One-to-many mappings, continuity constraints and latent variable models

Miguel Á. Carreira-Perpiñán

Dept. of Computer Science, University of Sheffield

M.Carreira@dcs.shef.ac.uk

IEE Colloquium on Applied Statistical Pattern Recognition, 20 April 1999, Birmingham, UK

Notation

Vectors are boldface lowercase, scalars are italics lowercase, sets are calligraphic uppercase.

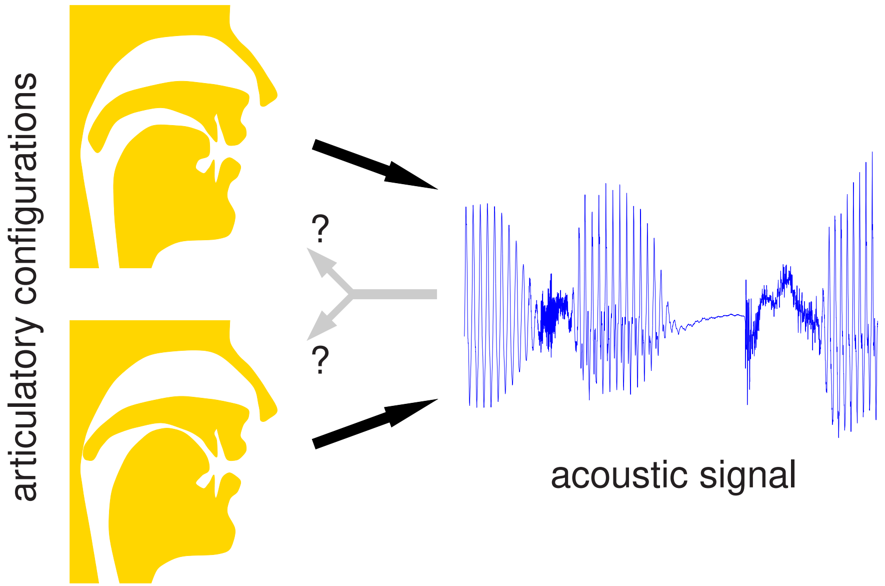
Symbol	Meaning
$\mathbf{t} = (t_1, \dots, t_D) \in \mathcal{T} \subset \mathbb{R}^D$	D-dimensional vector in observed space $\mathcal{T}$
$\mathbf{x} = (x_1, \dots, x_L) \in \mathcal{X} \subset \mathbb{R}^L$	L-dimensional vector in latent space $\mathcal{X}$
$\mathcal{I}, \mathcal{J} \in \{1, \dots, D\}$	Sets of indices for variable selection. Example: if $\mathcal{I} = \{1, 7, 3\}$ and $\mathcal{J} = \{2, 5\}$ then: $\mathbf{t}_\mathcal{I}$ is (t_1, t_7, t_3) $\mathbf{t}_\mathcal{J}$ is (t_2, t_5) $p(\mathbf{t}_\mathcal{I})$ is $p(t_1, t_3, t_7)$ $p(\mathbf{t}_\mathcal{J} \vert \mathbf{t}_\mathcal{I})$ is $p(t_2, t_5 \vert t_1, t_3, t_7)$ $p(\mathbf{t}_\mathcal{I}, \mathbf{t}_\mathcal{J})$ is $p(t_1, t_2, t_3, t_5, t_7)$
$\mathbf{t}_\mathcal{I} \to \mathbf{t}_\mathcal{J}$	$\mathbf{t}_\mathcal{J}$ as a function of $\mathbf{t}_\mathcal{I}$
$\{\mathbf{t}_n\}_{n=1}^N$	Set of N vectors not necessarily in sequence
$\{\mathbf{t}^{(n)}\}_{n=1}^N$	Set of N vectors in (temporal) sequence

Multivariate regression

- Multivariate regression is the problem of establishing a functional relationship or mapping (possibly nonlinear) between two groups of (real) variables: $\mathbf{x} \rightarrow \mathbf{y}$ (read: given $\mathbf{x}$, associate $\mathbf{y}$ to it).
- For our purposes, we classify mappings into:
  - One-to-one: given $\mathbf{x}$ there is a unique value of $\mathbf{y}$, for every $\mathbf{x}$ (e.g. $y = x^2$).
  - One-to-many: given $\mathbf{x}$ there may be several values of $\mathbf{y}$, for some $\mathbf{x}$ (e.g. $y = \pm\sqrt{x}$).
- Automatic learning of a mapping from data is relatively easy via linear regression, neural nets, generalised linear models, etc.
- But how to deal with the nonuniqueness inherent to one-to-many mappings?

We present an approach based in probabilistic models (namely, latent variable models) and continuity constraints.

