In: Proc. of the 1998 IEEE Signal Processing Society
Workshop on Neural Networks for Signal Processing
(NNSP98), pp.165–173, Cambridge, UK.
URL: http://www.dcs.shef.ac.uk/~miguel/papers/nnsp98.html

Experimental Evaluation of Latent Variable Models for Dimensionality Reduction

Miguel Á. Carreira-Perpiñán **Steve Renals**
Dept. of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{M.Carreira,S.Renals}@dcs.shef.ac.uk

Abstract

We use electropalatographic (EPG) data as a test bed for dimensionality reduction methods based in latent variable modelling, in which an underlying lower dimension representation is inferred directly from the data. Several models (and mixtures of them) are investigated, including factor analysis and the generative topographic mapping (GTM). Experiments indicate that nonlinear latent variable modelling reveals a low-dimensional structure in the data inaccessible to the investigated linear models.

1 Introduction

In latent variable modelling, a low-dimensional generative model is estimated from a data sample. A smooth mapping links the low-dimensional representation and the high-dimensional data, and dimensionality reduction is achieved via Bayes' theorem. Latent variable models include factor analysis, principal component analysis and the generative topographic mapping (GTM). In this paper, we apply the latent variable framework to electropalatographic data.

The technique of electropalatography (EPG) is well established as a relatively non-invasive, conceptually simple and easy-to-use tool for the investigation of lingual activity in both normal and pathological speech. Qualitative and quantitative data about patterns of lingual contacts with the hard palate during continuous speech may be obtained using EPG, and the technique has been used in studies of descriptive phonetics, coarticulation, and diagnosis and treatment of disordered speech (Hardcastle et al., 1991a). Typically, the subject wears an artificial palate moulded to fit

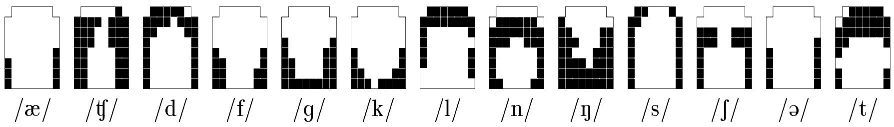

Figure 1: Representative EPGs for the typical stable phase of different phonemes. The 62-dimensional binary EPG vector is pictured rowwise from top to bottom resembling the human palate (top: alveoli; bottom: velum).

the upper palate with a number of electrodes mounted on the surface to detect lingual contact (62 in the Reading EPG system). The EPG signal is sampled at a frequency of 100 to 200 Hz and, for a given utterance, the sequence of raw EPG data consists of a stream of binary vectors with both spatial and temporal structure due to the constraints of the articulatory system. Note that the EPG signal alone is an incomplete articulatory description, omitting such details as nasalisation and vocalisation. Hence, the mapping from phonemes to EPG patterns is not one-to-one since certain phonemes (e.g. /æ/ and /ə/) can produce the same EPG patterns (fig. 1).

A number of studies suggest that tongue movements in speech may be appropriately modelled using a few elementary articulatory parameters (e.g. (Nguyen et al., 1996)). In this paper, we consider dimensionality reduction at the spatial level only. We compare the ability of several well-known latent variable models and mixture models to extract relevant structure from an EPG data set in an adaptive fashion. This contrasts with a number of widespread EPG data reduction techniques (Hardcastle et al., 1991b) which are based in a priori knowledge about electropalatography, typically in the form of fixed linear combinations of the EPG vector components. Such ad hoc methods are not robust and will not perform well in situations where the speech deviates from the standard (impaired speakers, different speech styles or unusual accents). We also show that nonlinearity is very advantageous for this real-world problem.

2 Generative modelling using latent variables

In latent variable modelling the assumption is that the observed high-dimensional data $\mathbf{t}$ is generated from an underlying low-dimensional process defined by a small number L of *latent variables* $\mathbf{x}$ (Bartholomew, 1987). The latent variables are mapped by a fixed transformation into a D-dimensional data space (measurement procedure) and noise is added there (stochastic variation). The aim is to learn the low dimensional generating process along with a noise model, rather than directly learning a dimensionality reducing mapping.

