has not been used before (Section 6.5), it may be helpful to study regression models at this point.

References

- Akerjord, M.-A. and Christophersen, N. (1996). Assessing mixing models within a common framework, *Environ. Sci. Technol.* **30**: 2105–2112.
- Osten, D. W. and Kowalski, B. R. (1984). Multivariate curve resolution in liquid chromatography, *Analytical Chemistry* **56**: 991–995.
- Paatero, P. and Tapper, U. (1994). Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* **5**: 111–126.
- Vanderginste, B., Essers, R., Bosman, T., Reijnen, J. and Kateman, G. (1985). Three-component curve resolution in liquid chromatography with multiwavelength diode array detection, *Analytical Chemistry* **57**: 971–985.
- Wolbers, M. (2002). *Linear Unmixing of Multivariate Observations*, PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.



Sep 8,00/6:47 | Example | contributions

Figure 2: Six series of eight consecutive observations

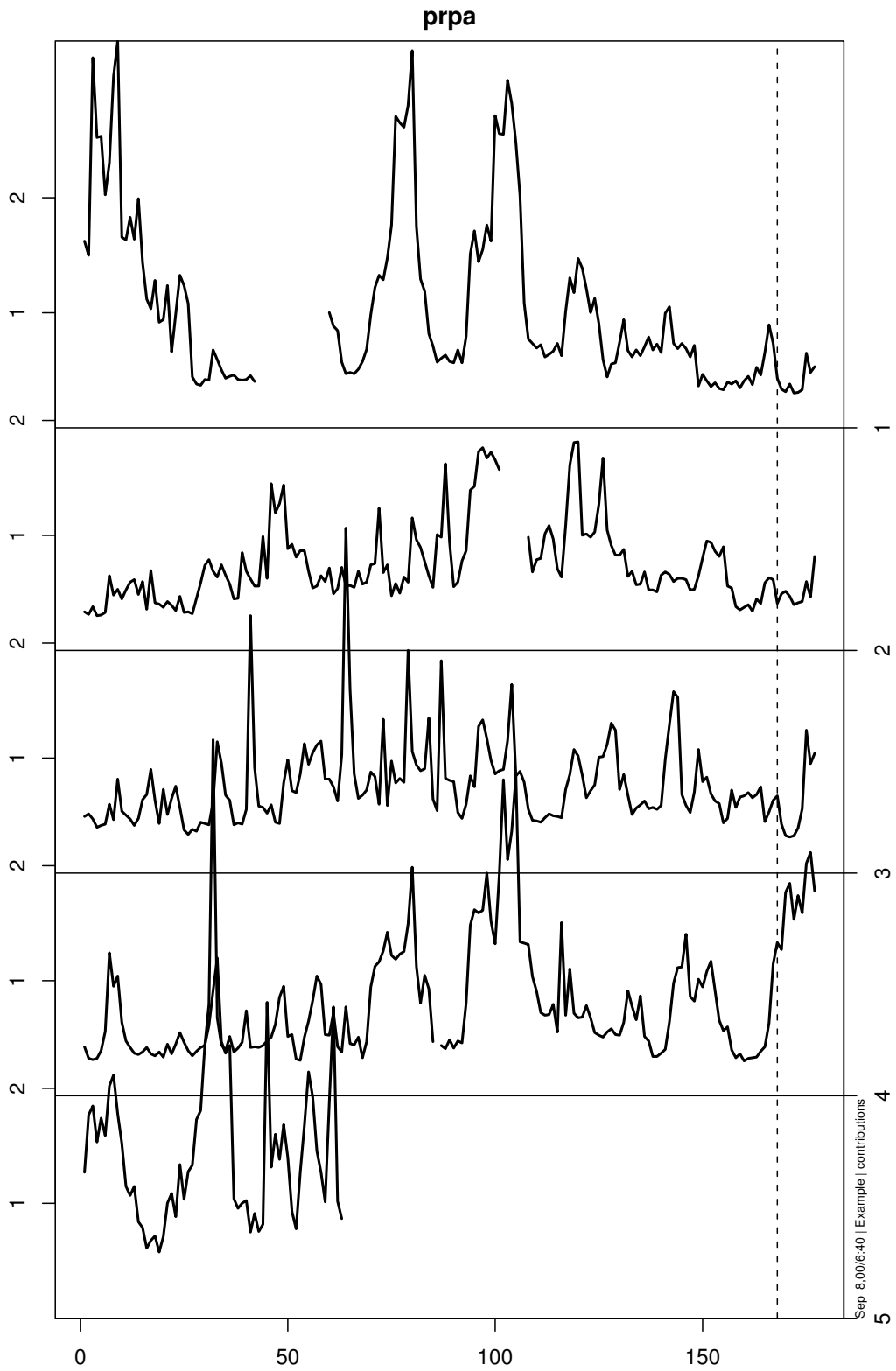
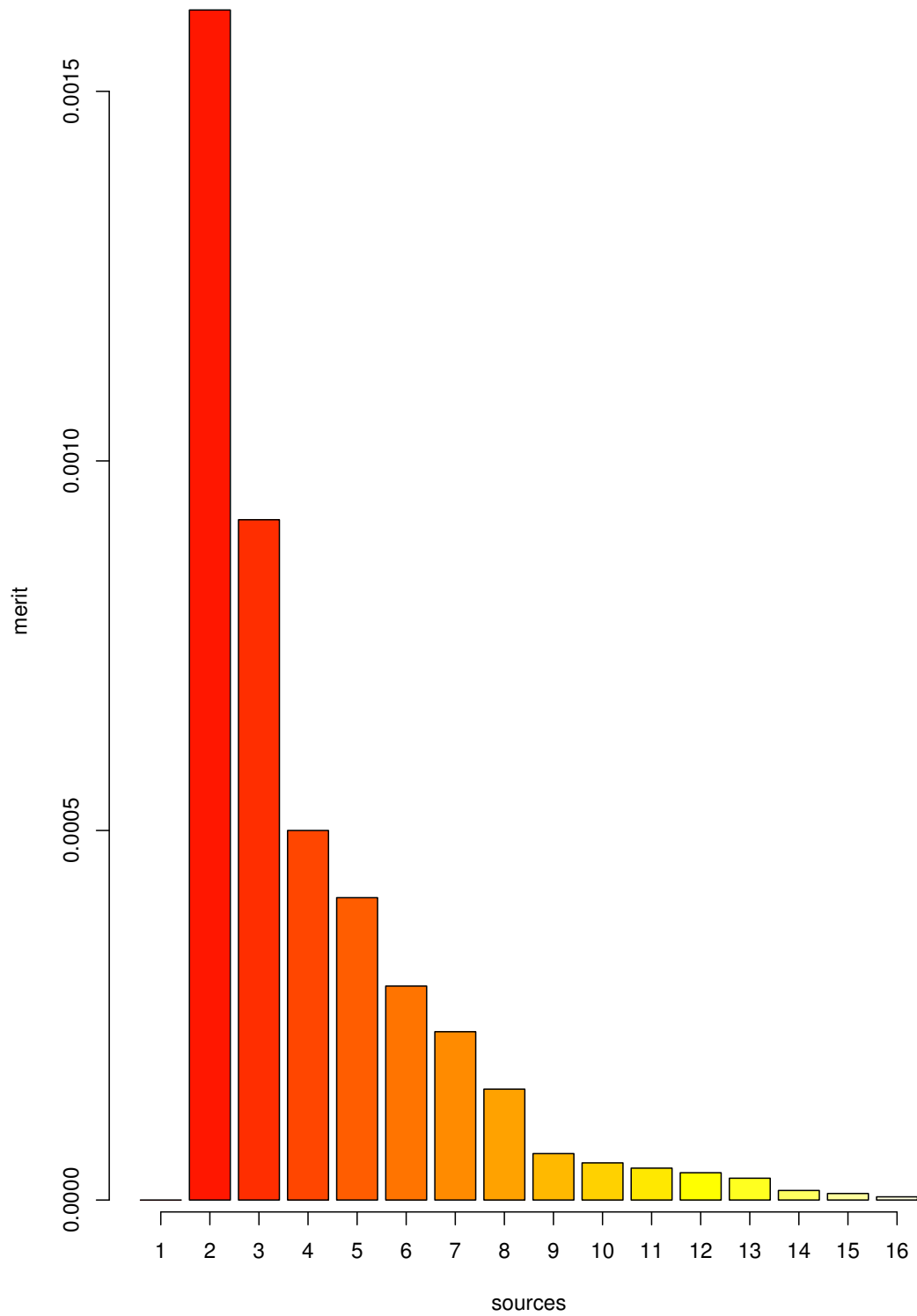


