17
Axiomatic geometries for text documents

Guy Lebanon

Abstract

High-dimensional structured data such as text and images is often poorly understood and misrepresented in statistical modelling. Typical approaches to modelling such data involve, either explicitly or implicitly, arbitrary geometric assumptions. In this chapter, we consider statistical modelling of non-Euclidean data whose geometry is obtained by embedding the data in a statistical manifold. The resulting models perform better than their Euclidean counterparts on real world data and draw an interesting connection between Čencov and Campbell's axiomatic characterisation of the Fisher information and the recently proposed diffusion kernels and square root embedding.

17.1 Introduction

Geometry is ubiquitous in many aspects of statistical modelling. During the last half century a geometrical theory of statistical inference has been constructed by Rao, Efron, Amari, and others. This theory, commonly referred to as information geometry, describes many aspects of statistical modelling through the use of Riemannian geometric notions such as distance, curvature and connections (Amari and Nagaoka 2000). Information geometry has been mostly involved with the geometric interpretations of asymptotic inference. Focusing on the geometry of parametric statistical families $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, information geometry has had relatively little influence on the geometrical analysis of data. In particular, it has largely ignored the role of the geometry of the data space X in statistical inference and algorithmic data analysis.

On the other hand, the recent growth in computing resources and data availability has lead to widespread analysis and modelling of structured data such as text and images. Such data does not naturally lie in $\mathbb{R}^n$ and the Euclidean distance and its corresponding geometry do not describe it well. In this chapter, we address the issue of modelling structured data using non-Euclidean geometries. In particular, by embedding data $x \in X$ into a statistical manifold, we draw a connection between

Algebraic and Geometric Methods in Statistics, ed. Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin and Henry P. Wynn. Published by Cambridge University Press. © Cambridge University Press 2010.

the geometry of the data space X and the information geometric theory of statistical manifolds.

We begin by discussing the role of the geometry of data spaces in statistical modelling and then proceed to discuss the question of how to select an appropriate geometry. We then move on to discuss the geometric characterisations due to Čencov, Campbell and Lebanon and their applications to modelling structured data. While much of this chapter is relevant for a wide variety of data, we focus on the specific case of text data.

17.2 The role of geometry in statistical modelling

Statistical modelling often involves making assumptions concerning the geometry of the data space X. Such assumptions are sometimes made explicitly as in the case of nearest neighbour classifiers and information retrieval in search engines. In these cases the model or learning algorithm makes direct use of a distance function d on X. In other cases, geometric assumptions are implicitly made, and are revealed only after a careful examination. For example, the choice of a parametric statistical family such as the Gaussian family carries clear geometrical assumptions. Other examples include the choice of a smoothing kernel in non-parametric smoothing and the parametric form of logistic regression

$$p(y|x\,;\theta) \propto \exp(-y\langle x,\theta\rangle), \quad x,\theta \in \mathbb{R}^n, y \in \{+1,-1\} \tag{17.1}$$

where careful examination reveals its dependence on the Euclidean margin

$$\langle x,\theta\rangle = \|\theta\|_2\langle x,\hat{\theta}\rangle = \|\theta\|_2\,(\|x\|_2 - d(x,H_\theta)) \tag{17.2}$$

where $d(x,H_\theta) = \inf_{y\in H_\theta}\|x-y\|_2$ is the Euclidean distance of x from the flat decision hyperplane orthogonal to the unit vector $\hat{\theta} = \theta/\|\theta\|_2$ (Lebanon and Lafferty 2004).

Before proceeding we pause to informally describe the geometric notions that we will use later on in this chapter. More details concerning basic Riemannian geometry may be found in general introduction to the field such as (Spivak 1975) or the statistically oriented monographs (Kass and Voss 1997, Amari and Nagaoka 2000).

A smooth manifold X is a continuous set of points on which differentiation and other smooth operations can take place. While a smooth manifold X by itself does not carry any geometrical properties, considering it in conjunction with a local inner product g turns the topological structure X into a geometric space (X,g) called a Riemannian manifold.

The local inner product or Riemannian metric g is defined as a smooth symmetric, bilinear and positive definite function $g_x(\cdot,\cdot)$, $g_x : T_xX \times T_xX \to \mathbb{R}$ where T_xX is the tangent space of a manifold X at $x \in X$. Assuming that X is a smooth surface in $\mathbb{R}^N$, the tangent space T_xX intuitively corresponds to the subspace of vectors in $\mathbb{R}^N$ that are centred at x and are tangent to the surface X at $x \in X$. The smoothness requirement refers to smoothness of $g_x(u,v)$ as a function of $x \in X$.

The inner product leads to the notion of lengths of parametrised curves $\alpha : I \to X$

$$l(\alpha) \overset{\text{def}}{=} \int_I \sqrt{g_{\alpha(t)}(\dot{\alpha}(t), \dot{\alpha}(t))}\, dt \tag{17.3}$$

where $\dot{\alpha}(t)$ is the tangent vector to the curve α at time t. Using the definition of curve lengths (17.3), the local metric g leads to a distance function d on X defined as the length of the shortest curve connecting the two points

$$d(x, y) \overset{\text{def}}{=} \inf_{\alpha : x, y \in \alpha} l(\alpha). \tag{17.4}$$

The simplest example of a Riemannian manifold is of course $(\mathbb{R}^n, \delta)$ where $\delta_x(u, v) = \sum u_i v_i$ is a metric that is constant in $x \in X$. Curve lengths (17.3) in this case become the Euclidean curve lengths from calculus and the distance function $d(x, y)$ in (17.4) becomes the Euclidean or L_2 distance $d(x, y) = \|x - y\|_2 \overset{\text{def}}{=} \sqrt{\sum (x_i - y_i)^2}$.

