Exploratory factor analysis of large data matrices

Exploratory factor analysis of large data matrices

July 18, 2018

Nickolay T. Trendafilov[1]
School of Mathematics and Statistics, Open University, UK

Sara Fontanella
Department of Medicine, Imperial College London, UK

Abstract

Nowadays, the most interesting applications have data with many more variables than observations and require dimension reduction. With such data, standard exploratory factor analysis (EFA) cannot be applied. Recently, a generalized EFA (GEFA) model was proposed to deal with any type of data: both vertical data (fewre variables than observations) and horizontal data (more variables than observations). The associated algorithm, GEFALS, is very efficient, but still cannot handle data with thousands of variables. The present work modifies GEFALS and proposes a new very fast version, GEFAN. This is achieved by aligning the dimensions of the parameter matrices to their ranks, thus, avoiding redundant calculations. The GEFALS and GEFAN algorithms are compared numerically with well-know data.

Keywords: Rank inequalities, alternating iterations, Procrustes problems.

[1]Corresponding author: Nickolay.Trendafilov@open.ac.uk

1 Introduction

The most popular dimension reduction techniques are principal component analysis (PCA) and exploratory factor analysis (EFA) [2].

Both techniques aim to find low dimension representation of the data by using a small number of linear combinations of the input variables. However, PCA and EFA differ in the way this is achieved and how the goodness-of-fit to the data is measured. PCA is much more popular than EFA, arguably because it enjoys: 1) a simple geometrical meaning, 2) stable and fast computational procedures to find its component loadings, and 3) no distributional assumptions. In contrast, EFA offers: 1) a complicated model, 2) comparatively slow computational procedures for finding its factor loadings, and 3) restrictive distributional assumptions that are usually difficult to justify.

However, EFA has one big advantage: it is capable of delivering a better fit to the data than PCA. That is the motivation for working towards a modernization of EFA that avoids its weaknesses. It is worth mentioning that EFA currently enjoys increasing popularity for building Bayesian computing models. Surprisingly, most modern work on EFA is connected with Bayesian methodology, both among statisticians (Carvalho *et al.*, 2008; West, 2003) and computer science/machine learning specialists (Chen *et al.*, 2010; Knowles and Ghahramani, 2011). Some modern 'big data' textbooks even define FA simply as a Bayesian model, e.g. Goodfellow *et al.* (2017). In contrast, classic EFA involving parameter estimation, well-known from multivariate analysis and psychometrics (Anderson, 1984; Mulaik, 2010), is hardly touched.

Recently, Trendafilov and Unkel (2011) proposed a new approach to EFA that yields in fast numerical procedures, considerably faster than the classic approaches to EFA estimation. In this paper, the aim is to further develop faster EFA procedures that are comparable with, and possibly outperform in some aspects, its main competitor, PCA.

Trendafilov and Unkel (2011) define EFA as a new data matrix factorization . Their idea was to make EFA a more sophisticated and precise dimension reduction technique, a kind of PCA generalization. In order to align EFA with PCA, and overcome the listed EFA weaknesses, the data and the factors are *not* considered random variables, which seems natural when factoring a data *matrix*.

The idea for this work stems from multiple experiments in which the GEFALS algorithms proposed in (Trendafilov and Unkel, 2011; Unkel and Trendafilov, 2013) was applied to large data sets. They showed that GEFALS becomes very slow when analysing data with tens of thousands of variables. Also, the estimated matrix parameters have an undesirable feature: they contain zero sub-matrices that hardly contribute to model-fit and overweight the computations. Indeed, the number of parameters in the EFA fundamental equation is inherited from the classical definition of EFA that assumes the data have more observations n than variables p, i.e. $n > p$. However, for modern data sets with $p \gg n$, some adjustments are required.

[2] EFA does not assume any prior knowledge about the relationships among original variables and factors. On the contrary, the confirmatory FA takes into account preliminary knowledge or hypothesis about those relationships (Mulaik, 2010, Ch15).

Thus, we propose here a new procedure, GEFAN, which works with parameters having effective dimensions rather than their 'initial' dimensions that are used in the EFA formulation. The benefit will be a considerable saving of storage and CPU time, making the analysis of large data sets feasible. It will be demonstrated that the new algorithm, GEFAN, is considerably faster and capable of analysing very large data matrices.

The paper is organized as follows. Section 2 defines the classical EFA and the generalized EFA (GEFA) models formulated as matrix decompositions (Trendafilov and Unkel, 2011). Then, Section 3 proposes a modification of GEFALS so that it can deal with very large data. Finally, Section 4 gives numerical illustration of the new algorithm GEFAN, and compares it with GEFALS. Results are briefly summarized in Section 5.