Figure 3: Observations on propane shown by week

PCA: Variance explained by additional dim.



Sep 4, 00/8:09 | Example | Subspace

Figure 4: Scree plot for principal components. The x axis is labeled by the number of sources, which is one higher than the number of principal components.

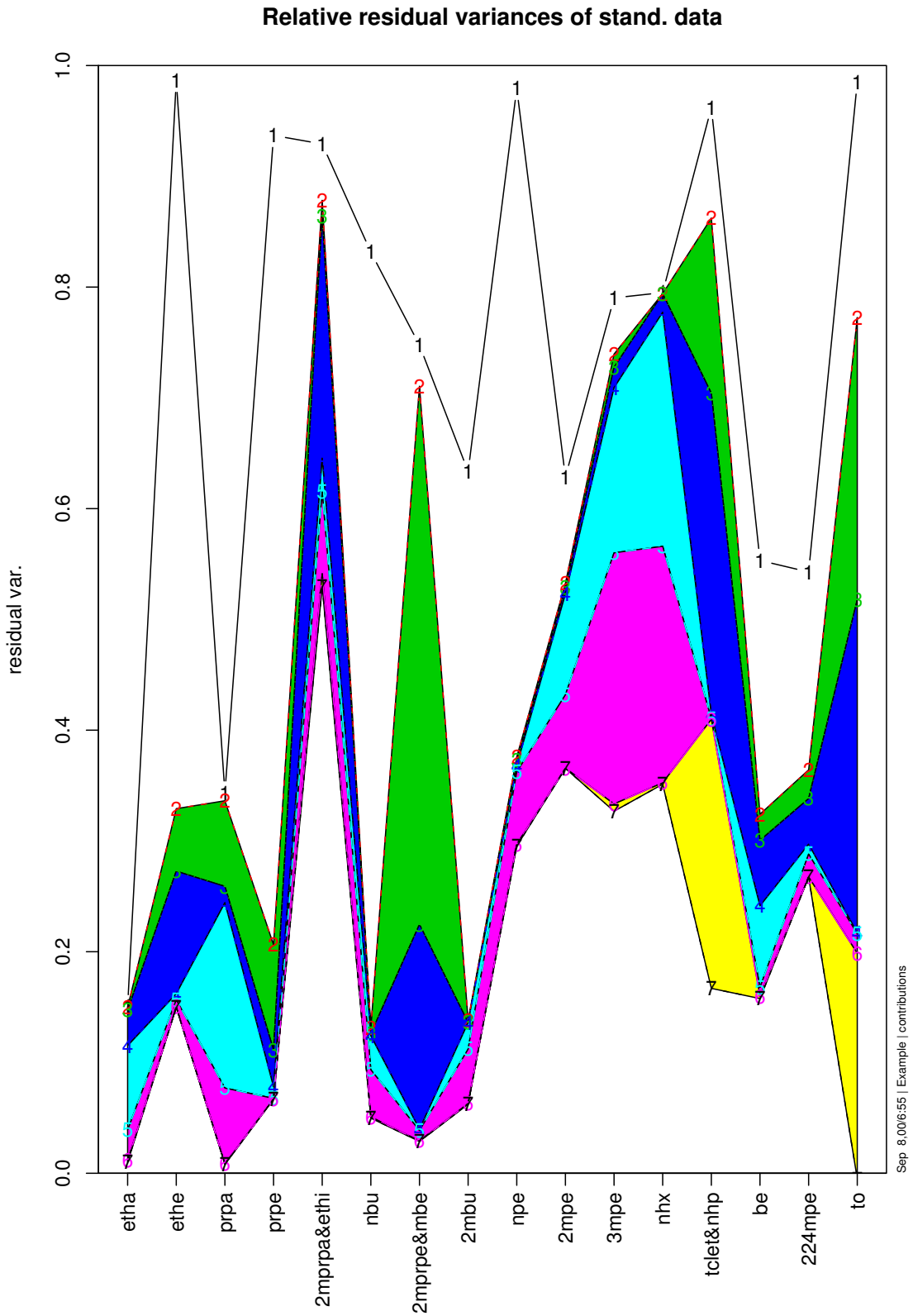


Figure 5: Residual variances for up to seven dimensions corresponding to 8 sources