In general, expressions (17.3), (17.4) do not have closed form expressions which may make their calculation slow and impractical, especially when the dimensionality of X is high. There are, however, parametric families of metrics $\mathcal{G} = \{g^\theta : \theta \in \Theta\}$ possessing efficient closed form expressions for (17.3), (17.4). Fortunately, such metric classes are often quite flexible and contain many popular distance functions e.g., (Lebanon 2006).

The local metric g associated with a Riemannian manifold (X, g) provides additional geometric structure beyond the concept of a distance function. This additional structure includes concepts such as curvature, flatness and angles and provides a full geometric characterisation of X.

Once a metric g on X has been identified it can be used in parametric modelling to define the parametric family under consideration. For example, the family

$$p(x\,; \mu, c) = \exp(-c\, d^2(x, \mu) - \log \psi(c, \mu)) \quad \mu \in X, c \in \mathbb{R}_{>0} \tag{17.5}$$

where d is given by (17.4) and $\psi(c, \mu) = \int \exp(-c\, d^2(x, \mu))\, dx$ generalises the Gaussian distribution to arbitrary Riemannian spaces (X, g). Inference on (X, g) using the family $\{p(\cdot\,; \mu, c) : \mu \in X, c \in \mathbb{R}_{>}0\}$ can then proceed according to standard statistical procedures such as maximum likelihood or Bayesian analysis.

The distribution (17.5) may also be used to define a geometric smoothing kernel for use in non-parametric smoothing (Wand and Jones 1995)

$$\hat{p}(x) = \frac{1}{m} \sum_{i=1}^{m} K_c(x, x_i) = \frac{1}{m} \sum_{i=1}^{m} p(x\,; x_i, c) \quad x_1, \ldots, x_m \in X. \tag{17.6}$$

Distributions such as (17.5) and the estimator (17.6) express an explicit dependence on the data manifold geometry (X, g) which may or may not be Euclidean.

The metric g can also be used in regression or classification where we estimate a conditional model $p(y|x), x \in X$. For example, following the reasoning in (17.1), (17.2) we can define the natural extension of logistic regression to (X, g) as

$$p(y|x\,; \theta, \eta) \propto \exp(\theta\, s(x, \eta)\, d(x, H_\eta)) \quad \theta \in \mathbb{R} \tag{17.7}$$

where H_η is a decision boundary that is a flat submanifold in (X, g) (parametrised by η) and $s(x, \eta) = +1$ or -1 depending on the location of x with respect to the decision boundary H_η (Lebanon and Lafferty 2004). The notation $d(x, H_\eta)$ refers to the geometric generalisation of the margin $d(x, A) \stackrel{\text{def}}{=} \min_{y \in A} d(x, y)$ with $d(x, y)$ defined in (17.4). Note that the metric g is expressed in this case through the distance function d and the definition of flat decision boundaries H_η. Similarly, the geometric analogue of non-probabilistic classifiers such as nearest neighbours or support vector machines (SVM) corresponding to (X, g) may be defined using the distance function (17.4) and the approximated geometric diffusion kernel $K_c(x, y) = \exp(-c\, d^2(x, y))$ (Lafferty and Lebanon 2005).

In many cases, the geometric models defined above and others reduce to well-known statistical procedures when the data space is assumed to have a Euclidean geometry. This emphasises the arbitrariness associated with the standard practice of assuming the data lies in $(X, g) = (\mathbb{R}, \delta)$. The non-Euclidean analogues mentioned above demonstrate the relaxation of this assumption in favour of arbitrary geometries.

In principle, the ideas above are not entirely new. The issue of which parametric family to select or which kernel to use in smoothing have been studied extensively in statistics. Our goal in this chapter is to examine these issues from a geometric perspective. This perspective emphasises the geometric assumptions on X which are often made implicitly and without much consideration. Bringing the geometry to the forefront enables us to discover new distances, parametric families and kernels that are more appropriate for data than their commonly used Euclidean counterparts. The benefit associated with the geometric viewpoint is particularly strong in the case of structured data such as text where it is often difficult to motivate the specific selection of distances, parametric families and kernels.

17.3 Geometry selection

We turn in this section to the problem of obtaining a suitable local metric g for a given data space X. Methods for obtaining the local metric may be roughly classified according to three categories: elicitation from a domain expert, estimation from data and axiomatic characterisation. We briefly describe these methods and then proceed to concentrate on the axiomatic characterisation category in the remainder of this chapter.

