In: Proc. of the 1999 IEE Colloquium on Applied
Statistical Pattern Recognition, Birmingham, UK.
URL: http://www.dcs.shef.ac.uk/~miguel/papers/iee-aspr99.html

ONE-TO-MANY MAPPINGS, CONTINUITY CONSTRAINTS AND LATENT VARIABLE MODELS

Miguel Á. Carreira-Perpiñán*

Abstract

We approach the problem of multivariate regression using latent variable models, which infer a low-dimensional representation of an observed, high-dimensional process. Defining functional relationships between variables may be conveniently done by picking informative points from the corresponding conditional distribution. However, this is problematic when this conditional distribution is multimodal, since there are in principle multiple candidates for the representative point, i.e., the mapping is one-to-many. We show, both with a toy example and with real-world data—the acoustic-to-articulatory mapping problem—that: 1) the modes of the conditional distribution contain information to potentially *invert* many-to-one as well as one-to-one mappings; 2) this information may be successfully used if some extra information is available, in particular continuity constraints for sequential data, for which we introduce a quantitative measure. We sketch algorithms for mode-finding in Gaussian mixtures and for performing smooth multivariate regression.

1 Introduction

Consider a system on which we can observe, or measure, several continuous variables, which we represent in vector form as $\mathbf{t} = (t_1, \dots, t_D) \in \mathbb{R}^D$. For example, the system could be a human speaker and $\mathbf{t}$ could be some representation of the acoustic and articulatory signals (at each moment in time), such as PLP coefficients and articulator positions. In many practical situations, we need to compute the values of several of these variables given values of other of the same variables. This is the problem of multivariate regression[1]. That is, we want to predict $\mathbf{t}_{\mathcal{J}} = \mathbf{f}(\mathbf{t}_{\mathcal{I}})$ where $\mathcal{I}, \mathcal{J} \in \{1, \dots, D\}$ are sets of indices and D is the dimensionality of the observed space. For example, if $\mathcal{I} = \{1, 7, 3\}$, then $\mathbf{t}_{\mathcal{I}} = (t_1 t_7 t_3)$. Therefore, we need to find a representation for the function $\mathbf{f}$. For simple cases with a low dimensionality D and where there is a mathematical model for the variables $t_1, \dots, t_D$ (say, a system of differential equations) this may be done analytically. But often the mathematical complications rule out this possibility and approaches that learn from data become necessary. Here, given a sample $\{\mathbf{t}_n\}_{n=1}^N$ of the observed variables, one estimates a (parametric) model, which can be deterministic (i.e., some powerful function approximator, such as multilayer perceptrons) or probabilistic. However, there is an additional difficulty: the mapping $\mathbf{f}$ may be one-to-many, in which case, given a value of $\mathbf{t}_{\mathcal{I}}$, the variables $\mathbf{t}_{\mathcal{J}} = \mathbf{f}(\mathbf{t}_{\mathcal{I}})$ may take several values, only one of which is realised at a given time. An example is the problem of the acoustic-to-articulatory mapping in speech. It is well known that, while given a time sequence of vocal tract configurations there is a unique output acoustic signal, the converse is not true: multiple vocal tract configurations can produce a given acoustic signal (Schroeter and Sondhi, 1994). One could imagine that there is an additional information $\Im$ that uniquely identifies the particular realisation $\mathbf{t}_{\mathcal{J}} = \mathbf{f}(\mathbf{t}_{\mathcal{I}}, \Im)$. Methods like multilayer perceptrons, oriented to estimating one-to-one mappings—in that they provide with a single, and same, value $\mathbf{t}_{\mathcal{J}}$ every time they are presented with a value $\mathbf{t}_{\mathcal{I}}$—, will usually give a compromise value of all the possible ones for $\mathbf{t}_{\mathcal{J}}$ and thus perform poorly[2]. Probabilistic methods, which can construct a conditional distribution $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$, can offer several values for $\mathbf{t}_{\mathcal{J}}$: a unimodal distribution indicates a one-to-one association while a multimodal one indicates a one-to-many association. Thus, we are potentially able to select the appropriate one if some additional information is available. In this work we use the temporal continuity of the signal to constrain the reconstructed values to give a trajectory as smooth as possible. Also, since the apparent high-dimensionality of the observed data is often due to noise, we use latent variable models to try to capture the low-dimensional, latent structure of the system.

