Metric Properties of Structured Data Visualizations through Generative Probabilistic Modeling

Peter Tiňo **Nikolaos Gianniotis**
University of Birmingham
School of Computer Science
Birmingham B15 2TT, UK
pxt,nxg@cs.bham.ac.uk

Abstract

Recently, generative probabilistic modeling principles were extended to visualization of structured data types, such as sequences. The models are formulated as constrained mixtures of sequence models - a generalization of density-based visualization methods previously developed for static data sets. In order to effectively explore visualization plots, one needs to understand local directional magnification factors, i.e. the extend to which small positional changes on visualization plot lead to changes in local noise models explaining the structured data. Magnification factors are useful for highlighting boundaries between data clusters. In this paper we present two techniques for estimating local metric induced on the sequence space by the model formulation. We first verify our approach in two controlled experiments involving artificially generated sequences. We then illustrate our methodology on sequences representing chorals by J.S. Bach.

1 Introduction

Topographic visualisation techniques have been an important tool in multi-variate data analysis and data mining. Generative probabilistic approaches [Bishop *et al.*, 1998b; 1998a; Kabán and Girolami, 2001; Nabney *et al.*, 2005] demonstrated advantages over non-probabilistic alternatives in terms of a flexible and technically sound framework that makes various extensions possible in a principled manner.Recently, Tino et al.[2004] introduced a general framework for visualizing sets of sequential data. Loosely speaking, a smooth two-dimensional manifold in the space of appropriate noise models (e.g. Hidden Markov models (HMM) of a given topology) is constructed so that HMMs constrained on the manifold explain the observed sequences well. Each sequence (data item) is then projected on the manifold by visualizing the relevant HMMs likely to be responsible for generating that sequence.

It is possible in *non-linear* projections that visualizations of two data items get close in the visualization space, even though they may be separated by a large distance in the data space (and vice-versa). Therefore, to be of practical use, non-linear projection methods must be equipped with representations in the visualization space of metric relationships among data items in the original data space. Otherwise, one would not be able to detect and understand the underlying cluster structure of data items. To that end, in the case of static vectorial data of fixed-dimensionality, Bishop et al. [1997] calculated magnification factors of the Generative Topographic Mapping (GTM) [Bishop *et al.*, 1998b] using tools of differential geometry.

However, in the case of structured data types (such as sequences or trees) one has to be careful when dealing with the notion of a metric in the data space. The generative probabilistic visualization models studied in this paper naturally induce a metric in the structured data space. Loosely speaking, 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 two-dimensional manifold of noise models. Due to non-linear parametrization of the manifold, projections of two quite different data items may end up being mapped close to each other. We emphasise, that in this framework, the 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. For example, if the noise models were 1st-order Markov chains, sequences would be naturally grouped with respect to their short-range subsequence structure. If, on the other hand, the noise models were stochastic machines specializing at latching special events occurring within the data sequences, the sequences would naturally group with respect to the number of times the special event occurred, irrespective of the length of temporal separation between the events.

In this paper, we quantify the extend to which small positional changes on manifold of local noise models explaining structured data (for visualization purposes the manifold is identified with visualization space) lead to changes in the distributions defined by the noise models. It is important to quantify the changes in a parametrization-free manner - we use approximations of Kullback-Leibler divergence. We present two techniques for estimating local metric induced on the structured data space by the model formulation. We first verify our approach in two controlled experiments involving artificially generated sequences. We then illustrate our methodology on sequences representing chorals by J.S. Bach.

2 A latent trait model for sequential data

Consider a set of sequences over the alphabet $\mathcal{S}$ of S symbols, $\mathcal{S} = \{1, 2, ..., S\}$. The n-th sequence is denoted by $\boldsymbol{s}^{(n)} = (s_t^{(n)})_{t=1:T_n}$, where $n = 1 : N$ and T_n denotes its length. The aim is to represent each sequence using a two-dimensional latent (visualization) space $\mathcal{V} = [-1, 1]^2$. In order to exploit the visualization space fully, a maximum entropy (uniform) prior distribution is imposed over the latent space. With each latent point $\boldsymbol{x}$, we associate a generative distribution over sequences $p(\boldsymbol{s}|\boldsymbol{x})$. One possibility (considered in this paper) is to use HMM with K hidden states [Rabiner and Juang, 1986]:

$$p(\boldsymbol{s}^{(n)}|\boldsymbol{x}) = \sum_{\boldsymbol{h}} p(h_1|\boldsymbol{x}) \prod_{t=2}^{T_n} p(h_t|h_{t-1}, \boldsymbol{x}) \prod_{t=1}^{T_n} p(s_t^{(n)}|h_t, \boldsymbol{x}), \tag{1}$$

where $\boldsymbol{h}$ is the set of all T_n-tuples over the K hidden states.