2 GEFA model and GEFALS algorithm

For clarity, we summarize here the main results from (Trendafilov and Unkel, 2011).

Let X be a standardized data matrix of n observations on p variables, i.e. $X^\top 1_{n\times 1} = 0_{p\times 1}$ and $X^\top X$ is the sample correlation matrix. Note that for some large data only centring is a plausible option. The rank of X is at most $\min\{n,p\}-1$. The m-factor EFA problem ($m \ll \min\{n,p\}$) can be formulated as a specific matrix decomposition/factorization of X in the following form (De Leeuw, 1994; Trendafilov and Unkel, 2011):

$$X \approx FA^\top + U\Psi = [F\ U][A\ \Psi]^\top\ , \tag{1}$$

where the parameters F, A, U and Ψ (diagonal) are unknown matrices of sizes $n \times m$, $p \times m$, $n \times p$ and $p \times p$ respectively. The fundamental EFA equation (1) means that EFA presents the data X as a linear combination of common and unique factors, F and U. The corresponding weights A are called factor loadings, Ψ is uniquenesses, and Ψ^2 contains the variances of U. The number of common factors, m, is generally unknown, but may be chosen before the analysis.

We stress that F and U are *not random* variables, but fixed unknown matrices. The classic EFA assumptions that the common and unique factors are uncorrelated (Mulaik, 2010, Ch6.2), can be adapted to the new EFA formulation (1) by simply requiring $F^\top F = I_m, U^\top U = I_p$ and $F^\top U = 0_{m\times p}$. From these assumptions, we find from (1) that the sample correlation matrix $X^\top X$ is modeled as:

$$X^\top X \approx C = AA^\top + \Psi^2\ . \tag{2}$$

The attractiveness of the new EFA approach is that *all* unknown parameters F, A, U and Ψ can be found simultaneously without any distributional assumptions. Classic EFA solves (2), which determines A and Ψ. Additional assumptions and effort are needed to find F and U (Unkel and Trendafilov, 2010). In this sense, (1) is an EFA model that can fit data, rather than correlations. As we will see later, another important benefit from this approach is that the EFA parameters are found by singular value decomposition (SVD), for which fast and reliable algorithms exist. This overcomes the algorithmic weakness of classical EFA, which depends on iterative procedures.

2.1 The classic case $n > p$

The EFA formulation (1) suggests that the unknown EFA parameters can be found by solving the following constraint least squares (LS) problem:

$$\min_{F,A,U,\Psi} \|X - [F\ U][A\ \Psi]^\top\|^2, \tag{3}$$

$$\text{subject to} \qquad F^\top F = I_m, U^\top U = I_p, F^\top U = 0_{m\times p} \text{ and } \Psi \text{ diagonal} . \tag{4}$$

This minimization problem is solved by alternating minimization over one unknown and keeping the remainder fixed. The resulting algorithm is called GEFALS.

First, note that (1) and (4) imply that the optimal Ψ (F, U and A fixed) is necessarily given by $\Psi = \text{diag}(X^\top U)$. In a similar way, one can establish that the optimal A (F, U and Ψ fixed) is given by $X^\top F$. The rotational indeterminacy of A can be avoided by taking the lower triangular part of $X^\top F$ (Trendafilov and Unkel, 2011).

Then, (3) is minimized over $[F,\ U]$, with A and Ψ kept fixed (known). Form an $n \times (m + p)$ block matrix $Z = [F,\ U]$ and an $p \times (m + p)$ block matrix $W = [A, \Psi]$. Straightforward calculations and the EFA constraints in (4) show that:

$$Z^\top Z = \begin{bmatrix} F^\top F & F^\top U \\ U^\top F & U^\top U \end{bmatrix} = \begin{bmatrix} I_m & 0_{m\times p} \\ 0_{p\times m} & I_p \end{bmatrix} = I_{m+p} . \tag{5}$$

Thus, the latter minimization reduces to a standard Procrustes problem:

$$\min_{Z^\top Z = I_{m+p}} \|X - ZW^\top\|^2, \tag{6}$$

for a given W and an orthonormal unknown Z. GEFALS alternates between finding (A, Ψ) and Z, until convergence.