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* $\mathbf{x} = (x_1, \dots, x_L)$ (Bartholomew, 1987). The latent variables are mapped by a fixed transformation into a D-dimensional data space and noise is added there. The aim is to learn the low dimensional generating process along with a noise model, rather than directly learning a dimensionality reducing mapping. Note that the low-

*Dept. of Computer Science, University of Sheffield, UK. Email: miguel@dcs.shef.ac.uk.

[1]Note that we do not deal here with the problem of predicting future values from past values, but with that of predicting in a given instant some variables given other variables.

[2]More precisely, one can easily prove that, in the limit of large samples and in the sense of the Euclidean norm, the best approximation using a one-to-one mapping is given by the conditional means: $\mathbf{f}(\mathbf{t}_{\mathcal{I}}) = \mathrm{E}\{\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}}\}$. This is the function that a universal approximator, such as a multilayer perceptron, will find (under certain conditions). The only improvement over this function comes from being able to assign different values to $\mathbf{t}_{\mathcal{J}}$ given the same $\mathbf{t}_{\mathcal{I}}$ at different times.

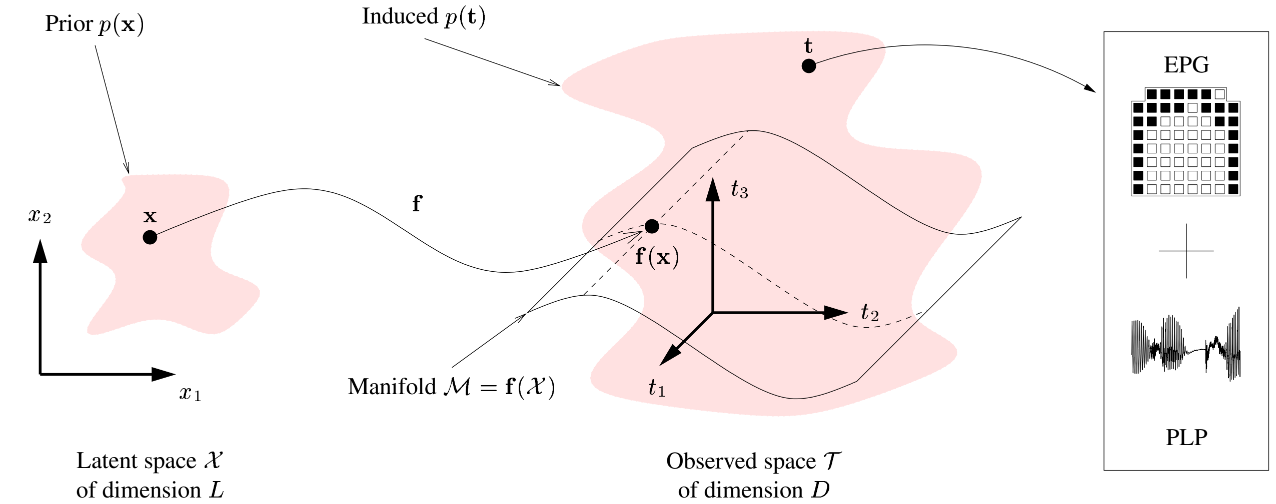

Figure 1: Schematic of a latent variable model where the observed data consists of EPG patterns and PLP coefficients.

dimensional representation is abstract and may not necessarily be interpretable in terms of any physical variables.

A latent variable model is specified by a prior distribution in latent space $p(\mathbf{x})$, a smooth mapping $\mathbf{f}$ from latent space to data space and a noise model in data space $p(\mathbf{t}|\mathbf{x})$. These three elements are equipped with parameters which we collectively call $\boldsymbol{\Theta}$. Integrating the joint probability density function $p(\mathbf{t}, \mathbf{x})$ over the latent space gives the marginal distribution in data space, $p(\mathbf{t})$. Figure 1 illustrates the idea. 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 the parameters $l(\boldsymbol{\Theta}) = \sum_{n=1}^N \log p(\mathbf{t}_n|\boldsymbol{\Theta})$, typically using an EM algorithm.