The acoustic-to-articulatory mapping problem



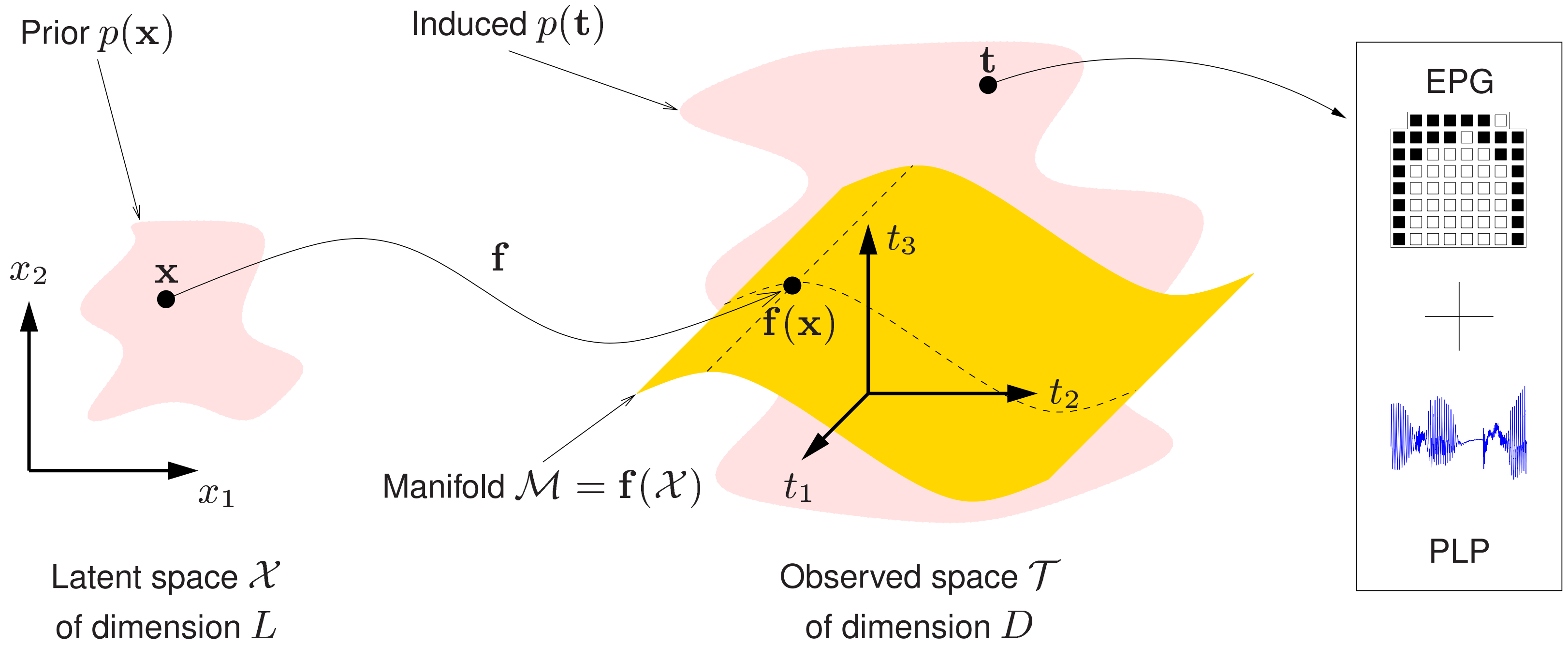
- ▷ A vocal tract configuration produces a unique acoustic signal. Thus the mapping articulatory $\longrightarrow$ acoustic is one-to-one.
- ▷ An acoustic signal can be produced by different vocal tract configurations. Thus the mapping acoustic $\longrightarrow$ articulatory is one-to-many.

Practical problems:

- ▷ Given an articulatory vector find the acoustic vector: easy.
- ▷ Given an acoustic vector find the articulatory vector: hard.

Latent variable models

The existence of functional relationships between the observed variables implies that $t_1, \dots, t_D$ live around a low-dimensional manifold $\mathcal{M}$ in observed space. Thus, it is convenient to model the data with a **latent variable model**, whose aim is to infer a low-dimensional representation of an observed, high-dimensional process.



Latent variable models (cont.)

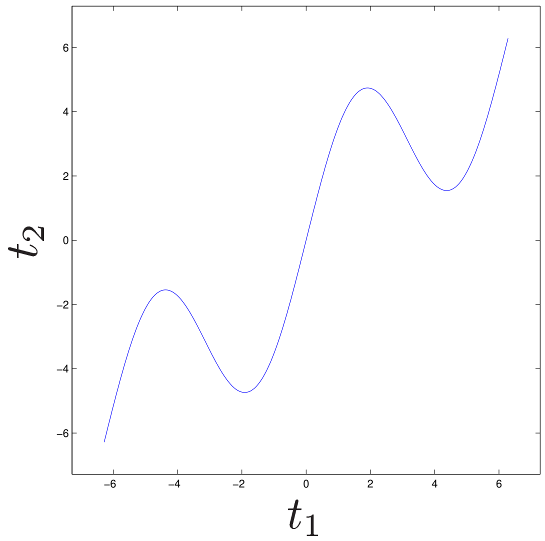
- The data distribution in observed space $\mathcal{T}$ is modelled using a low-dimensional representation in latent space $\mathcal{X}$. In the figure the observed space consists of 62-dimensional EPG patterns plus 13-dimensional PLP coefficients, thus $D = 75$.
- The prior in latent space $p(\mathbf{x})$, the mapping $\mathbf{f}(\mathbf{x})$ and the noise model in observed space $p(\mathbf{t}|\mathbf{x})$ are equipped with parameters $\boldsymbol{\Theta}$.
- Marginalisation in latent space: $p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$.
- Maximum likelihood parameter estimation from sample $\{\mathbf{t}_n\}_{n=1}^N$: $l(\boldsymbol{\Theta}) = \sum_{n=1}^N \log p(\mathbf{t}_n|\boldsymbol{\Theta})$.