2.2 The modern case $p \gg n$

Modern applications frequently require the analysis of data with more (often many more) variables than observations ($p \gg n$). Such data cause severe problems in many classic multivariate techniques, including EFA. Indeed, the classic EFA problem ($n > p$) is to fit a hypothetical correlation structure of the form in (2) to the sample correlation matrix $X^\top X$ (Mulaik, 2010, 6.2.2). However, $X^\top X$ is *singular* when $p \gg n$. Of course, one can still fit C to $X^\top X$, but this will differ from the original EFA problem. To see this, recall that the EFA model is initially defined by (1) and the assumptions for the involved unknowns, while (2) is derived from (1). Specifically, when $p \gg n$, the classic constraint $U^\top U = I_p$ can no longer be fulfilled. Thus, the classic EFA correlation structure (2) turns into

$$C = AA^\top + \Psi U^\top U \Psi . \tag{7}$$

The new correlation structure (7) coincides with the classic one (2) if the more general constraint $U^\top U\Psi = \Psi$ is introduced in place of $U^\top U = I_p$. In other words, a universal EFA definition, valid for any n and p, should impose the constraint $U^\top U\Psi = \Psi$. An

important consequence from this new assumption is that Ψ^2 is not positive definite (p.d.), when $p \gg n$ (Trendafilov and Unkel, 2011, Lemma 1).

The rest of the classic EFA constraints $F^\top F = I_m$ and $U^\top F = 0_{p\times m}$ remain valid. Thus, for $p \gg n$, EFA requires solution of the following constrained LS problem:

$$\min_{F,A,U,\Psi} \|X - ZW^\top\|^2, \tag{8}$$

$$\text{subject to} \qquad F^\top F = I_m, U^\top U\Psi = \Psi, F^\top U = 0_{m\times p} \text{ and } \Psi \text{ diagonal} \tag{9}$$

where, as before, $Z = [F\ U]$ and $W = [A\ \Psi]$. The solution is called generalized EFA (GEFA).

Trendafilov and Unkel (2011) use the *same* updating formulas for Ψ and A as in the classic case (Section 2.1). However, the update of Z is changed to reflect the presence of the new constraint $U^\top U\Psi = \Psi$. Then, the GEFA problem (8) – (9) reduces to:

$$\min_{ZZ^\top = I_n} \|X - ZW^\top\|^2 \ , \tag{10}$$

which is a standard Procrustes problem. GEFALS finds the GEFA parameters in the same alternating manner. The central GEFA result is summarized as:

Theorem 2.1 *(Trendafilov and Unkel, 2011) The matrix XW is always rank deficient, and the orthonormal Procrustes problems (6) and (10) have no unique orthonormal solution Z. When $p \gg n$, unlikely exception is possible, if rank$(XW) = n$.*

The phenomenon of factor indeterminacy is notorious for the problems it causes. Theorem 2.1 quantifies its origin, and thus, clarifies the existing philosophical explanations (Mulaik, 2010).

3 Modified GEFALS – GEFAN

The updating formulas for A and Ψ used by Trendafilov and Unkel (2011) are well-known from classic EFA, where $n > p$. They are easily obtained by pre-multiplication of equation (1) by $F^\top$ and $U^\top$. When $n > p$,

$$F^\top X \approx F^\top FA^\top + F^\top U\Psi = A^\top \tag{11}$$

$$U^\top X \approx U^\top FA^\top + U^\top U\Psi = \Psi \ . \tag{12}$$

As Ψ is assumed diagonal, (12) becomes $\Psi \approx \text{diag}(U^\top X)$. Simple calculations show that the first order necessary conditions for the minimum of (3) with respect to A and Ψ are $A = X^\top F$ and $\Psi = \text{diag}(U^\top X)$.

For $p \gg n$, we have $U^\top U \neq I_p$, and pre-multiplication of (1) by $U^\top$ leads to the following,

$$U^\top X \approx U^\top FA^\top + U^\top U\Psi = U^\top U\Psi \ . \tag{13}$$

which is different from (12). In this case, the first order necessary condition for minimization of (3) with respect to Ψ becomes $U^\top U\Psi = U^\top X$, instead of the classical condition $\Psi = U^\top X$ in (12).

The requirement for diagonal Ψ, and the fact that $U^\top U$ contains an identity submatrix and zeros elsewhere (see Lemma 3.1, Lemma 3.2), imply that the new optimality condition $U^\top U\Psi = U^\top X$ can be rewritten as $U^\top U\Psi = \text{diag}(U^\top X)$. Thus, we arrive at a new updating formula, $\Psi = (U^\top U)\text{diag}(U^\top X)$. This differs from GEFALS, which uses the same updating formula $\Psi = \text{diag}(U^\top X)$ for any data format, $n > p$ and $p \gg n$ (Trendafilov and Unkel, 2011; Unkel and Trendafilov, 2013). For data with $p \gg n$, we update Ψ through $\Psi = (U^\top U)\text{diag}(U^\top X)$. It has dramatic implications: large data, which *cannot* be analysed by GEFALS using the update $\Psi = \text{diag}(U^\top X)$, become perfectly manageable when the update is switched to $\Psi = \text{diag}(U^\top X)$.

