On Topographic Maps/Clustering of Structured Data

Peter Tiño
University of Birmingham
Birmingham B15 2TT, UK
www.cs.bham.ac.uk/~pxt

Acknowledgements:
Ata Kabán, Yi Sun, Nick Gianniotis, Steve Spreckley
**Vector quantization**

Take advantage of the cluster structure in the data $\mathbf{x}_i$, $i = 1, 2, \ldots, N$. To minimize the representation error, place the representatives $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M$, known as codebook vectors, in the center of each cluster.

- To transmit $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \ldots$, first transmit full information about the codebook $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$.
- Then instead of each point $\mathbf{x}_i$, transmit just the index of its closest representative codebook vector.
You can discover a hidden 1-dimensional structure of high-dimensional points by running a VQ on them, but constrain the codebook vectors \( \mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^M \) to lie on a one-dimensional ‘bicycle chain’.

‘bicycle chain’ represents the Channel noise
Two-dimensional grid of codebook vectors

Generalize the notion of ‘bicycle chain’ of codebook vectors: Take advantage of two-dimensional structure of the computer screen. Cover it with a 2-dimensional grid of nodes.
Constrained VQ - Placing the codebook vectors

1. Randomly place codebook vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M$ in $\mathbb{R}^n$.

2. Cycle through the set of data points and for each point $\mathbf{x}^i$ do:
   (a) Find the closest codebook vector $\mathbf{b}_{\text{win}(i)}$.
   (b) Move $\mathbf{b}_{\text{win}(i)}$ a bit closer to $\mathbf{x}^i$:
      
      $$
      \mathbf{b}_{\text{new}} = \mathbf{b}_{\text{old}} + \eta \cdot (\mathbf{x}^i - \mathbf{b}_{\text{old}}).
      $$

   (c) Push towards $\mathbf{x}^i$ also the codebook vectors $\mathbf{b}^j$ that are neighbors of $\mathbf{b}_{\text{win}(i)}$ on the bicycle chain (1-dimensional grid of codebook vectors).

   For each codebook vector $\mathbf{b}^j$, $j = 1, 2, \ldots, M$,

   $$
   \mathbf{b}^j_{\text{new}} = \mathbf{b}^j_{\text{old}} + \alpha \cdot \eta \cdot (\mathbf{x}^i - \mathbf{b}^j_{\text{old}}).
   $$
Structured data?

For vectorial data of fixed dimension constrained VQ is well-formulated - we have a metric (and hence a notion of Loss) in the data space.

What about topographic maps of sequences (EEG, DNA, documents etc.), or graphs (molecules etc.)?

Suggestions:

- Represent data through vectors of fixed dimension, then do the usual stuff.
- Add recursive feed-back connections to the usual models to allow for natural representation of recursive data types
- Model-driven topographic map construction
Recursive Self-Organizing Map - RecSOM

Circular argument: induced metric $\leftrightarrow$ topographic map of data!
Contractive dynamics

When the fixed-input dynamics for a fixed input $s \in \mathcal{A}$ is dominated by a unique attractive fixed point $y_s$, the induced dynamics on the map settles down in neuron $i_s$, corresponding to the mode of $y_s$,

$$i_s = \arg\max_{i \in \{1,2,\ldots,N\}} y_{s,i}.$$

The neuron $i_s$ will be most responsive to input subsequences ending with long blocks of symbols $s$.

Receptive fields of neurons on the map will be organized with respect to closeness of neurons to the fixed input winner $i_s$. 
Markovian organization of RFs

