Principal Curve Sonification

T. Hermann, P. Meinicke and H. Ritter
Faculty of Technology · University of Bielefeld · D-33615 Bielefeld
e-mail: {thermann,pmeinick,helge}@techfak.uni-bielefeld.de

ABSTRACT

This paper describes a new approach to render sonifications for high-dimensional data, allowing the user to perceive the "main" structure of the data distribution. This is achieved by computing the principal curve of the data set, which is a trajectory that passes through the "middle" of the data and allows to define a time order on the data points. The sonification can be imagined as the time-variant auditory scene, perceived while moving along the principal curve. In this paper a method for computing principal curves is presented, the sonification concept is introduced and some sonification examples are given.

Keywords

Sonification, Exploratory Data Analysis, Principal Curves

INTRODUCTION

The detection of hidden structures in large high-dimensional data sets is the goal of the research field *data mining*[3]. As machine learning methods here still perform poorly in comparison to the pattern recognition capabilities of the human brain and its sensory systems, a senseful strategy consists in providing an interface for inspection of high-dimensional data that reveals as much as possible of the data structure. This is the focus of research in Exploratory Data Analysis[15]. Methods like multidimensional scaling[2], principal component analysis (PCA) and projection pursuit[5] allow dimensionality reduction while maintaining the main structure of the data according to some optimality criterion. However, the results are mostly presented visually, neglecting that we also have a highly developed auditory perception channel. In this paper Principal Curve Sonification (PCS) is introduced as one technique to present multidimensional data acoustically, so that the user can listen to the "main" structure of the data set.

Principal Curves (PC) can be used to create nonlinear one-dimensional descriptions of data. They are well suited as a starting point for a general data presentation method as they can be defined for all types of multivariate data distributions. PCS provides an acoustic front end to inspect the data distribution in relation to its principal curve as well as the curve properties themselves. Since the PC is embedded in the data space, it is as difficult to visualize as the data itself. Here the PCS offers an alternative by giving information about its evolution in space, especially its curvature. The PCS is the auditory scene perceived while moving along the principal curve, integrating both data point attributes (low level) and local attributes like e.g. probability density estimations or local variance estimations (high level). Parameter Mapping[14] as well as Model-Based Sonification[8] are used to present the different types of information while the system user interactively controls the movement through the data space along the PC. As multiple auditory streams can be presented at the same time, a variety of information sources can be integrated to form a rich auditory display.

This paper is structured as follows: the next section presents principal curves and a method to compute them for general data sets. The following section summarizes the sonification design, the presented information and sound synthesis issues. Then some sound examples are presented and application fields of the PCS are pointed out. The paper ends with a conclusion, where also some possible extensions are proposed.

PRINCIPAL CURVES

Regarding data sets which are sampled from high-dimensional distributions in 'real world situations', we often realize that the intrinsic dimensionality is much lower than that of the embedding space. Therefore often principal component analysis (PCA) is the starting point to achieve some dimensionality reduction. This approach can be motivated from a minimization of a cost function which measures the squared distance of the data w.r.t. a linear manifold. The data are projected onto this manifold to achieve a dimensionality reduced representation, which captures the main variation of the data. However, while maintaining the dimension of the projection space, often the data set can be much better approximated by allowing a higher flexibility of the manifold, which can be achieved by a nonlinear model. Principal curves are continous one-dimensional manifolds that approximate the data in this sense and thus pass through the "middle" of a d-dimensional data set. PC's have been successfully applied to solve practical problems, like the alignment of magnets of the Stanford linear collider [6] or Ice Floe Identification in Satellite Images [1]. Recently PC's have also been presented within the framework of statistical learning theory [16].

