Combining Visual and Auditory Data Exploration for finding structure in high-dimensional data

Thomas Hermann
Faculty of Technology
Bielefeld University, Germany
thermann@techfak.uni-bielefeld.de

Mark H. Hansen
Bell Laboratories
Murray Hill, New Jersey
cocteau@bell-labs.com

Helge Ritter
Faculty of Technology
Bielefeld University, Germany
helge@techfak.uni-bielefeld.de

Abstract

We consider the combined use of visualization and sound for uncovering important structure in high-dimensional data. Our approach is based on Markov chain Monte Carlo (McMC) simulations. McMC is a popular computational tool for making analytical inferences from complex, high-dimensional probability densities. Given a particular target density p, we simulate a Markov chain that has p as its stationary distribution. We propose a new tool for exploratory data analysis based on an audio representation of McMC output. Several audio streams provide us with information about both the behavior of the Markov chain as well as characteristics of the target density p. We apply this method to the task of identifying structures in high-dimensional data sets by taking p to be a nonparametric density estimate.

In this paper, we present a detailed description of our sonification design and illustrate its performance on test cases consisting of both synthetic and real-world data sets. Sound examples are also given.

1 Introduction

The need of methods for "intelligent" processing of data is rapidly growing with the increasing accumulation of data sets of often very high dimension and size. One chief goal of is the detection of structures or patterns in such data. A very similar task is routinely solved by the sensory processing systems of animals and humans: our ability to segment the high-dimensional and very complex spatio-temporal signals from our sensors into stable "objects" and to effortlessly discern many kinds of causal relationships among the so-derived entities constitutes a superb example of a highly sophisticated "datamining system" evolved by nature over millions of years.

This has motivated the approach of exploratory data analysis, in which the goal is to render the original data in a suitably transformed way so that we can invoke our natural pattern recognition capabilities in order to search for regularities and structures. To date, this approach has been elaborated with an almost exclusive focus on our visual modality and there have been developed a plethora of data visualization techniques, such as Multidimensional Scaling, Projection Pursuit or Self-Organizing Maps, which all aim at creating a low-dimensional image of the original data [**?**] that is better matched to the capabilities of our vision system which has been evolved to detect objects in a three- or lower dimensional space.

Here, we argue that exploratory data analysis may benefit significantly from combining these visualization techniques with methods for the construction of data mappings *in the auditory domain*. Similarly as in visualization, such mappings must be designed in a way to maximally match the transformed data to the natural perception capabilities of our auditory system, which is capable of very fine discriminations in natural sounds and, similar to vision, can detect even

subtle "auditory objects" which can be very hard to discriminate for current automated methods (the extreme difficulty to achieve the goal of speaker-independent speech recognition in the presence of noise provides perhaps the most telling example of that point).

This approach can build on methods developed in the discipline of *sonification* which has been emerging during recent years and which is concerned with the investigation of techniques to render auditory presentations from data [3]. In this paper, we propose an approach which creates one or several auditory streams that encode "'interesting" local features of a high-dimensional data distribution in a form that can be very flexibly "tailored" to match the sensory discrimination cababilities of our auditory system. The basis for our approach is the method of Monte Carlo sampling, which generates a sequence of sampling points whose density approximates the density of the sampled data set. Since this method yields directly a "'serial representation" of the underlying data density in the form of a Markov chain, it provides a more natural and convenient starting point for sonification than, e.g., the widely used kernel density estimation methods. We exploit this feature with the concepts of "Auditory Grains" which provide a direct sonification of the Monte Carlo steps, and "Auditory Information Buckets", which collect the local statistics of the Markov chain and transform it into a convenient auditory format, allowing to make an interesting range of local distributional features of the underlying data set audible and to implement an analog of a zooming mechanism for perceptual resolution control in the auditory domain.

We combine the resulting auditory display with standard visualization methods, such as PCA-based multidimensional scaling and the Sammon map. The spatial display allows to control and explore the auditory display in several interesting ways, e.g., by positioning a "virtual listener" in the projected space, or by spatial clipping of the McMC sampling to a vicinity of a selected location. Conversely, the auditory information can be used to change the view point for the visual projection. In this way, the simultaneous availability of the spatial (visual) and the auditory display can mutually support their interactive exploration. While at this point this is an initial demonstration only, it clearly shows the promise of combining vision and audition for exploratory data analysis.