<table>
<thead>
<tr>
<th>baaaab</th>
<th>aaaaab</th>
<th>aaaaab</th>
<th>bab</th>
<th>aabab</th>
<th>bbbaa</th>
<th>bbaaa</th>
<th>bbaaa</th>
</tr>
</thead>
<tbody>
<tr>
<td>aabaab</td>
<td>aaaaab</td>
<td>aaaaab</td>
<td>baaa</td>
<td>bbaaab</td>
<td>babab</td>
<td>bbaaa</td>
<td>babaa</td>
</tr>
<tr>
<td>ababbb</td>
<td>b</td>
<td>bbaab</td>
<td>bbbab</td>
<td>aaaaab</td>
<td>aaaaab</td>
<td>aaaa</td>
<td>aabaa</td>
</tr>
<tr>
<td>baabbb</td>
<td>aaaa</td>
<td>baba</td>
<td>aaba</td>
<td>baaaab</td>
<td>baaaab</td>
<td>baaaab</td>
<td>aaaaab</td>
</tr>
<tr>
<td>aabbbb</td>
<td>aabbb</td>
<td>aabbb</td>
<td>baaa</td>
<td>aaaaab</td>
<td>aaaaab</td>
<td>aaaaab</td>
<td>aaaaab</td>
</tr>
<tr>
<td>bbbb</td>
<td>bbbb</td>
<td>bbbb</td>
<td>bbaa</td>
<td>abbaa</td>
<td>aaaa</td>
<td>aaaa</td>
<td>aaaa</td>
</tr>
<tr>
<td>bbbbbb</td>
<td>bbbbbb</td>
<td>bbbbbb</td>
<td>bbbbbb</td>
<td>bbbbbb</td>
<td>bbbbbb</td>
<td>bbbbbb</td>
<td>bbbbbb</td>
</tr>
<tr>
<td>babbbb</td>
<td>bbbba</td>
<td>bbbba</td>
<td>bbbba</td>
<td>bbbba</td>
<td>bbbba</td>
<td>bbbba</td>
<td>bbbba</td>
</tr>
<tr>
<td>aabbbb</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
</tr>
<tr>
<td>aabbbb</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
<td>babaa</td>
</tr>
</tbody>
</table>

P. Tiño
Markovian suffix-based RF organization

Assuming a unimodal character of the fixed point $y_s$, as soon the symbol $s$ is seen, the mode of the activation profile $y$ will drift towards the neuron $i_s$.

The more consecutive symbols $s$ we see, the more dominant the attractive fixed point of $F_s$ becomes and the closer the winner position is to $i_s$.

In this manner, a Markovian suffix-based RF organization is created.
Generative Probabilistic Model - Advantages?

- principled formulation
- coping with missing data
- consistent building of visualization hierarchies
- understanding hierarchies through model responsibilities
- semi-supervised mode possible (automatic initialization of child plots (e.g. MML))
Building mixtures constrained along low-dimensional manifolds

many possible ways of doing it ...

- **Break symmetry in positioning mixture components** (Gaussians) by introducing **channel noise**. Some Gaussians must be ”similar”, because they are likely to be swapped by a noisy communication channel. Vector Quantization through noisy channel.

- **Explicit non-linear embedding of low-dim latent space into high-dim model space** (e.g. centers of Gaussians in high-dim data space). Force noise models to live only on the low-dimensional embedding.
Let’s build a probabilistic model of data...

... that respects our 1-dim assumptions about the data organization.

This is a constrained mixture of noise models (Gaussians) - (Spherical) Gaussians (of the same width) are forced to have their means organized along a smooth 1-dim manifold.
Smooth embedding of continuous latent space

(low−dim latent space (continuous)

(smooth) non−linear embedding in high−dim model space

constrained mixture
Generative Topographic Mapping

GTM (Bishop, Svensén and Williams) is a latent variable model with a non-linear RBF $f_M$ mapping a (usually two dimensional) latent space $\mathcal{H}$ to the data space $\mathcal{D}$. This is a generative probabilistic model.

Latent space

Centres $x_i$

Data space $t_n$

Projection manifold

RBF net
GTM - differential geometry on projection manifold

Magnification Factors:
We can measure the stretch in the sheet using magnification factors, and this can be used to detect the gaps between data clusters.