For tractability reasons, we discretize the latent space into a regular grid of C points[1] $\boldsymbol{x}_1, ..., \boldsymbol{x}_C$. In order to have the HMMs topologically organized, we (in the spirit of [Bishop *et al.*, 1998b; Kabán and Girolami, 2001]) constrain the flat mixture of HMMs,

$$p(\boldsymbol{s}) = \frac{1}{C} \sum_{c=1}^{C} p(\boldsymbol{s}|\boldsymbol{x}_c),$$

by requiring that the HMM parameters be generated through a parameterised *smooth* nonlinear mapping from the latent space $\mathcal{V}$ into the HMM parameter space:

$$p(h_1 = k|\boldsymbol{x}) = g_k(\boldsymbol{A}^{(\boldsymbol{\pi})}\boldsymbol{\phi}(\boldsymbol{x})), \tag{2}$$

$$p(h_t = k|h_{t-1} = l, \boldsymbol{x}) = g_k(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x})), \tag{3}$$

$$p(s_t = s|h_t = k, \boldsymbol{x}) = g_s(\boldsymbol{A}^{(\boldsymbol{B}_k)}\boldsymbol{\phi}(\boldsymbol{x})), \tag{4}$$

where indexes k, l run from 1 to K; index s ranges from 1 to S, and

- the function $g(.)$ is the softmax function[2] and $g_k(.)$ denotes the k-th component returned by the softmax, i.e.

$$g_k\left((a_1, a_2, ..., a_q)^T\right) = \frac{e^{a_k}}{\sum_{i=1}^{q} e^{a_i}}, \quad k = 1, 2, ..., q,$$

- $\phi(.) = (\phi_1(.), ..., \phi_M(.))^T, \phi_m(.) : \mathcal{R}^2 \to \mathcal{R}$ is an ordered set of M non-parametric nonlinear smooth basis functions (typically RBFs),
- the matrices $\boldsymbol{A}^{(\boldsymbol{\pi})} \in \mathcal{R}^{K \times M}$, $\boldsymbol{A}^{(\boldsymbol{T}_l)} \in \mathcal{R}^{K \times M}$ and $\boldsymbol{A}^{(\boldsymbol{B}_k)} \in \mathcal{R}^{S \times M}$ are free parameters of the model.

Assuming the sequences $\boldsymbol{s}^{(n)}$, $n = 1 : N$, were independently generated, the model likelihood reads

$$\mathcal{L} = \prod_{n=1}^{N} p(\boldsymbol{s}^{(n)}) = \prod_{n=1}^{N} \frac{1}{C} \sum_{c=1}^{C} p(\boldsymbol{s}^{(n)}|\boldsymbol{x}_c). \tag{5}$$

[1] these sample (grid) points are analogous to the nodes of a Self Organising Map

[2] which is the canonical inverse link function of multinomial distributions

Maximum likelihood estimates of the free parameters can be obtained via Expectation-Maximization (EM) algorithm [Tino *et al.*, 2004].

Given a sequence $\boldsymbol{s}$, some regions of the latent visualization space are better at explaining it than the others. A natural representation (visualization) of $\boldsymbol{s}$ is the mean of the posterior distribution over the latent space, given that sequence:

$$Proj(\boldsymbol{s}) = \sum_{c=1}^{C} \boldsymbol{x}_c \, p(\boldsymbol{x}_c|\boldsymbol{s}).$$

3 Quantifying metric properties of the visualization space

In this section, we present two approaches to quantifying metric properties of the visualization space $\mathcal{V}$ as sensitivities (measured by Kullback-Leibler divergence) of local noise models (HMM) addressed by the points in $\mathcal{V}$ to small perturbations in $\mathcal{V}$.