In addition, numerical experiments with GEFALS show that, when $p \gg n$, the estimated parameters F and U contain large zero sub-matrices which hardly contribute to the model fit and overweight the computations. One consequence of this undesirable feature is that GEFALS' performance is unnecessarily slowed down.

In order to address these issues, we propose here an algorithm that is partially inspired by the approach adopted in Unkel and Trendafilov (2013). The new algorithm is called GEFAN. For data with $n > p$, GEFAN coincides with GEFALS. For data with $p \gg n$, GEFAN works only with the non-zero submatrices of the GEFA matrix parameters to avoid redundant calculation. Some algebraic features of the effective sizes of the GEFA parameters are discussed in Trendafilov and Unkel (2011). Others are considered below.

Considering A and Ψ fixed, write the GEFA problem (8) for the case $p \gg n$ as:

$$\min_{\substack{F^\top F = I_m \\ U^\top F = O_{p\times m}}} \|X - FA^\top - U\Psi\|^2 \ , \tag{14}$$

keeping in mind the new constraint $U^\top U\Psi = \Psi$ from (9). The problem in (14) can be solved by alternately solving the following two problems: for fixed U,

$$\min_{F^\top F = I_m} \|(X - U\Psi) - FA^\top\|^2 \ , \tag{15}$$

and for fixed F,

$$\min_{\substack{U^\top U\Psi = \Psi \\ U^\top F = O_{p\times m}}} \|(X - FA^\top) - U\Psi\|^2 \ . \tag{16}$$

The problem in (15) is a standard orthonormal Procrustes problem. The solution is $F = VW^\top$, where $VDW^\top$ is the SVD of $(X - U\Psi)A$. However, the problem in (16) needs more attention. For this reason, we first transform the objective function in (16) as follows. Let $F_\perp$ be the $n \times (n - m)$ matrix containing an orthonormal basis of the nullspace of F in $\mathbb{R}^n$, so that the block matrix $[F \ F_\perp]$ is $n \times n$ orthogonal. Then, the

objective function in (16) can be rewritten as follows:

$$\|[F\ F_\perp]^\top (X - FA^\top - U\Psi)\|^2 = \left\| \begin{bmatrix} F^\top(X - FA^\top - U\Psi) \\ F_\perp^\top(X - FA^\top - U\Psi) \end{bmatrix} \right\|^2 = \tag{17}$$

$$\left\| \begin{bmatrix} F^\top X - A^\top \\ F_\perp^\top X - F_\perp^\top U\Psi \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} 0_{r\times p} \\ F_\perp^\top X - F_\perp^\top U\Psi \end{bmatrix} \right\|^2 = \|F_\perp^\top X - F_\perp^\top U\Psi\|^2 \ ,$$

making use of the EFA optimality condition (11), which is valid for any format of data.

Further simplification of the problem can be achieved by taking into account the structure of U and Ψ that follows from the EFA constraints. We start with the following two results that help to avoid redundant calculations.

Lemma 3.1 *$F^\top F = I_m$ and $U^\top F = O_{p\times m}$ imply that rank$(U) \leq n - m$.*

PROOF : $\text{rank}(U^\top) + \text{rank}(F) - n \leq \text{rank}(U^\top F) = 0$ ((Horn and Johnson, 1985, 0.4.5.(c))) ■

Lemma 3.1 implies that U can be rewritten (possibly after reordering of its columns) as a block matrix $[U_1\ U_2]$ with U_1 of size $n \times (n - m)$ and $U_2 = 0_{n\times(p-n+m)}$. Note, that U_2 can be huge and it seems reasonable to exclude it from the calculations.

We already know from Section 2.2 that Ψ^2 cannot be positive definite, i.e. $\Psi^2 \geq 0$.

Lemma 3.2 *Suppose Ψ has r zero diagonal entries. Then rank$(\Psi) = p - r \leq n - m$.*

PROOF : The proof is based on the constraint $U^\top U\Psi = \Psi$ and Lemma 3.1. Assume, for simplicity, that the r variables with zero unique variances are the last r variables. Then, $\text{rank}(\Psi) = p - r$ and Ψ can be partitioned as $\Psi = \begin{bmatrix} \Psi_1 & 0_{(p-r)\times r} \\ 0_{r\times(p-r)} & 0_{r\times r} \end{bmatrix}$, where Ψ_1 is $(p - r) \times (p - r)$ diagonal with nonzero entries. Similarly, U can be partitioned as $U = [U_1^*\ U_2^*]$, where U_1^* and U_2^* are block matrices with sizes $n \times (p - r)$ and $n \times r$, respectively. Then, the general constraint $U^\top U\Psi = \Psi$ gives ${U_2^*}^\top U_1^*\Psi_1 = 0_{r\times(p-r)}$ and ${U_1^*}^\top U_1^*\Psi_1 = \Psi_1$. As Ψ_1 is non-singular diagonal matrix, they are equivalent to ${U_2^*}^\top U_1^* = 0_{r\times(p-r)}$ and ${U_1^*}^\top U_1^* = I_{p-r}$. This implies that $\text{rank}(U_1^*) = p - r$. As U_1^* is a submatrix of U, it follows from Lemma 3.1 that $p - r \leq n - m$. ■