Let $\vec{f}(\lambda)$ be a parametrization of a curve. All vector elements $f_i(\lambda)$ are continous functions of a single variable λ, which parametrizes the curve. A unique and natural parametrization can be given in terms of the arc length.
A projection index $\lambda_f : R^d \to R$ can be defined as

$$\lambda_f(\vec{x}) = \sup_{\lambda} \left\{ \lambda : \|\vec{x} - \vec{f}(\lambda)\| = \inf_{\nu} \|\vec{x} - \vec{f}(\nu)\| \right\}, \tag{1}$$

which is the largest value of λ for which the curve is closest to the point $\vec{x}$.
Now Hastie and Stuetzle defined the principal curve of a continous data distribution $p(\vec{x})$ by the self-consistency property[6], which says that the mean of all data projecting on a point $\vec{f}(\lambda)$ just is $\vec{f}(\lambda)$:

$$E(X \mid \lambda_f(X) = \lambda) = \vec{f}(\lambda) \;\; \forall \;\; \lambda. \tag{2}$$

Normally, we don't know the continous probability distribution, since we only have a finite sample of that distribution. Unfortunately the definition of the principal curves cannot be taken for discrete data sets without modification, as there is usually only one data point in the data set projecting to a certain λ and thus a curve, that passes through all data points will minimize the squared distance. This problem can be avoided by introducing a regularization constraint, which allows to control the complexity of the principal curve and thus prevent the curve from overfitting the data. This can e.g. be done by restricting the length of the curve [9] or by a smoothness constraint[6], which can be implemented by considering all data points in a range σ of $\vec{f}(\lambda)$ to compute the local mean.

The Principal Curve Algorithm

We have used a similar algorithm like Hastie and Stuetzle[6] to compute the PC of a n-point data set. For computational simplicity, we only consider polygonal lines with a limited number of vertices and thus represent a PC by the ordered set of vertex coordinates. Figure 1 illustrates a PC for a 2d data distribution. Obviously, the PC is capable to catch the main variation of the data.

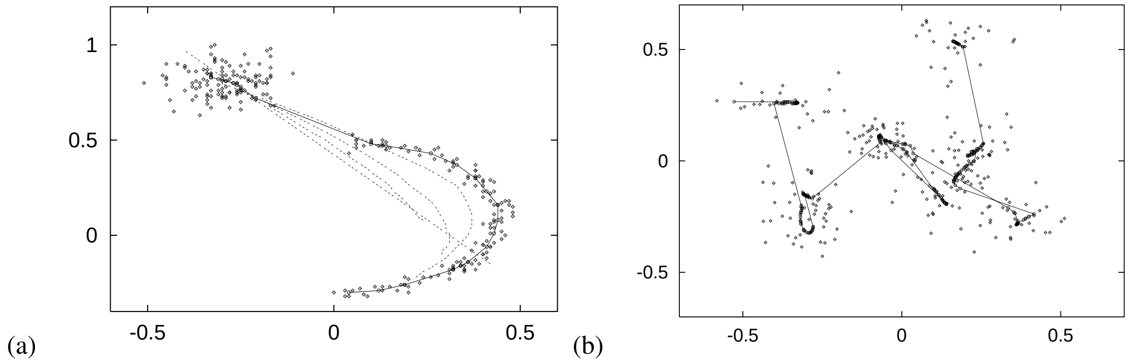


Figure 1: (a) Principal Curve of a 2d toy data set, where the points are sampled from a 2d Gaussian and a noisy half-circle. The solid line shows the PC. The polygon vertices are marked with circles. The dashed curves show the progress of the adaptation process for various values of σ, the straight line is the first principal axis of the data. (b) Principal Curve of a 9d data set with clusters on the 10 vertices of an 9-dimensional tetrahedron. The plot shows the first two principal components. The PC passes once through the middle of each cluster (note that the 2d projection cannot resolve all 10 clusters).

Initialization of the PC is done by setting the vertices equally spaced along the first principal component, between the minimal and maximal projection position of the data. For computation of the PC, the vertex coordinates are iteratively improved doing projection steps and expectation steps.

The **projection step** computes the projection parameter $\lambda_f(\vec{x_i})$ for all data points $x_i, i = 1 \ldots n$ according to the definition (1). Furthermore the minimal distance and the projection point coordinates are calculated. To monitor the degree of approximation, the mean squared distance to the PC is computed for the training data set and an independent test set of the same distribution.

In the **expectation step**, the reference points are adjusted to reduce the empirical error. This is done by shifting all reference points to the mean of the subset of data points that project to the neighborhood of each reference point. The necessary computations are implemented as a local kernel regression on the projection indices using a Gaussian kernel with width σ.