3.1 Approximating Kullback-Leibler divergence through observed Fisher information matrix

Consider the visualization (latent) space $\mathcal{V} = [-1, +1]^2$ and the two-dimensional manifold $\mathcal{M}$ of local noise models $p(\cdot|\boldsymbol{x})$ (e.g. HMM with 3 states, emissions over 7 symbols) parametrized by the latent space through (1) and (2-4). The manifold is embedded in manifold $\mathcal{H}$ of all noise models of the same form (e.g. all HMMs with 3 states and emissions over 7 symbols). Consider a latent point $\boldsymbol{x} \in \mathcal{V}$. If we displace $\boldsymbol{x}$ by an infinitesimally small perturbation $d\boldsymbol{x}$, the Kullback-Leibler divergence $D_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x} + d\boldsymbol{x}))$ between the corresponding noise models $p(\cdot|\boldsymbol{x}), (p(\cdot|\boldsymbol{x} + d\boldsymbol{x}) \in \mathcal{M}$ can be determined via Fisher information matrix

$$\boldsymbol{F}(\boldsymbol{x}) = -E_{p(\cdot|\boldsymbol{x})}[\nabla^2 \log p(.|\boldsymbol{x})]$$

that acts like a metric tensor on the Riemannian manifold $\mathcal{M}$ [Kullback, 1959]:

$$D_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x} + d\boldsymbol{x})) = d\boldsymbol{x}^T \boldsymbol{F}(\boldsymbol{x}) \, d\boldsymbol{x}.$$

The situation is illustrated in figure 1.

Local noise models used in this paper (HMMs) are latent variable models and there is no closed-form formula for calculating the Fisher information matrix. However, Lystig and Hughes[2002] presented a framework for efficient calculation of the *observed* Fisher information matrix of HMMs based on forward calculations related to those used in maximum likelihood parameter estimation via EM algorithm. We adapt the framework to a special kind of parametrization of HMM used in the latent trait model of section 2. Each HMM has two parameters – latent space coordinates $\boldsymbol{x} = (x_1, x_2)$ (these can be thought of as coordinates on the visualization screen).

Consider a set $\mathcal{S}(\boldsymbol{x})$ of N sequences independently generated by HMM $p(\cdot|\boldsymbol{x})$. All sequences have equal length T. Given the n-th sequence, we start recursion from the beginning of the sequence:

$$\lambda_1^{(n)}(k) = p(s_1^{(n)}|h_1 = k, \boldsymbol{x}_c)p(h_1 = k|\boldsymbol{x})$$

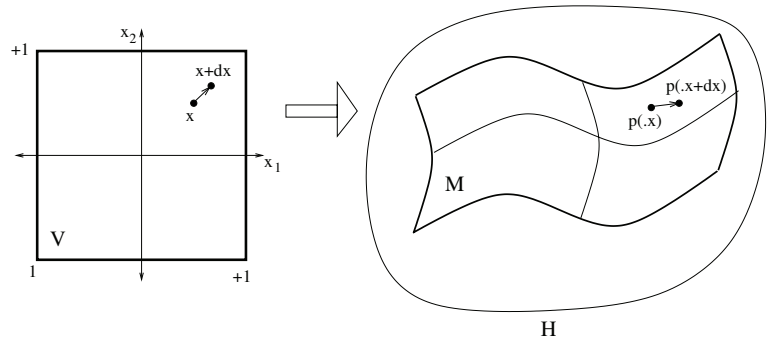


Figure 1: Two-dimensional manifold $\mathcal{M}$ of local noise models $p(\cdot|\boldsymbol{x})$ parametrized by the latent space $\mathcal{V}$ through (1) and (2-4). The manifold is embedded in manifold $\mathcal{H}$ of all noise models of the same form. Latent coordinates $\boldsymbol{x}$ are displaced to $\boldsymbol{x} + d\boldsymbol{x}$. Kullback-Leibler divergence $D_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x} + d\boldsymbol{x})$ between the corresponding noise models $p(\cdot|\boldsymbol{x}), (p(\cdot|\boldsymbol{x} + d\boldsymbol{x}) \in \mathcal{M}$ can be determined via Fisher information matrix $\boldsymbol{F}(\boldsymbol{x})$ that acts like a metric tensor on the Riemannian manifold $\mathcal{M}$.