To simplify the derivation of the new GEFA algorithm (GEFAN) we assume that Ψ and U are partitioned as in the proof of Lemma 3.2. Note, that these assumptions are not restrictive. In practice, GEFAN counts the number r of zero diagonal entries in Ψ at each iteration step, which also changes the size of Ψ_1. Thus, GEFAN works with the actual sizes $n \times (p - r)$ and $(p - r) \times (p - r)$ of U_1 and Ψ_1, respectively.

Now, write $X - FA^\top$ as the block matrix $[X_1\ X_2]$ with X_1 of size $n \times (p - r)$. Note, that X_1 is composed of those columns of X that correspond to non-zero entries in Ψ_1. Then, with this notation, the problem in (16) becomes:

$$\min_{U_1^\top U_1 = I_{p-r}} \|F_\perp^\top X_1 - F_\perp^\top U_1\Psi_1)\|^2 + \text{constant} \ , \tag{18}$$

where the constant does not depend on U_1.

In general, a problem like (18) does not have a closed form solution. It is known as the weighted orthonormal Procrustes problem or the Penrose regression problem (Chu and Trendafilov, 2001). Fortunately, the problem in (18) can be simplified and solved as an ordinary Procrustes problem. Indeed, let $Q = F_\perp^\top U_1$ be a new unknown. Then $Q^\top Q = U_1^\top F_\perp F_\perp^\top U_1 = U_1^\top (I_{n-m} - FF^\top) U_1 = U_1^\top U_1 = I_{p-r}$, so the unknown Q is orthonormal $(n-m) \times (p-r)$ and the problem in (18) reduces to:

$$\min_{Q^\top Q = I_{p-r}} \|F_\perp^\top X_1 - Q\Psi_1\|^2 \ , \tag{19}$$

which is a standard Procrustes problem whose solution is given by the SVD of $F_\perp^\top X_1 \Psi_1$. When Q is found, U_1 is obtained from $U_1 = F_\perp Q$. If $p - r = n - m$ (which generally seems the case in numerical experiments), then Q is square, and it is orthogonal. Then, $U_1 U_1^\top = F_\perp Q Q^\top F_\perp^\top = F_\perp F_\perp^\top$, which implies:

$$UU^\top + FF^\top = U_1 U_1^\top + FF^\top = F_\perp F_\perp^\top + FF^\top = I_n \ .$$

These considerations show that the parameters obtained by GEFAN possess equivalent features to those by GEFALS (Section 2.2).

To summarize, a solution to the EFA problem for data with $p \gg n$ is obtained by solving (15) for F and (19) for U, followed by updating A and Ψ. Finding such $\{F, U, A, \Psi\}$ is repeated until convergence. The orthogonal complement $F_\perp$ is not uniquely determined, but the solution of (19) is unique for a given $F_\perp$, provided that $\text{rank}(X_1) \geq p - r$. This implies that A and Ψ are uniquely determined.

GEFAN is an alternating least squares procedure, alternately finding$[F, U]$ and $[A, \Psi]$. In general, alternating procedures may get stuck while solving one of the sub-problems. However, GEFAN uses SVD to find $[F, U]$, while the new $[A, \Psi]$ is a simple update involving matrix multiplications only. Thus, convergence problems have never arisen.

The GEFAN algorithm, the modified GEFALS, can be summarized as follows:

$F \leftarrow \texttt{rand}(n, m) - .5$, $U \leftarrow \texttt{rand}(n, p) - .5$, $A \leftarrow X^\top F$, $\Psi \leftarrow \text{diag}(U^\top X)$

$f_{old} = ||X||_F^2, f = ||X - FA^\top||_F^2$

while $|f_{old} - f| > 10^{-6}$

$\mathcal{I} \leftarrow \{i \ : \ |\Psi(i,i)| > 10^{-7}\}$, $n_\mathcal{I} \leftarrow \#(\mathcal{I})$

$U_1 \leftarrow U(:, \mathcal{I})$, $\Psi_1 \leftarrow \Psi(\mathcal{I})$

For fixed A, U_1 and Ψ_1

$$\min_{F^\top F = I_m} \|[X(:, \mathcal{I}) - U_1 \Psi_1] - FA^\top\|^2$$

$A \leftarrow X^\top F$, $X_1 \leftarrow X - FA^\top$, $F_\perp \leftarrow [F \ F_\perp]$ ($n \times n$ orthogonal)