A latent variable model is specified by a prior in latent space $p(\mathbf{x})$, a smooth mapping $\mathbf{f}$ from latent space to data space and a noise model in data space $p(\mathbf{t}|\mathbf{x})$ (fig. 2). These three elements are equipped with parameters which we collectively call $\mathbf{\Theta}$. Integrating the joint probability density function $p(\mathbf{t}, \mathbf{x})$ over latent space gives the marginal distribution in data space, $p(\mathbf{t})$. Given an observed sample in data space $\{\mathbf{t}_n\}_{n=1}^N$ of N D-dimensional real vectors that has been generated by an unknown distribution, a parameter estimate can be found by maximising the log-likelihood of

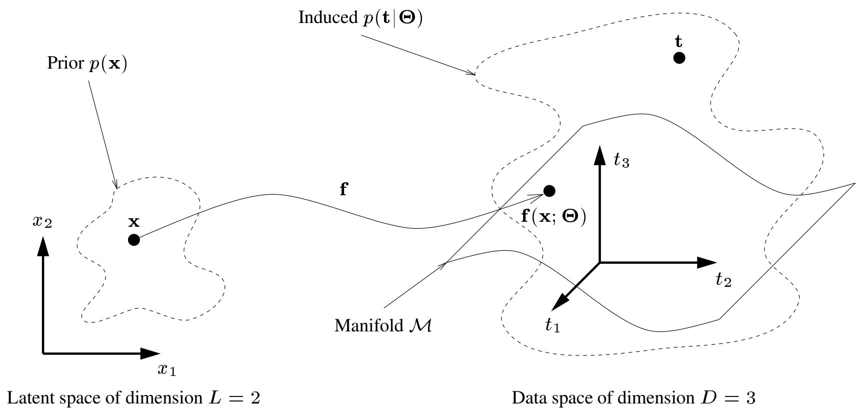

Figure 2: Schematic of a latent variable model with a 3D data space and a 2D latent space.

the parameters $l(\mathbf{\Theta}) = \sum_{n=1}^{N} \log p(\mathbf{t}_n|\mathbf{\Theta})$, typically using an EM algorithm.

Once the parameters $\mathbf{\Theta}$ are fixed, Bayes' theorem gives the posterior distribution in latent space given a data vector $\mathbf{t}$, i.e. the distribution of the probability that a point $\mathbf{x}$ in latent space was responsible for generating $\mathbf{t}$:

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})} = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}}. \tag{1}$$

Summarising this distribution in a single latent space point $\mathbf{x}^*$ (typically the mean or the mode) results in a *reduced-dimension representative* of $\mathbf{t}$. This defines a corresponding dimensionality reducing mapping $\mathbf{F}$ from data space onto latent space, $\mathbf{x}^* = \mathbf{F}(\mathbf{t})$, which will be most successful when the posterior distribution $p(\mathbf{x}|\mathbf{t})$ is unimodal and sharply peaked. Applying the mapping $\mathbf{f}$ to the reduced-dimension representative we obtain the reconstructed data vector $\mathbf{t}^* = \mathbf{f}(\mathbf{x}^*)$. In the usual way, we define the squared reconstruction error for the sample as $E = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{t}_n - \mathbf{t}_n^*\|^2$ using the Euclidean norm.

We consider the following latent variable models, for which EM algorithms are available:

Factor analysis (Bartholomew, 1987; Rubin and Thayer, 1982), in which the mapping is linear, the prior in latent space is unit Gaussian and the noise model is diagonal Gaussian. The marginal in data space is then Gaussian with a constrained covariance matrix.

Principal component analysis (PCA), which can be considered a particular case of factor analysis with isotropic Gaussian noise model (Tipping and Bishop, 1997).

The generative topographic mapping (GTM) (Bishop et al., 1998) is a nonlinear latent variable model, where the mapping is a generalised linear model, the prior in latent space is discrete uniform and the noise model is isotropic

Gaussian. The marginal in data space is then a constrained mixture of Gaussians.