The plan of the paper is as follows. In Sec. 2 we explain the Markov chain Monte Carlo (McMC) approach underlying our method. Sec. 3 explains the different sonification techniques building on that approach. In Sec. 4, we discuss the formation of auditory streams, explain the construction of McMC-based "sonification maps", provide some examples (Sec. 6). Sec. 7 contains the conclusions and prospects for future work.

2 McMC simulation

Traditional Monte Carlo techniques use an independent sample of data drawn from a target density p to estimate various features of p. In many practical settings, the density under study is too complex to be used directly and we are forced to rely on asymptotic approximations, numerical integration or McMC methods. The idea behind McMC is that we generate a sequence $\{x_0, x_1, x_2, \dots\}$ that has as its stationary distribution the target density p. In this section, we present a simple McMC scheme known as the Metropolis algorithm [1]. We also illustrate how the output from this simulation can be used to infer properties of p. In the statistics literature, most applications of McMC are associated with so-called Bayesian models. In this case, the variable x is a vector of parameters in a probability model and p is a *posterior distribution* for x. The characteristics of p relate directly to the uncertainty present in the components of x. McMC can be applied more generally, however, and throughout this section we refer somewhat generically to a density p.

The Metropolis algorithm generates a sequence of points x_t by drawing propositions from a *jumping distribution*, $J(x_b|x_a)$, starting with x_a. J is required to be symmetric; or $J(x_b|x_a) =$

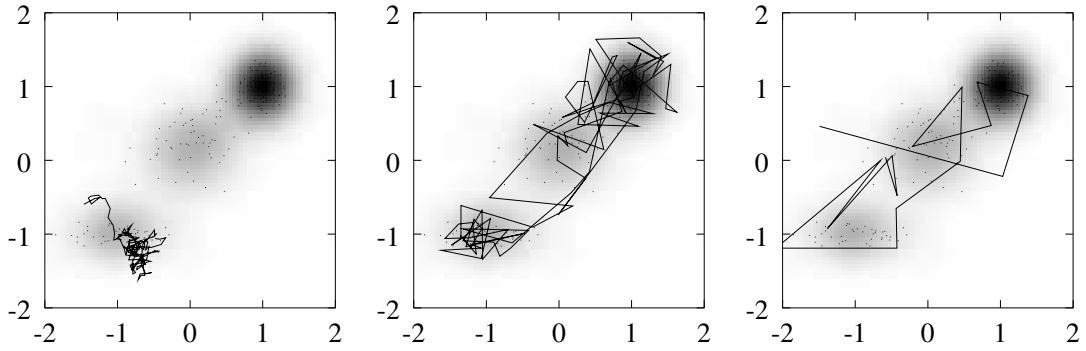

Figure 1: McMC random walk in a 2d distribution with 3 clusters. Grey values represent probability density, data points are plotted with points. 200 McMC steps are shown as line for jumping distribution with variance (a) 10 %, (b) 80 % , (c) 400% of the data set variance. (a) shows low mixing, (c) has only few accepted moves.

$J(x_a|x_b)$ for all possible arguments. (see [5] for details). Consider we have drawn a proposition x^* from $J(x^*|x_{t-1})$ Now, acceptance of the proposition is made dependent upon the *acceptance ratio*

$$r = \frac{p(x^*)}{p(x_{t-1})} \tag{1}$$

such that x_t is set to x^* only with probability $\min(r, 1)$; otherwise, the proposition is rejected and we remain at $x_t = x_{t-1}$. Such a sequence of samples converges to p [1].

The qualitative properties of this Markov chain depends on J. For example, suppose we let J be a Gaussian distribution with variance-covariance matrix $\sigma^2 I_d$, where I_d is the $d \times d$ identity matrix. Then, if σ^2 is small compared to the variance of p, the probability that our chain moves between the different modes of p is small; hence we remain near the same mode for a long time. On the other hand, if σ^2 is very large, the acceptance ratio for each proposed move tends to be small and we rarely leave our current position. Therefore, while convergence is guaranteed at least theoretically for many choices of J, the jumping distribution has considerable influence on the finite-sample properties of the chain. Figure 1 shows the output from several runs of the Metropolis algorithm for a two-dimensional target density. The jumping distributions are Gaussians with small, medium and large variances.

