

Mathematical Challenges in Cross-Classified Factor Analysis

Eric Slud, Statistics Program

Abstract. We survey the problem of choosing an orthonormal basis for representing waveforms observed under varying experimental conditions. When the distribution of waveform ordinates is Gaussian, this is a variant of classic Principal Components and Factor Analysis statistical models. Interesting mathematical issues arise in

- formulating models in terms of identifiable parameters, which may lie in manifolds rather than open Euclidean-space regions,
- finding an appropriate asymptotic framework, due to the comparable numbers of cross-classifying variables and replicate waveforms within each category, and
- numerically computing model estimates.

These problems will be illustrated in terms of real data involving transverse cross-sectional ultrasound pictures of the human tongue during speech.

RESEARCH JOINT WITH MY (FORMER) STUDENTS
YANG CHENG AND SOPHIE (HSIAO-HUI) TSOU.

Formal Data Structure

Fix grid of x -coordinates $\mathbf{x} = x_1, \dots, x_p$. Observe

$$\mathbf{Y}^{(r,j)} \in \mathbf{R}^p, \quad r = 1, \dots, R, \quad 1 \leq j \leq J$$

corresponding to discretized waveforms $(x_i, Y_i^{(r,j)})$.

Indices r for pure independent replicates.

Indices j for categorical cross-classifying labels.

Problem

To use data to represent all labelled curves with respect to an orthonormal basis consisting of constant level $\mathbf{1}$ plus q additional columns formed into $p \times q$ matrix Λ .

Objectives :

- dimension-reduction ($q \ll p$),
- identification of interpretable basis
(columns of Λ below),
- assessment of necessary category-specific differences
(nonconstancy of $\mathbf{a}^{(j)}$, $\mathbf{b}^{(j)}$ below, over j).

Model error-variances may or may not be *nuisance parameters*.

Models & Parameterization

Generalized factor-analysis model

$$\begin{aligned} \mathbf{Y}^{(r,j)} &= \mu^{(j)} \mathbf{1} + \Lambda \mathbf{f}^{(r,j)} + \epsilon^{(r,j)} \\ &\sim \mathcal{N}(\mu^{(j)} \mathbf{1} + \Lambda \mathbf{a}^{(j)}, \Lambda \text{diag}(\mathbf{b}^{(j)}) \Lambda' + \text{diag}(\psi^{(j)})) \end{aligned}$$

$$\mathbf{1}, \epsilon^{(r,j)}, \psi^{(j)} \in \mathbf{R}^p$$

$$\mathbf{f}^{(r,j)}, \mathbf{a}^{(j)}, \mathbf{b}^{(j)} \in \mathbf{R}^q$$

$$f^{(r,j)} \stackrel{ind}{\sim} \mathcal{N}(\mathbf{a}^{(j)}, \text{diag}(\mathbf{b}^{(j)})), \quad \epsilon^{(r,j)} \stackrel{ind}{\sim} \mathcal{N}(\mathbf{0}, \text{diag}(\psi^{(j)}))$$

Columns $\Lambda^{(k)}$, $k = 1, \dots, q$ of Λ are **basis** for expanding all centered wave-forms

$$(I_p - \mathbf{1}\mathbf{1}') \mathbf{Y}^{(r,j)}$$

Scaling coefficients $f_k^{(j,r)}$ have a systematic constant part $a_k^{(j)}$ and a random $\mathcal{N}(0, b_k^{(j)})$ part (*iid* over r)

Unknown parameters

$$\vartheta \equiv (\Lambda, \{\mu^{(j)}, \mathbf{a}^{(j)}, \mathbf{b}^{(j)}, \psi^{(j)}\}_{j=1}^J)$$

high-dimensional. In tongue-data example:

$$p = 100, \quad q = 2, \quad J = 66, \quad R = 30$$

Why Random Coefficients & Independent Errors ?

Classic Factor Model: $J = 1$, R large, $\mathbf{a} \equiv \mathbf{0}$, $\mu = 0$.

Measurements were p psychological test scores (e.g., IQ battery); many replicates r for different test subjects. Objective was to justify single ($q = 1$ or as small as possible) composite score on which to project ‘general intelligence’. Coefficients $\mathbf{f}^{(r)}$ have meaning for individuals, but psychometric model describes the *population*.

Independence of model-errors $\epsilon_i^{(j,r)}$ across i is a restrictive assumption, needed for *model identifiability* (ie unique specification) of parameters.