17.3.1 Geometry elicitation

The most straightforward way to obtain g is to have the statistician or a domain expert define it explicitly. Unfortunately, a complete specification of the geometry is a difficult task as the inner product function g_x is local and needs to be specified at each point $x \in X$ in a smooth manner. Another source of difficulty is that it is not always easy for non-experts to understand what is the meaning or role of the local inner product g_x and specify it accordingly.

The problem of eliciting a geometry is similar to prior elicitation in subjective Bayesian analysis. In order to successfully use domain knowledge in specifying a geometry, a statistician or a geometer needs to interact with a domain expert.

The two experts work as a team with the statistician posing carefully thought-out questions to the domain experts. The responses made by the domain expert are used to obtain a relatively small class of appropriate geometries for use later on in the modelling process.

The following example makes this process more concrete. For simplicity we assume that $X = \mathbb{R}^n$ making the metric g_x a symmetric bilinear positive definite function $g_x : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Through interaction with the domain expert, the metric search can start with a very simple metric and progressively consider more complicated forms. For example, the search may start with a constant diagonal metric $g_x(u, v) = g(u, v) = \sum_{j=1}^{n} g_j u_j v_j$ whose parameters $g_1, \ldots, g_n$ represent the importance of different dimensions and are set by the domain expert.

After specifying the constants $g_1, \ldots, g_n$ we can consider a more complicated forms by eliciting the need for non-diagonal entries in g representing the coupling of different dimensions. Finally, extending the elicitation to non-constant metrics, we can start with a base metric form $g'(u, v)$ and modulate it as necessary in different dimensions according to its position e.g., $g_x(u, v) = \prod_{j=1}^{n} h_j(x_j) g'(u, v)$. The choice of simple modulation functions facilitate their characterisation by the domain expert. For example, modulation functions such as $h_j(z) = \exp(c_j z)$ represent monotonic increase or decrease and can be characterised by eliciting the constants $c_1, \ldots, c_n$. Note that the elicitation process described here results in a well-defined metric i.e. symmetric bilinear positive definite $g_x(u, v)$ that is smooth in x.

It is important to ensure that the elicited geometry lead to efficient computation of (17.3) and (17.4). This can be achieved by limiting the classes of metrics under consideration to include only metrics g leading to closed form expressions (17.3), (17.4). We examine such classes of metrics in Section 17.5.

17.3.2 Estimating geometry from data

An alternative approach to elicitation is to estimate the geometry from data. We start by discussing first the unsupervised learning scenario where the available data $\{x_i : i = 1, \ldots, m\} \subset X$ is unlabelled and then proceed to the case of supervised learning where the available data is labelled $\{(x_i, y_i) : i = 1, \ldots, m\} \subset X \times \{-1, +1\}$.

It is often the case that while $X \subset \mathbb{R}^N$, the space X itself is of much lower dimensionality. For example, images of size 100 × 100 are embedded in $\mathbb{R}^{10^5}$ by vectorising the array of 100 × 100 pixels. Assuming that the images share a certain characteristic such as describing natural scenes or faces, the set X of possible data is a relatively small subset of $\mathbb{R}^{10^5}$. In fact, for many classes of images such as face images or handwritten digits X can be shown to be a smooth low-dimensional subset of $\mathbb{R}^{10^5}$.

Manifold learning is the task of separating the lower-dimensional X from the higher-dimensional embedding space $\mathbb{R}^N$ (Saul and Roweis 2003). Assuming no further information it is customary to consider the metric $g = \delta$ on X that is inherited from the embedding Euclidean space i.e. $g_x(u, v) = \sum u_i v_i$ where $u, v \in$

T_xX are expressed in coordinates of the embedding tangent space $T_p\mathbb{R}^N \cong \mathbb{R}^N$. The resulting distance $d(x, y)$ between two points $x, y \in X$ is the Euclidean length of the shortest curve α connecting x, y that lies completely within X. In contrast to the Euclidean distance, this distance is customised to the submanifold $X \subset \mathbb{R}^N$ and does not consider curves passing through $\mathbb{R}^N \setminus X$ in (17.4).

An alternative approach is to select a metric g from a parametric family $\mathcal{G} = \{g^\theta : \theta \in \Theta\}$ based on maximisation of normalised data volume (Lebanon 2006)

$$\prod_{i=1}^{m} \frac{\mathrm{dvol}(g_{x_i})}{\int_X \mathrm{dvol}(g_x)\, dx} \quad \text{where} \quad \mathrm{dvol}(g_x) = \sqrt{\det g_x}.$$

In contrast to manifold learning, this approach has the advantage that by carefully selecting the metric family $\mathcal{G}$ it is possible to ensure that the obtained metric leads to efficient computation of the distance function and other related quantities (Lebanon 2006).

In the supervised case, the presence of data labels y_i can be used to obtain a geometry that emphasises certain aspects of the data. For example, since the task of classifying or estimating $p(y|x)$ requires significant geometric separation between the two classes $U = \{x_j : y_j = 1\}$, $V = \{x_j : y_j = -1\}$, it makes sense to obtain a metric $g \in \mathcal{G}$ that realises a high degree of separation between u and V. The selected metric g can then be used in constructing a conditional model $p(y|x)$ or a classifier. As in the previous case, careful selection of $\mathcal{G}$ can ensure efficient computing of the distances and other geometric quantities.

