Optimizing principal components analysis (PCA) methodology for ERP component identification and measurement: Theoretical rationale and empirical evaluation

Jürgen Kayser and Craig E. Tenke

Department of Biopsychology, New York State Psychiatric Institute, New York

NYSPI psychophysiology

http://psychophysiology.cpmc.columbia.edu

Introduction

- Although PCA is widely used to determine "data-driven" ERP components, it is unclear if and how specific methodological choices may affect factor extraction. We report here the effects of three variations when applying temporal PCA (tPCA) to ERP data:
  1) Type of association matrix (correlation / covariance)
  2) Varimax rotation (scaled / unscaled)
  3) Number of components extracted and rotated

Theoretical Rationale

- The usefulness of the extracted factors can be evaluated by specific knowledge about the variance distribution of ERPs, which are characterized by the removal of baseline activity. The variance should be small for sample points before and shortly after stimulus onset (across and within cases), but large near the end of the recording epoch and at ERP component peaks.
- As a covariance matrix preserves this information, it is lost with a correlation matrix that assigns equal weights to each sample point, yielding the possibility that small but systematic variations may form a factor.
- These considerations were evaluated and confirmed with simulated ERP data (see Figures 1–3).

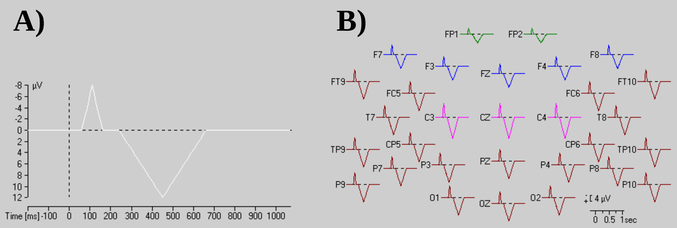

Figure 1. A) Invariant waveform template (128 sample points, 100 samples/sec, 200 ms baseline) used to generate two pseudo ERP data sets for 30 electrode 'sites' and 20 'subjects.' A 'topography' was introduced by scaling the template for selected sites with a factor of 0.5 (Fp1/2), 0.8 (F7/8, F3/4, Fz), or 1.2 (C3/4, Cz). For the second data set, random noise (range ±0.25 µV, uniform distribution) was added to each sample point. **B)** ERP 'group' average of noise data set.

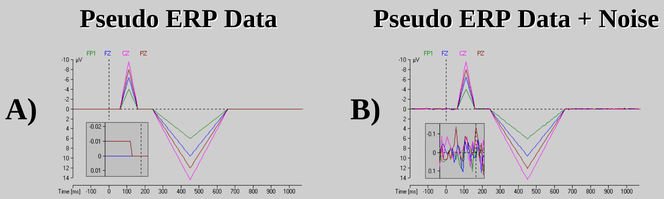

Figure 2. A) Pseudo ERPs at four electrode sites (Fp1, Fz, Cz, Pz). A constant, low-level voltage offset (-0.01 µV) was systematically applied to the pre-stimulus baseline (-200 .. – 50 ms) at every other electrode (e.g., see Pz in inset). **B)** Pseudo ERPs as in A), but with random noise added. Note that the low-level offset at Pz is lost (see inset).

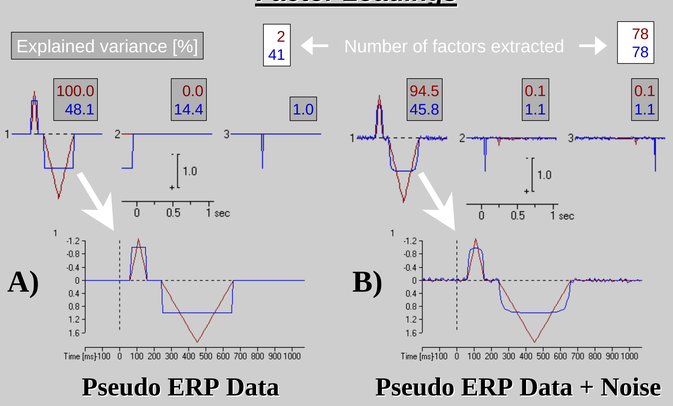

Figure 3. Time course of factor loadings for the first PCA factors extracted from the covariance or correlation matrix for pseudo ERP data with (**B**) and without noise (**A**). The covariance-based PCA extracted a component (factor 1), that accurately reflected the introduced variance shape for both data sets. The correlation-based PCA only produced a component (factor 1) that indicated the direction, but not the size of variations from zero (i.e., from baseline). Similarly, the constant low-level offset was disproportionally reflected in another component (factor 2) for the noise-free data.