For fixed $X_1, F_\perp$ and Ψ_1

$$\min_{Q^\top Q = I_{n_\mathcal{I}}} \|F_\perp^\top X_1(:, \mathcal{I}) - Q\Psi_1\|^2$$

$U_1 \leftarrow F_\perp Q$, $\Psi_1 \leftarrow \text{diag}(U_1^\top U_1)\text{diag}[U_1^\top X_1(:, \mathcal{I})]$

$U(:, i) \leftarrow U_1(:, i)_{i \in \mathcal{I}} \cup 0_{i \in \complement \mathcal{I}}$, $\Psi(i, i) \leftarrow \Psi_1(i, i)_{i \in \mathcal{I}} \cup 0_{i \in \complement \mathcal{I}}$

$f_{old} = f, f = ||X - FA^\top - U\Psi||_F^2$

end while

4 Numerical examples

In this Section, we compare the solutions obtained by GEFALS with those obtained by GEFAN . As noted earlier, they are some data set that GEFAN can analyse while GEFALS cannot produce a solution.

Factor analysis of data with more variables than observations is a relatively new subject. The following well-known example illustrates that the new procedures produce sensible solutions that are compatible with the ones given by classic EFA and/or PCA. We will see that the GEFALS and GEFAN solutions coincide and simply differ in the way the zero entries in Ψ are handled.

The example is Thurstone's 26-variable data set on boxes that is widely used in factor analysis, it was used by Trendafilov and Unkel (2011) and a great number of authors in classic EFA. Thurstone collected a random sample of 20 boxes and measured their three dimensions x (length), y (width) and z (height) (Thurstone, 1947, p.141). The boxes are the observations, and the variables are 26 functions of these dimensions listed below in Table 1. Thus, the data set has $n = 20$ and $p = 26$. As the first three eigenvalues of the sample correlation matrix are considerably greater than the rest, three common factors will be sought, i.e. $m = 3$. Note, that the rank of this (standardized) data matrix is 18. Thus, according to Lemma 3.2, the number of zeros (r) in Ψ for this particular example should be between 9 and 11.

The first GEFALS solution is depicted in the first four columns of Table 1. It was obtained with the classical update $\Psi = \text{diag}(U^\top X)$. The table shows that Ψ has 11 elements that are 0 to four decimal places, which agrees with theory, although only 3 of the elements are 0 to 15 decimals.

Then, we applied a second form of GEFALS by replacing the classical updating formula with the new one, $\Psi = (U^\top U)\text{diag}(U^\top X)$, which is suited to data with $n > p$. The solution is given in the second (middle) four columns of Table 1. Now, Ψ has three exact zeros according to the floating point arithmetic (0.00000000000000e+000), but the loadings are nearly the same as with the first form of GEFALS. The fit (0.5919) is unchanged.

The GEFAN solution is given in the last four columns of Table 1. Again, the fit and loadings are virtually the same. Now, Ψ has $r = 11$ exact zeros. The benefit from the new procedure is that a considerable amount of redundant calculation is avoided. GEFAN explicitly works with smaller matrices than GEFALS by reducing matrices to their effective dimensions. For example, the active part of U has size 20×16, which is significantly smaller than its "model" size 20×26. This reduces calculation and,

in particular, SVDs are calculated for much smaller matrices. Finally, we stress that the GEFALS/GEFAN loadings are close to the classic EFA solution (Thurstone, 1947, p.370-371).

Table 1: Two solutions for Thurstone's 26-variable box data.