Finite mixtures of latent variable models may be constructed, and the EM algorithm used to obtain a maximum likelihood parameter estimate (Everitt and Hand, 1981). A reduced-dimension representative can be obtained from the mixture component with the highest responsibility, or as the average of the per-component representatives weighted by the responsibilities. We also consider mixtures of multivariate Bernoulli distributions, where each component is parameterised by a D-dimensional probability vector; note that this does not define a latent space, but only a discrete collection of components.

3 Experiments

We used a subset of the EUR-ACCOR database (Marchal and Hardcastle, 1993), consisting of $N = 11338$ 62-bit EPG frames sampled at 200 Hz from 14 different utterances by an English speaker. The data set was split into a training (75%) and a test set (25%). All the data used were unlabelled. Using the training set, we found maximum likelihood estimates for the following probability models: factor analysis, PCA, GTM, mixtures of factor analysers and mixtures of multivariate Bernoulli distributions. Figure 3 shows the prototypes found by several of these methods. Figure 4 gives the log-likelihood and reconstruction error curves for each method on both the training and test sets. Performing the same experiments on data sets where all repeated EPG frames were removed did not change significantly the shape of these curves.

Factor analysis and principal component analysis were performed on the EPG data followed by varimax rotation to improve factor interpretability without altering the model. Each 62-dimensional factor loadings vector or principal component vector may be considered as a dimensionality reduction index, in the sense that projecting an EPG frame on the vector is equivalent to a linear combination of its components. The resultant basis vectors are shown in the first two rows of fig. 3 for 9th-order models. Several of these factors can be associated to well-known EPG data reduction indices or to linear combinations of them; e.g. $\boldsymbol{\lambda}_1$ to a velar index or $\boldsymbol{\lambda}_3$ to an alveolar one. But note that the derived factors indicate adaptive structure (e.g. asymmetry) which a priori derived indices cannot capture.

Several of the principal components are similar to some of the factor loading vectors due to the fact that for this data set the uniquenesses matrix was relatively isotropic (i.e. a multiple of the identity), in which case factor analysis is equivalent to PCA.

We used the probabilistic PCA model of (Tipping and Bishop, 1997) to compute the log-likelihood curves of fig. 4. PCA outperforms factor analysis in reconstruction error and factor analysis outperforms PCA in log-likelihood. Thus in terms of generative modelling, factor analysis is superior to PCA, but in terms of reconstruction error PCA is a better linear method.

Generative topographic mapping (GTM) Both factor analysis and PCA can only extract linear structure from the data. For factor analysis, the null hypothesis that "the data sample can be explained with L factors" was rejected for all values of