The Metropolis algorithm is perhaps the simplest technique for quickly generating a Markov chain with the correct stationary distribution. Many other schemes exist in the statistics literature that extend this approach to much more elaborate modeling contexts. In general, the samples $\{x_0, x_1, x_2, \dots\}$ are used to estimate properties of p like its mean and variance. When p represents a statistical model, understanding its mode structure is an important component in making inferences about the system under study. While this introduction to McMC has been brief, it is sufficient to motivate our sonification model.

3 McMC Sonification

The most spreaded approach to render sonifications from high-dimensional data is Parameter Mapping [4]: this technique uses instruments with a set of attributes which characterize or control its sound. Common attributes are frequency, amplitude, modulation rates, duration, and certainly the onset time in an score. Parameter Mapping sonifications are an analogon to plotting: for each data point, a tone is played with a given instrument whose attribute values are computed from the data values by a set of more or less simple mapping rules. The

sonification then is just the superposition of all these tones. Although this technique seemed to be quite promising for the presentation of high-dimensional data (as an instrument can have many attributes), Parameter Mapping shows some severe problems. The most important are, that the attributes are not independent (e.g. the perception of loudness and pitch are coupled), perception of features is nonlinear related to attribute values, there is no canonical way to map data components to attributes and a complicated mapping must be refered to for interpreting the sound.

These drawbacks are partly overcome by Model-Based Sonification, which was proposed recently in [2].

3.1 Model-Based Sonification

Model-Based Sonification can be motivated from the observation that our auditory system is well-optimized for the interpretation of sounds occurring in our physical environment. The source of such sounds is a physical material, whose dynamics allows for motions which can be perceived as sound. Most materials are in a state of equilibrium without interaction and thus do not produce sound. Often, humans themselves excite acoustic systems by hitting, striking or shaking the physical system and thus put energy into it. Therefore, sound is an important feedback which communicates properties of the material. Arguing that our brain is tuned to draw information about material properties from such sounds, the goal of Model-Based Sonification is to carry this cycle over to data sonification. A sonification model is a way to build up a kind of virtual *data material* from the data set by identifying dynamical elements, providing a dynamics and interaction types. The user then might explore the data set by shaking, plucking of hitting the virtual data material. The main advantages of this Model-Based Sonification approach are as follows:

- Sonification models need only few control parameters which can easily be understood.
- Sonification models can be applied to arbitrary high-dimensional data.
- Knowledge about the model facilitates or even enables understanding the sound.
- Interaction with a sounding material by excitation is something human users are familiar with.

Some sonification models like data sonograms, data set spring meshes and sonification of particle trajectories in a data potential have been considered recently [2]. In the last mentioned model, the sound was rendered by computing a set of particle trajectories through data space, where data points are imagined as attractive point masses. Introducing a friction force, the particles converge into local minima of the potential function which correspond to clusters in the data set and by this encodes local properties of the data set into sound. This model inspired the McMC sonification model

3.2 McMC Sonification Model

The McMC sonification model defines an process in data space similar to that explained in the particle trajectory model explained above. Whereas in the particle model, a potential function is derived by the superposition of contribing mass kernels, the potential is here replaced by a given density $p(x)$, which is - in the case of a given data set - computed by kernel density estimation with a Gaussian kernel. In this sonification model, the stochastic McMC steps are taken as start points for deterministic processes which explore properties of the local environment, encoding them into auditory representants which superimpose to the final sonification. As we very mainly interested into the mode structure of p, corresponding to a clustering in a data set, we took a

local mode search algorithm as the deterministic step. More specifically, we used an adaptive step size gradient ascent algorithm to find the nearest mode. A series of information is stored during this local optimization step for later usage within different sonification streams:

- function value $p(x_i)$ at the start position of step i
- mode center coordinates x_i^m
- function value $p(x_i^m)$ at the nearest mode
- distance $d_{im} = \|x_i - x_i^m\|$
- number of steps c required until convergence

Besides that quantitative values we can also identify symbolic information by making explicit the mode or attractor of the current deterministic step. Therefore we administer a list of mode structures which hold the position of each mode, the number of so far attracted McMC steps, density value at mode center, average density of all attracted particles and an associated bucket structure (see below). Parts of the information is not presented on each step, but summarized in *Auditory Information Buckets* explained below.