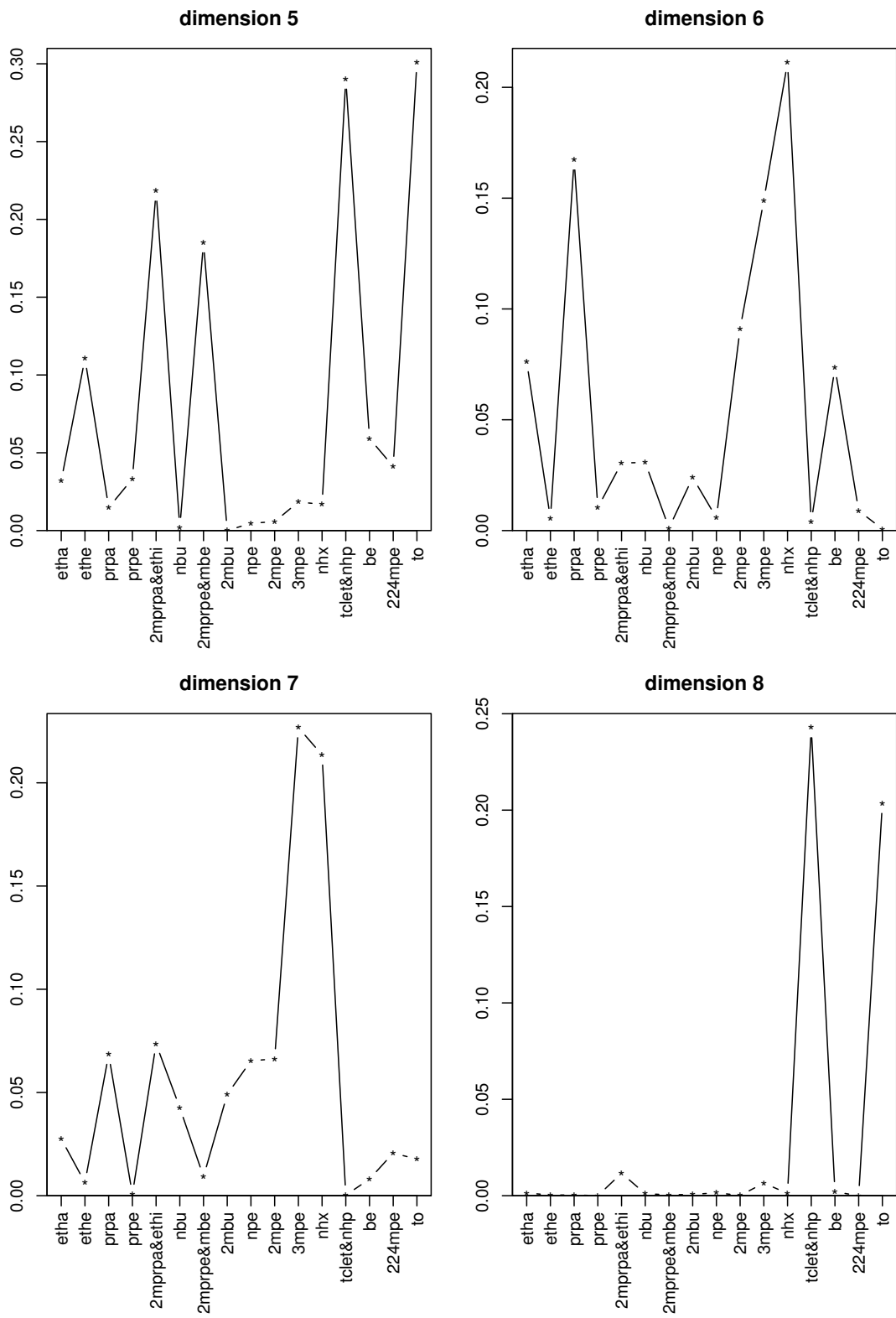


Figure 6: Increments of residual variances

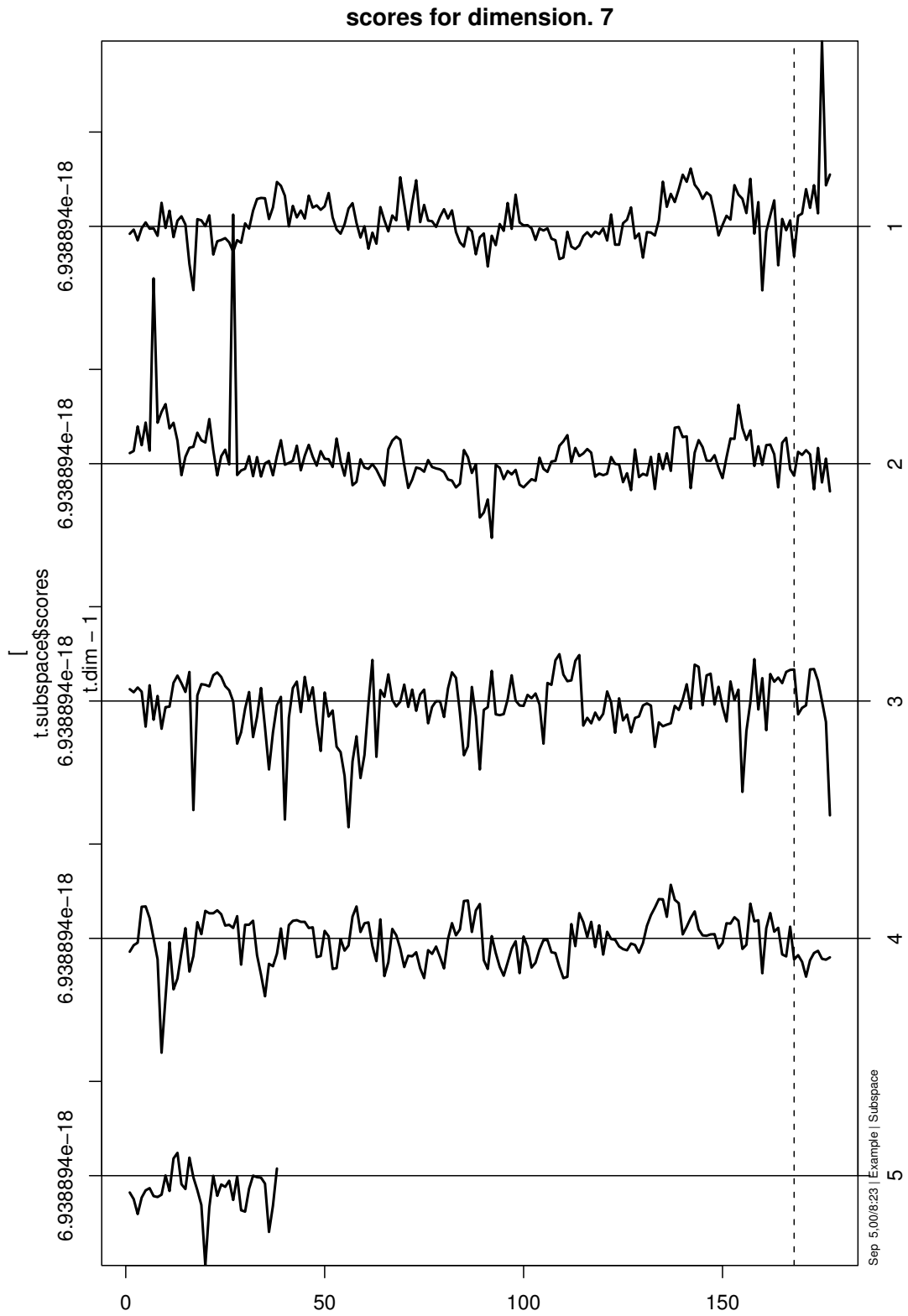
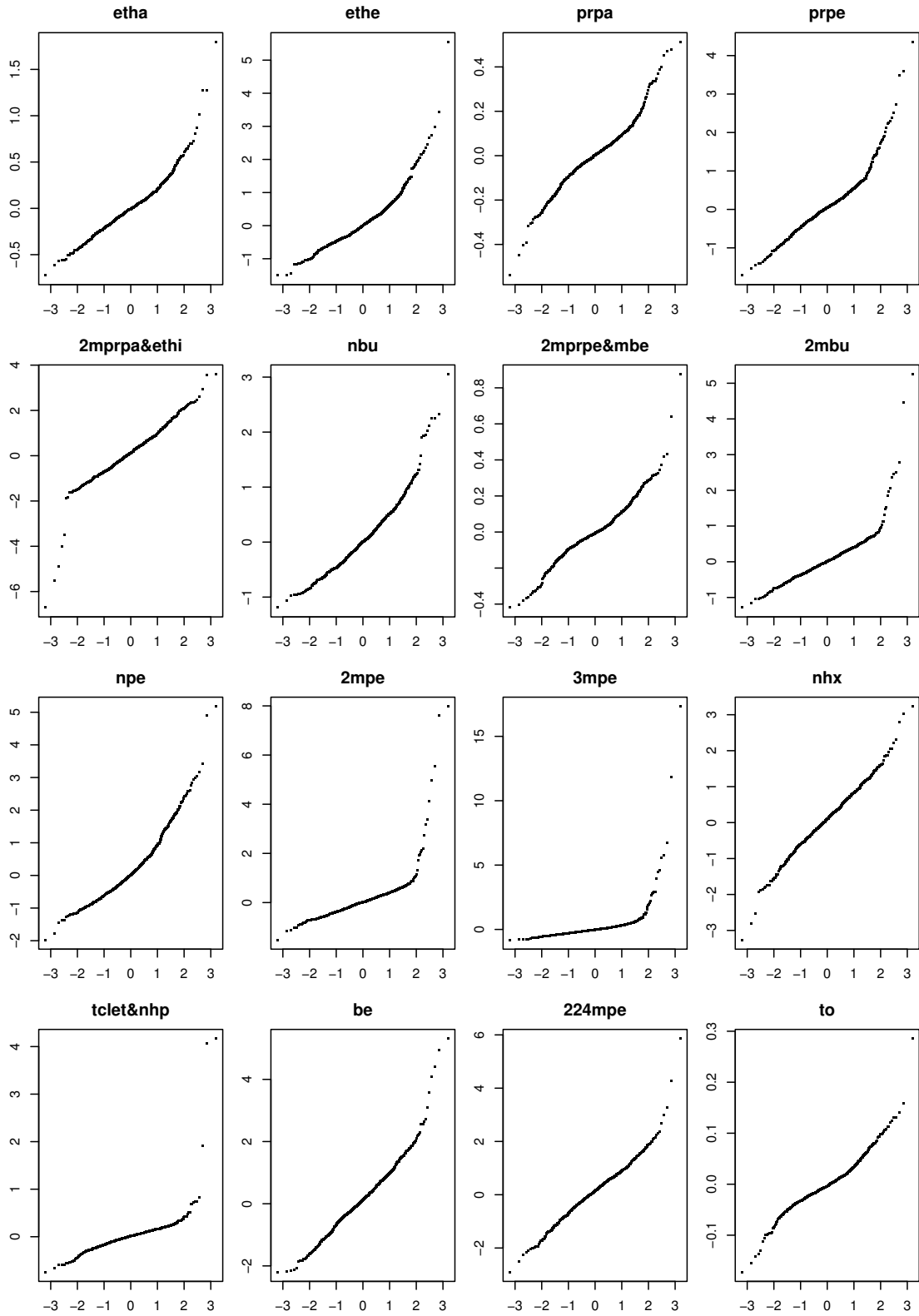