To our experience, if the curve is fitted for too small σ, in some cases the data is badly approximated in a way that can be described as topological disorder. This can be avoided by applying a deterministic annealing technique[13]: starting with a high value of σ, the smoothing parameter is gradually decreased during optimization. A suitable start value for σ^2 is given by the largest eigenvalue of the data covariance matrix. For large values of σ, the curve is contracted to a single point at the data set mean. With decreasing σ, the curve gradually adapts to the data. After each expectation step, σ is reduced by a factor of 0.95 until the mean squared distance in the test set begins to increase due to overfitting. To apply the deterministic annealing, we have slightly modified the projection step of Hastie and Stuetzle, allowing the data also to project onto a linear continuation of the PC beyond the first and last vertex. This keeps the PC's length in the order of magnitude of the data variance and thus avoids a contraction of the PC, especially for large values of σ.

The learning speed can be accelerated drastically by starting with a small random sub-sample of the data set, and therefore with only few polygon vertices of the PC and increasing the sample size with decreasing σ. This makes the algorithm well suited also for large data sets.

PRINCIPAL CURVE SONIFICATION

Many sonifications in literature are a kind of parameter mapping sonification: analogous to a scatter plot, data attributes are mapped to sound event (tone) attributes like time stamp, volume, pitch, timbre, etc[10]. The evolution in time, which is besides pitch the best perceived quality, is given an arbitrary role. Therefore, even the same data sonified using a different mapping may not be recognized as having the same structure. Principal Curve Sonification aims to define a "natural" usage of the time axis, in order to perceive the main structure. As the principal curve is a one-dimensional manifold it maps directly to the one-dimensional time axis. Defining the model of data points lying in an euclidean vector space and thinking of the sonification as the time-variant data "soundscape" while moving along the principal curve, a helpful key for interpretation of the sonification is given. Taking this approach, we will now discuss the kind of information integrated into PCS and how it is sonified using different auditory streams. For sonification, both parameter mapping and model-based sonification are used.

Model-Based Sonification

Model-Based Sonification was proposed as an extension to prior used sonification methods[8]. Our auditory system is optimized to extract valuable information from the sound we perceive in our real world. This sound is always a consequence of physical processes. Therefore the main idea is to take this over to data sonification by defining a data material or a scenery from a data set by establishing "physical laws" that permit a vibrational reaction of the data material to excitations. The rendering of the sonification is done by exciting the data material and audification of the dynamic reaction. This yields a form of interaction which is familar to every human user who strikes, hits or shakes a physical object. It especially exploits our capabilities to relate our actions with acoustic feedback of the data material. Thus the data more or less directly becomes the sounding instrument, which is examined, excited, or played by the user. Knowledge about the setup furthermore makes interpretation of the sound easier, because sound properties are correlated with a situation context. In PCS, the sonification model consists in a moving-in-space scenery. The data is not taken to instantiate a data material, but it controls the layout of the scenery and its acoustic properties.

Moving in Virtual Spaces

In PCS, the scenery is a high-dimensional euclidean space in which the listener moves along a path given by the principal curve. Moving gives us the chance to listen for relations between data from different perspectives. As different 2d views of a visual scene may improve our perception of structures in three dimensions, moving in acoustic scenes can enrich our understanding of data relations in high-dimensional spaces. Therefore, an authentic virtual soundscape offering acoustic properties, that we are familiar with, should be aimed with PCS. This is of course a demanding task with currently available computation power. A spatial audio synthesis engine is currently in work, so the demonstrations only use the left/right pan to add localization cues. When moving on a trajectory, we define the viewing direction to be parallel to the tangent. In addition, we have to define the orientation of the listener. This can be done using the curvature of the principal curve, which can be defined by the second derivative (acceleration) of the curve. Tangent and acceleration vector are used to span the hearing plane. The localization of the data points is done by projecting them onto this plane. The distance r between listener and point source determines the perceived volume of the source sound, using a $1/(r+\epsilon)^2$ law. The ϵ prevents the volume from diverging for small r. Our limited computational power prevents us from modeling all point sources for the whole sonification. Instead we trigger each data point, when the listener reaches its projection index $\lambda_f(\vec{x})$ while tracing the curve. Next we will resume what information should be made available by the sonification.