Summarizing, the sonification model is a discrete stochastic process with interleaved deterministic steps which explore local attributes of $p(x)$. The sonification superimposes relevant information about this process in data space in form of auditory markers whose time coordinate corresponds to the McMC step counter.

This basic sonification can be easily extended by adding further auditory streams. Our first extension is an auditory stream which summarizes the collected information on a larger time scale and thus can be thought of as a kind of zoom out. For this purpose we introduced the concept of Auditory Information Buckets (AIB).

Information buckets provide a way to choose the granularity of information which is presented acoustically. A bucket can be thought of as a place where information is collected about certain time steps. Both a counter for the number of items in the bucket and the value(s) of the data are stored. Furthermore, a threshold is defined to limit the bucket size. Reaching the threshold triggers a flushing of the bucket which results in the synthesis of a sound element for the bucket content. Thus the rate and complexity of these bucket sonifications can be easily chosen by adjusting the threshold. Within the McMC sonification model, buckets are used to summarize the characteristics of the modes. Each particle contributes to its respective mode bucket. On a flushing event, the data covariance matrix Σ of all contributed points is computed and a complex granular sound is rendered using the eigenvalues of Σ. Thus bucket sonification allows to derive conclusions about the shape and intrinsic dimensionality of the attraction basin.

3.3 Self-Organizing Pitch Maps

A very important property of the auditory grains is its pitch. We design the auditory grains so that its pitch corresponds to the probability density of the nearest mode. Doing this, two modes with very similar values cannot be distinguished acoustically. Self-organizing pitch maps solve this problem using a neural network (or better, a neuron chain) which optimizes a nonlinear monotonic mapping function in order to amplify pitch differences without changing the order. A detailed explanation of Self-Organizing Pitch Maps is shifted to our final version of the paper due to time problems.

4 Auditory Streams of the McMC Sonification

The sonification currently uses 3 auditory streams whose volume can be controled: (i) a stream containing auditory grains to present the McMC random walk in data space. (ii) a stream to provide information about rejected propositions of the McMC process and thus to inform the listener about the McMC efficiency, (iii) a stream containing auditory information bucket sounds which summarize information about the modes in the sense of an acoustic *zoom out*.

4.1 McMC monitoring stream

This stream provides information about the current McMC step. For each step, an auditory grain is added whose time stamp (start time of event) is propotional to the step index i, as this is the natural time axis in this process. The speed (time between sucessive steps) may be chosen by the user, specifying the time T between MC steps.

However, as a series of grains onto a regular time grid leads to either a monotonic rhythmical pattern or the perception of pitch at frequency $1/T$, we add a random time jitter of $T/4$.

The grains consist of nonharmonic periodic functions, multiplied with an envelope function with smooth attack and decay. The parameters are pitch, duration, decay time, volume and spectral composition (which is a multidimensional parameter). Actually, we drive pitch mapping $g(f(x_m))$, duration mapping $f(x_m)$ and keep the other parameters constant. $g(\cdot)$ is a nonlinear monotonic function, adaptively learned by the self-organizing pitch map, which assures that modes of similar density can be easily discerned. The amplitude of the grains is used to present the relevance of the step: The first step that converges into a new mode leads to a loud grain, as this is an interesting information. The more frequently McMC steps converge to this mode, the less information contains such an event and so the volume decreases. The amplitude information thus allows to conclude how well the distribution converged to equilibrium: if there is no change of volume while the McMC random walk stays in a certain modes' attraction basin, convergence is reached.

4.2 McMC propositions stream

This stream aims to provide more insight into the McMC simulation. Again, auditory grains are used here, allocating a different frequency band, shifted 2 octaves up. The given information is about the propositions that were refused, the distance between the last and the actual McMC step, the acceptance ratio (1) and the current distance to the mode. We still optimize this stream and will explain its composition in the final version of the paper.