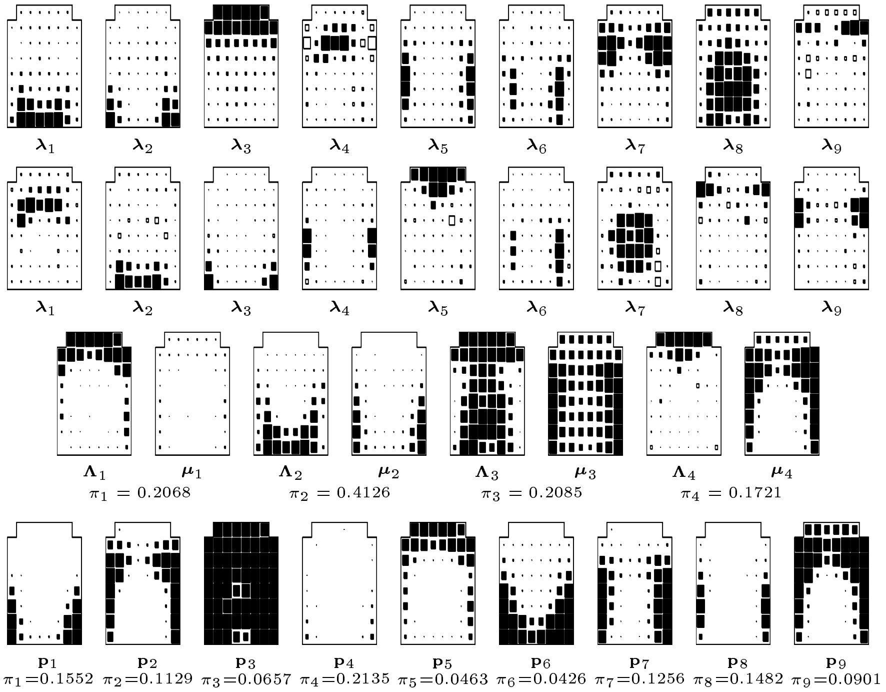

Figure 3: Prototypes and factors found by several methods. Each 62-dimensional vector is pictured as the EPG frames of fig. 1, but now each pixel is represented by a rectangle whose area is proportional to the magnitude of the pixel value and whose colour is black for positive values and white for negative ones. Row 1: factors after varimax rotation. Row 2: principal components after varimax rotation. Row 3: means $\boldsymbol{\mu}_m$ and factor loadings $\boldsymbol{\Lambda}_m$ for a mixture of factor analysers. Row 4: prototypes for a mixture of multivariate Bernoulli distributions.

$L < 45$ at a significance level of 95%. Using an EM algorithm (Bishop et al., 1998), we trained a two-dimensional GTM model with the following parameters: 20×20 grid in two-dimensional latent space ($K = 400$ points), scaled to the $[-1, 1] \times [-1, 1]$ square, and $\sqrt{F} \times \sqrt{F}$ grid in the same square of F Gaussian basis functions of width equal to the separation between basis functions centres; $\sqrt{F}$ varied from 3 to 14. For each data point, we took as reduced-dimension representative the mode of its posterior distribution (which was unimodal and sharply peaked for over 90% of the data points). The log-likelihood curve for the test set as a function of the number of basis functions used F shows a maximum for $F = 49$, indicating that overfitting occurs for $F > 49$ (but observe that the reconstruction error keeps decreasing steadily past that limit). Comparison with the other methods shows that GTM, using a latent space of only $L = 2$ dimensions, outperforms all the other methods in log-likelihood and reconstruction error in a wide range of latent space dimensions. PCA needs $L = 10$ principal components to attain the same reconstruction error as GTM with $F = 49$ basis functions, and all $L = 62$ components to surpass its

