Chapter 4

Modeling of relevance

Samuel Kaski, Jaakko Peltonen, Kai Puolamäki, Janne Sinkkonen, Jarkko Venna, Arto Klami, Jarkko Salojärvi, Eerika Savia

4.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. The key concept is *modeling of relevance*: data are usually full of patterns but the extracted ones should obviously be relevant to the analyst. An explicit definition of what is relevant is usually not known, and relevance needs to be inferred indirectly.

We have developed methods that use the structure of data in constraining which kinds of regularities are considered relevant. The structure here means several data sources or data sets. To make the task more concere, we have divided the ways of using the structure of data into three subtypes:

- *Relevance through data fusion* can mean two principal things: *dependency mining* and *supervised mining*, which are applicable in different settings. In both, several sources are combined with the goal of identifying relevant *aspects*, features or feature combinations, of data.

  In *dependency mining* or exploration, the aim is to decompose variation in each data source into source-specific and shared components. The within-source variation is assumed irrelevant, "noise", and only the shared effects are relevant. An example is measurement of several noisy signals from a common source, when characteristics of the noise are not known. More examples are given in Sections 5 and 6.

  While dependency mining is symmetric, in *supervised mining* a supervising auxiliary data set supervises the mining of primary data. Otherwise the methods are similar. If the supervising set consists of class labels of the primary data samples, the setup is *supervised unsupervised learning.* Our earlier research topic *learning metrics* was one suitable method for supervised unsupervised learning.

- *Relevant subtask learning* is a new research topic we introduced for addressing the problem of having too little representative or known-to-be-relevant training data. Given that other, partly or wholly irrelevant data sets are available, the relevant small data set is used as a "query" to retrieve more relevant data. At the same time, a model is built using all relevant data.

  This work can be seen as a special kind of asymmetric multi-task learning, or as combining information retrieval with multi-task learning.

- For *modeling of networks* we develop scalable models capable of dealing with uncertainty in network data. Networks are the simplest kinds of relational data, where the relations give hints of relevance.

These two general topics are useful in most of the modeling tasks above:

- *Discriminative generative modeling* describes how to use rigorous statistical modeling machinery for learning what is relevant to classes, and for making inference.

- *Information visualization* is a central subproblem in exploratory analysis and mining. We have introduced new very competitive nonlinear projection methods particularly suitable for projection to small dimensions for visualization.

4.2 Relevance through data fusion

Unsupervised data exploration or mining is defined as search for systematic properties, statistical structures or patterns from data. The findings need to be *relevant* as well, and typically relevance has been defined implicitly by selecting which kinds of patterns to find, which distance measures or features to use, and which model family to use. In general, relevance is defined by bringing in prior knowledge or assumptions to the task.

We have introduced methods for bringing in the prior information in a data-driven way, by choosing additional data sources and defining relevance through statistical dependencies between the sources. The underlying assumption is that aspects of data that are visible in one source only are "noise", whereas aspects visible in several sources describe the common thing of which all sources have different views. This will become clearer in the detailed descriptions below.

A straightforward way of finding the shared view is to build representations of data from each source, by maximizing the statistical dependency of the representations of different sources. We have developed both theory and practical methods for this task, and applied the methods in particular in neuro- and bioinformatics (Chapters 6 and 5).

Probabilistic models for detecting dependencies

Above the general approach was formulated in terms of maximizing a chosen dependency measure for mappings, that is, representations of the observations. We have previously introduced various methods for this task, including *associative clustering* and a linear projection method maximizing a non-parametric estimate of mutual information.

Recently we have studied an alternative formulation for the same task. One of the central problems in data analysis is overlearning, which means that models estimated with small data sets do not generalize well to new observations. One common solution to overlearning is to apply *Bayesian analysis* that allows treating uncertainties and choosing model complexity in a justified manner. Prior information can be rigorously incorporated to improve learning from small data sets, it is straightforward to extend models by changing distributional assumptions, and it is easy to construct larger models by combining submodels, at least in principle.

Bayesian tools can be applied to probabilistic models providing a generative description of the observed data. The methods for detecting dependencies between data sets are not, however, formulated as such models, and hence the Bayesian approach has not been possible for this task. We have introduced new theory on how such models can be built [3], and presented example models derived from the theory.