Methods

- Real ERP data, collected from healthy, right-handed adults uring a visual half-field study (see Figure 4), were repeatedly submitted to tPCA using BMDP statistical software (4M; Dixon, 1992). Columns of the data matrix represented time (110 sample points from –100 to 1,000 ms), and rows consisted of subjects (16), conditions (4), and electrode sites (30).
- tPCAs were performed for three extraction / rotation criteria:
  1) Covariance matrix / Varimax rotation on raw data
  2) Correlation matrix / Varimax rotation
  3) Covariance matrix / Varimax rotation on standardized variables

  This is the default in SPSS for the covariance matrix!
- 110 tPCAs were computed for each extraction / rotation condition, by systematically increasing the number of components to be extracted from 1 to 110 (= number of variables)

VARIABLES = 128.
CASES = 1920. /
VARIABLE USE = 11 to 120. /
FACTOR METHOD = PCA.
NUMB = {Factors to be extracted}.
{Extraction Method} /
ROTATE METHOD = VMAX. /

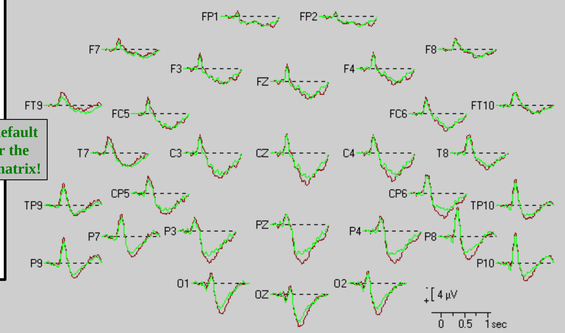

Figure 4. Grand average ERPs for 16 healthy adults for neutral and negative visual stimuli at 30 recording sites, averaged across hemifield of presentation (250 ms exposure in visual half-field paradigm). Data from Kayser et al (2000).

Results

- Limiting the number of components changed the morphology of some components considerably (see Figures 5B and 6B).
- However, more liberal or unlimited extraction criteria did ***not*** degrade or change high-variance components. Instead, their interpretability was improved by more distinctive time courses with narrow and unambiguous peaks (i.e., low secondary loadings; see Figures 5A and 6A).
- Some physiologically meaningful ERP components that are small in amplitude and/or topographically localized (e.g., P1) were found to have a PCA counterpart (e.g., Factor 130; see Figure 8A), that were lost with restricted solutions due to their low overall variance contributions.
- Covariance-based factors had more distinct time courses (i.e., lower secondary loadings) than the corresponding correlation-based factors (Figures 5B and 6B), thereby allowing a better interpretation of their electrophysiological relevance.
- Correlation-based solutions were likely to produce artificial factors that merely reflected small but systematic variations when the ERP waveform intersected the baseline (i.e., zero; cf. Factors –70, 10, and 50 in Figures 6A and 8B).
- Scaling covariance-based PCA factors before rotation approximated correlation-based solutions, and ultimately yielded the same coefficients (factor loadings) when all components were rotated (see Figure 6A).

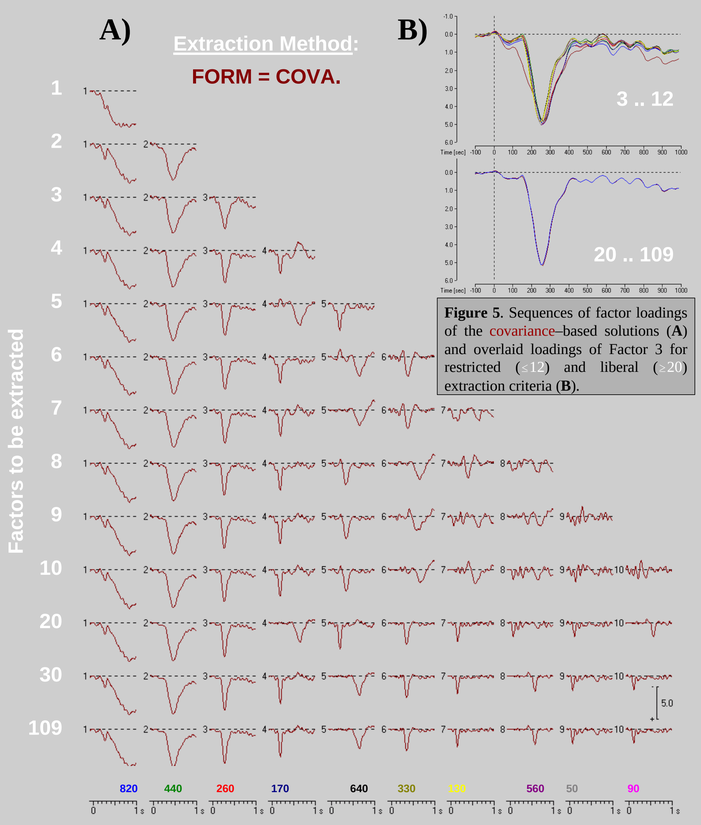

Figure 5. Sequences of factor loadings of the covariance–based solutions (**A**) and overlaid loadings of Factor 3 for restricted (≤12) and liberal (≥20) extraction criteria (**B**).