Recursive step:

$$\begin{aligned}\lambda_t^{(n)}(k) &= p(s_t^{(n)}, h_t = k|s_1^{(n)}, \ldots, s_{t-1}^{(n)}, \boldsymbol{x}) \\ &= \sum_{i=1}^{K}[\lambda_{t-1}^{(n)}(i)p(s_t^{(n)}|h_t = k, \boldsymbol{x}) \\ &\quad p(h_t = k|h_{t-1} = i, \boldsymbol{x})](\Lambda_{t-1}^{(n)})^{-1},\end{aligned}$$

where $\Lambda_t^{(n)} = \sum_{j=1}^{K} \lambda_t(j)^{(n)}$.

Starting again at the beginning of the sequence, we recursively evaluate 1st-order derivatives with respect to latent coordinates x_1, x_2. Letting $q \in \{1, 2\}$, we have:

$$\psi_1^{(n)}(k; x_q) = \frac{\partial}{\partial x_q} p(s_1^{(n)}|h_1 = k, \boldsymbol{x})p(h_1 = k|\boldsymbol{x}),$$

$$\psi_t^{(n)}(k; x_q) = \frac{\frac{\partial}{\partial x_q} p(s_1^{(n)}, \ldots s_t^{(n)}, h_t = k|\boldsymbol{x})}{p(s_1^{(n)}, \ldots, s_{t-1}^{(n)}|\boldsymbol{x})},$$

and $\Psi_t^{(n)}(x_q) = \sum_{j=1}^{K} \psi_t^{(n)}(j; x_q)$.

The calculations recursively employ $\psi_{t-1}^{(n)}$, $\lambda_{t-1}^{(n)}$ and $\Lambda_{t-1}^{(n)}$. We also need 1st-order derivatives of initial state, state transition and state-conditional emission probabilities with respect to latent coordinates. We present only the equation for state transitions, other formulas can be obtained in the same manner:

$$\frac{\partial}{\partial x_q} p(h_t = k|h_{t-1} = l, \boldsymbol{x}) = g_k(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x}))$$

$$\left(\boldsymbol{A}_k^{(\boldsymbol{T}_l)} \frac{\partial \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q} - \sum_{i=1}^{K}[g_i(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x}))\boldsymbol{A}_i^{(\boldsymbol{T}_l)} \frac{\partial \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q}]\right),$$

where $\boldsymbol{A}_i^{(\cdot)}$ denotes the i-th row of the parameter matrix $\boldsymbol{A}^{(\cdot)}$.

2nd-order derivatives are obtained in a recursive manner as well:

$$\omega_1^{(n)}(k; x_q, x_r) = \frac{\partial^2}{\partial x_q \partial x_r} p(s_1^{(n)}|h_1 = k, \boldsymbol{x})p(h_1 = k|\boldsymbol{x}),$$

$$\omega_t^{(n)}(k; x_q, x_r) = \frac{\frac{\partial^2}{\partial x_q \partial x_r} p(s_1^{(n)}, \cdots, s_t^{(n)}, h_t = k|\boldsymbol{x})}{p(s_1^{(n)}, \cdots, s_{t-1}^{(n)}|\boldsymbol{x})}$$

and $\Omega_t^{(n)}(x_q, x_r) = \sum_{j=1}^{K} \omega_t^{(n)}(j; x_q, x_r)$.

The calculations recursively employ $\psi_{t-1}^{(n)}$, $\omega_{t-1}^{(n)}$, $\lambda_{t-1}^{(n)}$ and $\Lambda_{t-1}^{(n)}$. We also need both 1st- and 2nd-order derivatives of initial state, state transition and state-conditional emission probabilities with respect to latent coordinates. Again, we present only the equation for state transitions:

$$\begin{aligned}\frac{\partial^2}{\partial x_q \partial x_r} p(h_t = k|h_{t-1} = l, \boldsymbol{x}) &= \frac{\partial}{\partial x_r} g_k(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x})) \\ &\left(\boldsymbol{A}_k^{(\boldsymbol{T}_l)} \frac{\partial \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q} - \sum_{i=1}^{K}[g_i(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x}))\boldsymbol{A}_i^{(\boldsymbol{T}_l)} \frac{\partial \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q}]\right) \\ &+ g_k(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x})) \\ &\left(\boldsymbol{A}_k^{(\boldsymbol{T}_l)} \frac{\partial^2 \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q \partial x_r} - \sum_{i=1}^{K}[\frac{\partial}{\partial x_r} g_i(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x}))\boldsymbol{A}_i^{(\boldsymbol{T}_l)} \frac{\partial \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q} + \right. \\ &\left. g_i(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x}))\boldsymbol{A}_i^{(\boldsymbol{T}_l)} \frac{\partial^2 \boldsymbol{\phi}(\boldsymbol{x})}{\partial x_q \partial x_r}]\right)\end{aligned}$$