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

Factor analysis (Bartholomew, 1987), 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.

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.

3 Regression with latent variables

Once the latent variable model has been trained using data from the observed space, we have a probabilistic model $p(\mathbf{x}, \mathbf{t})$ for all the variables of interest. For simplicity of notation, we omit the dependence on the parameters and the model. Using the standard operations of marginalisation and conditioning, it is possible to obtain the distributions of any

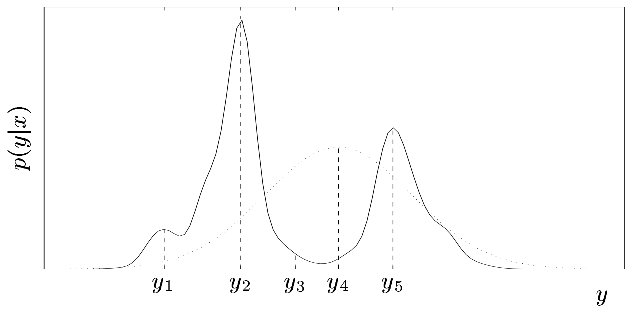

Figure 2: A unimodal (dotted line) and a multimodal conditional distribution (solid line). The vertical, dashed lines mark the modes and the means.

variable(s) with respect to any other variable(s). For example, to find the distribution in latent space, we compute the posterior distribution of the latent variables with respect to the observed ones,

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})},$$

which leads to dimensionality reduction and has been investigated for electropalatographic data in (Carreira-Perpiñán and Renals, 1998). Here, we construct conditional distributions of the form $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ where $\mathcal{I}, \mathcal{J} \in \{1, \dots, D\}$ are sets of indices and D is the dimensionality of the observed space. From a conditional distribution $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ it is possible to construct a functional relationship $\mathbf{t}_{\mathcal{J}} = \mathbf{f}(\mathbf{t}_{\mathcal{I}})$ provided that the entropy of this conditional distribution is low. That is, given $\mathbf{t}_{\mathcal{I}}$, only a small region of the space of $\mathbf{t}_{\mathcal{J}}$ should have nonnegligible probability mass: $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ is sharply peaked. As mentioned in section 1, to derive a functional relationship $y = f(x)$ from a conditional distribution $p(y|x)$, one can take a point that conveniently summarises the information contained in $p(y|x)$, e.g., the mean or the mode(s). If $p(y|x)$ is unimodal (like the dotted curve in fig. 2), the mean will usually be near the mode (value y_4). But if $p(y|x)$ is multimodal (like the solid curve in fig. 2), then each mode is potentially a valid solution (values y_1, y_2,

y_5), while the mean (value y_3) may be a misleading estimate if it lies in a low-probability area. For the latent variable models investigated here, the distribution $p(\mathbf{t})$ in observed space is either Gaussian (factor analysis) or a mixture of isotropic Gaussians (GTM), and so $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ is again Gaussian or a Gaussian mixture, respectively.

3.1 Exhaustive mode finding

The Gaussian case offers no problem as the mean coincides with the mode and the distribution is unimodal. For Gaussian mixtures, it is possible to find all the modes efficiently by using a maximisation algorithm starting from each centroid, such as gradient ascent combined with quadratic optimisation or a fixed-point iteration (Carreira-Perpiñán, 1999). Spurious modes may be discarded if their probability is lower than a suitable threshold; this accelerates the regression algorithm and may make it more robust. Additionally, it is possible to obtain error bars (i.e., a confidence interval) at each mode by locally approximating the density function by a normal distribution. However, if the dimensionality of $\mathbf{t}_{\mathcal{J}}$ is high, the error bars become very wide due to the curse of the dimensionality.