4.3 AIB stream

This stream contains auditory information buckets, which give a summary about the modes. Each bucket is introduced by a pitched tone, whose pitch corresponds to the pitch of the McMC process stream shifted by an octave down. The rest of the bucket sound is a time-variant oscillation, where pitch and amplitude are controlled by the current content of the mode structure. They summarize the probability mass of the mode, the eigenvalues of the data covariance matrix, the average distance of all particles that contribute. The eigenvalues are taken as strength of the harmonics of a periodic waveform which is used within the grain sound. Thus brightness of the bucket sounds correspond to intrinsic dimensionality. We currently investigate how amplitude modulation can be used to increase perceptual resolution for the task of intrinsic dimensionality estimation from the sound.

5 Interactive Data Exploration by McMC Sonification Maps

McMC sonification has shown to be an interesting technique to get an auditory summary of a density p which might be also derived from a high-dimensional data set. However, one problem we encounter in this pure auditory inspection of data is the missing selectivity: in a visual data plot, it is easy to select parts of the data by focussing to them - simply by viewing onto certain parts in the plot. We can also communicate easily about visualizations because we can point to individual elements. Both a data selection and pointing into auditory information displays is more difficult than with visualization, because auditory displays are inherently dynamical. Time is an neccessary component of auditory information presentation. However, by combining sonification and visualization, we can provide techniques both for the selection and the pointing task.

In this section we present auditory maps, which allow to interactively select a focus for auditory data inspection. This is done by browsing a suited data plot with a mouse pointer. A coordinate vector is computed from the mouse position onto the two-dimensional projection space. In the case of linear projections like PCA or projection pursuit, this can be done by kernel regression for the subspace orthogonal to the projection space. In nonlinear data visualizations like SOM's or MDS plots, a coordinate vector can be computed by kernel regression as well, e.g. by averaging the k nearest neighbors in the projection.

Next, we present different implementations of such an audio-visual display using McMC sonification:

Listener Location: In this setup, the mouse pointer locates a listener in data space: the McMC process continues through the whole data space, but the distance of the current McMC step now determines how loud an event can be perceived in an intuitive manner: volume decays with increasing distance d by $v \propto 1/d^q$, where q is a small float ($q = 2$ is suited for 3-dimensional data). Thus a spatial focus is added to the sonification without disrupting the McMC flow between modes. As the auditory elements corresponding to a certain location are amplified with this setup, a kind of pointing into audio strems is realized: auditory pointing is transformed to pointing into corresponding points of a map.

McMC sphere clipping: In this setup, the McMC process steps are forced to stay within a sphere around a selected position in the plot Allowing the user to select the radius of the sphere, local properties of the high-dimensional density function can be exploited more directly. Of course, the auditory information can be related very directly to spots in the plot. Furthermore, doing this, the McMC process can be forced to explore certain parts of the data space.

McMC region comparison: This setup extends the former by allowing to mark two positions in the plot. Every t seconds, the McMC simulation is forced to change between these regions. This allows to compare two parts of the data or function at hand by their corresponding acoustic patterns.

We just started to investigate how to enhance McMC sonifications with visualization techniques and therefore only provide some few sound examples rendered using this interactive scheme. However, the usefulness of interaction can unfortenately not be presented within a written paper. Nonetheless, a more detailed investigation into usefullness is neccessary to proof its utility.