Common Principal Components, Fleury 1986:

same as factor model except $J > 1$, to assess adequacy of same basis matrix Λ for all categories j .

Waveform Models: $\mu^{(j)}$, $\mathbf{a}^{(j)}$, Λ of direct interest.

Perhaps also the comparison of the size of category non-random versus random effect sizes ($\mathbf{a}^{(j)}$ vs. $\mathbf{b}^{(j)}$).

Models with \mathbf{a} , \mathbf{b} first studied in Yang Cheng thesis (2004); maybe because of different motivation from psychometrics, or because computational obstacles have only recently been easy to overcome.

Identifiability of Parameters

Model: $\mathbf{Y}^{(r,j)} = \mu^{(j)} \mathbf{1} + \Lambda \mathbf{f}^{(r,j)} + \epsilon^{(r,j)}$

Since $\mathbf{f}_k^{(r,j)} \sim \mathcal{N}(a_k^{(j)}, b_k^{(j)})$, directions $\Lambda^{(k)}$ enter mean signal $E(\mathbf{Y}^{(r,j)})$ **and** covariance $\text{Var}(\mathbf{Y}^{(r,j)})$.

Parameter Space: assume $\vartheta \in \Theta$, i.e.

$$\vartheta = (\Lambda, \{\mu^{(j)}, \mathbf{a}^{(j)}, \mathbf{b}^{(j)}, \psi^{(j)}\}_{j=1}^J) \text{ satisfies:}$$

$$\Lambda' \Lambda = I_q, \quad \Lambda' \mathbf{1} = \mathbf{0}, \quad \text{1st nonzero elt} > 0 \text{ in each } \Lambda^{(k)}$$

$$\mu^{(j)} \in \mathbf{R}, \quad \mathbf{a}^{(j)} \in \mathbf{R}^q, \quad \mathbf{b}^{(j)} \in \mathbf{R}_+^q, \quad \psi^{(j)} \in \mathbf{R}_+^p$$

$$\sum_{j=1}^J b_k^{(j)} \text{ strictly } \searrow \text{ in } k$$

Proposition 1 (Identifiability, Cheng 2004, Tsou 2005)

There is a 1-to-1 correspondence between parameters $\theta \in \Theta$ and probability laws for $(\mathbf{Y}^{(j,r)}, 1 \leq j \leq J)$.

The case $J = 1, \mathbf{a} = \mathbf{0}$ is the hardest one. Third Θ condition is just one possible way to order the columns of Λ uniquely.

Still some unsolved issues of uniqueness of $(\Lambda, \mathbf{b}, \psi)$ being determined from $\Lambda \text{diag}(\mathbf{b}) \Lambda' + \text{diag}(\psi)$.

Profile Likelihood Estimation

Restrict for simplicity to case $J=1$, suppress index j , and take $\psi = \sigma^2 I_p$. Parameters $\vartheta \equiv (\vartheta_1, \vartheta_2) = (\Lambda, (\mu, \mathbf{a}, \mathbf{b}, \sigma^2))$ estimated by maximizing log-likelihood:

$$\begin{aligned} & -\frac{R(p-q)}{2} \log \sigma^2 - \frac{1}{2} \sum_{r=1}^R \sum_{k=1}^q \frac{1}{b_k + \sigma^2} (\Lambda^{(k)'} Y^{(r)} - a_k)^2 \\ & - \frac{R}{2} \sum_{k=1}^q \log(b_k + \sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^R (\|Y^{(r)} - \mu \mathbf{1}\|^2 - \|\Lambda' Y^{(r)}\|^2) \end{aligned}$$

Can maximize first, explicitly and uniquely, over

$$\vartheta_2 = (\mu, \mathbf{a}, \mathbf{b}, \sigma^2)$$

as function $\hat{\vartheta}_2(\Lambda)$. **When $J > 1$, maximization decouples over different parameters $(\mu^{(j)}, \mathbf{a}^{(j)}, \mathbf{b}^{(j)}, \sigma_j^2)$.**