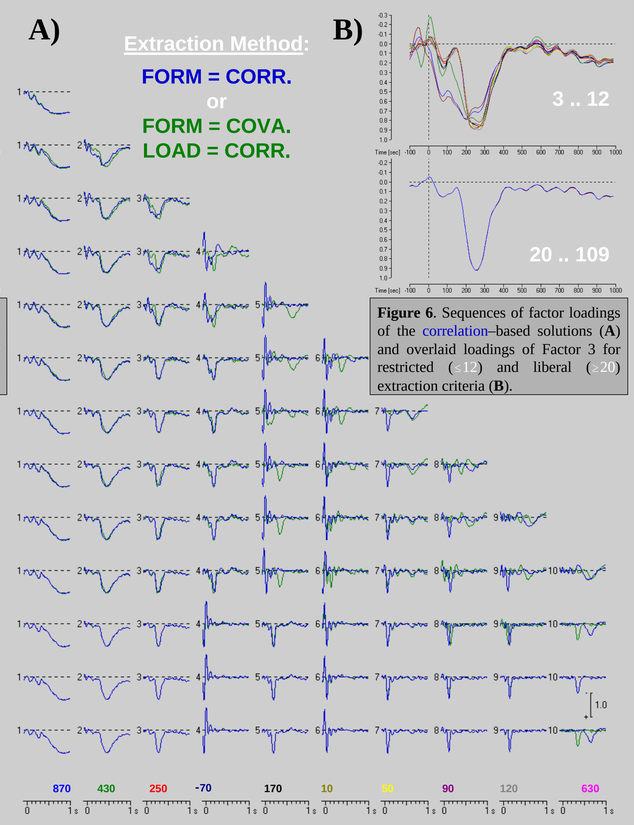

Figure 6. Sequences of factor loadings of the correlation–based solutions (**A**) and overlaid loadings of Factor 3 for restricted (≤12) and liberal (≥20) extraction criteria (**B**).

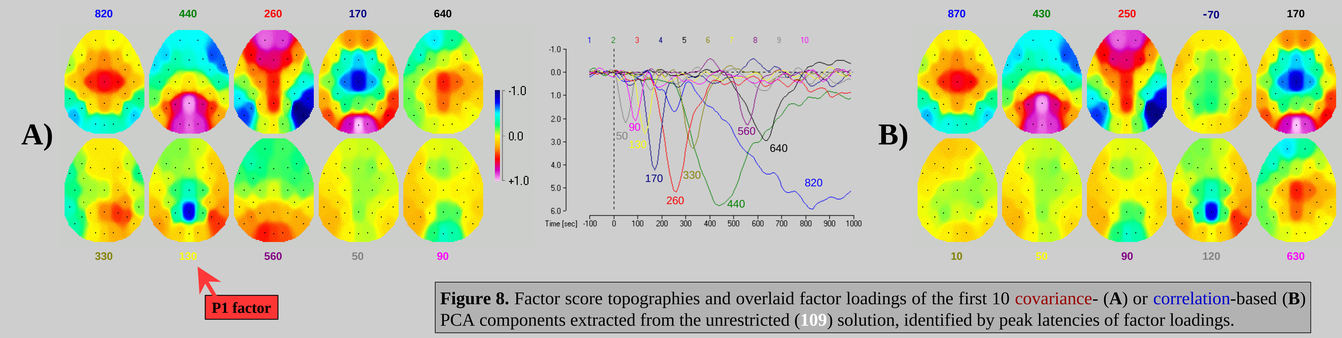

Figure 8. Factor score topographies and overlaid factor loadings of the first 10 covariance- (**A**) or correlation-based (**B**) PCA components extracted from the unrestricted (109) solution, identified by peak latencies of factor loadings.

Conclusions

- **Factor extractions of the unscaled covariance matrix are preferable to correlation- / scaled covariance-based PCA solutions.**
- **For ERP data, there is no reason to restrict the number of factors to be extracted.**

References

Dixon, W.J. (Ed.) (1992). *BMDP Statistical Software Manual (Vol. 2).* Berkeley, CA: University of California Press.

Kayser, J., Bruder, G.E., Tenke, C.E., Stewart, J.E., & Quitkin, F.M. (2000). Event-related potentials (ERPs) to hemifield presentations of emotional stimuli: differences between depressed patients and healthy adults in P3 amplitude and asymmetry. *International Journal of Psychophysiology*, 36(3), 211-236.

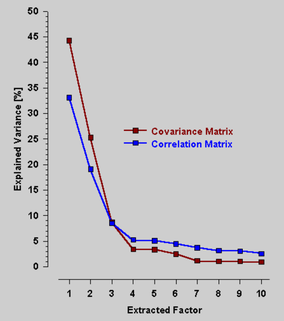

Figure 7. Plots of eigenvalues (percentage of overall variance) for the first 10 factors extracted from the unrestricted (109) covariance or correlation solution.