Examples of latent variable models (all of which can be trained via an EM algorithm):

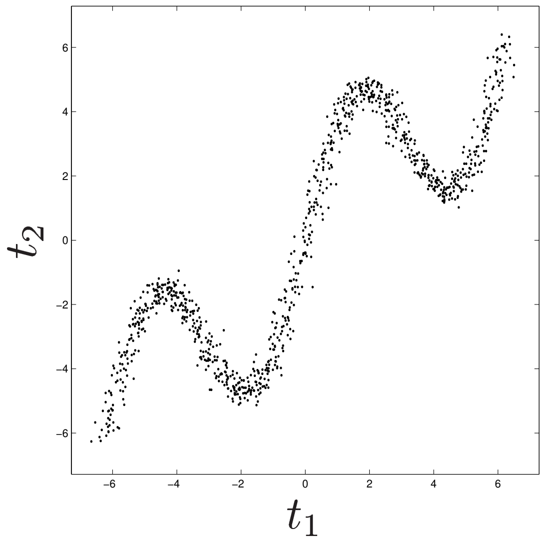
- *Factor analysis*: prior is normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$, mapping is linear, noise model is normal with diagonal covariance matrix.
- *Generative topographic mapping (GTM)*: prior is uniform over discrete latent grid, mapping is a generalised linear model, noise model is normal with isotropic covariance matrix.

Deriving mappings from conditional distributions

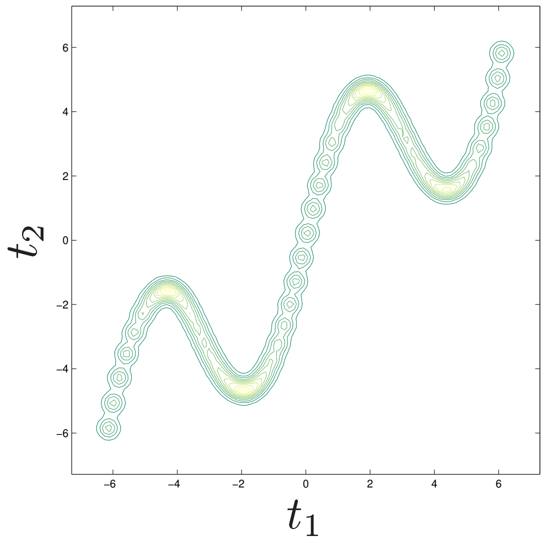
Consider one-dimensional data ($L = 1$) observed in $D = 2$ dimensions, i.e., $\mathbf{t} = (t_1, t_2)$:



Curve $\mathbf{t} = (x, x + 3\sin x)$ for $x \in [-2\pi, 2\pi]$.



Point cloud $\{\mathbf{t}_N\}_{n=1}^{N}$, sampled from the curve with additive $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ noise.