4 Using continuity constraints

Since the bare conditional distribution $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ is just an account of the relative proportions of the data independent of the time, we need some extra information at a given instant to decide which of the modes to choose. A property of most natural phenomena is their temporal continuity, i.e., the fact that the signals involved do not change abruptly, due to physical constraints. For example, for speech, the movement of the articulators (production organs) is relatively slow due to their inertia and to forces of friction, tension, etc. If the regression method provides with multiple choices at each time frame, such as GTM (as we saw in section 3), we can choose the mode that gives the smoothest reconstructed trajectory. That is, consider a sequence[3] of consecutive observations of vectors $\{\mathbf{t}_{\mathcal{I}}^{(n)}\}_{n=1}^{N}$. At time frame n, $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}}^{(n)})$ will have several modes, of which we will choose one; let us call it $\mathbf{t}_{\mathcal{J}}^{(n)}$. Then the reconstructed sequence will be $\{(\mathbf{t}_{\mathcal{I}}^{(n)}, \mathbf{t}_{\mathcal{J}}^{(n)})\}_{n=1}^{N}$. Now let us define a *smoothness measure* $\mathcal{L}$ for a polygonal trajectory $\{\mathbf{p}^{(n)}\}_{n=1}^{N} \subset \mathbb{R}^D$ as the sum of the Euclidean distances between consecutive points:

$$\mathcal{L}\left(\{\mathbf{p}^{(n)}\}_{n=1}^{N}\right) \stackrel{\text{def}}{=} \sum_{n=1}^{N-1} \delta\left(\mathbf{p}^{(n)}, \mathbf{p}^{(n+1)}\right)$$

for $\delta(\mathbf{u},\mathbf{v}) \stackrel{\text{def}}{=} \|\mathbf{u}-\mathbf{v}\|_2$ and $\|\mathbf{u}\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbf{u}^T\mathbf{u}}$. That is, $\mathcal{L}\left(\{\mathbf{p}^{(n)}\}_{n=1}^{N}\right)$ is the length of the trajectory. Then, from all the possible trajectories, we select the shortest one. This leads to the following minimisation problem:

$$\min_{\{\mathbf{t}_{\mathcal{J}}^{(n)}\}_{n=1}^{N}} \mathcal{L}\left(\{(\mathbf{t}_{\mathcal{I}}^{(n)}, \mathbf{t}_{\mathcal{J}}^{(n)})\}_{n=1}^{N}\right).$$

[3] In our notation, $\{\mathbf{t}^{(n)}\}_{n=1}^{N}$ means a sequence of temporally ordered vectors, while $\{\mathbf{t}_n\}_{n=1}^{N}$ means an arbitrary collection of vectors, not necessarily consecutive in time.

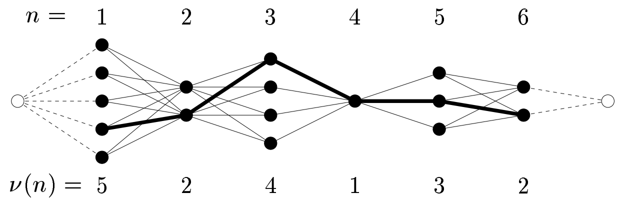

Figure 3: A layered graph. The layers are represented as vertical columns of nodes, each node being a point in D dimensional space. An edge between two nodes is labelled with a cost equal to the Euclidean distance between the nodes; the white nodes in the left and right ends are fictitious and have zero-cost edges. The shortest path between the end nodes is the shortest trajectory passing through all layers (such as the one in thick line).

4.1 Smoothest trajectory search

Call $\nu(n)$ the number of possible choices at time step n, i.e., the number of modes of $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}}^{(n)})$. Then, the search space contains $\prod_{n=1}^{N} \nu(n)$ different trajectories, and thus depends exponentially on N. However, there are efficient algorithms to search this space. The problem can be formulated as finding the shortest path in a layered graph (fig. 3) between the leftmost and the rightmost nodes.

A full search can be carried out by a divide-and-conquer algorithm:

- select the centre layer (i.e., for $n = \frac{N}{2}$), with $\nu\left(\frac{N}{2}\right)$ nodes;
- for each one of its nodes, solve recursively the layered graphs to its left and to its right (which contain approximately $\frac{N}{2}$ layers each) and join the resulting shortest sub-paths with the node;
- of the $\nu\left(\frac{N}{2}\right)$ solutions computed, choose the shortest one.

If all the layers have the same number of nodes ν, this algorithm has a polynomial complexity of $\mathcal{O}(N^{1+\log_2 \nu})$ approximately.