Figure 7: Scores for dimension 7 of the subspace, shown for a weekly period

Distributions of stand. residuals for 7 dimensions



Sep 4,00/9:00 | Example | Subspace

Figure 8: QQ-Plots of the distributions of residuals for 7 sources

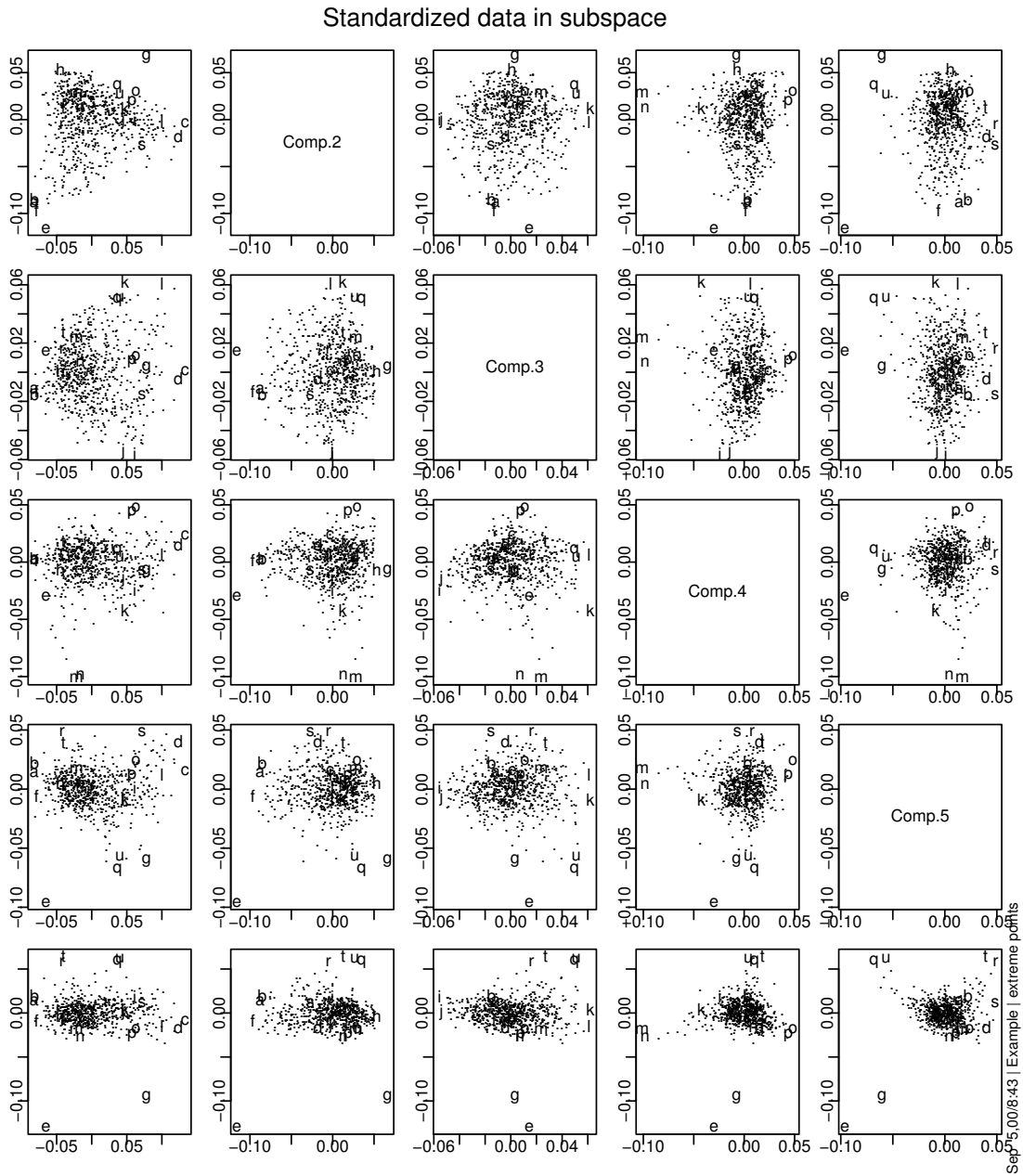


Figure 9: Scatterplot matrix of scores for the original basis of the fitted subspace (a robust version of the principal components)