Sonification Design

Various kinds of information can be presented using sonification, simply by using different acoustic elements. The presented information can be organized into distinct categories: complex observables, attributes of an individual data point and properties of the PC. Table 1 shows an overview.

This information is presented acoustically using different sonification methods. It must be mentioned here, that there are

	Attribute	**Method**	**Sonification**
Observables:	Local probability density estimations	kernel density estimation	time-variant oscillator
	Local intrinsic data dimensionality	local PCA	-
	Average distance from PC	locally weighted distance mean	time-variant oscillator pitch
Data Points:	Relative orientation to the listener	projection on hearing space	spatialization of tick sounds
	Distance from PC	distance law	volume of tick sounds
	Data Features (e.g. class label)	scaling	frequency of tick sounds
PC properties:	Velocity of Listener on PC	user adjusted	wavetable sound volume
	Spatial orientation of the listener	hearing plane	sound localization
	Local curvature of PC	acceleration vector	sound pitch

Table 1: Information presented with PCS.

multiple ways of choosing the sonification. We also thought about a user-adjustable graphical interface to link information types with sound elements. However, our experience is, that the user is overloaded with this flexibility and further, that learnability of the sonification decreases. Therefore, we select a suitable sonification design, which surely is a subjective choice, but that allows the PCS user to familiarize with the usual soundscape and to develop a higher sensibility. The user should be able to influence only very few parameters like the volumes of independent acoustic streams, which are reset to default values on each start of the PCS.

The selection of the acoustic elements can be motivated with the following reasoning: the properties of the data distribution can be sorted by their relevance for understanding the structure. The acoustic properties once chosen can be sorted by their perceptibility. Connecting corresponding entries of these two lists suggests an assignment. However, the design of the sonification should also consider user expectations. For instance, to represent a higher distance by a lower volume would match our expectations and thus make interpretation easier.

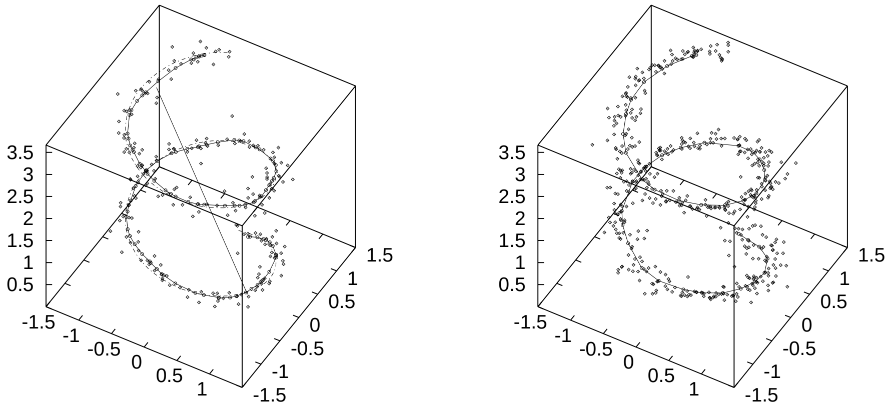
The individual data points are presented using the moving-in-space scenery. One way to sonify the data is to generate a continous sound for each data point. Thus, the relative position to the PC is perceived by a change of volume when passing the data point. Additionally, while moving towards the sound source, the Doppler effect would influence pitch and would be a useful cue for the perception of spatial relations. However, for each position on the PC the distance to the complete data set must be computed, which demands a lot of computation power. So for a first step, as an alternative, we use a kind of “Geiger tick”, which is emitted whenever the listeners position passes a projecting point onto the PC. The Geiger tick’s acoustic properties are determined by the relative position of the data point to the listener, similar to real sound propagation: with increasing distance r, volume diminishes with $1/(r+\epsilon)^2$, and the spatial orientation is chosen w.r.t. the plane spanned by tangent and acceleration vector. Further information can be put into the sound of the tick: the pitch can be driven by a user selected data point feature. Thus it can be perceived, e.g. if classes mix in a cluster or how an attribute varies while moving along the PC. The tick is realized by an exponentially decreasing sine wave.