A heuristic search can be carried out by a greedy algorithm:

- select any layer n, with $\nu(n)$ nodes (ideally $\nu(n) = 1$);
- work backwards (from n down to 1) and forwards (from n to N) picking at each new layer the closest node to the current one;
- of the $\nu(n)$ solutions computed, choose the shortest one.

This algorithm has linear complexity, $\mathcal{O}(N)$, and is thus very fast, but it is sensitive to the starting layer and can perform poorly when it only finds a suboptimal solution, as the experiments below show.

At the time of writing this paper, only the greedy algorithm was implemented. However, the experiments below show that the desired solutions present a smaller value of $\mathcal{L}$, which suggests that the full search may find a solution close to the desired one.

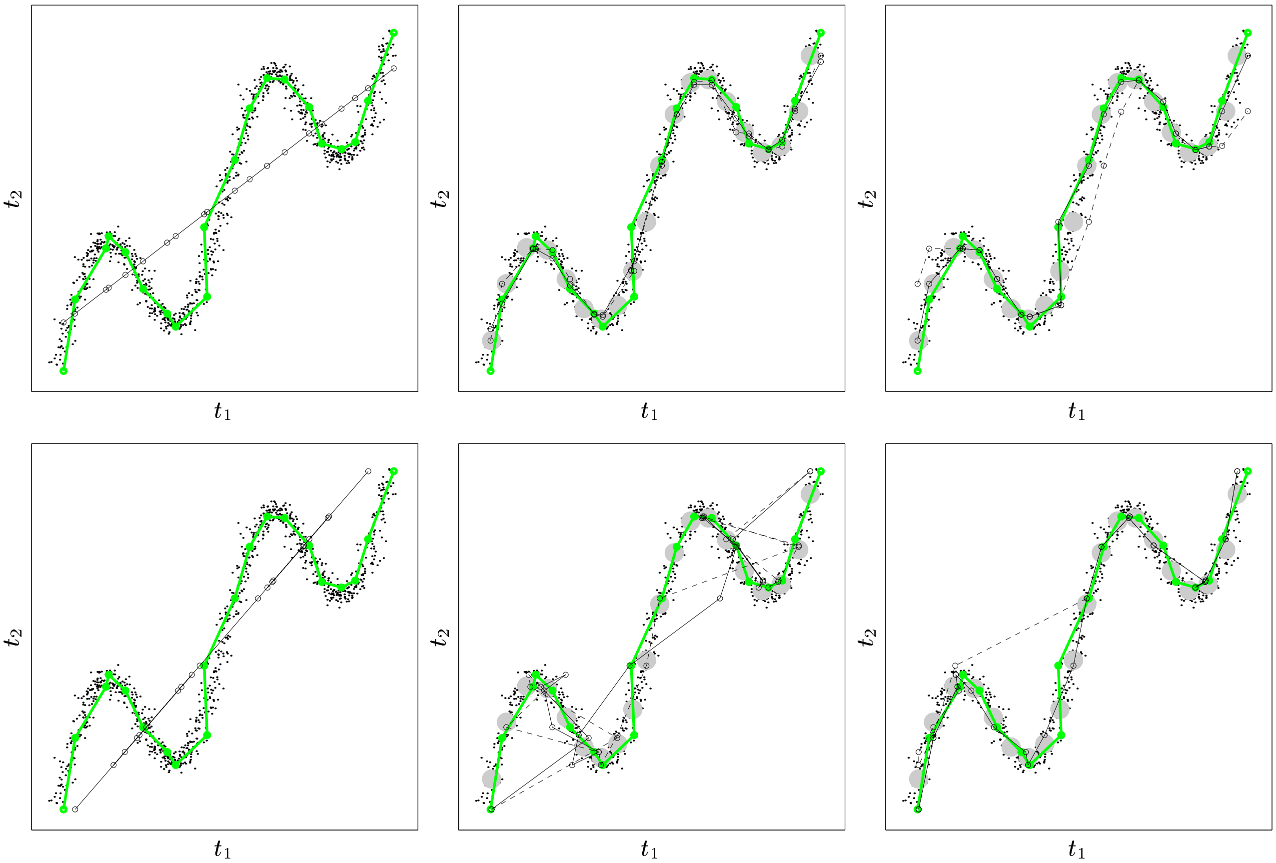

Figure 4: Trajectory reconstruction for a 2D problem, where the observed data fall in the curve $(t_1, t_2) = (x, x + 3\sin(x))$ for $x \in [-2\pi, 2\pi]$, with normal isotropic noise added. For all graphs, the thick line indicates the true trajectory, the dots the training data and the circles the Gaussian components for the GTM model. Top row: t_2 given t_1. Bottom row: t_1 given t_2. Left column: factor analysis (solid line). Centre column: GTM with conditional mean (solid line) and conditional global mode (dashed line). Right column: GTM with conditional closest mode (solid line) and conditional modes found with the greedy algorithm (dashed line).

5 Experiments with a toy data set

Using a nonlinear data set in an observed space of dimension $D = 2$, we trained a factor analysis model and a GTM model, both with a one-dimensional latent space. For GTM we used a grid of 20 latent points, which gives a Gaussian mixture of 20 components in observed space. Figure 4 shows the results of reconstructing a two-dimensional trajectory, i.e., t_2 as a function of t_1 (which is easy, since the mapping is one-to-one) and t_1 as a function of t_2 (which is difficult, since the mapping is one-to-many).