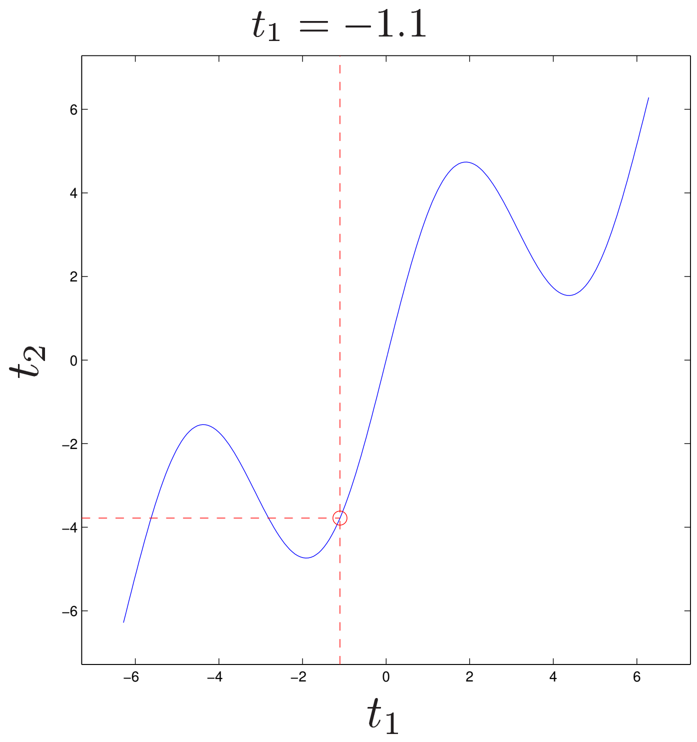


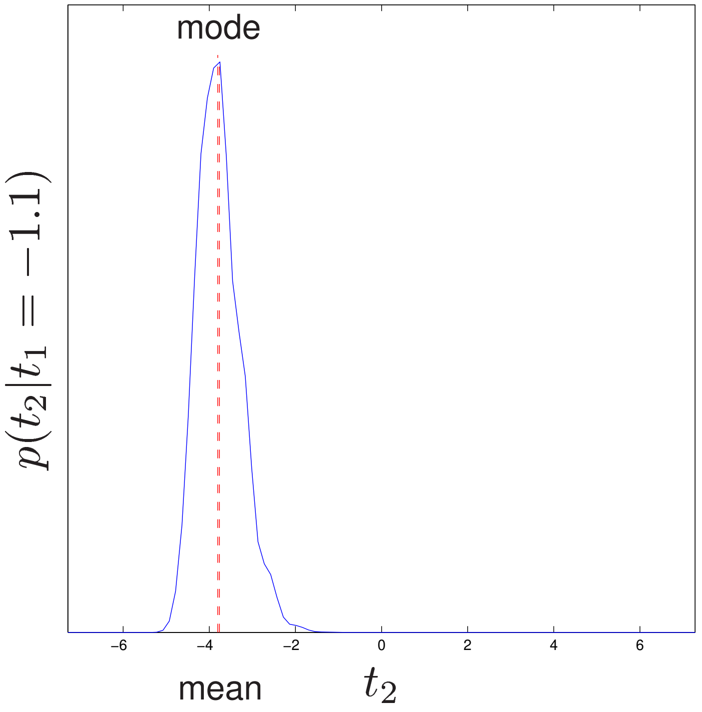
Density model for the distribution $p(\mathbf{t})$ obtained from the sample (contour plot).

Thus $t_2 = f(t_1) = t_1 + 3\sin t_1$.

Deriving mappings from conditional distributions (cont.)

If the mapping $t_2 = f(t_1)$ is unknown, we can use $p(t_2|t_1)$ to derive it:

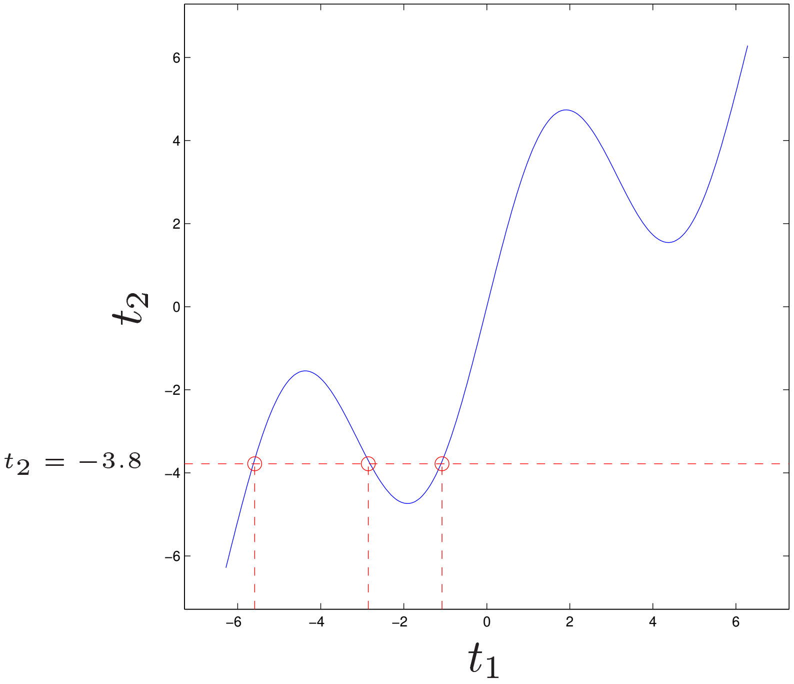


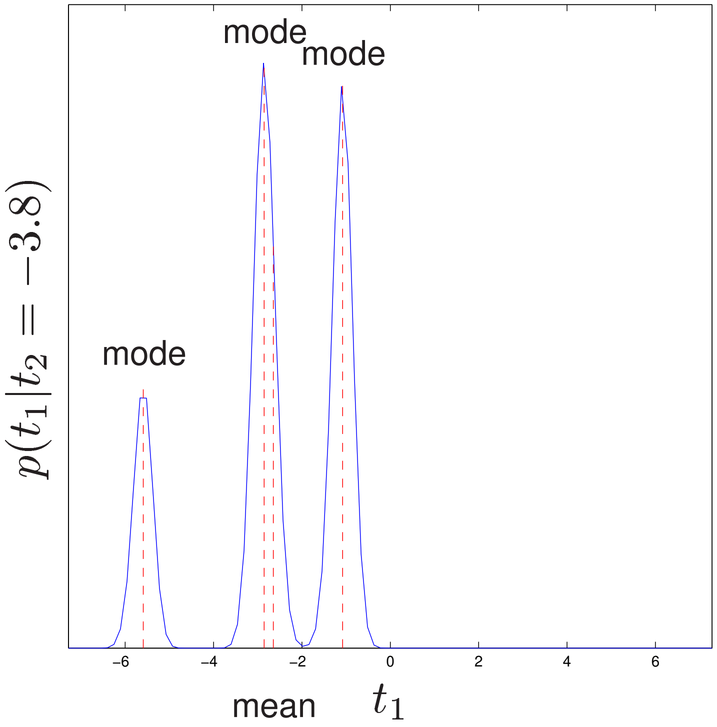


Unimodal distribution: both the mean and the mode are good approximations to the real value.