Time-variant oscillators are used to present locally averaged density and distance observables. The oscillators frequency is given by $f = f_0 + f_d p(\vec{x})$ where $p(\vec{x})$ is the estimated probability density. Additionally the amplitude of the oscillator is modulated by the local average distance. The PC-specific properties are presented in an acoustic stream that resembles the sound of a moving vehicle. Large curvatures result in a higher pitched sound, while the volume of the sound corresponds to the velocity of the listener.

EXAMPLES

For the first example, a noisy spiral data set in a three dimensional space is taken. The embedding dimension is still so low that a visualization (see fig. 2,(a)) allows to perceive the structure. The principal curve obviously follows the data. The sonifications are available at [7]. For illustration, short time fourier transforms (STFT) of some sonifications are shown in figure 3. In this example it is easy to hear that the local density (pitch of the time-variant oscillator and tick rate) is rather constant during the whole sonification. The nearly constant curvature lets expect a kind of circular trajectory.

The next two examples show that the clustering of data can be perceived as well. The first data set consists of 10 clusters embedded in a 9d data space and located at the 10 vertices of a 9d hyper-tetrahedron. The PCS (examples at [7], STFT in fig. 3,(b)) gives very directly the information about the number and relative size of the clusters. This is easier perceived than from the 2d plot in figure 1,b). The second example is the iris data set [4], a classic 4d bench-mark data set, which consists of three clusters, cluster 1 being well separated from the others, clusters 2 and 3 having a small overlap. We computed the PC without the class attribute and let the class attribute control the pitch of the Geiger ticks. The separation of class 1 and 2 as well as the small overlap between class 2 and 3 becomes audible, demonstrating that PCS is a highly promising pre-stage for cluster analysis.



(a) (b)

Figure 2: Principal Curve of noisy spiral example: computed PC (solid line)) with reference points positions (circles) after 5 adaption steps. The generator curve is shown in a dotted line. (b) modified noisy spiral: the noise variance is periodically modified along the spiral. This substructure is easily overseen in the visual display.

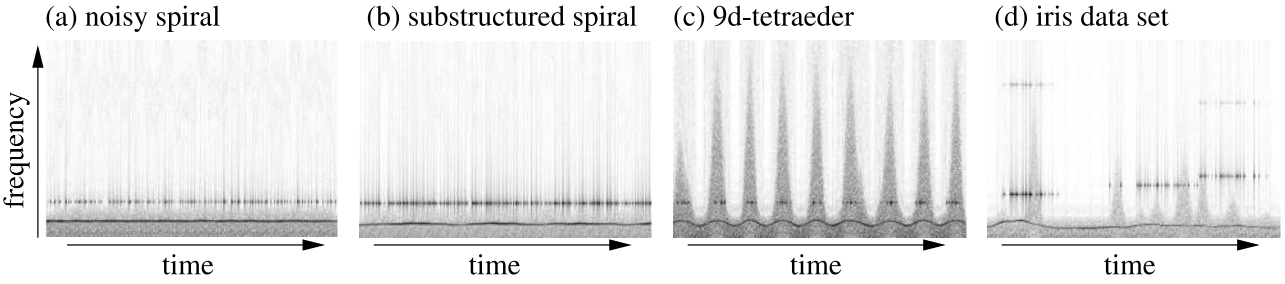


Figure 3: Short Time Fourier Transforms (STFT) of the PCS examples.

The last example is a modified noisy spiral, where the noise variance varies periodically along the spiral. This can hardly be perceived from a 3d plot (see fig. 2, (b)) and is thus an example of a regularity that is likely to remain unnoticed in a visual display. In contrast, it is easily detected with the sonification: the variation is very clearly perceived from the pitch and level modulation of the time variant oscillator which sonifies the local density.

To check the hypothesis that PCS can facilitate the detection of such structures in data, we carried out an experiment with human subjects. The task consisted in the detection of the number of noise variance modulations on the noisy spiral data (shown in fig 2, (b)). Controled variables were the presentation type (visual/auditive/both) and the intensity of the modulation. We evaluated for 15 subjects their performance on 90 data sets, measured by the relative error, which is the number of wrong answers under each condition, divided by the total number of wrong answers of a participant. We found a significant reduction of the relative error on addition of PCS. Additionally the average processing time decreased by 31 %.