Result after substituting back into log-likelihood is

Profile log-Likelihood $l_P(\Lambda, \mathbf{Y}) =$

$$-\frac{R(p-q)}{2} \log \hat{\sigma}^2 - \frac{R}{2} \sum_{k=1}^q \log(\Lambda^{(k)'} S_Y \Lambda^{(k)}) - \frac{Rp}{2}$$

where (using notation $\mathbf{v} = \mathbf{v}^{\otimes 2}$)

$$\bar{Y} = \frac{1}{R} \sum_{r=1}^R Y^{(r)} \quad , \quad S_Y^2 = \frac{1}{R} \sum_{j=1}^R (Y^{(r)} - \bar{Y})^{\otimes 2}$$

$$\hat{\sigma}^2 = \frac{1}{p-q} \{tr((S_Y + \bar{Y}^{\otimes 2})(I_p - \Lambda \Lambda')) - p^{-1}(\mathbf{1}' \bar{Y})^2\}$$

Computational Issues

Approaches to maximization:

- direct optimization of $\log Lik$ over ϑ does not work well in realistically large problems;
- maximization of $l_P(\Lambda, \mathbf{Y})$ almost as difficult ($\dim = pq - q(q - 1)/2$);
- EM algorithm (Cheng 2004 thesis, $\mathbf{a}^{(j)} \neq \mathbf{0}$, general $J > 1$, but $\psi^{(j)} = \sigma_j^2 \mathbf{1}$)

idea is to maximize iteratively by finding

$$\vartheta_{k+1} = \arg \max_{\vartheta} E_{\vartheta_k}(\log Lik(\mathbf{Y}, \mathbf{f}, \vartheta) | \mathbf{Y})$$

easy to implement analytically but converges slowly;

- profile $\log Lik$ approach solving for Λ as function of general ψ when $\mathbf{a} = \mathbf{0}$.

Implemented in Tsou 2005 thesis, for $J = 1$; Newton-Raphson optimization over ψ , $\dim = p$

- slow convergence in problems with $\mathbf{a}^{(j)} = \mathbf{0}$ found in Tsou 2005 thesis to be associated with ‘boundary’ solutions (some $\psi_k^{(j)} = 0$)
- convergence can be speeded up by projecting from \mathbf{R}^p down to eigenspace for largest singular values of

$$S_Y^2 \equiv \frac{1}{JR} \sum_{j=1}^J \sum_{r=1}^R (\mathbf{Y}^{(j,r)} - \bar{Y}^{(j)})^{\otimes 2}$$

Maximum Likelihood Asymptotics, $R \rightarrow \infty$

Standard statistical theory shows in regular finite-dim *Euclidean* parameter problems that as sample-size (*iid* replicate number R) goes to ∞ MLE satisfies

$$\sqrt{R}(\hat{\vartheta} - \vartheta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, I^{-1})$$

as long as $I = -E_{\vartheta_0}(\nabla^{\otimes 2} \log \text{Lik}(\vartheta_0))$ is non-singular.

In Factor models, orthonormality constraints on Λ columns make the nontrivial verification take the form (Cheng 2004): for some $c > 0$,

$$\frac{1}{R} E_{\vartheta_0} \{l_P(\mathbf{Y}, \Lambda_0) - l_P(\mathbf{Y}, \Lambda)\} \geq c \cdot \|\Lambda - \Lambda_0\|^2$$

This is a positive-definite Information verification for the Profile Likelihood based on parameter with values in a manifold.

‘Two-Index Asymptotics’

Previous slide relates to standard asymptotics in which J is fixed but $R \rightarrow \infty$. This is artificial in real examples. Both indices get large in large projects !

Two features of large cross-classified datasets — the subject of my most recent RIT’s — which lead to new phenomena:

- independent data $Y^{(j,r)} \sim f(y, \Lambda, \vartheta_2^{(j)})$ distributed with parameter Λ common to all, but $\vartheta_2^{(j)}$ decoupled (for fixed Λ) in different groups j ;

and

- both $R, J \rightarrow \infty$ with $R = \mathcal{O}(J)$

Since overall dataset size is of order $N = RJ$ this says replication numbers R are $\mathcal{O}(\sqrt{N})$.

Paper of Li, Lindsay, & Waterman (2003) just discussed in RIT has result which says in this two-index asymptotic context:

MLE’s are consistent but no longer asymptotically efficient (i.e., no longer have minimum possible variance); and known ‘projected score estimating equation’ correction restores efficiency, as long as $J = \mathcal{O}(R^{1+\delta})$.

Tongue Data Example

Data example drawn from long-term NIH project of Dr. Maureen Stone of UM Dental School, on which Prof. Paul Smith and students Yang Cheng and Sophie Tsou also worked.

Data consisted of cross-sectional ultrasound-derived images of human tongue during speech, with about 100 (x, y) coordinate pairs for each of 6 speakers and 11 vowel sounds ($J = 11 \times 6 = 66$) with replications based on separate sessions, images within session, and (S vs L) consonant bracketing ($R = 3 \times 5 \times 2 = 30$).