	GEFALS (0.5919)				**GEFALS** (0.5919)				**GEFAN** (0.5919)			
Variable	A			Ψ	A			Ψ				
x	1.0	0	0	.0000	1.0	0	0	.0000	1.0	0	0	0
y	.25	.97	0	-.0000	.25	.97	0	.0000	.25	.97	0	0
z	.10	.24	.96	.0000	.10	.23	.96	-.0000	.10	.23	.96	0
xy	.68	.73	-.01	-.0000	.68	.73	-.00	.0000	.68	.73	-.00	0
xz	.49	.20	.84	-.0000	.49	.20	.84	.0000	.49	.20	.84	0
yz	.19	.60	.77	.0000	.20	.59	.77	-.0000	.20	.59	.77	0
x^2y	.82	.55	-.00	-.1383	.82	.55	-.00	.1382	.82	.54	-.00	.1383
xy^2	.52	.84	-.03	-.0066	.52	.84	-.03	-.0124	.52	.84	-.03	0
x^2z	.68	.16	.69	.1407	.68	.15	.69	-.1406	.68	.15	.69	.1406
xz^2	.33	.25	.90	-.0001	.33	.24	.90	-.0000	.33	.24	.90	-.0012
y^2z	.24	.73	.60	.1726	.24	.73	.60	-.1726	.25	.73	.60	.1727
yz^2	.15	.46	.85	-.0000	.15	.46	.85	.0000	.16	.46	.85	0
x/y	.45	-.87	-.04	.1670	.45	-.87	-.04	.1669	.44	-.87	-.05	.1670
y/x	-.47	.86	.01	-.1702	-.47	.86	.01	.1701	-.46	.87	.02	.1703
x/z	.31	-.16	-.88	-.2848	.31	-.15	-.88	-.2848	.31	-.15	-.89	.2848
z/x	-.36	.20	.87	.2181	-.36	.20	.87	.2181	-.36	.20	.88	.2181
y/z	.05	.39	-.88	-.2379	.05	.40	-.87	-.2379	.05	.40	-.87	.2379
z/y	-.04	-.37	.88	-.2552	-.04	-.38	.88	.2552	-.04	-.38	.88	.2553
$2x+2y$	.79	.61	.00	.0000	.79	.61	.00	0	.79	.61	.00	0
$2x+2z$	.74	.16	.65	-.0000	.74	.16	.65	0	.74	.15	.65	0
$2y+2z$	.22	.76	.60	.0000	.22	.76	.61	0	.23	.76	.61	0
$(x^2+y^2)^{1/2}$	.87	.49	-.00	.0019	.87	.49	-.00	-.0021	.87	.49	-.01	-.0022
$(x^2+z^2)^{1/2}$	.90	.11	.40	.0097	.90	.11	.40	.0097	.91	.10	.39	-.0098
$(y^2+z^2)^{1/2}$	.24	.86	.43	.0023	.24	.86	.44	-.0019	.24	.86	.44	.0015
xyz	.46	.54	.68	.0414	.46	.54	.68	-.0414	.47	.54	.68	.0414
$(x^2+y^2+z^2)^{1/2}$	.80	.52	.28	-.0088	.80	.52	.28	.0088	.80	.52	.28	.0088

The previous example was too small to yield a noticeable difference between the performances of GEFALS and GEFAN. To appreciate the advantage of using the new algorithm, the different algorithms were applied to a large horizontal data matrix of size 72×12582. The data are available from `http://research.dfci.harvard.edu/korsmeyer/MLL.htm` and were studied by Armstrong *et al.* (2002) to select gene profiles to one kind of leukaemia. The following experiments use the standardized data matrix.

First, two-factor solutions were determined. We start with the original GEFALS as proposed in (Trendafilov and Unkel, 2011), i.e. using only the classic update $\Psi = \mathrm{diag}(U^\top Z)$. The reported results in Table 2 (line GEFALS orig) are obtained from 10 runs and include: the minimal fit achieved (among the 10 runs), the average fit and its standard deviation. These same features are also reported for the CPU in seconds. The average fit for the original GEFALS over 10 runs was 94.392 with standard deviation .002. This small standard deviation shows that the algorithm produced stable results. In contrast, the CPU times are quite diverse, ranging from about 3 seconds for the fastest run, and up to 23 seconds for the slowest.

The remaining results are also obtained with 10 runs of the corresponding algorithm. The second line of Table 2 gives the results obtained by GEFALS (new), when the new update $\Psi = \text{diag}(U^\top U)\text{diag}(U^\top Z)$ is used (appropriate for these data as $p \gg n$). The improvement in terms of CPU time is obvious: this algorithm is much faster, and the standard deviation is rather small suggesting similar performance in all 10 runs. The difference in the fit is less than 1%. Now, the third line of Table 2 gives the results obtained by the new algorithm GEFAN. The fit is virtually identical to that of GEFALS, but GEFAN is three times faster. The small CPU standard deviation also indicates a uniform time in each run.

Further on, Table 2 provides the same information for GEFALS (new) and GEFAN, found for an increasing number of factors. The general conclusion is that the algorithms deliver a stable fit: all fit standard deviations are very small. As expected, GEFAN is always faster, by about three times.