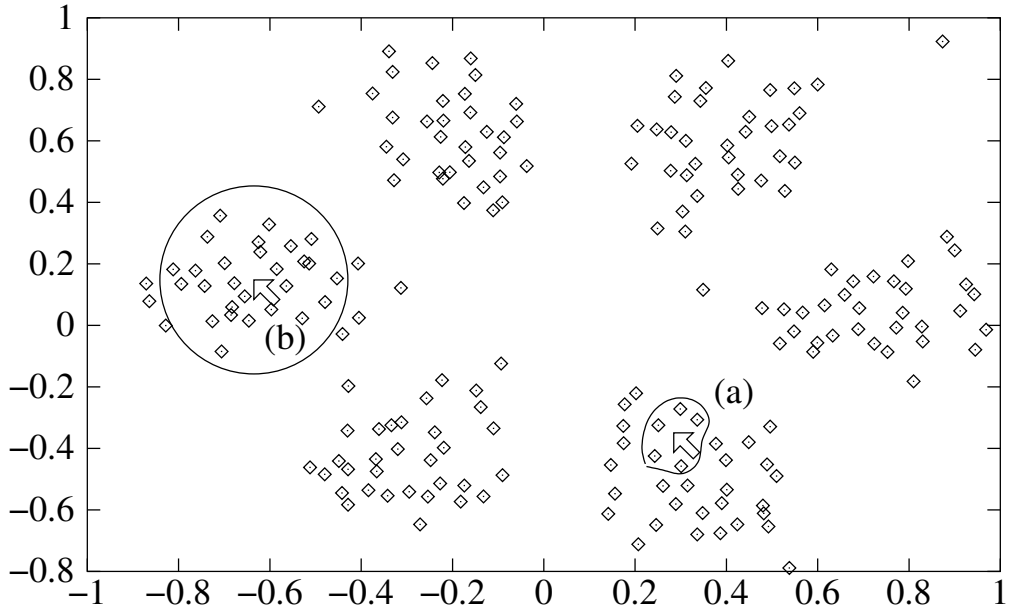

Figure 2: 2d Sammon map of 6 clusters at the corners of an 5-dim. simplex. (a) illustrates k-nearest neighbor projection: clicking the mouse restarts the McMC simulation onto the weighted mean of the 5 nearest neighbor in projection space. (b) illustrates the McMC listener: steps close to the mouse pointer result in louder events.

6 Examples

In this section, we present examples of our sonification method applied to both synthetic and real world problems. Sound files for these examples can be found at the web site [**?**]. Given time constraints, however, we have uploaded only files for selected examples; a complete set will be available prior to KDD 2001.

6.1 McMC Sonification for Cluster-Analysis

In this example, McMC sonification is used to explore clusters in a synthetic six-dimensional data set. The data were drawn from a mixture of some spherical Gaussian distributions in 6d space. The probability masses, the covariance matrices, the cluster centers and the number of clusters are chosen randomly.

In this simple example, the clustering can be easily depicted from a 2d-plot, as shown in Figure 3. Our target function is a kernel density estimate using a Gaussian kernel with covariance matrix $0.2V$, where V is the sample 6×6 covariance matrix estimated from the data. In sound example 1 we present a sonification of the first 1000 steps of the McMC simulation. We initially present only the McMC process stream. In this audio stream, the number of clusters can be perceived from the number of differently pitched tones. As discussed above, the relative size (in terms of data points) associated with each cluster as well as the convergence of the chain are also easily identifiable. In sound example 2, we add the second stream, and immediately observe that the cluster with the highest density has the smallest variance. Adding the buckets stream in sound example 3, we further can derive that this specific cluster has a low intrinsic dimensionality.