PICTURES AND TABLE TAKEN FROM TSOU 2005 THESIS.

Pictures consist of plotted Principal Components (eigenvectors of Σ_Y) overlaid with estimated columns ($q = 2$) of Λ .

Table shows estimated scale-factors α_j in speaker and sound estimated variances σ_j^2 , except that Table has j doubly indexed through a for vowel sound and s for speaker.

First Principal Direction

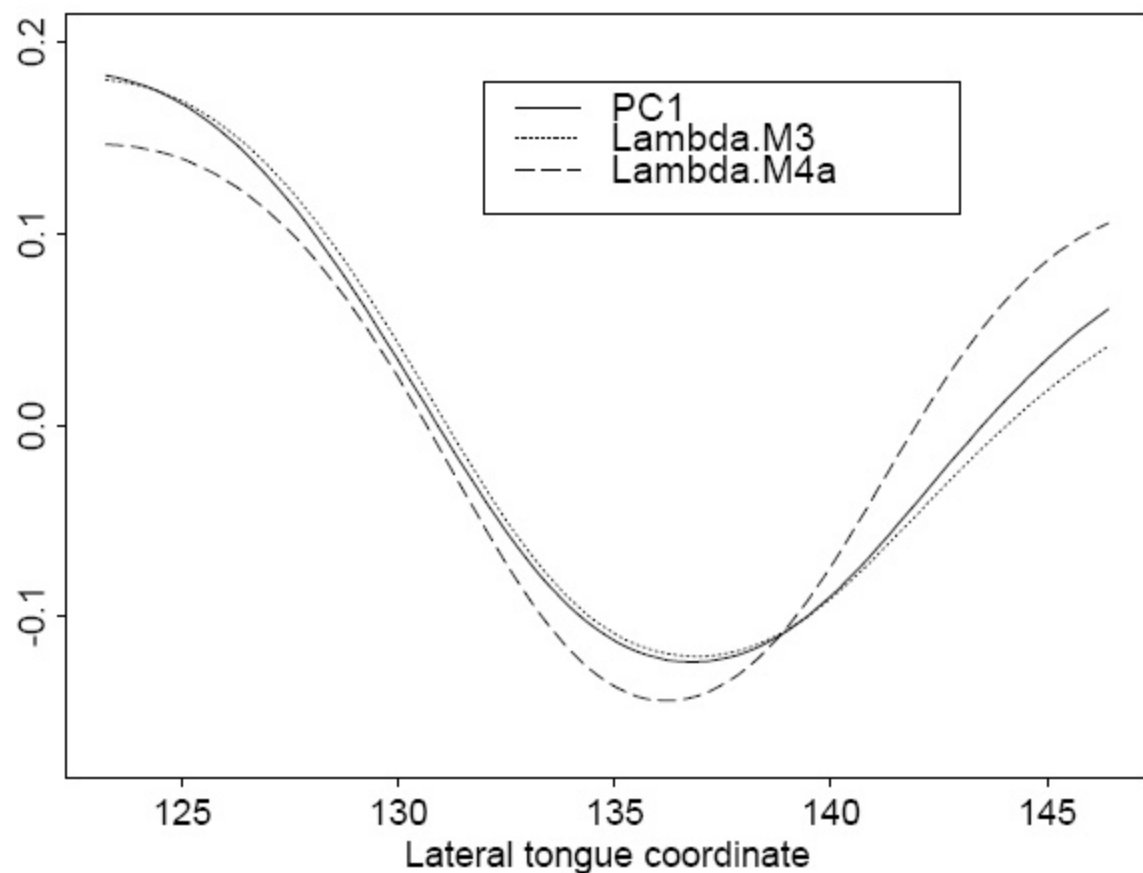


Figure 5.2: First Principal Direction for coronal tongue data based on (PCA), (M3) and (M4a).

