Mathematical Statistics
Stockholm University
Research Report **2013:10**,
http://www.math.su.se

Distribution-free factor analysis—Estimation theory and applicability to high-dimensional data

Rolf Sundberg, Stockholm University
Uwe Feldmann, University of Saarland*

December 2013

Abstract

We here provide a distribution-free approach to the random factor analysis model. We show that it leads to the same estimating equations as for the classical ML estimates under normality, but more easily derived, and valid also in the case of more variables than observations ($p > n$). For this case we also advocate a simple iteration method. In an illustration with $p = 2000$ and $n = 22$ it was seen to lead to convergence after just a few iterations. We show that there is no reason to expect Heywood cases to appear, and that the factor scores will typically be precisely estimated/predicted as soon as p is large. We state as a general conjecture that the nice behaviour is not despite $p > n$, but because $p > n$.

Key words: EFA; FA; fixed point iterations; likelihood equations; more variables than observations; SVD.

1 Introduction

In this paper we consider parameter estimation in a distribution-free version of the standard (Gaussian) factor analysis (FA) model, with special emphasis on the case of more variables than observations. The FA model means describing a sample $x_1, \ldots, x_n$ of p-dimensional vectors as

$$x_i = \mu + \Lambda f_i + e_i \,, \quad i = 1, \ldots, n. \tag{1}$$

Here μ is the mean value vector, Λ is a $p \times k$ coefficients (loadings) matrix, $k < \min(n, p)$, and the f_is are mutually independent latent k-vectors (factor scores), standardized to zero mean and unit covariance matrix I_k (for identifiability). The e_is are assumed mutually independent p-vectors with uncorrelated components and diagonal covariance matrix Ψ^2. Also, f_i and e_i should be mutually independent. In matrix form we write (1) as $X = \mu\mathbf{1} + F\Lambda^T + E$, with the vectors of (1) as rows.

Usually, normality of f and e in (1) is assumed, and more observations than variables, that is $n > p$. Then Gaussian maximum likelihood methods can be used, and are more or less standard. However, in recent years interest has increased both in more robust methods and in methods for the case of more variables than observations, $p > n$. Among papers having appeared after the comprehensive review by Bartholomew & Knott (1999, ch. 3), we mention Robertson & Symons (2007), who study extension of Gaussian maximum likelihood to the case $p > n$, and a number of

[*] Corresponding author: Rolf Sundberg; Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@math.su.se. Website: www.math.su.se/∼rolfs