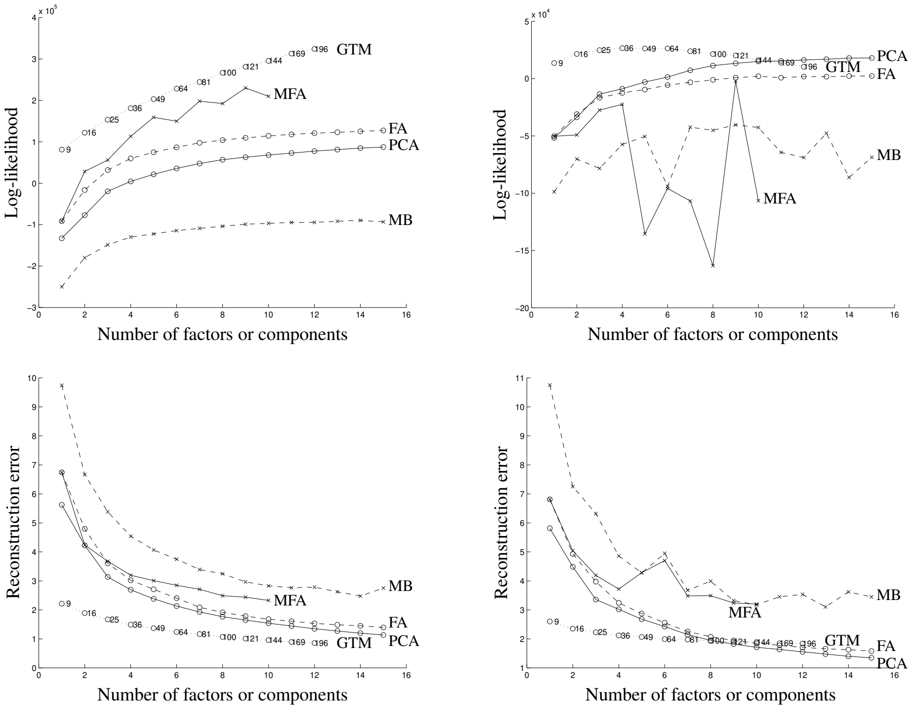

Figure 4: Comparison between methods in terms of log-likelihood (top) and reconstruction error (bottom) (left: training set; right: test set): factor analysis (FA), principal component analysis (PCA), generative topographic mapping (GTM), mixtures of factor analysers (MFA) and mixtures of multivariate Bernoulli distributions (MB). Note that the x axis refers to the order of the factor analysis or principal component analysis, the number of mixture components in the case of mixture models and the square root of the number of basis functions in the case of GTM.

log-likelihood.

Mixtures of factor analysers of $L = 1$ factor per factor analyser[1] were estimated using an EM algorithm (Ghahramani and Hinton, 1996) with random starting points. Fig. 3 (row 3) shows the loading vectors and means for a mixture of $M = 4$ components. The effect of the model is to place factors ($\mathbf{\Lambda}_m$) in different regions of data space ($\boldsymbol{\mu}_m$); the factors found coincide with some of the factors found in factor analysis or with linear combinations of them and the means coincide with some of the typical EPG patterns of fig. 1. In the training set, the log-likelihood and reconstruction error of the mixture are always better than that of factor analysis, but there is not much improvement in the test set. The log-likelihood space has a number of local maxima of different log-likelihood value and that explains the ragged appearance of the curves (where each point corresponds to a single estimate, not to an average of

[1]This kind of mixture has a number of log-likelihood singularities in parameter space (Heywood cases), and for $L > 1$ EM failed to converge to a proper solution.

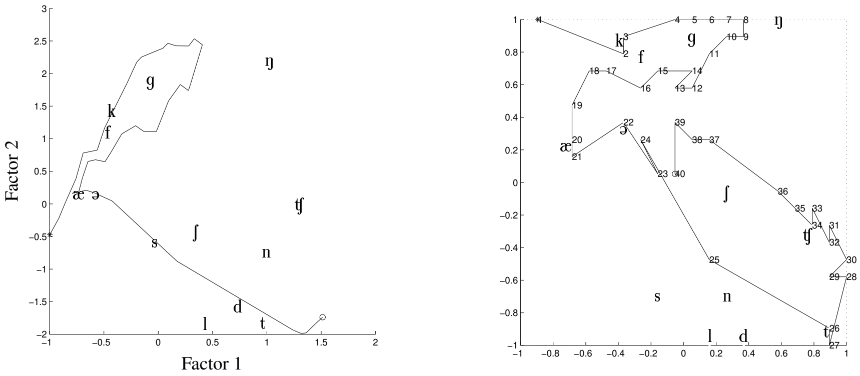

Figure 5: Two-dimensional plot of the trajectory of the highlighted utterance fragment "I prefer **Kant to** Hobbes for a good bedtime book," with phonemic transcription /aɪ prɪˈfɜ ˈkænt tə ˈhɒbz fər ə ˈgʊd ˈbedtaɪm ˈbʊk/. Left: using factor analysis (latent space of factors 1 and 2). Right: using GTM with $F = 49$ basis functions and a 20×20 latent space grid (points are numbered correlatively). The start and end points are marked as $*$ and $\circ$, respectively. The phonemes are those of figure 1.

several estimates).

Mixtures of multivariate Bernoulli distributions were estimated using an EM algorithm (Everitt and Hand, 1981) with random starting points. Fig. 3 (row 4) shows the prototypes $\mathbf{p}_m$ for a 9-component mixture. Note that each p_{md} value is in $[0, 1]$ and thus is readily interpretable as an EPG vector, unlike loading vectors which can have positive and negative values. Again, most of the prototypes coincide with some of the typical EPG patterns of fig. 1. However, the log-likelihood value is less than that of any of the other methods and the reconstruction error is also greater. The reason is that each component of the mixture lacks a latent space and can only reconstruct a data vector as its prototype $\mathbf{p}_m$.