Denoting

$$\mathcal{Q}_{q,r}^{(n)} = \frac{\Omega_T^{(n)}(x_q, x_r)}{\Lambda_T^{(n)}} - \frac{\Psi_T^{(n)}(x_q)\Psi_T^{(n)}(x_r)}{(\Lambda_T^{(n)})^2},$$

we calculate elements of the observed Fisher information matrix $\hat{\boldsymbol{F}}(\boldsymbol{x})$, given the set of sequences $\mathcal{S}(\boldsymbol{x})$, as

$$\hat{\boldsymbol{F}}(\boldsymbol{x})_{q,r} = -\frac{1}{N}\sum_{n=1}^{N} \mathcal{Q}_{q,r}^{(n)}.$$

To illustrate metric properties of the visualization space, we calculate the observed Fisher information matrices $\hat{\boldsymbol{F}}(\boldsymbol{x_c})$ in all latent centers $\boldsymbol{x}_c$, $c = 1, 2, ..., C$, and detect through eigen-analysis of $\hat{\boldsymbol{F}}(\boldsymbol{x_c})$ directions of dominant local change in Kullback-Leibler divergence between the HMM parametrized by $\boldsymbol{x}_c$ and its perturbed version parametrized by $\boldsymbol{x}_c + d\boldsymbol{x}$. In the visualization space, we signify magnitude of the dominant change (dominant eigenvalue of $\hat{\boldsymbol{F}}(\boldsymbol{x_c})$) by local brightness of the background as well as mark the direction of the dominant change by a piece of line.

3.2 Direct recursive approximation of Kullback-Leibler divergence

Another alternative approach to probe metric properties of the visualization space is the estimate $D_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x}+\Delta\boldsymbol{x}))$ directly. Do [2003] presented an efficient algorithm for approximating Kullback-Leibler (K-L) divergence between two HMM of the same topology. The approximation is based on backward calculations related to those used in maximum likelihood parameter estimation via EM algorithm.

With each hidden state $k \in \{1, 2, ..., K\}$ we associate an auxiliary process $\beta_k(t)$. Given a sequence $\boldsymbol{s} = (s_t)_{t=1:T}$, the likelihood $p(\boldsymbol{s}|\boldsymbol{x})$ can be efficiently calculated by

- starting at the end of the sequence:
$$\beta_k(T; \boldsymbol{x}) = p(s_T|k, \boldsymbol{x})$$
- Recursive step:
$$\beta_k(t; \boldsymbol{x}) = p(s_t|k, \boldsymbol{x}) \sum_{i=1}^{K} p(h_{t+1} = i|h_t = k, \boldsymbol{x})\beta_i(t+1; \boldsymbol{x})$$
- Final step:
$$p(\boldsymbol{s}|\boldsymbol{x}) = \sum_{k=1}^{K} p(h_1 = k|\boldsymbol{x})\beta_k(1; \boldsymbol{x})$$

For $u, v \in (0, 1]$, define

$$\kappa(u, v) = u \log \frac{u}{v}.$$

The Kullback-Leibler (K-L) divergence between two HMMs $p(\cdot|\boldsymbol{x})$ and $p(\cdot|\boldsymbol{x'})$, can be approximated as follows:

- Recursion is initiated at end of the sequence:
$$D_k(T; \boldsymbol{x}, \boldsymbol{x'}) = \kappa(p(s_T|k, \boldsymbol{x}), p(s_T|k, \boldsymbol{x'}))$$
- Recursive step:
$$D_k(t; \boldsymbol{x}, \boldsymbol{x'}) = \kappa(p(s_t|k, \boldsymbol{x}), p(s_t|k, \boldsymbol{x'})) + D_{KL}(p(h_{t+1}|h_t = k, \boldsymbol{x})||p(h_{t+1}|h_t = k, \boldsymbol{x'})) + \sum_{i=1}^{K} p(h_{t+1} = i|h_t = k, \boldsymbol{x})D_i(t + 1; \boldsymbol{x}, \boldsymbol{x'})$$
- Final step: the empirical K-L divergence, given the sequence $\boldsymbol{s}$, is bounded by
$$\mathcal{K}(\boldsymbol{s}, \boldsymbol{x}, \boldsymbol{x'}) = D_{KL}(p(h_1|\boldsymbol{x})||p(h_1|\boldsymbol{x'})) + \sum_{k=1}^{K} p(h_1 = k|\boldsymbol{x})D_k(1; \boldsymbol{x}, \boldsymbol{x'}).$$

