Dimension Reduction of Articulatory Data

Miguel Á. Carreira-Perpiñán\textsuperscript{a}

Dept. of Computer Science, University of Sheffield

M.Carreira@dcs.shef.ac.uk

NATO ASI on *Generalization in Neural Networks and Machine Learning*
4–15 August, Isaac Newton Institute, Cambridge, U.K.

\textsuperscript{a}This work has been partially supported by the Spanish Ministry of Education.
Electropalatography (EPG)

- A plastic pseudopalate fitted to a person’s mouth detects the presence or absence of contact between the tongue and the palate in 62 different locations.
- During an utterance the artificial palate signal is sampled at 200 Hz.
- The result is a long sequence of EPG frames, each of which is a 62-dimensional binary vector.
- Some data reduction is needed to extract useful information from this enormous amount of data.
- The ACCOR-II database contains EPG sequences synchronised with the speech signal for different sentences/speakers/styles/languages.
- Note that the mapping phoneme-to-EPG is not one-to-one, e.g. [æ]/[ə] or [d]/[t].

Figure 1: Representative EPGs for the typical stable phase of different phonemes.
Mechanical constraints of the vocal tract suggest it has a limited number of degrees of freedom.

A low-dimensional representation of EPG data would give insight on the acoustic-to-articulatory mapping problem (finding the vocal tract configuration that produced a given speech signal).

Most EPG reduction methods are based in a priori assumptions.

Seek automatic methods that learn from data for adaptation to different speakers, styles, etc.

Figure 2: Ad hoc indices for EPG reduction (= basis vectors of an orthogonal, linear projection).
Latent Variable Models

- Given a sample \( \{ t_n \}_{n=1}^N \subset \mathbb{R}^D \) we want to model the probability distribution \( p(t) \) that generated it in terms of a small number \( L \) of latent variables \( x \in \mathbb{R}^L \). Define:
  - Prior distribution in latent space \( p(x) \).
  - Smooth mapping \( y(x; \Theta) \) from latent space onto \( L \)-manifold \( M \) with parameters \( \Theta \).
  - Noise model \( p(t|x; \Theta) \) for \( y(x; \Theta) \) that extends \( M \) to \( \mathbb{R}^D \).

The distribution of \( t \) in data space under this model is \( p(t|x) = \int p(t|x; \Theta)p(x) \, dx \).

- Parameter estimation by maximum likelihood: \( \hat{\Theta} = \arg \max_{\Theta} \log \prod_{n=1}^N p(t_n|\Theta) \).

- Posterior distribution in latent space (inverse mapping): \( p(x|t; \hat{\Theta}) = \frac{p(t|x; \hat{\Theta})p(x)}{p(t|\hat{\Theta})} \) (Bayes’ th.).
Factor Analysis (FA)

- Covariance structure model that tries to explain linear correlations of a large set of variables $t$ in terms of a small number of underlying factors $x$:

$$ t = \Lambda x + u + \mu $$

$$ \begin{cases} 
  \mathbb{E}\{x\} = 0, \text{cov}\{x\} = I \\
  \mathbb{E}\{u\} = 0, \text{cov}\{u\} = \Psi \text{ diagonal} \\
  \text{cov}\{x, u\} = 0 
\end{cases} \implies \begin{cases} 
  \mathbb{E}\{t\} = \mu \\
  \text{cov}\{t\} = \Lambda \Lambda^T + \Psi 
\end{cases} $$

- FA can be seen as a latent variable model with linear mapping and normal distributions:

$$ y(x; \Lambda) = \Lambda x + \mu \quad p(x) \sim \mathcal{N}(0, I) \quad p(t|x; \Lambda, \Psi) \sim \mathcal{N}(y(x; \Lambda), \Psi) $$

- The FA model must be identified: unique solution up to orthogonal rotations. **Varimax rotation** often produces an interpretable set of loadings.

- Iterative maximum likelihood estimation of parameters $\Lambda$ and $\Psi$ assuming normal data.

- Dimension reduction by Thomson scores: $x = \mathbb{E}\{x|t\} = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} \Lambda^T \Psi^{-1} t$. 

Miguel A. Carreira-Perpiñán
Results with Factor Analysis

Figure 3: Loadings (= basis vectors) for the first 9 factors.

- The loadings associated to the first few factors justify empirically the ad hoc indices and also reflect the idiosyncrasy of the data set.

- Orthogonal rotation of the eigenEPGs found by principal component analysis (PCA) produces a similar set of basis vectors in this case.

- Reconstruction error in the least-squares sense using the first two factors is 35% worse than that of PCA (which is optimal for linear mappings).
Results with Factor Analysis (cont.)

- Visualisation in 2D latent space by projecting the data on the subspace spanned by 2 factors:
  - 2D-map with regions corresponding to different phonemes classes.
  - Plot of the trajectory followed during an utterance.
- Discrimination depends on the particular projection direction chosen, e.g. F2–F4 is better than F1–F2. Since the projection is linear, this is a particular case of projection pursuit.
Results with Mixtures of Factor Analysers (MFA)

- Mixture $p(t|\Theta) = \sum_{i=1}^{M} \pi_i p(t|\pi_i, \mu_i, \Lambda_i, \Psi_i)$ of factor analysers $p(t|\Lambda_i, \Psi_i)$ centred in means $\mu_i$, trained by EM algorithm for maximum likelihood.

- Preliminary results with $L = 1$ factor per component show that a local representation is found in which:
  - The means $\mu_i$ coincide with typical EPG patterns such as those of fig. 1.
  - The loadings $\Lambda_i$ are linear combinations of those obtained by standard FA, fig. 3.

This suggests that only certain factors are relevant for a given phoneme.

\[\begin{align*}
\mu_1 &\approx \mu \approx [n] \\
\Lambda_1 &\approx 1.5 \lambda_3 + \lambda_7 \\
\mu_2 &\approx [\text{v}] \\
\Lambda_2 &\approx \lambda_5 \\
\mu_3 &\approx [\theta] \\
\Lambda_3 &\approx \lambda_1 + 0.2 \lambda_2
\end{align*}\]
The Generative Topographic Mapping (GTM)

- GTM is a latent variable model with:
  1. Mapping implemented with *generalised linear model*: \( y(x; W) = W \Phi(x) \) for fixed basis function vector \( \Phi \).
  2. Discrete, uniform prior in latent space \( p(x) = \frac{1}{K} \sum_{i=1}^{K} \delta(x - x_i) \) (usually in the hypercube \([-1, 1]^L\)): Monte Carlo approximation of a continuous, uniform prior.
  3. Normal, isotropic noise model \( p(t|x; W, \sigma) \sim \mathcal{N}(y(x; W), \sigma^2 I) \).

So \( p(t|W, \sigma) \) is a *constrained* mixture of Gaussians.

- Maximisation of the log-likelihood through EM algorithm.
- Disadvantage: choice 1 gives exponential complexity in \( L \) (*curse of the dimensionality*), which in practice prevents \( L > 2 \).
The posterior distribution $p(x|t; W, \sigma)$ is approximately unimodal and sharply peaked for most patterns: single responsibility for each EPG (left graph: plot of the posterior distribution mean and mode linked by a line for a random subset of the data points).

- Discriminating ability better than that of the best factor pair.
- Reconstruction error in the least-squares sense is half of that of PCA using the first two eigenvectors (and equal to that of the first 9 eigenvectors).