17.3.3 Axiomatic characterisations

Axiomatic characterisations employ geometrical tools to single out a single metric, or a family of metrics that enjoy certain desirable properties or axioms. It is remarkable that the characterised geometries are often related to well-known statistical procedures and distances. As a result, the axiomatic characterisation may be used to motivate these procedures from a geometrical perspective. On the other hand, modifying or augmenting the axioms results in new geometries that may be more appropriate for the specific space X under consideration. The next section contains more details on this topic and Section 17.5 discusses it in the context of text data.

17.4 Congruent embeddings and simplicial geometries

The n-dimensional simplex

$$\mathbb{P}_n = \left\{ x \in \mathbb{R}^{n+1} : \forall i \; x_i > 0, \sum_{i=1}^{n} x_i = 1 \right\} \subset \mathbb{R}^{n+1}$$

represents the set of all positive probability distributions, or alternatively multinomial parameters, over $n + 1$ items. In the case of $n = 2$ it is easy to visualise the simplex $\mathbb{P}_2$ as a 2-D triangle shaped surface in $\mathbb{R}^3$ or $\mathbb{R}^2$ (see Figure 17.1). Closely

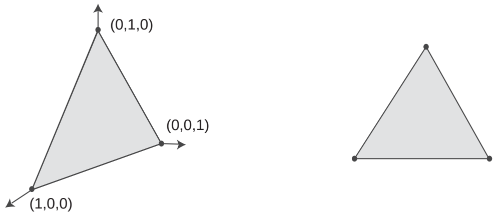

Fig. 17.1 The 2-simplex $\mathbb{P}_2$ may be visualised as a surface in $\mathbb{R}^3$ (left) or as a triangle in $\mathbb{R}^2$ (right).

related to the simplex is

$$\mathbb{R}^n_{>0} = \{x \in \mathbb{R}^n : \forall i \; x_i > 0\}$$

representing non-normalised non-negative measures.

We consider the simplex $\mathbb{P}_n$ and $\mathbb{R}^n_{>0}$ rather than their closures $\overline{\mathbb{P}_n}, \overline{\mathbb{R}^n_{>0}}$ which contain zero components to ensure that they are smooth manifolds. The discussion concerning $\mathbb{P}_n$ and $\mathbb{R}^n_{>0}$ presented here applies to the above closures in their entirety through the use of limiting arguments such as the ones described in (Lafferty and Lebanon 2005).

At first glance, the simplex seems to describe probabilities or statistical models rather than the data space itself. However, many types of structured data can be represented as points in $\overline{\mathbb{P}_n}, \overline{\mathbb{R}^n_{>0}}$ or their products $\overline{\mathbb{P}^k_n}, \overline{\mathbb{R}^{nk}_{>0}}$ using embedding arguments (Lebanon 2005b). We elaborate on this in the next section where we demonstrate various embedding techniques of text documents as distributions and conditional distributions. As a result, an axiomatic characterisation of the geometry underlying the above spaces is directly applicable to modelling the embedded data using the ideas mentioned in Section 17.2.

Čencov's characterisation of the simplex geometry makes use of invariance under congruent embedding by Markov morphisms. We start by informally defining the necessary geometric concepts. Our presentation is based on the relatively simple exposition given by Campbell (Campbell 1986) rather than Čencov's original formulation (Čencov 1982).

A bijective and smooth mapping between two Riemannian manifolds $f : (M, g) \to (N, h)$, defines the push-forward transformation $f_* : T_x M \to T_{f(x)} N$ which maps tangent vectors in M to the corresponding tangent vectors in N. Since tangent vectors correspond to differentiation operators, f_* generalises the well-known Jacobian mapping from real analysis. Using the push-forward map we define the pull-back metric $f^* h$ on M defined as

$$(f^* h)_x(u, v) = h_{f(x)}(f_* u, f_* v).$$

If $f^* h = g$ we say that the mapping f is an isometry between (M, g) and (N, h). In this case, the two manifolds may be considered geometrically equivalent as all their geometrical content including distances, volume, angles and curvature are in perfect agreement.

Definition 17.1 A Markov morphism is a matrix $Q \in \mathbb{R}^{n \times l}$, with $n \leq l$, having non-negative entries such that every row sums to 1 and every column has precisely one non-zero element.

Markov morphisms $Q \in \mathbb{R}^{n \times l}$ are linear transformations which map $\mathbb{P}_{n-1}$ injectively into $\mathbb{P}_{l-1}$. Referred to as congruent embeddings by Markov morphism, these mappings are realised by $x \mapsto xQ$ where $x \in \mathbb{P}_{n-1}$ and $xQ \in \mathbb{P}_{l-1}$ are considered as row vectors. A close examination of the mapping $x \mapsto xQ$ shows that it corresponds to probabilistic refining of the event space $\{1, \ldots, n\} \mapsto \{1, \ldots, l\}$ where the refinement $i \to j$ occurs with probability Q_{ij}.