Directional Curvatures:
We can also measure the directional curvature of the 2-D sheet. Visualize the magnitude and direction of the local largest curvatures to see where and how the manifold is most folded.
Magnification Factors (detect clusters)

On Topographic Maps/Clustering of Structured Data

P. Tiño
**Other data types?**

- Easy extension to count/histogram data by changing the noise distribution (independent Bernoulli/multinomial, binomial) - Latent Trait Model.
  - Can be used to visualize large document collections, discussion groups, etc.
  - Based on ‘bag-of-words’

- For sequential data
  - Need noise models that take into account temporal correlations within sequences, e.g. Markov chains, HMMs, etc.
  - The same latent space organization as before
  - Constrained mixture of noise models corresponding to latent centers living on the computer screen
Hidden Markov Model

Stationary emissions conditional on hidden (unobservable) states. Hidden states represent basic operating ”regimes” of the process.
Latent Trait HMM (LTHMM) - constrained mixture of HMM

Use HMM as the noise model

For each HMM (latent center) we need to parametrise several multinomials

- initial state probabilities
- transition probabilities
- emission probabilities (discrete observations)

Multinomials are parametrised as in LTM.
LTHMM - training

Constrained mixture of HMMs is fitted by Maximum likelihood using an E-M algorithm

Two types of hidden variables:

- which HMM generated which sequence (responsibility calculations is in mixture models)
- within a HMM, what is the state sequence responsible for generating the observed sequence (forward-backward-like calculations)
2-dim manifold $\mathcal{M}$ of local noise models (HMMs) $p(\cdot|x)$ parametrized by the latent space $\mathcal{V}$ through a smooth non-linear mapping. $\mathcal{M}$ is embedded in manifold $\mathcal{H}$ of all noise models of the same form.
**LTHMM - metric properties**

Latent coordinates $\mathbf{x}$ are displaced to $\mathbf{x} + d\mathbf{x}$. How different are the corresponding noise models (HMMs)? Need to answer this in a parametrization-free manner...

**Local Kullback-Leibler divergence** can be estimated by

$$D[p(\mathbf{s}|\mathbf{x})||p(\mathbf{s}|\mathbf{x} + d\mathbf{x})] \approx d\mathbf{x}^T J(\mathbf{x}) d\mathbf{x},$$

where $J(\mathbf{x})$ is the **Fisher Information Matrix**

$$J_{i,j}(\mathbf{x}) = -E_{p(\mathbf{s}|\mathbf{x})} \left[ \frac{\partial^2 \log p(\mathbf{s}|\mathbf{x})}{\partial x_i \partial x_j} \right]$$

that acts like a metric tensor on the Riemannian manifold $\mathcal{M}$.
LTHMM - Fisher Information Matrix

HMM is itself a latent variable model. \( J(x) \) cannot be analytically determined.

There are several approximation schemes and an efficient algorithm for calculating the observed Fisher Information Matrix.
LTHMM - Induced metric in data space

Structured data types - careful with the notion of a metric in the data space.

LTHMM naturally induces a metric in the structured data space.

Two data items (sequences) are considered to be close (or similar) if both of them are well-explained by the same underlying noise model (e.g., HMM) from the 2-dimensional manifold of noise models.

Distance between structured data items is implicitly defined by the local noise models that drive topographic map formation.

If the noise model changes, the perception of what kind of data items are considered similar changes as well.
LTHMM - experiments

- **Toy Data**
  400 binary sequences of length 40 generated from 4 HMMs (2 hidden states) with identical emission structure (the HMMs differed only in transition probabilities). Each of the 4 HMMs generated 100 sequences.

- **Melodic Lines of Chorals by J.S. Bach**
  100 chorales. Pitches are represented in the space of one octave, i.e. the observation symbol space consists of 12 different pitch values.
Bach chorals

Latent space visualization

no sharps no flats

sharps

flats

\( g-f#-e-d#-e \)

\( g-f#-bb-a \)

\( g-f#-g-a-bb-a \)
Latent organization regularizes the model

Evolution of negative log-likelihood per symbol measured on the training (o) and test (*) sets (Bach chorals experiment).
Eclipsing Binary Stars

Line of sight of the observer is aligned with orbit plane of a two star system to such a degree that the component stars undergo mutual eclipses.

Even though the light of the component stars does not vary, eclipsing binaries are variable stars - this is because of the eclipses.

The light curve is characterized by periods of constant light with periodic drops in intensity.