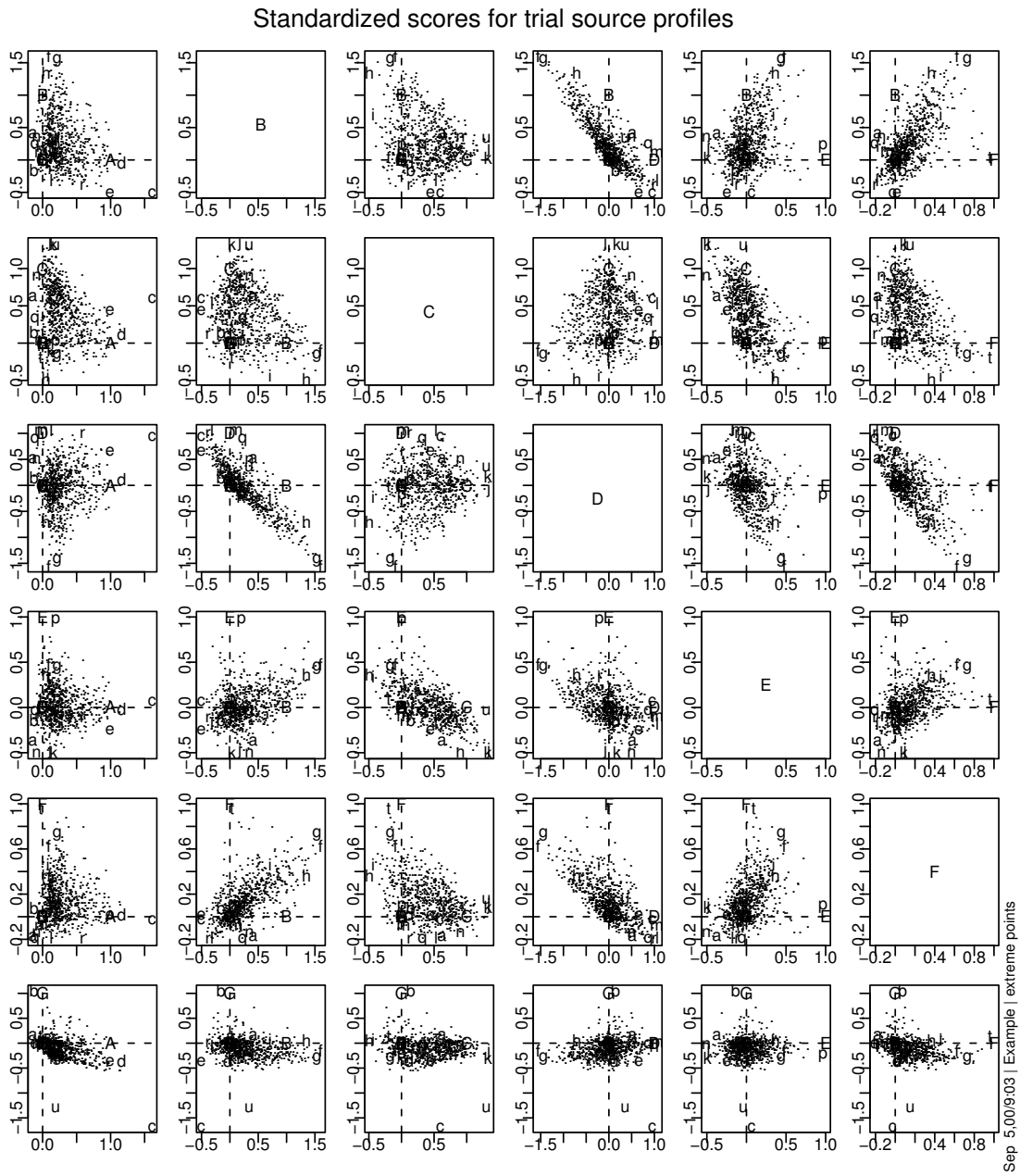


Figure 10: Scatterplot matrix of scores with respect to seven selected “corner points”