Second Principal Direction

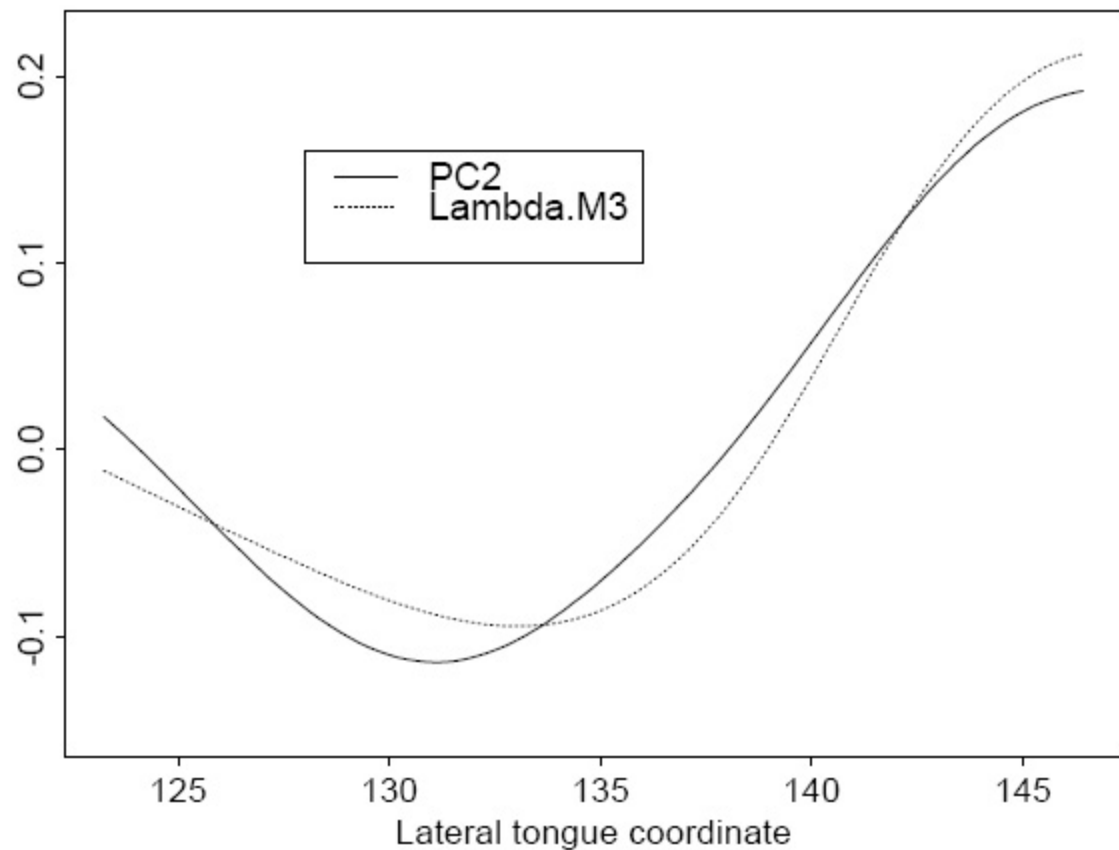


Figure 5.3: Second Principal Direction for coronal tongue data based on (PCA) and (M3).

	M.S.	M.D.	S.G.	C.S.	G.W.	L.G.
ae	1.033	0.955	0.793	0.788	1.161	0.443
ah	0.583	0.663	1.073	0.901	0.843	0.996
aw	0.683	0.370	0.922	0.703	0.554	1.312
e	1.421	1.418	0.832	0.400	1.543	1.307
eh	0.738	0.962	1.002	0.650	1.199	0.691
ih	1.828	1.003	0.920	0.735	1.119	0.498
iy	1.147	3.082	2.784	0.304	3.615	2.241
o	0.548	0.530	0.767	0.634	0.804	1.092
uh	0.593	0.372	1.081	0.714	0.539	1.218
uu	2.096	0.976	1.569	0.722	1.110	0.677
uuh	0.480	0.480	1.240	0.655	0.643	1.564

Table 5.1: The estimated values of the scaled parameters α_{as} .

References

- Anderson, T.W. (1984) *Intro. to Multiv. Analysis*
Flury (1984, 1988) *Common Principal Components*
H. Hotelling (1933)
Anderson, T.W. and Rubin, H. (1956)
D. Lawley (1940), K. Jöreskog (1969)
Yang Cheng (2004) *UMCP Stat Thesis.*
Hsaio-Hui Tsou (2005) *UMCP Stat Thesis.*

Mental Test references

- Spearman (1904), Thurstone (1947) et al.
Gould, S.J. (1981) *The Mismeasure of Man*

Tongue references

- Slud, Stone, Smith, and Goldstein (2002) *Phonetica.*
Smith, Stone, Slud, and Tsou (2005) Preprint.

Two-index asymptotics reference

- Li, Lindsay, and Waterman (2003) *JRSSB*