Proposition 17.1 ((Čencov 1982)) *Let $\{(\mathbb{P}_n, g^{(n)}) : n = 1, 2, 3, \ldots\}$ be a sequence of Riemannian manifolds. Then, any congruent embedding by a Markov morphism acting on this sequence $Q : (\mathbb{P}_k, g^{(k)}) \to (\mathbb{P}_l, g^{(l)})$ is an isometry onto its image if and only if*

$$g_x^{(n)}(u, v) \propto \sum_{i=1}^{n+1} \frac{u_i v_i}{x_i}, \quad x \in \mathbb{P}_n \tag{17.8}$$

where $u, v \in T_x\mathbb{P}_n$ are expressed in coordinates of the embedding $T_p\mathbb{R}^{n+1}$ i.e. $u, v \in \{z \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} z_i = 0\}$.

The metric (17.8) coincides with the Fisher information

$$g_\theta^{(n)}(u, v) = \mathrm{E}_{p_\theta}\{(D_u \log p_\theta)[D_v \log p_\theta]\}$$

where p_θ is the multinomial distribution parametrised by θ. D_u, D_v are the partial differentiation operators corresponding to the tangent vectors u, v. As a result, the metric (17.8) is commonly referred to as the Fisher metric for $\mathbb{P}_n$ and the resulting geometry is certainly the most important example of information geometry.

The axioms underlying the characterisation theorem are sometimes referred to as invariance under sufficient statistics transformations (Amari and Nagaoka 2000). The name comes from the fact that the inverses of Markov morphisms correspond to extracting statistics which are sufficient by definition under the multinomial associated with the rougher event space $\mathbb{P}_n$. The above proposition implies that under Markov morphisms any Riemannian metric different from (17.8) will necessarily transform to a different functional form. This makes it difficult to know that a metric different from (17.8) is the appropriate one since its precise shape depends on the granularity of the event space.

An interesting way to visualise the Fisher metric is to consider the isometry ν between the Fisher information inner product on the simplex and the Euclidean inner product on the positive orthant of the sphere. This isometry $\nu : (\mathbb{P}_n, g^{(n)}) \to (\mathbb{S}_m^+, \delta)$ is defined by $\nu(x_1, \ldots, x_{n+1}) = (\sqrt{x_1}, \ldots, \sqrt{x_{n+1}})$ where

$$\mathbb{S}_n^+ = \left\{ x \in \mathbb{R}^{n+1} : \forall i \ x_i > 0, \ \sum_i x_i^2 = 1 \right\}$$

and δ is as before the metric inherited from the embedding Euclidean space $\delta_x(u, v)$

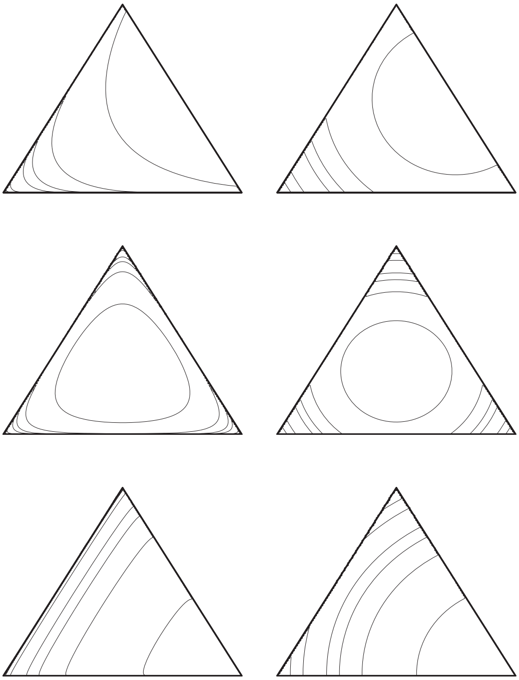

Fig. 17.2 Equal distance contours on $\mathbb{P}_2$ from the upper right edge (top row), the centre (centre row), and lower right corner (bottom row). The distances are computed using the Fisher information (left) and Euclidean (right) metrics.

$= \langle u, v\rangle = \sum_i u_i v_i$. In other words, transforming the probability vector by taking square roots maps the simplex to the positive portion of the sphere where the Fisher metric $g^{(n)}$ becomes the standard Euclidean inner product. As a result, the distance function $d(x, y), x, y \in \mathbb{P}_n$ corresponding to the Fisher metric may be computed as the length of the shortest curve connecting $\nu(x), \nu(y)$ on the sphere

$$d(x, y) = \arccos\left(\sum_i \sqrt{x_i y_i}\right). \tag{17.9}$$

Figure 17.2 illustrates (17.9) on the simplex $\mathbb{P}_2$ and contrasts it with the Euclidean distance function $\|x - y\|_2$ resulting from $(\mathbb{P}_n, \delta)$.