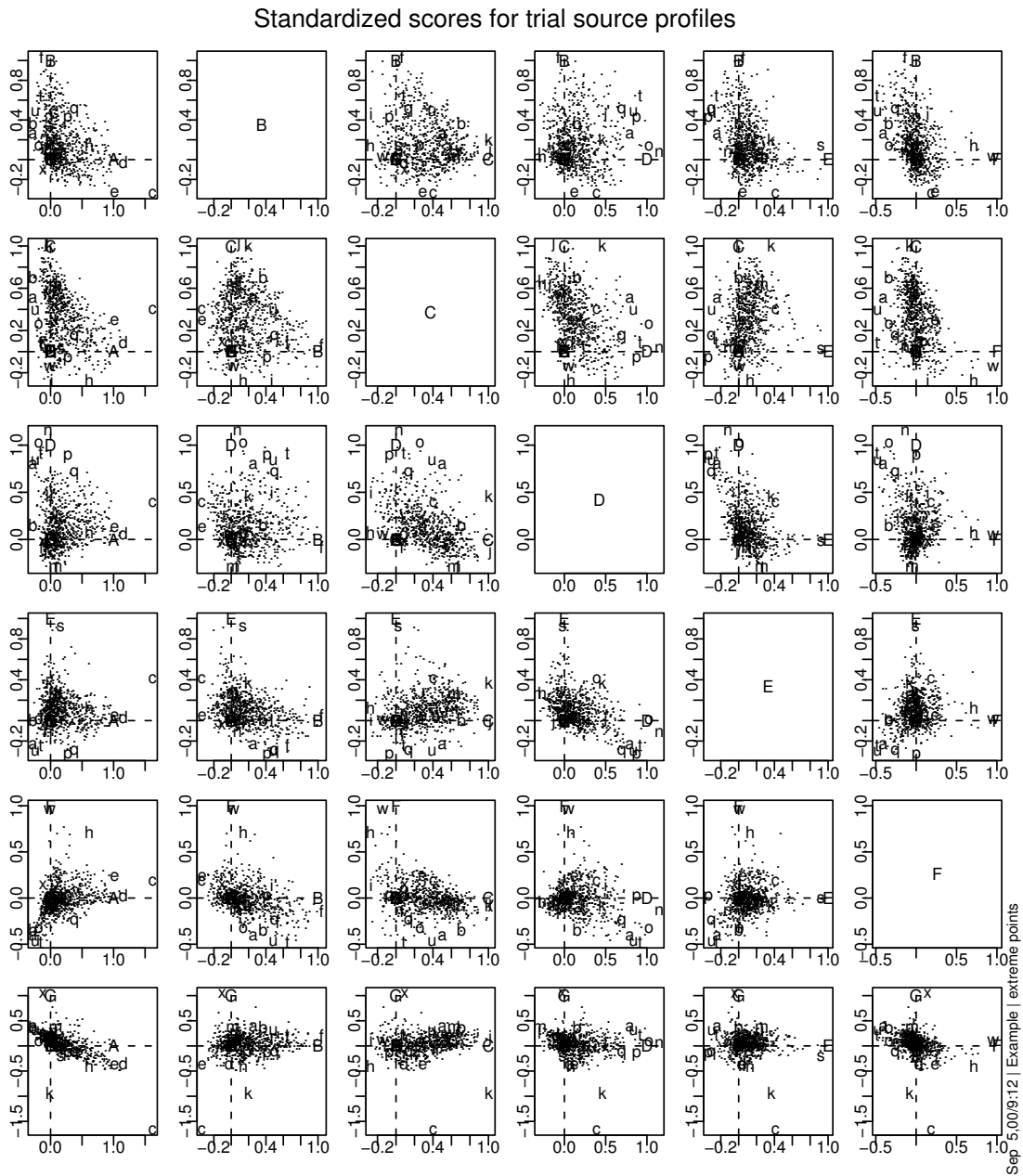


Figure 11: Scatterplot matrix of scores with respect to an improved selection of “corner points”

New and old trial profiles

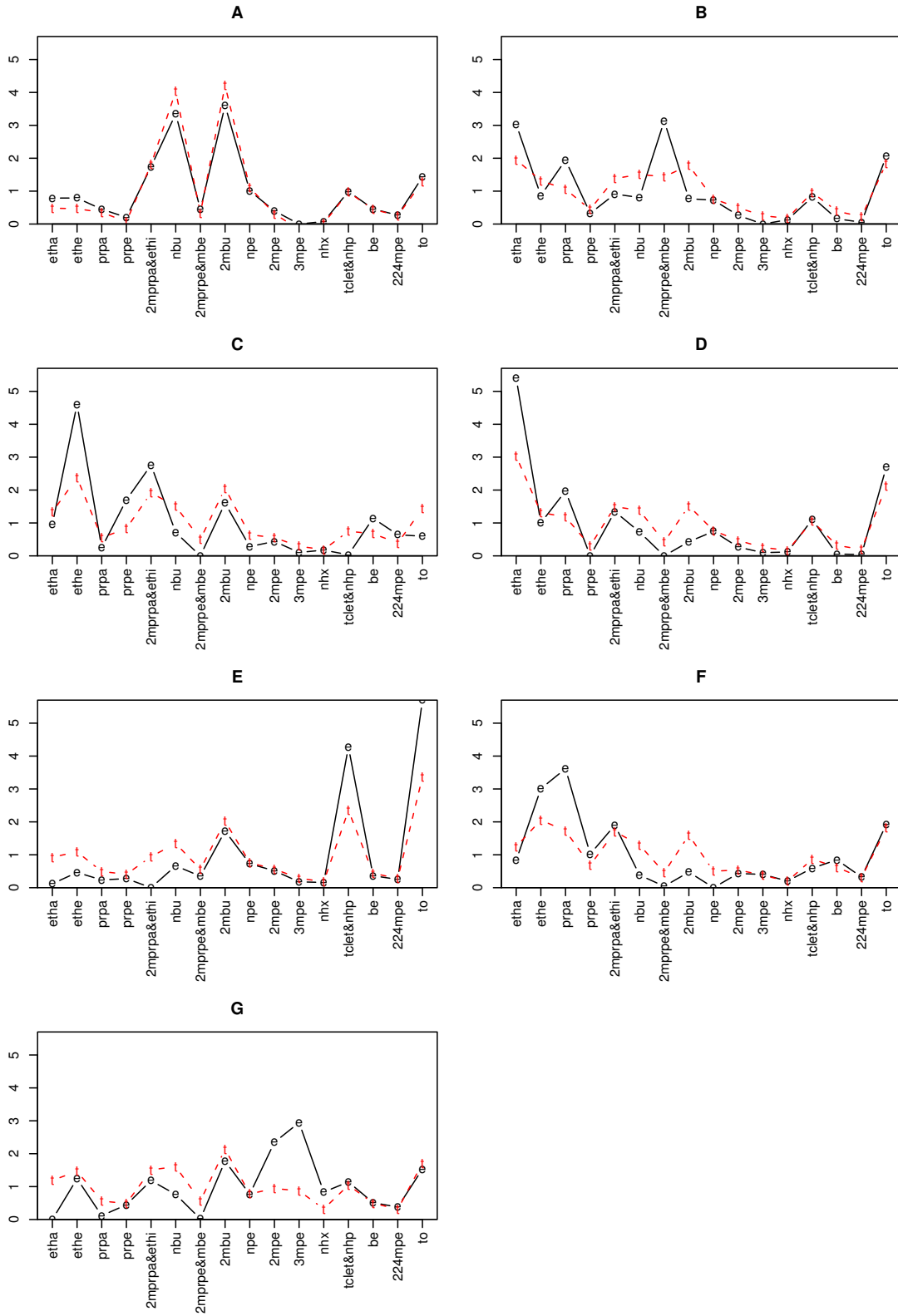
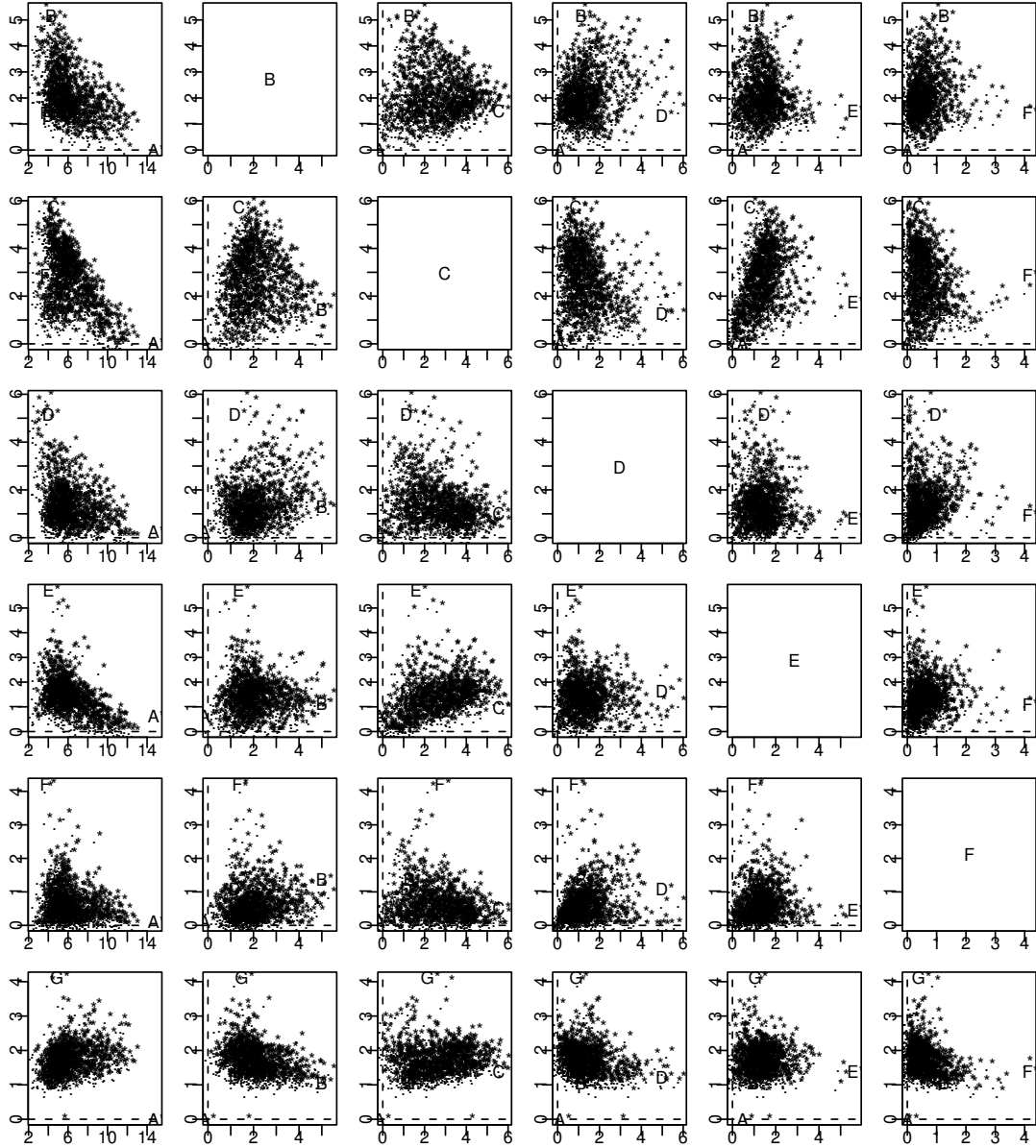