Table 2: Comparison of three GEFA algorithms

		Fit			CPU (sec)		
Factors	Method	Minimum	Average	StD	Minimum	Average	StD
2	**GEFALS** orig	94.392	94.397	.002	2.44	16.06	6.20
	GEFALS	94.760	94.765	.004	.75	1.34	.48
	GEFAN	94.763	94.768	.004	.34	.48	.09
3	**GEFALS**	90.628	90.635	.005	1.13	2.58	1.96
	GEFAN	90.632	90.636	.004	.37	.47	.10
4	**GEFALS**	87.607	87.614	.004	1.12	1.91	.43
	GEFAN	87.603	87.610	.005	.39	.50	.07
5	**GEFALS**	84.991	84.997	.005	.89	1.29	.29
	GEFAN	84.988	84.997	.007	.36	.49	.09
6	**GEFALS**	82.933	82.937	.003	1.46	2.99	.88
	GEFAN	82.929	82.935	.006	.56	.88	.24

We now consider another leukaemia-related data set, this time collected and studied by St. Jude Research[3]. In this example, we applied GEFALS (new) and GEFAN to a centred and normalized 248×12558 data matrix. Table 3 gives the results, reporting the quantities as in Table 2 and also providing information on the number of iterations used by each algorithm.

As with the first example, GEFAN is three-four times faster than GEFAN. Again, we see that both algorithms produce quite close fits, and are very stable: the fits' standard deviations for each algorithm is small. The number of iterations use by both algorithms are reasonably similar. Clearly, working with a smaller matrix parameters pays off, and each GEFAN iteration is computationally much cheaper.

All numerical experiments were performed using Matlab R2018a on a MacBook Pro 2017, 2.8 GHz Intel i7 (quad-core), 16 GB RAM.

[3] `http://datam.i2r.a-star.edu.sg/datasets/krbd/Leukemia/Stjude.html`

Table 3: More comparisons of GEFALS vs. GEFAN

		Fit			CPU (sec)			Iterations		
Factors	Method	Min	Av	StD	Min	Av	StD	Min	Av	StD
2	**GEFALS**	.322	.323	0.000	5.50	7.49	1.75	18	27	7.67
	GEFAN	.322	.323	0.001	1.06	1.68	.35	16	29	6.98
3	**GEFALS**	.292	.293	0.001	7.48	9.83	1.05	27	37	4.22
	GEFAN	.292	.293	0.001	1.31	1.79	.40	21	30.1	7.75
4	**GEFALS**	.281	.282	0.001	5.59	9.29	2.13	18	34.6	9.52
	GEFAN	.281	.282	0.000	1.09	2.17	.73	17	40.5	15.67
5	**GEFALS**	.270	.271	0.001	7.34	10.14	1.51	26	38.2	6.60
	GEFAN	.270	.271	0.000	1.49	2.06	.36	25	34.4	7.40
6	**GEFALS**	.261	.262	0.001	5.28	9.23	2.03	18	34.0	8.67
	GEFAN	.261	.262	0.001	1.43	2.23	.45	19	39.9	10.66

5 Discussion

This paper proposed a modification of the EFA model to make it applicable to data with more variables than observations. The most important practical aspect of this new development is that the notorious factor (scores) indeterminacy is quantified and related to the ranks of the parameter matrices and the number of EFA parameters. The resulting EFA algorithm is efficient because it works with parameter matrices whose dimensions are aligned with their ranks, thus avoiding redundant calculations. This greatly reduces the CPU time required to fit the EFA model – by two third for the large data sets in our experiments. The numerical solutions follow precisely the theoretically prescribed features of the unknown parameters.

Acknowledgements

This work is supported by a grant RPG-2013-211 from The Leverhulme Trust, UK. The authors thank Professor Paul Garthwaite for his help to polish the text.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, **30**, 41–47.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High-

dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, **35**, 1438–1456.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance. *IEEE Trans Signal Process*, **58**, 6140–6155.

Chu, M. T. and Trendafilov, N, T. (2001). The orthogonally constrained regression revisited. *Journal of Computational and Graphical Statistics*, **10**, 746–771.

De Leeuw, J. (1994). Block relaxation algorithms in statistics. In H.-H. Bock, W. Lensi, and M. M. Richter, editors, *Information Systems and Data Analysis*, pages 308–324. Springer: Berlin, GE.

Goodfellow, I., Bengio, Y., and Courville, A. (2017). *Deep Learning*. The MIT Press, Cambridge, MA.

Horn, R. A. and Johnson, C. A. (1985). *Matrix Analysis*. Cambridge University Press, Cambridge, UK.

Knowles, D. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, **5**, 1534–1552.

Mulaik, S. A. (2010). *The Foundations of Factor Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. University of Chicago Press, Chicago, IL.

Trendafilov, N. T. and Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, **20**, 874–891.

Unkel, S. and Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review*, **78**, 363–382.

Unkel, S. and Trendafilov, N. T. (2013). Zig-zag routine for exploratory factor analysis of data matrices with more variables than observations. *Computational Statistics*, **28**, 107–125.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, pages 723–732. Oxford University Press.