Given the set of N sequences, $\mathcal{S}(\boldsymbol{x})$, generated independently by the HMM addressed by $\boldsymbol{x}$, an estimate of $D_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x}+\Delta\boldsymbol{x}))$ is calculated as

$$\hat{D}_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x}+\Delta\boldsymbol{x})) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}(\boldsymbol{s}^{(n)}, \boldsymbol{x}, \boldsymbol{x}+\Delta\boldsymbol{x}).$$

Again, to illustrate metric properties of the visualization space, we perturb latent centers $\boldsymbol{x}_c$ in regular intervals around small circle centered at $\boldsymbol{x}_c$ and for each perturbation $\Delta\boldsymbol{x}$ calculate $\hat{D}_{KL}(p(\cdot|\boldsymbol{x})\|(p(\cdot|\boldsymbol{x}+\Delta\boldsymbol{x}))$. As before, in the visualization space, we signify magnitude of the dominant change by local brightness of the background as well as mark the direction of the dominant change by a piece of line.

4 Experiments

In this section we demonstrate techniques introduced in sections 3.1 and 3.2 to provide additional metric information when visualizing sequential data.

In all experiments the latent space centres $\boldsymbol{x}_c$ were positioned on a regular 10×10 square grid ($C = 100$) and there were $M = 16$ basis functions ϕ_i. The basis functions were spherical Gaussian functions of the same width $\sigma = 1.0$. The basis functions were centred on a regular 4×4 square grid, reflecting uniform distribution of the latent classes. We account for a bias term by using an additional constant basis function $\phi_{17}(\boldsymbol{x}) = 1$.

Free parameters of the model were randomly initialized in the interval $[-1, 1]$. Training consisted of repeating EM iterations [Tino *et al.*, 2004]. Typically, the likelihood levelled up after 30-50 EM cycles.

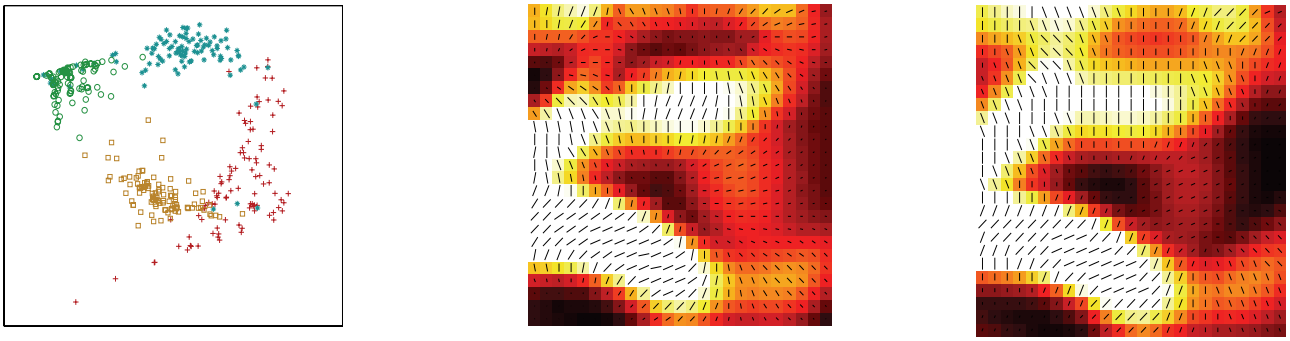
When calculating metric properties of the visualization space, each set $\mathcal{S}(\boldsymbol{x}_c)$ consisted of 100 generated sequences of length 40.