Two-dimensional visualisation of EPGs We can represent graphically a sequence of EPG frames (sample from an utterance) by plotting their projections in a two-dimensional latent space (joining consecutive points with a line). Figure 5 shows such a representation for factor analysis using factors 1 and 2 (left) and GTM (right). For linear projections (such as the ones provided by factor analysis and PCA), finding the best projection (e.g. in the sense of being as nongaussian as possible) is called *projection pursuit*. However, this is not the criterion optimised by factor analysis, which, in general, is then not a good method for this aim. The two-dimensional latent space produced by GTM gives a qualitatively better visualisation of the articulatory space than that of two-dimensional linear projections.

4 Discussion

Latent variable models can be useful to present the relatively high-dimensional information contained in the EPG sequence in a way which is easier to understand and handle. This is a significant advantage over general probability models and mixtures of them, which are suitable for classification or clustering but not for dimensionality reduction. Finite mixtures offer the possibility of fitting, in a soft way, different models in different data space regions and thus offer potential to model complex data. However, training is slow and often difficult due to the log-likelihood surface being plagued with singularities.

For the cases studied, overfitting in the log-likelihood can appear if the number of parameters of the model is large enough, but the reconstruction error presents a steady decrease in both the training and test sets for any number of parameters. Unidentified models (in which different, nontrivial combinations of parameter values produce exactly the same distribution) can produce different estimates from the same data set. This is not the case for factor analysis and PCA, and seems not to have much importance for mixtures of multivariate Bernoulli distributions (Carreira-Perpiñán and Renals, 1998), but may pose problems of interpretation for the other models.

GTM has proven to be the best method both in log-likelihood and reconstruction error, despite being limited to a two-dimensional space due to its computational complexity. This suggests that the intrinsic dimensionality of the EPG data may be substantially smaller than that suggested (5 to 10) by other studies (e.g. (Nguyen et al., 1996)).

All the methods we have studied require knowledge of the latent space dimension or the number of mixture components; a possible way to determine the optimal ones is by model selection.

Acknowledgments

This work was supported by ESPRIT Long Term Research Project SPRACH (20077), by a scholarship from the Spanish Ministry of Education and Science and by an award from the Nuffield Foundation. The authors acknowledge Markus Svensén and Zoubin Ghahramani for the Matlab implementations of GTM and the mixtures of factor analysers, respectively, and Alan Wrench for providing them with the ACCOR data. Matlab implementations of principal component analysis, factor analysis (and varimax rotation) and mixtures of multivariate Bernoulli distributions are available from the first author at URL `http://www.dcs.shef.ac.uk/~miguel/`.

References

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. Charles Griffin & Company Ltd., London.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234.

Carreira-Perpiñán, M. Á. and Renals, S. J. (1998). On finite mixtures of multivariate Bernoulli distributions. Submitted to *Neural Computation*.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York.

Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Hardcastle, W. J., Gibbon, F. E., and Jones, W. (1991a). Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders. *Brit. J. of Disorders of Communication*, 26:41–74.

Hardcastle, W. J., Gibbon, F. E., and Nicolaidis, K. (1991b). EPG data reduction methods and their implications for studies of lingual coarticulation. *J. of Phonetics*, 19:251–266.

Marchal, A. and Hardcastle, W. J. (1993). ACCOR: Instrumentation and database for the cross-language study of coarticulation. *Language and Speech*, 36(2, 3):137–153.

Nguyen, N., Marchal, A., and Content, A. (1996). Modeling tongue-palate contact patterns in the production of speech. *J. of Phonetics*, 24:77–97.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Tipping, M. E. and Bishop, C. M. (1997). Mixtures of principal component analysers. In *Proceedings of the IEE Fifth International Conference on Artificial Neural Networks*. London:IEE.