Figure 12: Source profiles before and after expanding them to the limits of non-negativity

Standardized scores for final source profiles



Sep 8, 007/21 | Example | contributions

Figure 13: Scatterplot matrix of scores for the final solution

Trial and final profiles

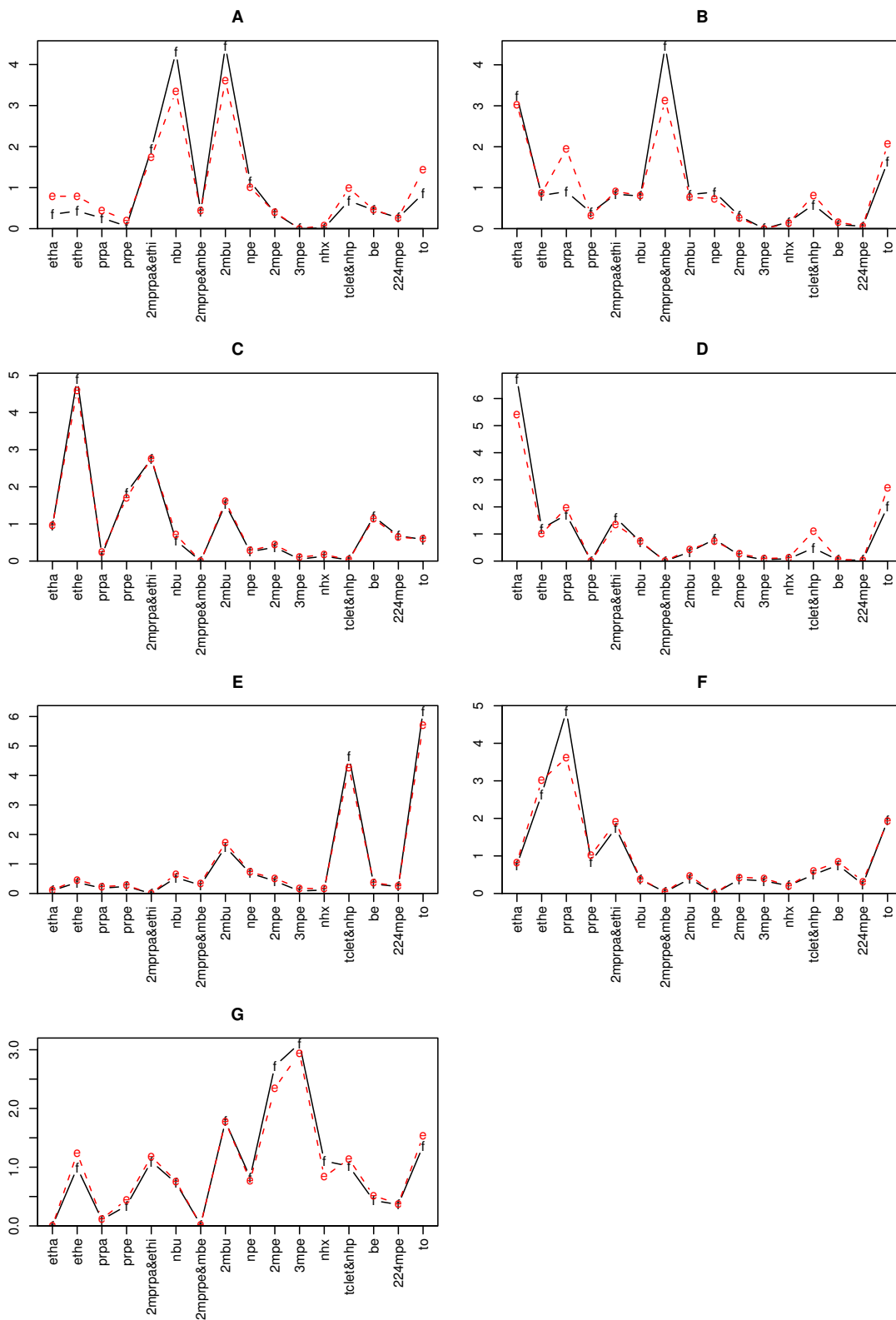


Figure 14: Source profiles for the final solution, along with starting values for the non-negative least squares algorithm