4.1 Toy data

We first generated a toy data set of 400 binary sequences of length 40 from four HMMs ($K = 2$ hidden states) with identical emission structure, i.e. the HMMs differed only in transition probabilities. Each of the four HMMs generated 100 sequences. Local noise models $p(\cdot|\boldsymbol{x})$ were imposed as HMMs with 2 hidden states. Visualization of the sequences is presented in figure 2(a). Sequences are marked with four different markers, corresponding to the four different HMMs used to generate the data set. We stress that the model was trained in a completely unsupervised way. The markers are used for illustrative purposes only. Representations of induced metric in the local noise model space based on Fisher Information matrix (section 3.1) and direct K-L divergence estimations (section 3.2) can be seen in figures 2(b) and 2(c), respectively. Dark areas signify homogeneous regions of local noise models and correspond to possible clusters in the data space. Light areas signify abrupt changes in local noise model distributions (as measured by K-L divergence) and correspond to boundaries between data clusters. The visualization plot reveals that the latent trait model essentially discovered the organization of the data set and the user would be able to detect the four clusters, even without help of the marking scheme in figure 2(a). Of course, the latent trait model benefited from the fact that the distributions used to generate data were from the same model class as the local noise models. There were few atypical sequences generated by the HMM marked with '+', and this is clearly reflected by their projections in the lower left corner.

4.2 Melodic lines of chorals by J.S. Bach

Next we subjected the latent trait model to a set of 100 chorales by J.S. Bach from UCI repository of Machine Learning Databases. We extracted the melodic lines – pitches are represented in the space of one octave, i.e. the observation symbol space consists of 12 different pitch values. Local noise models had $K = 3$ hidden states. Figure 3 shows choral visualizations, while representations of induced metric in the



(a) (b) (c)

Figure 2: (a) Visualization of 400 binary sequences of length 40 generated by four HMMs with 2 hidden states and with identical emission structure. Sequences are marked with four different markers, corresponding to the four HMMs. Also shown are representations of induced metric in the local noise model space based on Fisher Information matrix (b) and direct K-L divergence estimations (c).

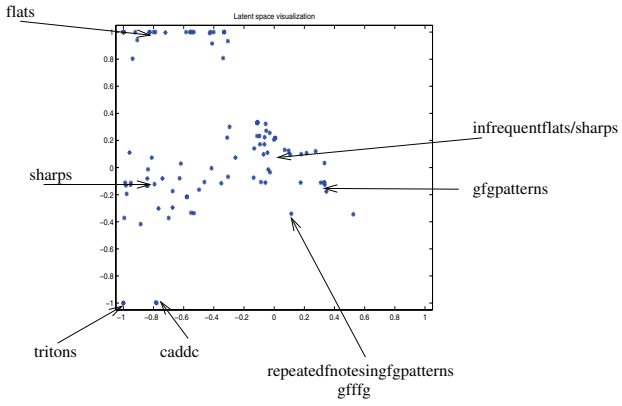


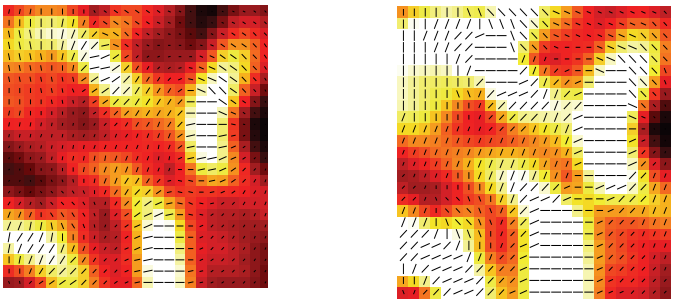
Figure 3: Visualization of 100 melodic lines of chorales by J.S. Bach.

local noise model space based on Fisher Information matrix and direct K-L divergence estimations can be seen in figures 4(a) and 4(b), respectively. The method discovered natural topography of the key signatures, corroborated with similarities of melodic motives. The melodies can contain sharps and flats other than those included in their key signature due to both modulation and ornaments. The upper region contains melodic lines that utilise keys with flats. Central part of the visualisation space is taken by sharps (left) and almost no sharps/flats (center). The lower region of the plot contains chorals with tense patterns (e.g containing tritons) and is quite clearly strongly separated from other chorals. Again, the overall clustering of chorals is well-matched by the metric representations of figures 4(a) and 4(b). Flats, sharps and tense patterns are clearly separated, as are sharps and infrequent sharps/flats. There are more interesting features to this visualization (e.g. enharmony patterns) that cannot be addressed due to space limitations.