Due to the nonlinear character of the data set, factor analysis performs poorly: $p(t_2|t_1)$ and $p(t_1|t_2)$ provide linear mappings passing through the global mean of the data set.

GTM approximates much better the data distribution. To transform a conditional distribution into a mapping, we try four different strategies: the mean; the global mode, i.e., the mode with highest probability; the modes selected using the greedy algorithm of section 4; and the *closest* mode, which is the mode closest to the true target value. This latter case gives a lower bound in the reconstruction error, and cannot be achieved in general, since the target values are unknown in a real situation. In the $t_1 \to t_2$ case the conditional distribution $p(t_2|t_1)$ is unimodal for all values of t_1 and all methods (either mean- or mode-based) perform approximately equally well. However, in the $t_2 \to t_1$ case the conditional distribution $p(t_1|t_2)$ is multimodal for a number of values of t_2, and the only method that adequately recovers the trajectory is the *closest* modes one—the other strategies producing very jagged trajectories. This shows that the conditional distribution has actually captured the information about the correct predicted values. We note that the trajectory length $\mathcal{L}$ for each method is 29.4, 39.5, 22.2 and 19.2, in the order mentioned above. Thus, the *closest* modes provide with the shortest trajectory, which suggests that a global search would find a trajectory similar to the desired one.

6 Real-world problem: prediction of PLP coefficients and EPG patterns

To demonstrate the potential ability of the method described for regression in a real-world problem, we trained latent variable models with both acoustic and articulatory speech data and computed conditional distributions of the acoustic