Deriving mappings from conditional distributions (cont.)

Now for $p(t_1|t_2)$:

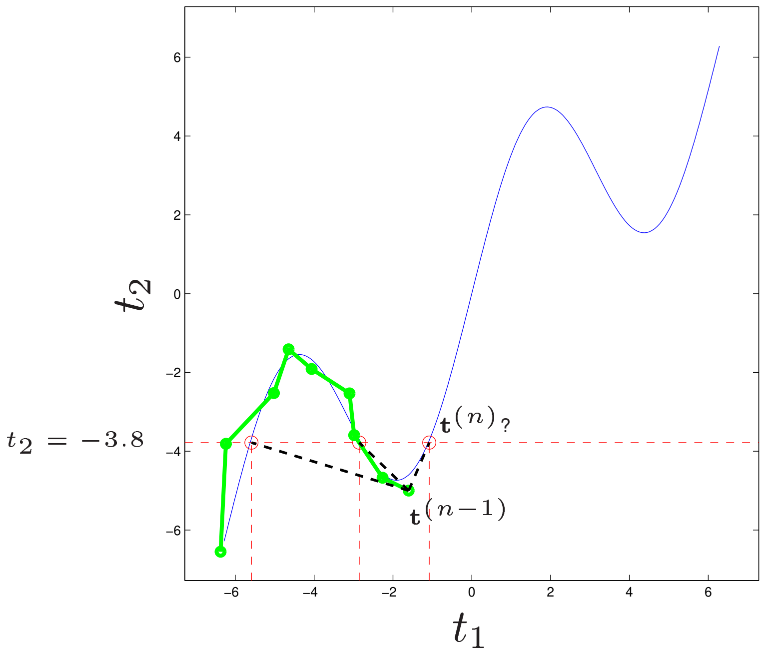




Multimodal distribution: one of the modes is the right one, but which one? We need extra information.

Continuity constraints

Consider a continuous trajectory (a curve) in observed space, sampled at times $\tau^{(1)} < \tau^{(2)} < \cdots < \tau^{(N)}$: $\{\mathbf{t}^{(n)}\}_{n=1}^{N}$ are N points in data space. By the very definition of continuity, $\mathbf{t}^{(n)}$ should be *close* to $\mathbf{t}^{(n-1)}$ and $\mathbf{t}^{(n+1)}$ (if $\tau^{(n-1)}$ and $\tau^{(n+1)}$ are small enough). We can use this to predict, at each time $\tau^{(n)}$, the values of $\mathbf{t}_{\mathcal{J}}$ given those of $\mathbf{t}_{\mathcal{I}}^{(n)}$.



Reconsider the case of the previous slide with $t_2^{(n)} = -3.8$. If we knew that $(t_1^{(n-1)}, t_2^{(n-1)}) = (-1.6, -5)$, then the most reasonable possibility would be $t_1^{(n)} = -1.1$, i.e., the closest one to $t^{(n-1)}$, rather than $t_1^{(n)} = -2.9$ or -5.6.

More generally: of all the possible combinations (of reconstructed trajectories), pick the smoothest one.

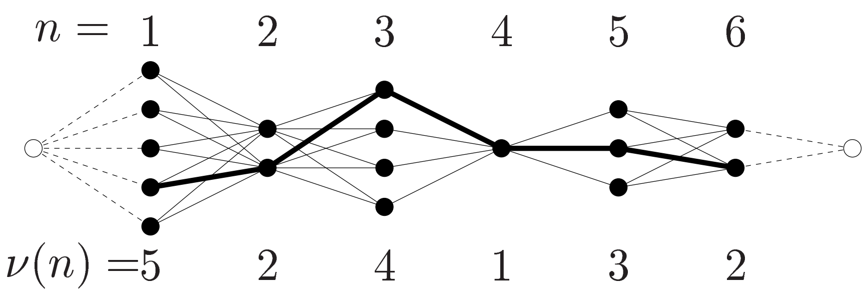
Search for the smoothest trajectory

Define the **smoothness** of a polygonal trajectory $\{\mathbf{t}^{(n)}\}_{n=1}^{N}$ as

$$\mathcal{L}\left(\{\mathbf{t}^{(n)}\}_{n=1}^{N}\right) = \sum_{n=1}^{N} \left\|\mathbf{t}^{(n+1)} - \mathbf{t}^{(n)}\right\|$$

where $\|\cdot\|$ is the Euclidean distance, or some other convenient distance.