As mentioned in Section 17.2, the metric contains additional information besides the distance function that may be used for statistical modelling. For example, flat surfaces in $(\mathbb{P}_n, g^{(n)})$ are curved in $(\mathbb{P}_n, \delta)$ and vice versa. An interesting visualisation of this can be found in Figure 17.3 which contrasts the standard definition of logistic regression (17.1) which assumes Euclidean geometry with its Fisher information analogue (17.7). The decision boundaries in the non Euclidean case correspond to flat surfaces in the Fisher geometry which are the correct geometric analogue of linear hyperplanes. A similar demonstration may be found in Figure 17.4 which contrasts the decision boundaries obtained by support vector machines (SVM) with

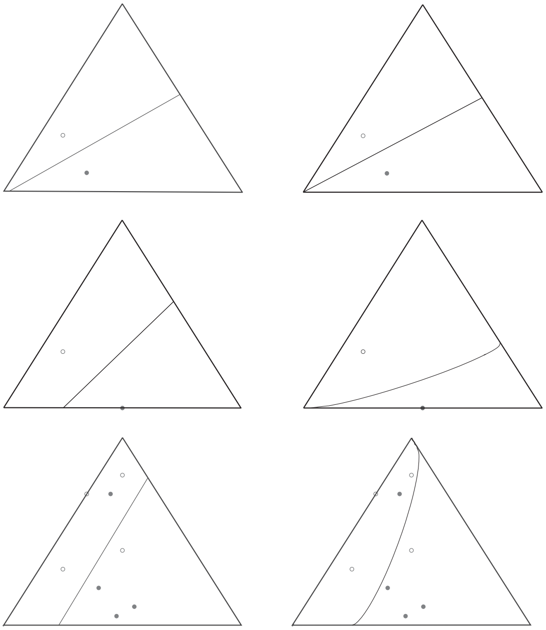

Fig. 17.3 Experiments contrasting flat decision boundaries obtained by the maximum likelihood estimator (MLE) for Euclidean logistic regression (left column) with multinomial logistic regression (right column) for toy data in $\mathbb{P}_2$.

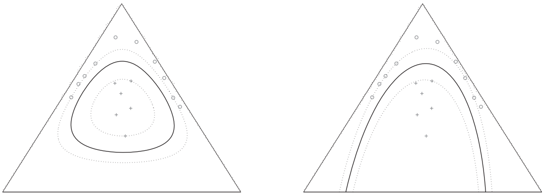

Fig. 17.4 Decision boundaries obtained by SVM trained on synthetic data using the Euclidean heat kernel (right) and the information geometry heat kernel (left).

the Euclidean diffusion kernel (also known as radial basis function or RBF kernel) and with the Fisher geometry diffusion kernel (Lafferty and Lebanon 2005).

Proposition 17.2 ((Campbell 1986)) *Let* $\{(\mathbb{R}^n_{>0}, g^{(n)}) : n = 2, 3, \ldots\}$ *be a sequence of Riemannian manifolds. Then, any congruent embedding by a Markov morphism* $Q : (\mathbb{R}^n_{>0}, g^{(n)}) \to (\mathbb{R}^l_{>0}, g^{(l)})$ *is an isometry onto its image if, and only if,*

$$g^{(n)}_x(u, v) = A(|x|) \sum_i \sum_j u_i v_j + |x| B(|x|) \sum_i \frac{u_i v_i}{x_i} \tag{17.10}$$

where $|x| = \sum x_i$ *and* $A, B : \mathbb{R} \to \mathbb{R}$ *are smooth functions.*

The restriction of $\mathbb{R}^n_{>0}$ to the simplex results in $x \in \mathbb{P}_{n-1} \subset \mathbb{R}^n_{>0}$, $\sum u_i = \sum v_i = 0$

making the choice of A immaterial as the first term in (17.10) zeros out. Similarly, in this case, $|x| = 1$ making the choice of B immaterial as well and reducing Proposition 17.2 to Proposition 17.1.

The extension of Proposition 17.1 to products $\mathbb{P}_n^k$ and $\mathbb{R}_{>0}^{nk}$ corresponding to spaces of conditional distributions and non-negative measures is somewhat more complicated as the definition of Markov morphisms need to be carefully formulated. The appropriate extension characterises the invariant metric on $\mathbb{R}_{>0}^{kn}$ as

$$\begin{aligned} g_x^{(k,n)}(u,v) = A(|x|) \sum_{a,b,c,d} u_{ab} v_{cd} + |x| B(|x|) \sum_{a,b,d} \frac{u_{ab} v_{ad}}{|x_a|} \\ + |x| C(|x|) \sum_{a,b} \frac{u_{ab} v_{ab}}{x_{ab}}, \quad x \in \mathbb{R}_{>0}^{kn} \end{aligned} \tag{17.11}$$

where $u, v \in T_x \mathbb{R}_{>0}^{kn} \cong \mathbb{R}^{k \times n}$, $|x| \overset{\text{def}}{=} \sum_a |x_a| \overset{\text{def}}{=} \sum_{a,b} x_{ab}$, and $A, B, C : \mathbb{R} \to \mathbb{R}$ are smooth functions. See (Lebanon 2005a) for further details and for the analogue expression corresponding to spaces of conditional distributions $\mathbb{P}_n^k$.

17.5 Text documents

Documents are most accurately described as time series $y = \langle y_1, \ldots, y_N \rangle$ containing words or categorical variables $y_i \in V$. We assume that the vocabulary or set of possible words V is finite and with no loss of generality, define it to consist of integers $V = \{1, \ldots, |V|\}$.