If one of the stars is larger than the other (primary star), one will be obscured by a total eclipse while the other will be obscured by an annular eclipse.
Eclipsing Binary Star - normalized flux

Original lightcurve

Shifted lightcurve

Phase-normalised lightcurve
Eclipsing Binary Star - the model

Parameters:
Primary mass: \( m \) (1-100 solar mass)
mass ration: \( q \) (0-1)
eccentricity: \( e \) (0-1)
inclination: \( i \) (\(0^o - 90^o\))
argument of periastron: \( \alpha p \) (\(0^o - 180^o\))
log period: \( \pi \) (2-300 days)
Empirical priors on parameters

\[ p(m, q, e, i, ap, \pi) = p(m)p(q)p(\pi)p(e|\pi)p(i)p(ap) \]

Primary mass density:
\[ p(m) = a \times m^b \]
where

\[ a = \begin{cases} 
0.6865, & \text{if } 0.5 \times M_{\text{sun}} \leq m \leq 1.0 \times M_{\text{sun}} \\
0.6865, & \text{if } 1.0 \times M_{\text{sun}} < m \leq 10.0 \times M_{\text{sun}} \\
3.9, & \text{if } 10.0 \times M_{\text{sun}} < m \leq 100.0 \times M_{\text{sun}} 
\end{cases} \]

\[ b = \begin{cases} 
-1.4, & \text{if } 0.5 \times M_{\text{sun}} \leq m \leq 1.0 \times M_{\text{sun}} \\
-2.5, & \text{if } 1.0 \times M_{\text{sun}} < m \leq 10.0 \times M_{\text{sun}} \\
-3.3, & \text{if } 10.0 \times M_{\text{sun}} < m \leq 100.0 \times M_{\text{sun}} 
\end{cases} \]
Mass ratio density

\[ p(q) = p_1(q) + p_2(q) + p_3(q) \]

where

\[ p_i(q) = A_i \times \exp(-0.5 \frac{(q-q_i)^2}{s_i^2}) \]

with

\[ A_1 = 1.30, A_2 = 1.40, A_3 = 2.35 \]
\[ q_1 = 0.30, q_2 = 0.65, q_3 = 1.00 \]
\[ s_1 = 0.18, s_2 = 0.05, s_3 = 0.10 \]

Log-period density

\[ p(\pi) = \begin{cases} 
1.93337\pi^3 + 5.7420\pi^2 - 1.33152\pi + 2.5205, & \text{if } \pi \leq \log_{10}18 \\
19.0372\pi - 5.6276, & \text{if } \log_{10}18 < \pi \leq \log_{10}300 
\end{cases} \]

etc.
GTM for topographic organization of fluxes from eclipsing binary stars

Smooth parametrized mapping $F$ from 2-dim latent space into the space where 6 parameters of the eclipsing binary star model live.

Model light curves are contaminated by an additive observational noise (Gaussian). This gives a local noise model in the (time,flux)-space.

Each point on the computer screen corresponds to a local noise model and "represents" observed eclipsing binary star light curves that are well explained by the local model.

MAP estimation of the mapping $F$ via E-M.
Outline of the model (1)

Latent space (computer screen)

Apply smooth mapping $\Gamma$

Apply physical model

Parameter space

Coordinate vector $[x_1, x_2]$

Parameter vector $\Theta$

$x$

$V$

$M$

$\Omega_M$
Outline of the model (2)

- **Distribution space**
  - \( p(O|x) \)
  - \( \Omega_H \)

- **Regression model space**
  - \( J \)
  - \( f_{\Pi(x)} \)
  - \( \Omega_J \)

- **Apply physical model**

- **Apply Gaussian observational noise**
Artificial fluxes - projections
Artificial fluxes - model
Real fluxes - projections + model
Final comments

• Natural and principled formulation of a visualization technique for structured data.

• Generative nature of the model – can deal with missing data, hierarchical plots, model selection issues etc.

• The framework can be extended to more complicated noise models, e.g. coupled HMMs for visualizing multivoice musical textures.

• Can naturally operate with prior knowledge.

• Approximation and speed-up techniques are needed as scaling is an issue (E-step).