5 Discussion and conclusion

When faced with the problem of non-linear topographic data visualization, one needs additional information putting relative distances of data projections in proportion to local stretchings/contractions of the visualization manifold in the data space. It is important to base representations of manifold stitchings/contractions on the particular notion of distance or similarity between data items that drives formation of the topographic mapping. If one visualizes structured data, such as sequences, this can be a highly non-trivial task, especially for the family of neural-based topographic mappings of sequences [Koskela *et al.*, 1998; Horio and Yamakawa, 2001; Voegtlin, 2002; Strickert and Hammer, 2003]. However, when the visualization model is formulated as a latent trait model, such calculations follow naturally from the model formulation. Two sequences are considered to be similar if both of them are well-explained by the same underlying noise model (in our particular case - HMM). Noise models are organized on a smooth two-dimensional manifold and we can study changes of the local metric (based on Kullback-Leibler divergence) due to non-linear parametrization. We have presented two approaches to metric analysis of the visualization space and experimentally verified that additional information provided by metric representations in the visualization space can be useful for understanding of the overall organization of data. We stress. that the data organization revealed is always with respect to the notion of similarity between data items imposed in the model construction.

The results can be easily generalized e.g. to tree-structured data using, for example, Hidden Markov Tree models (HMTM) [Durand *et al.*, 2004]). We stress that the presented framework is general and can be applied to visualizations through latent trait models of a wide variety of structured data, provided a suitable noise model is used. For example, we are currently working on topographic organizations (and metric properties of such) of fluxes form binary star



(a) (b)

Figure 4: Representations of induced metric in the local noise model space based on Fisher Information matrix (a) and direct K-L divergence estimations (b) for visualizations of choral by J.S. Bach.

complexes, where the noise models are based on real physical models of binary stars.

References

[Bishop *et al.*, 1997] C.M. Bishop, M. Svensén, and C.K.I. Williams. Magnification factors for the GTM algorithm. In *Proceedings IEE Fifth International Conference on Artificial Neural Networks*, pages 64–69. IEE, London, 1997.

[Bishop *et al.*, 1998a] C.M. Bishop, M. Svensén, and C.K.I. Williams. Developments of the Generative Topographic Mapping. *Neurocomputing*, 21:203–224, 1998.

[Bishop *et al.*, 1998b] C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–235, 1998.

[Do, 2003] Minh N. Do. Fast approximation of kullback–leibler distance for dependence trees and hidden markov models,. *Signal Processing Letters, IEEE*, 10(4):115–118, 2003.

[Durand *et al.*, 2004] J.-B Durand, P. Goncalves, and Y. Guedon. Computational methods for hidden markov tree models-an application to wavelet trees. *IEEE Transactions on Signal Processing*, 52(9):2552–2560, 2004.

[Horio and Yamakawa, 2001] K. Horio and T. Yamakawa. Feedback self-organizing map and its application to spatio-temporal pattern classification. *International Journal of Computational Intelligence and Applications*, 1(1):1–18, 2001.

[Kabán and Girolami, 2001] A. Kabán and M. Girolami. A combined latent class and trait model for the analysis and visualization of discrete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):859–872, 2001.

[Koskela *et al.*, 1998] T. Koskela, M. Varsta znd J. Heikkonen, and K. Kaski. Recurrent SOM with local linear models in time series prediction. In *6th European Symposium on Artificial Neural Networks*, pages 167–172, 1998.

[Kullback, 1959] S. Kullback. *Information Theory and Statistics*. Wiley, New York, NY, 1959.

[Lystig and Hughes, 2002] Theodore C. Lystig and James P. Hughes. Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689, 2002.

[Nabney *et al.*, 2005] I. Nabney, Y. Sun, P. Tiňo, and A. Kabán. Semisupervised learning of hierarchical latent trait models for data visualisation. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), 2005.

[Rabiner and Juang, 1986] R.L. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3:4–16, 1986.

[Strickert and Hammer, 2003] M. Strickert and B. Hammer. Neural gas for sequences. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 53–57, 2003.

[Tino *et al.*, 2004] P. Tino, A. Kaban, and Y. Sun. A generative probabilistic approach to visualising sets of symbolic sequences. In W. DuMouchel J. Ghosh. R. Kohavi, J. Gehrke, editor, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 701–706. ACM Press, 2004.

[Voegtlin, 2002] T. Voegtlin. Recursive self-organizing maps. *Neural Networks*, 15(8–9):979–992, 2002.