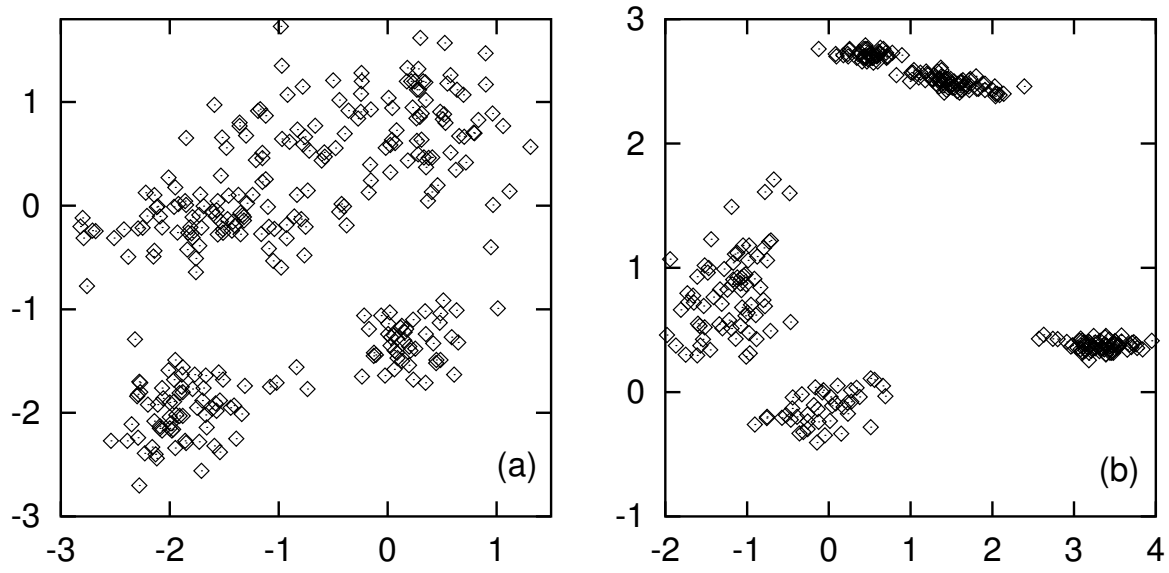

Figure 3: clustered data in 6d space. The plots show 5 clusters, projected onto (a) axis x_0 and x_1, (b) first two principal axis. Plots depend strongly upon the selection of the axis. McMC sonification is independant upon rotations.

6.2 Dimensionality Estimation

In this example, a data set consisting of a chain of clusters with different dimensionalities (rang of the covariance matrix) is taken as the basis for McMC sonification. The different intrinsic dimensionality can be perceived in the AIB streams. This example stresses the utility of the interaction scheme: using the McMC Listener mode, the volume of the sound emphasizes which sounds and steps contribute to the cluster at the current mouse position. On the other side, the changes between neighboring modes can still be perceived. If the McMC simulation is only slowly mixing (e.g. because of an unsuited choice of the jumping distribution), the mouse can be invoked to force the McMC simulation to restart at a specific cluster in the chain. The audio example demonstrates a sequence of McMC sonifications with small variance of the jumping distribution, where each cluster is activated in a sequence. It can be heard that one of the clusters contains a significant substructurization. This cluster is itself a mixture of several modes. Such an structure can get lost in data visualizations but is emphasized in an McMC sonification.

7 Conclusion

We have presented a new tool for monitoring McMC simulations and a new technique to combine visual displays with auditory displays for an exploratory analysis of high-dimensional functions and data sets. By experimenting with this sonification model, we found that our method also provides considerable insight into the structure of the target density p. For example, we can easily hear the number of modes of p, as well as their density and probability mass. This kind of insight is extremely helpful in high-dimensional settings. When p is a function of a small number of variables ($d \leq 4$), one can use traditional visualization methods to understand the important features of p. When $d > 4$, however, visual techniques begin to fail, and the resulting plots are more difficult to interpret. Carrying these experiences one step further, we have also applied this sonification scheme to high-dimensional data sets. Here, the target distribution p

is taken to be a nonparametric density estimate. In our examples we have used a simple kernel estimator, but any nonparametric technique will work. The auditory streams that track modes in p now provide direct information about clusters in the data.

The presented McMC sonification is designed to make usage of different human listening capabilities: we are able to process many auditory streams simultaneously without effort, we can concentrate and thus filter out certain acoustic information which we prefer and we can discern even subtle changes in dynamic auditory patterns.

In this way, the output from a converged McMC simulation provides valuable insights about p that are difficult to capture by purely visual means.

Additional new elements in this sonification are the concept of *Auditory Information Buckets*, which provide a way to *zoom out* acoustically, and nonlinear frequency maps using 1d-SOMs to facilitate mode comparisons, which were introduced in section 4.

Another important new element in this data exploration technique is the combination of visualization and sonification. It allows a new kind of browsing data by giving an auditory feedback on the users actions. The user can more easily compare characteristics of parts of the data and easier memorize characteristic sound patterns as they can be anchored to spots in the plot. Furthermore, the combination of plots and sonification enables a search for specific auditory patterns and thus allows to stress auditory structures for the purpose of communication about them. This might develop into tools for auditory pointing. As McMC sonification is based onto a sonification model, it has fixed model parameters which remain fixed if the target density or data set is exchanged. Thus the McMC sonification can be used for data of arbitrary dimensionality and must be learned only once.

However, our impression is that the coupling between data, sound and user actions could be even stronger: a real time visualization of all data points which are close to the current McMC position and thus a shift towards dynamical visualizations and sonifications. To more fully exploit these still largely untapped synergies of combining visual and auditory data exploration methods will be the subject of future research.

References

[1] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, 1995.

[2] 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.

[3] G. Kramer, editor. *Auditory Display - Sonification, Audification, and Auditory Interfaces.* Addison-Wesley, 1994.

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

[5] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *the Annals of Statistics*, 22(2):1701–1727, 1994.