Suppose we are given the part $\mathcal{I}$ of a continuous trajectory, $\{\mathbf{t}_{\mathcal{I}}^{(n)}\}_{n=1}^{N}$, and we want to reconstruct it, i.e., to predict the value of $\mathbf{t}_{\mathcal{J}}^{(n)}$ at each n. Call $\nu(n)$ the number of modes of $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}}^{(n)})$ at each n. This gives a **search space** $\mathcal{S}$ containing $\prod_{n=1}^{N} \nu(n)$ trajectories. The optimisation problem is $\min_{\text{traj}\in\mathcal{S}} \mathcal{L}(\text{traj})$, which is equivalent to finding the shortest path in a layered graph.



Search for the smoothest trajectory (cont.)

▷ **Local** search by a *greedy* algorithm:

1. select any layer n, with $\nu(n)$ nodes (ideally $\nu(n) = 1$);
2. 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;
3. of the $\nu(n)$ solutions computed, choose the shortest one.

This heuristic is fast, but prone to finding suboptimal solutions.

▷ **Global** search by *dynamic programming*, based on the optimality principle: *regardless of the policy adopted in previous layers, the remaining decisions must constitute an optimal policy*.

This is (slightly) slower, visiting all links in the graph exactly once.

NB: in this work, only the local search was implemented.

Exhaustive mode finding

We want to find all the modes of $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$. We consider two forms of p of practical interest:

▷ **Gaussian**: trivial (the mode coincides with the mean).

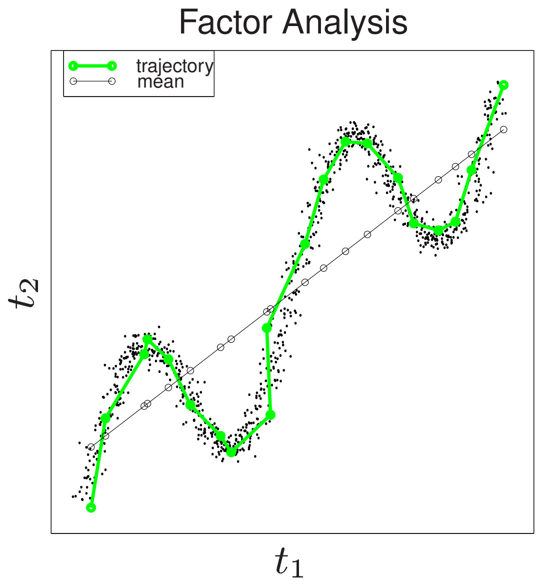
▷ **Mixture of Gaussians**: starting from each centroid, use an iterative algorithm to locate stationary points (where the gradient vanishes); then remove minima, saddle points and coincident maxima. Several algorithms are available:

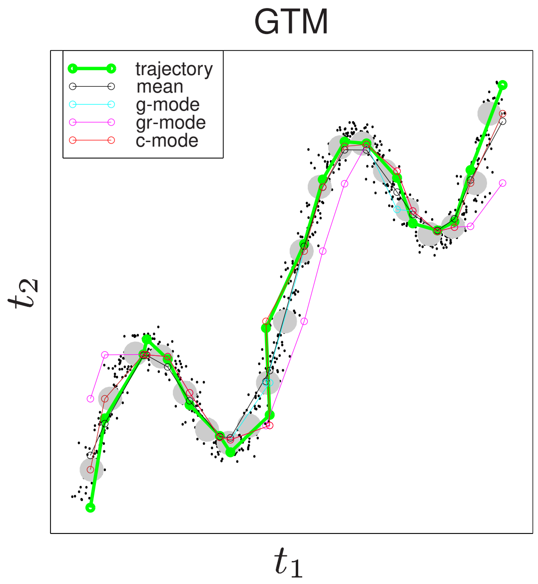
- Gradient ascent combined with quadratic optimisation (the gradient and the Hessian can be computed analytically).
- Fixed-point iteration.

It is possible to compute *error bars* for each mode by locally approximating the density by a normal distribution (using the Hessian at the mode).

Toy experiment

The thick green line is a two-dimensional trajectory for the previous example. The figures show the reconstructed trajectories when only t_1 was given.