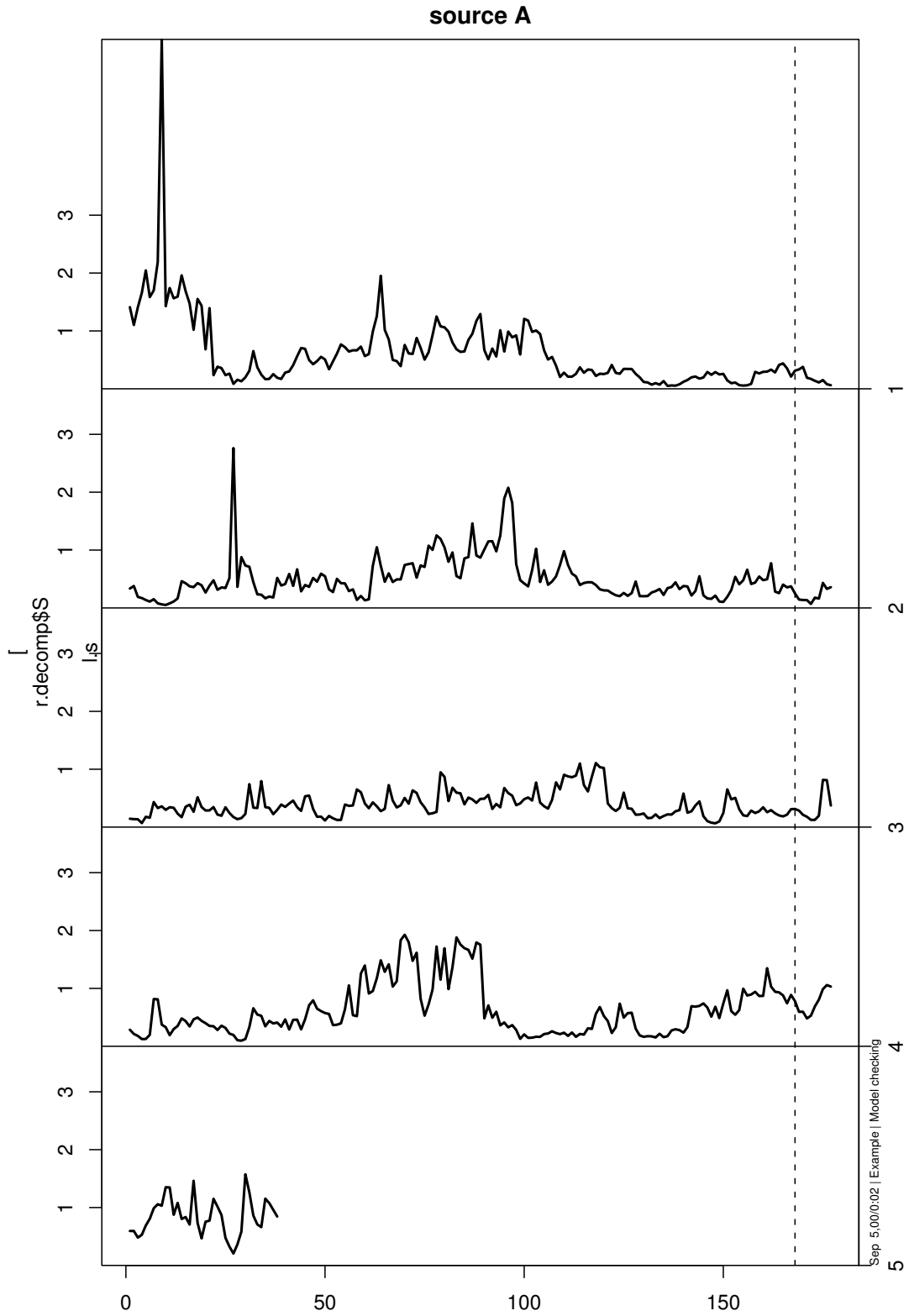


Figure 15: Scores for the first source, shown by week

Mean contributions of the sources

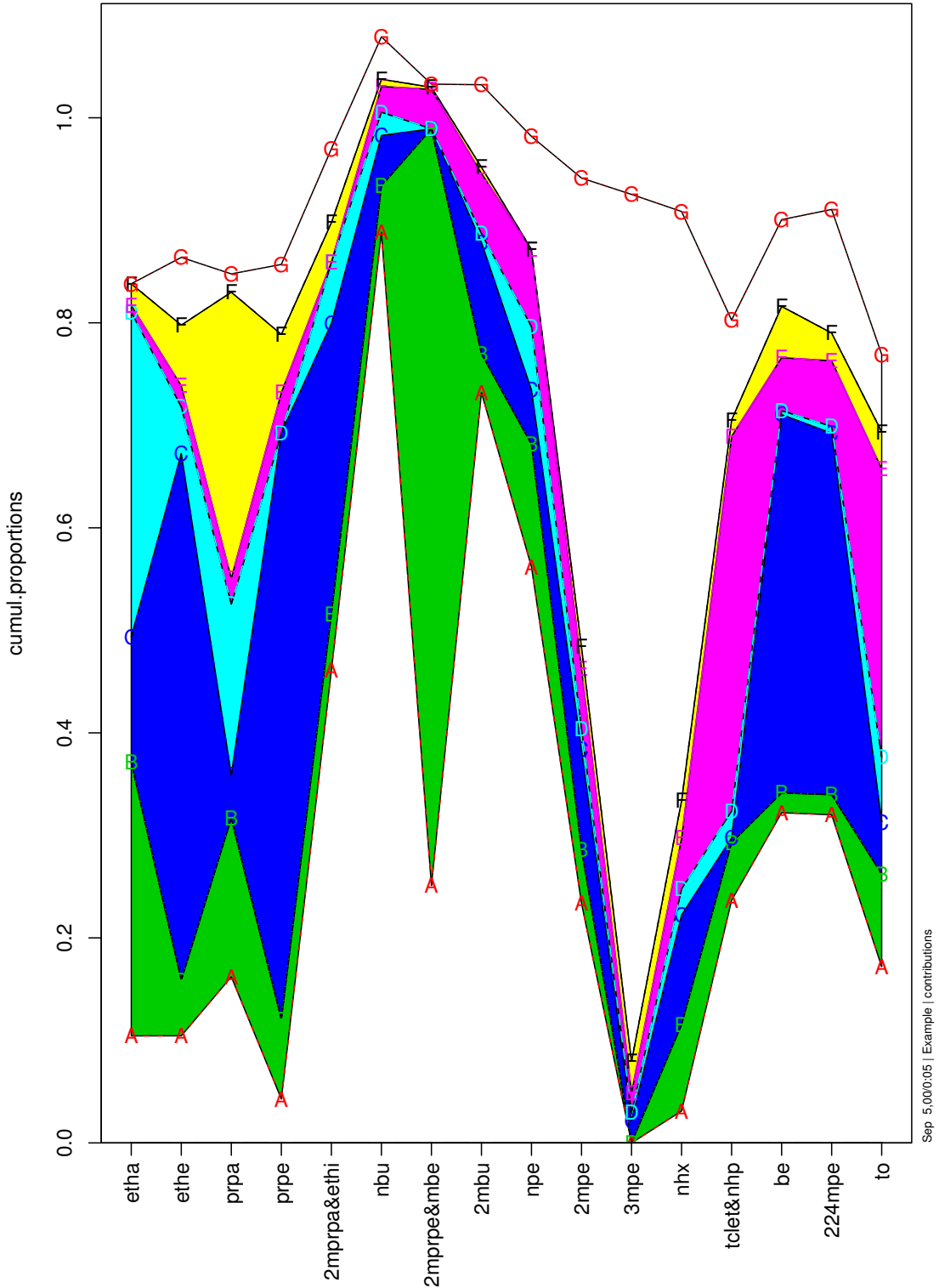


Figure 16: Contributions of the sources to the total concentrations