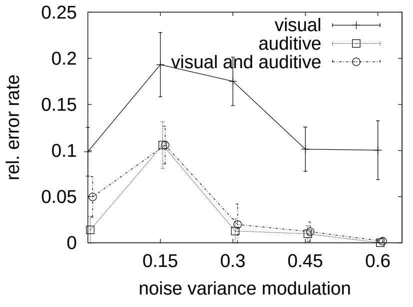


Figure 4: Results of the experiment: the relative error decreased significantly ($\alpha = 1\%$) on addition of PCS. The error bars show the standard deviation.

CONCLUSION

We have presented the new concept of Principal Curve Sonification as a means to perceive the structure of multidimensional data sets. The main motivation was to find a natural meaning for the time axis while presenting data acoustically. This was achieved by developing a moving-in-space model where the time corresponds to the position on a trajectory. The Principal Curve therefore is an intuitively sensible choice for this trajectory as it is a 1d-manifold. In comparison to other choices, like the first principal component axis as the time axis, it is more flexible and thus it can extract more structure from the data.

This sonification idea, to find a natural representation for the time axis, can be carried over to other data approximation methods like 1d self-organizing maps (SOM)[12]. Since it can be shown that the 1d SOM approximates a principal curve[11], the above sonification principles can be applied with only minor modifications.

The experiments show that PCS can provide a useful presentation of the data: the clustering of the data is easily perceived and structures that are not easily recognized visually can be detected. However, we believe that the potential of PCS goes well beyond this. With PCS, data distributions in multidimensional spaces can be sonified in a rather short time of some seconds, and thus they can be compared by a trained listener. We currently apply this idea to medical data analysis, where structures must be searched in very long time series. With this data, a simple time compression for direct playback destroys the relevant information. So we produce a high-dimensional data distribution from several slices of the time series for each patient, which is then sonified using PCS. This allows the medical expert to learn to distinguish the patient's state from the acoustic patterns of the PCS and to draw diagnosis relevant information from that.

Our current sonification is still open for additional acoustic streams which may further enrich the auditory display, whereas visual displays would become overloaded. For instance, local intrinsic dimensionality estimations can be integrated into the auditory scenery. This can be done by computing the spectrum of the local covariance matrix for the neighbourhood of each data point and use this to drive the time-evolution of the ticks' sound. The explorations of such extensions and their evaluation within further real world applications will be the focus of future research.

REFERENCES

1 J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87:7–16, 1992.

2 D. R. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

3 U. M. Fayyad et al., editor. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

4 R. A. Fisher. UCI repository of maschine learning databases. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/.

5 J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23:881–890, 1974.

6 T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

7 T. Hermann. Principal Curve Sonification - Examples http://www.techfak.uni-bielefeld.de/∼thermann/projects/index.html, 1999.

8 T. Hermann and H. Ritter. Listen to your Data: Model-Based Sonification for Data Analysis. In M. R. Syed, editor, *Advances in intelligent computing and mulimedia systems*. Int. Inst. for Advanced Studies in System Research and Cybernetics, 1999.

9 B. Kegl, A. Krzyzak, T. Linder, and K. Zeger. Learning and design of principal curves. *submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. URL=http://magenta.mast.queensu.ca/∼linder/psfiles/KeKrLiZe97.ps

10 G. Kramer, editor. *Auditory Display - Sonification, Audification, and Auditory Interfaces*. Addison-Weslay, 1994.

11 Filip Mulier and Vladimir Cherkassky. Self-organization as an iterative kernel smoothing process. *Neural Computation*, 7(6):1165–1177, 1995.

12 H. Ritter, T. Martinetz, and K. Schulten. *Neural Computation and Self-Organizing Maps. An Introduction*. Addison-Wesley, Reading, MA, 1992.

13 K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.

14 C. Scaletti. Sound synthesis algorithms for auditory data representations. In G. Kramer, editor, *Auditory Display*. Addison-Wesley, 1994.

15 D. W. Scott. *Multivariate Density Estimation*. Wiley & Sons, 1992.

16 A. J. Smola, S. Mika, and B. Schölkopf. Quantization functionals and regularized principal manifolds. *NeuroCOLT2 27150*, 1989.