The representation of y as categorical time series is problematic due to its high dimensionality and since the representation depends on the document length which makes it hard to compare documents of varying lengths. A popular alternative is to represent the document using its word histogram, also known as the bag of words (bow) representation

$$\gamma^{\text{hist}}(y) \overset{\text{def}}{=} \left(\frac{1}{N} \sum_{j=1}^{N} \delta_{1,y_j}, \ldots, \frac{1}{N} \sum_{j=1}^{N} \delta_{k,y_j} \right) \in \mathbb{R}^{|V|}. \tag{17.12}$$

For example, assuming $V = \{1, \ldots, 5\}$ we have

$$\gamma^{\text{hist}}(\langle 1,4,3,1,4 \rangle) = \gamma^{\text{hist}}(\langle 4,4,3,1,1 \rangle) = \left(\frac{2}{5}, 0, \frac{1}{5}, \frac{2}{5}, 0 \right). \tag{17.13}$$

The histogram representation (17.12) maps documents to the simplex $\mathbb{P}_{V-1}$ but the embedding γ^{hist} is neither injective nor onto. The lack of injectivity is not a serious problem since by definition the histogram representation ignores word order and identifies two documents with the same word contents as the same document. The image of the histogram representation is a strict subset of the simplex containing only vectors with rational coefficients $\text{image}(\gamma^{\text{hist}}) = \mathbb{P}_{V-1} \cap \mathbb{Q}^V$. However, since the $\text{image}(\gamma^{\text{hist}})$ is a discrete set, it makes sense to consider instead the interior of its completion $\text{int}(\overline{\text{image}(\gamma^{\text{hist}})})$ which coincides with the simplex $\mathbb{P}_{V-1}$.

The histogram embedding of text in $\mathbb{P}_{V-1}$ has a clear statistical interpretation. Assuming that text documents are generated by unknown multinomial distributions

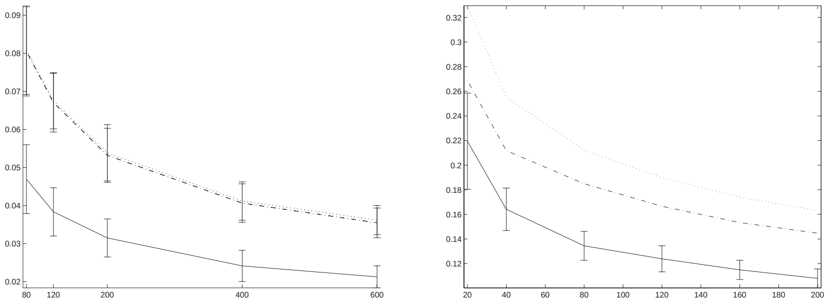

Fig. 17.5 Error rate over a held out set as a function of the training set size (WebKB data). Left: SVM using Fisher diffusion kernels (solid), Euclidean diffusion kernel (dashed), and linear kernel (dotted). Right: Logistic regression using the Fisher geometry (solid), Euclidean geometry (dashed) and Euclidean geometry following L_2 normalisation (dotted). Error bars represent one standard deviation.

$y \sim \text{Mult}(\theta_y)$, we have that the histogram representation is the maximum likelihood estimator for the multinomial parameter $\gamma^{\text{hist}}(y) = \hat{\theta}_y^{\text{mle}}$. Viewed in this way, $\gamma^{\text{hist}}(y)$ is but one possible embedding in the simplex. Other estimators such as the maximum posterior under a Dirichlet prior $\gamma^{\text{map}}(y) \propto \gamma^{\text{hist}}(y) + \alpha$ and empirical Bayes would result in other embeddings in the simplex.

Since the embedded documents represent multinomial distributions, it is natural to invoke Čencov's theorem and to use the Fisher geometry in modelling them. Experiments on a number of real-world text classification datasets indicate that the resulting classifiers perform significantly better than their Euclidean versions. Figure 17.5 contrasts the error rates for the Fisher and Euclidean based SVM (using diffusion kernels) and logistic regression. Further details and additional results may be found in (Lebanon and Lafferty 2004, Lafferty and Lebanon 2005, Lebanon 2005b).

In the case of embedded text documents it is beneficial to consider metrics g that are not symmetric i.e. $g_x(u,v) \neq g_{\pi(x)}(\pi(u),\pi(v))$ where $\pi(z)$ permutes the components of the vector z. Intuitively, the different components correspond to different vocabulary words which carry a non-exchangeable semantic meaning. For example, stop words such as '`the`' or '`a`' usually carry less meaning than other content words and their corresponding component should influence the metric $g_x(u,v)$ less than other components. Similarly, some words are closely related to each other such as '`often`' and '`frequently`' and should not be treated in the same manner as two semantically unrelated words. Some progress along these lines is described in (Lebanon 2006) where the invariance axioms in Proposition 17.1 are extended in a way that leads to characterisation of non-symmetric metrics.