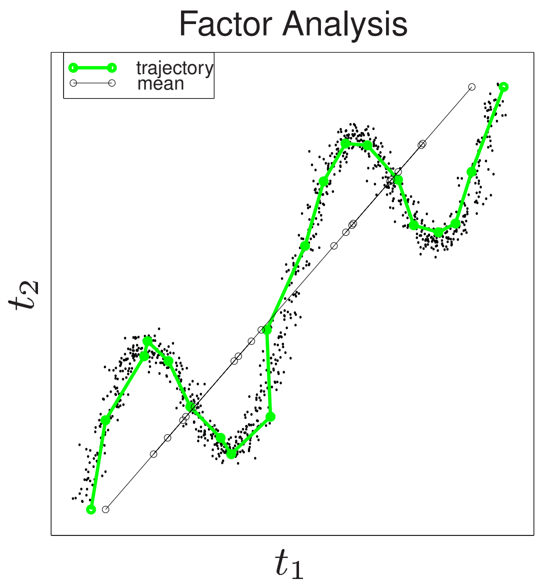


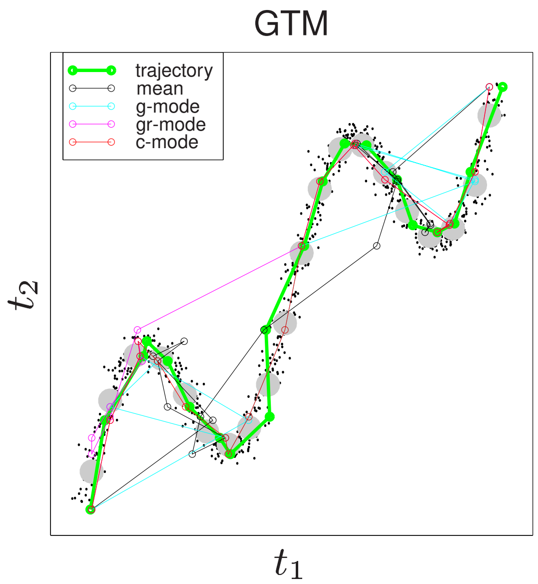


Method (for GTM)	Trajectory length $\mathcal{L}$
mean	25.2
g-mode	26.2
c-mode	26.5
gr-mode	23.1
full search	N/A

Toy experiment (cont.)

Now, the figures show the reconstructed trajectories when only t_2 was given.

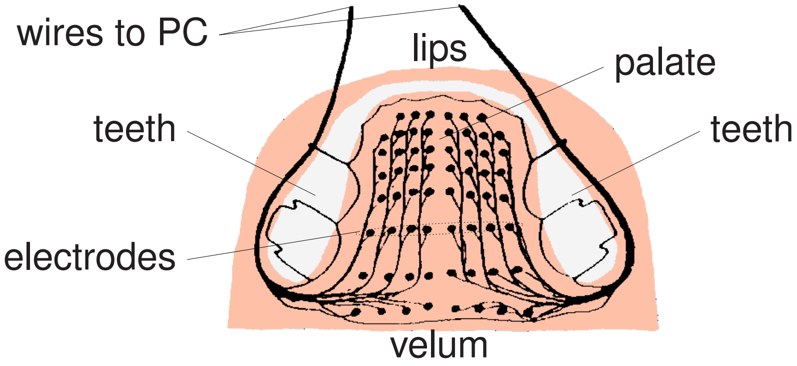


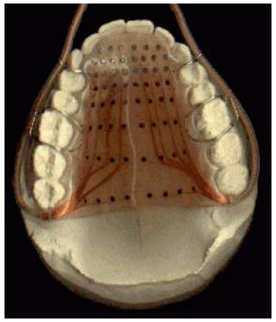


Method (for GTM)	Trajectory length $\mathcal{L}$
mean	35.9
g-mode	45.6
c-mode	29.2
gr-mode	31.4
full search	N/A

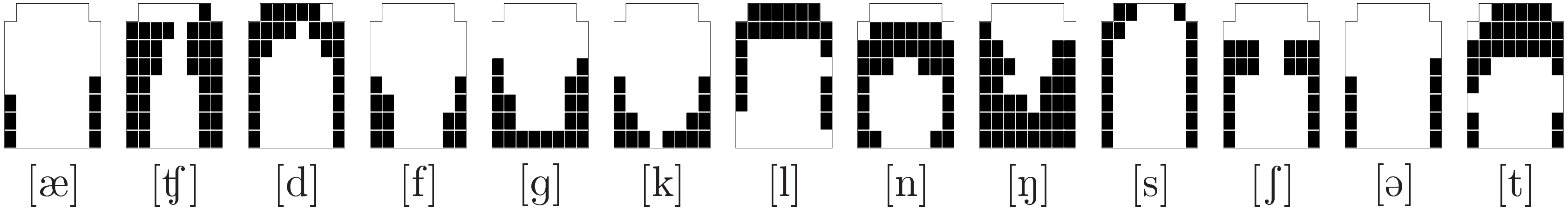
Electropalatography

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.





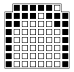
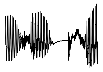
The Reading pseudopalate



Representative EPG frames for the typical stable phase of different phonemes, pictured rowwise from alveoli (top) to velum (bottom).