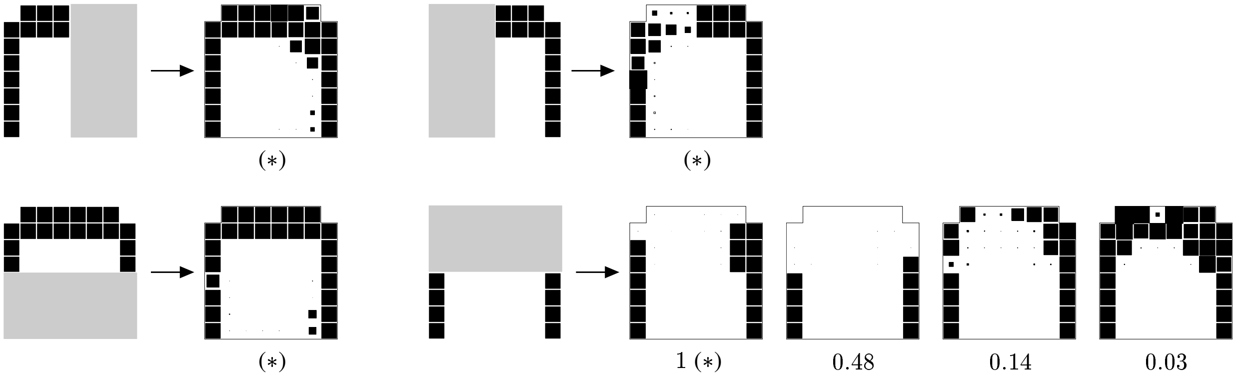

Figure 5: Use of the conditional distribution modes to predict, or reconstruct, variables in observed space. Here, we use the GTM model to compute the distribution of the EPG part greyed out (the unknown values) conditional on the EPG part which is not greyed out (the known values). The modes are given to the right of the arrow, labelled with their normalised probability if there is more than one mode. In all four cases, the mean (marked $*$) coincided approximately with one of the modes. Note the left-right asymmetry. As is customary in the electropalatography literature, the EPG vectors are pictured rowwise from top to bottom resembling the human palate (top: alveoli; bottom: velum).

Data set	EPG pattern given the PLP coefficients					PLP coefficients given the EPG pattern				
	Factor analysis	GTM				Factor analysis	GTM			
		mean	g-mode	gr-mode	c-mode		mean	g-mode	gr-mode	c-mode
Training	3.7635	2.2736	2.8681	3.6111	0.9462	0.8870	0.5777	0.6221	0.7435	0.4206
Test	3.5060	2.7667	3.5012	4.3522	1.4809	0.8632	0.7967	0.9061	0.8436	0.6102
Utterance	2.6172	1.4398	1.7046	1.6785	0.8061	0.6723	0.7778	0.8103	0.6228	0.5865

Table 1: Average quadratic reconstruction error of the EPG patterns given the PLP coefficients and vice versa.

variables given the articulatory ones and vice versa. Our articulatory variables here are electropalatographic (EPG) frames, rather than the positions of the different articulators, due to the unavailability of a more appropriate data set. Each EPG frame is a binary pattern indicating the presence or absence of contact between the tongue and the hard palate in a fixed set of locations in the palate. In this case, the mapping EPG $\rightarrow$ acoustic signal is one-to-many, since different phonemes may correspond to the same EPG frame.

The data were obtained from the ACCOR database (Marchal and Hardcastle, 1993). This database, designed for the cross-language study of coarticulation, contains electropalatographic and acoustic measurements (among other measurements) for utterances in different European languages and varying speech styles (slow, fast, etc.). We selected the utterance "Put your hat on the hatrack and your coat in the cupboard" for speaker FG and computed from its acoustic waveform 12th-order PLP coefficients (Hermansky, 1990) plus the log-energy, all at 200 Hz. The EPG data consists of 62-bit frames sampled at 200 Hz, which we consider as 62-dimensional vectors of real numbers, with components indexed from 1 (top left) to 62 (bottom right). Thus, the resulting sequence consisted of over 600 75-dimensional real vectors. We constructed a training set by picking, in random order, 80% of these vectors, and a test set with the remaining 20%. Thus, these two sets have lost the temporal continuity present in the original utterance. We constructed an additional set with temporal continuity by selecting 100 consecutive frames from the utterance. All the data used were unlabelled.

The models trained were factor analysis with 9 factors (= dimensionality of latent space) and the generative topographic mapping (GTM) with a latent space of dimension 2 and a 20×20 grid.

In fig. 5 we used GTM to reconstruct parts of the EPG frame given other parts of it. Note how the reconstructed pattern is slightly different when the left half is given than when the right half is given, revealing asymmetry in the tongue movement. When the bottom half is given, the distribution for the top half happens to be multimodal, with several patterns (corresponding to open vowels and alveolars) becoming possible.

Table 1 shows the results for the average quadratic reconstruction error of the EPG pattern given the PLP coefficients and vice versa, defined as $E_2 = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{\mathbf{t}}_n - \mathbf{t}_n\right\|_2^2$, where $\mathbf{t}_n$ is the true vector, $\hat{\mathbf{t}}_n$ the reconstructed one and N the total number of vectors in the set under consideration. For the linear-normal model (factor analysis), the conditional distribution is always normal and so the only point to consider to reconstruct the vector is the conditional mean (equal to the conditional mode). As with the toy data set, we tried four possibilities for GTM: the conditional mean; the global conditional mode (g-mode); the modes selected using the greedy algorithm of section 4 (gr-mode); and the conditional mode closest to the vector to be reconstructed (c-mode).