While the histogram embedding provides a convenient document representation and achieves reasonable accuracy in text classification, it is less suitable for more sequential tasks. Since it completely ignores word ordering e.g., (17.13), it is not suitable for modelling the sequential progression of semantics throughout documents. A reasonable alternative is to assume that different words y_s, y_t in the document y are generated by different multinomials θ_s, θ_t, where $\theta_s \to \theta_t$ as $s \to t$ i.e., close

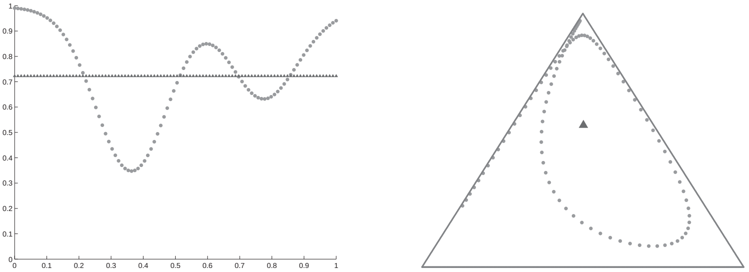

Fig. 17.6 Documents over $V = \{1, 2\}$ and $V = \{1, 2, 3\}$ can be represented as curves in the simplex $\mathbb{P}_1$ (left) and $\mathbb{P}_2$ (right). The horizontal line (left) and triangle (right) represent the global histogram representation of the document.

words are generated by similar multinomials. The local likelihood estimator for this semi-parametric model uses local smoothing to estimate the locally weighted bag of words (lowbow) or multinomial models (Lebanon *et al.* 2007). Replacing the discrete location parameter $1, \ldots, N$ within documents by a continuous interval I the local estimator provides a smooth curve in the simplex $\mathbb{P}_{V-1}$ representing the smooth transition of the local multinomial models $\{\theta_t : t \in I\}$ throughout the document. For example, the curves corresponding to the documents $z = \langle 1,1,1,2,2,1,1,1,2,1,1 \rangle$ and $w = \langle 1,3,3,3,2,2,1,3,3 \rangle$ over $V = \{1, 2\}$ and $V = \{1, 2, 3\}$ (respectively) are illustrated in Figure 17.6.

The resulting curve $\gamma(y) : I \to \mathbb{P}_{V-1}$ embeds documents in an infinite product of simplices $\mathbb{P}^I_{V-1}$. Probabilistically, the curve $\gamma(y) \in \mathbb{P}^I_{V-1}$ represents a conditional multinomial distribution. Using the characterisation (17.11) we obtain a geometry for use in sequential modelling of the curves $\gamma(y)$ (Lebanon *et al.* 2007, Mao *et al.* 2007). Experiments reported in (Lebanon *et al.* 2007) confirm the practical benefit of using the sequential embedding in $\mathbb{P}^I_{V-1}$ using the characterised geometry.

17.6 Discussion

Modelling high-dimensional structured data is often poorly understood. The standard approach is to use existing models or classifiers as black boxes without considering whether the underlying assumptions are appropriate for the data. In particular many existing popular models assume, either explicitly or implicitly, that the data space X is well characterised by the Euclidean geometry. Explicitly obtaining a geometry for the data space X through elicitation, learning from data, or axiomatic characterisation, enables the construction of more accurate and data-specific models.

In this chapter, we discussed the role of data geometry in statistical modelling and described several approaches to obtaining a geometry for the data space. Using the embedding principle and Čencov's theorem we describe several axiomatic characterisations of the geometry of X. These geometries are closely related to the Fisher information and provide an interesting connection to the theory of information geometry and its relation to asymptotic inference. Furthermore, experimental evidence demonstrates that the characterised geometries lead to geometric generalisations of popular classifiers which provide state-of-the-art performance in modelling text documents.

References

Amari, S.-I. and Nagaoka, H. (2000). *Methods of Information Geometry* (American Mathematical Society, Oxford University Press).

Campbell, L. L. (1986). An extended Čencov characterization of the information metric. In *Proc. of the American Mathematical Society* **98**(1), 135–41.

Čencov, N. N. (1982). *Statistical Decision Rules and Optimal Inference* (Providence, RI, American Mathematical Society).

Kass, R. E. and Voss, P. W. (1997). *Geometrical Foundation of Asymptotic Inference* (New York, John Wiley & Sons).

Lafferty, J. and Lebanon, G. (2005). Diffusion kernels on statistical manifolds, *Journal of Machine Learning Research* **6**, 129–63.

Lebanon, G. (2005a). Axiomatic geometry of conditional models, *IEEE Transactions on Information Theory* **51**(4), 1283–94.

Lebanon, G. (2005b). Riemannian geometry and statistical machine learning. PhD thesis, School of Computer Science, Carnegie Mellon University.

Lebanon, G. (2006). Metric learning for text documents, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(4), 497–508.

Lebanon, G. and Lafferty, J. (2004). Hyperplane margin classifiers on the multinomial manifold. In *Proc. of the 21st International Conference on Machine Learning* (ACM press).

Lebanon, G., Mao, Y. and Dillon, J. (2007). The locally weighted bag of words framework for document representation, *Journal of Machine Learning Research* **8**, 2405–41.

Mao, Y., Dillon, J. and Lebanon, G. (2007). Sequential document visualization, *IEEE Transactions on Visualization and Computer Graphics* **13**(6), 1208–15.

Saul, L. and Roweis, S. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds, *Journal of Machine Learning Research* **4**(2), 119–55.

Spivak, M. (1975). *A Comprehensive Introduction to Differential Geometry*, Vol 1–5, (Publish or Perish).

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing* (Boca Raton, Chapman & Hall/CRC).