Prediction of PLP coefficients and EPG patterns

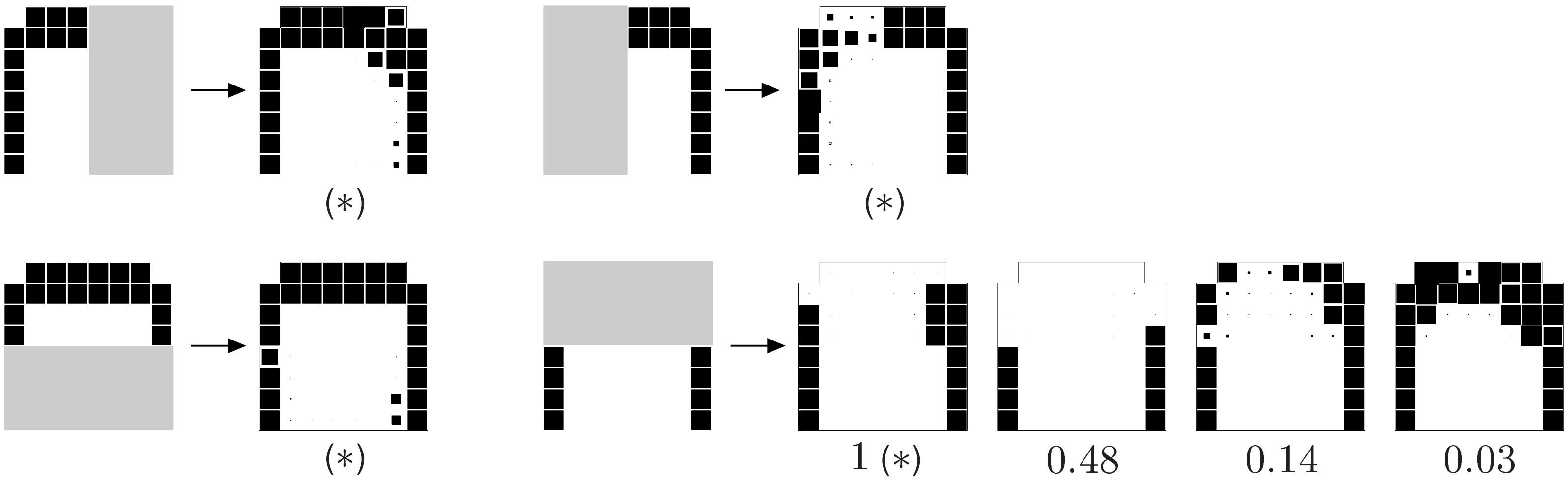
The dataset

- ▷ 62-dimensional EPG patterns  plus 13-dimensional PLP coefficients  sampled at 200 Hz. Both mappings, EPG $\rightarrow$ PLP and PLP $\rightarrow$ EPG, are one-to-many.
- ▷ ACCOR database utterance “Put your hat on the hatrack and your coat in the cupboard” for speaker FG: over 600 75-dimensional vectors. This data set was shuffled and split into a training set (TR) and a test set (TE), with 80% and 20% of the vectors, respectively. The shuffling destroys the continuity of the sequence.
- ▷ A third, continuous data set consisting of 100 consecutive vectors of the utterance.

The models

- ▷ Factor analysis with a latent space of dimension $L = 9$ (total: 825 parameters).
- ▷ GTM with a latent space of dimension $L = 2$, a 20×20 latent grid and a 7×7 RBF grid (total: 3751 parameters).

Reconstruction of single EPG frames



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.

Reconstruction error for the data sets

Data set	EPG pattern given the PLP coefficients					PLP coefficients given the EPG pattern				
	Factor analysis	GTM mean	GTM g-mode	GTM gr-mode	GTM c-mode	Factor analysis	GTM mean	GTM g-mode	GTM gr-mode	GTM 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

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

▷ **mean**: the conditional mean (the theoretically optimal one-to-one mapping).

▷ **g-mode**: the global mode, i.e., the mode with highest probability.

▷ **gr-mode**: the mode selected using the greedy algorithm.

▷ **c-mode**: the mode closest to the true target value (lower bound for any choice of modes).

The consistenly large improvement of the **c-mode** over the **mean** indicates that the modes contain information to achieve a low reconstruction error. It remains to be proven whether a full search over the space of smooth trajectories can extract that information.

Discussion

- ▷ We have proposed an approach to the difficult problem of inverting many-to-one mappings based in the combination of:
  - latent variable models, which can stochastically capture the low-dimensional structure of the data;
  - 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;
- ▷ The modes of the conditional distribution contain the *potentially* correct values. The continuity constraint—in those cases where it is applicable—may recover the *actually* correct ones.
- ▷ Problems of latent variable models:
  - The latent dimension L must be fixed in advance—could use model selection.
  - The complexity of GTM is $\mathcal{O}(e^L)$.
- ▷ Application to missing data imputation: no restriction for $\mathcal{I}$ and $\mathcal{J}$ to be fixed across points.

Further work

- ▷ Does a global search guarantee an improvement over the conditional mean regression?
- ▷ Use other smoothness measures, e.g. the curvature instead of the length.
- ▷ Compare with standard function approximators: MLPs, recurrent MLPs and MLP committees.
- ▷ What is the robustness against poor density models (e.g. a mixture with too few components).
- ▷ How to deal with . . .
  - Very short sequences? Unique reconstruction may not be possible.
  - Unbounded sequences ($N \to \infty$)? This may require assuming stationarity.

Papers and Matlab code available at:

`http://www.dcs.shef.ac.uk/~miguel`