Regarding method comparison, the table shows that GTM attains a smaller error than factor analysis in almost all cases. Thus, even though GTM assumes an isotropic noise model (and our variables have different ranges and variances, e.g. the EPG variables are in $[0, 1]$ while the log-energy is in $[1, 5]$ approximately) and it uses a latent space of dimension 2, its nonlinear mapping compensates enough to outperform a linear method using a diagonal noise model and a latent space of dimension 9.

Regarding the use of the mean or a mode, the average reconstruction errors for GTM show that, in all cases, the mean performs better than the global mode (in agreement with the theory) but worse than the closest mode. We employed an additional strategy, not shown in the table, where the mean is used if the conditional distribution is unimodal and the global mode if it is multimodal. This gave an error virtually equal to that of the global mode, which indicates that the divergence occurs in multimodal conditional distributions. Thus, the best predictor is one of the modes, but not necessarily the mode with the highest probability. Looking now at the results using continuity constraints with the greedy algorithm, we see a worse performance over that of the mean or the global mode in the training and test sets. This is reasonable since we deliberately eliminated the continuity from these sets. For the utterance fragment, in the case of the regression EPG $\rightarrow$ PLP we observe a performance improvement over that of the mean, but not in the case PLP $\rightarrow$ EPG. This suggests that this suboptimal algorithm may be able to take advantage of the continuity of the data in some cases. However, since—as in the toy data set—the length of the *closest* modes trajectory was the smallest one, it is likely that a global search could improve the reconstruction performance over that of the other methods, approaching the optimal one as bounded by that of the *closest* modes.

7 Discussion

Our claim is that the combination of:

- latent variable models, which can capture the low-dimensional structure of the data in a stochastic way;
- the probabilistic nature of the model, which allows to compute in practice several candidates for predicting the values of some variable(s) given other variable(s);
- the use of continuity constraints to select those candidates that give the smoothest reconstructed trajectory in observed space;

provides with an approach to the difficult problem of inverting many-to-one mappings. Regarding its practical implementation, we have also sketched algorithms for computing all the modes of a Gaussian mixture and for enforcing continuity by minimising the trajectory length in observed space. Our experimental results confirm that the information about the *potentially* correct values may be captured by the probabilistic model and that the continuity constraint—in those cases where it is applicable—may help to recover the values that are *actually* correct at a given time. A heuristic search for smooth trajectories was seen to find suboptimal solutions. Thus, more work is necessary to determine whether a global search for the smoothest trajectory guarantees an improvement over the conditional mean regression.

Some issues that require further investigation include:

- A poor probabilistic model may give rise to conditional distributions with too many or too few modes. Thus, it is necessary to test how robust the approach is in such a situation.
- Depending on the data distribution, a set $\{\mathbf{t}_{\mathcal{I}}^{(n)}\}_{n=1}^{N}$ may contain too few points to uniquely determine a smooth trajectory.
- Comparison with standard function approximators, such as multilayer perceptrons.

A potential problem of latent variable models is that the dimension of the latent space has to be fixed in advance, although an optimal one could be found by model selection. An additional problem of methods that sample the latent space, like GTM, is that their computational cost grows exponentially with the dimension of the latent space.

Finally, a regression can be seen as missing data imputation, where given the present values $\mathbf{t}_{\mathcal{I}}$, the missing values $\mathbf{t}_{\mathcal{J}}$ are to be filled in using the knowledge of the distribution $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$. Thus, the same formalism applies to missing data imputation. Note that $\mathcal{I}$ and $\mathcal{J}$ may be different for each point in the data set to be reconstructed.

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. (1999). Mode-finding in Gaussian mixtures. Technical Report CS–99–03, Dept. of Computer Science, University of Sheffield.

Carreira-Perpiñán, M. Á. and Renals, S. (1998). Dimensionality reduction of electropalatographic data using latent variable models. *Speech Communication*, 26(4):259–282.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *J. Acoustic Soc. Amer.*, 87(4):1738–1752.

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.

Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech and Audio Process.*, 2(1):133–150.