The proposed model family consists of latent variable models (see Fig. 4.1), where the observed data of each source is assumed to be an additive composition of two sources: one that is shared with the other data sources, and one that is specific to that particular data source. We have shown [3] that such models can extract the statistical dependencies to the shared latent source if a particular requirement is satisfied: the part of the model describing the data-source-specific variation in the observed data should be accurate enough.

Based on this basic principle we have re-derived an earlier probabilistic interpretation of canonical correlation analysis [1], and provided two novel models. In [3] a Bayesian clustering model for detecting dependencies is solved with variational Bayes approximation. The model is illustrated graphically in Figure 4.1. In [2] a Bayesian version of canonical correlation analysis is introduced, this time using Gibbs sampling for inference. Besides introducing a way of analyzing small-sample data to CCA the method lifts a critical restriction of classical CCA: the requirement of global linear dependency. It is overcome by

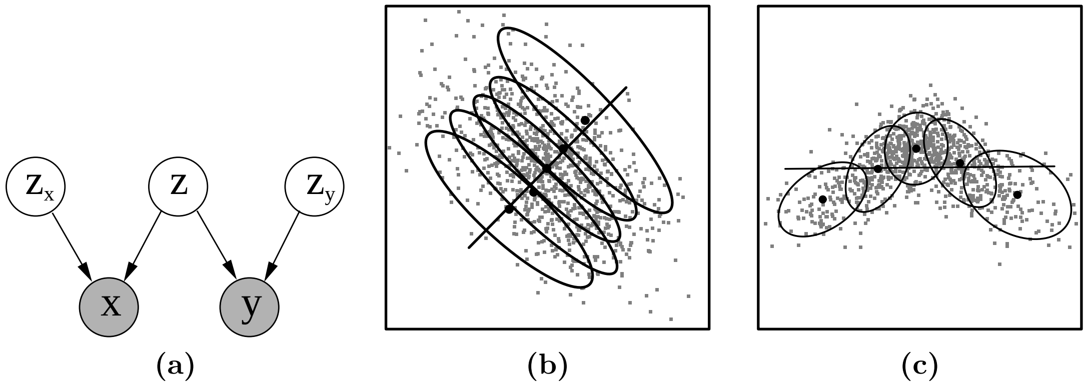

Figure 4.1: **(a):** A general latent-variable model structure for detecting statistical dependencies between $\mathbf{x}$ and $\mathbf{y}$. **(b-c):** Illustration of a clustering version of the general model. The two panels show scatter-plots of two data sets having co-occurring samples. The lines depict linear dependency found by canonical correlation analysis. The clusters found by the clustering model have aligned according to the dependency, while still capturing nonlinear structure of the data in panel **(c)**.

introducing a Dirichlet process mixture model, allowing different kinds of dependencies in different parts of the data space.

Dependency with class variables

A common case of two data sources is class labels coupled with feature vectors. Standard classifiers use the dependencies between the sources to predict class labels; other applications include visualization, discriminative clustering or discriminative feature extraction. These tasks use the labels to guide unsupervised analysis of the features; we call them *supervised unsupervised learning.*

Recently we have studied a particular application of supervised unsupervised learning: fast learning of a class-discriminative subspace of data features. The subspace is defined by a linear transformation, and the features in the class-discriminative subspace are *discriminative components* of data. The subspace is useful for visualization, dimensionality reduction, feature extraction, and for learning a regularized distance metric.

Earlier we had learned such transformations with nonparametric estimation [5] which is accurate but slow; the computational complexity is $O(N^2)$ per iteration; here N is the number of samples. We now introduced a method that learns the linear transformation in a fast, semisupervised way [4], by optimizing a mixture model for classes in the subspace. The new method (Fig. 4.2) is fast ($O(N)$ per iteration) and semi-supervised, that is, can use unlabeled and pairwise-constrained data as well as labeled data.

References

[1] Fransic R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Tech. Rep 688, Department of Statistics, University of California, Berkeley, 2005.

[2] Arto Klami and Samuel Kaski. Local dependent components. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432, 2007.

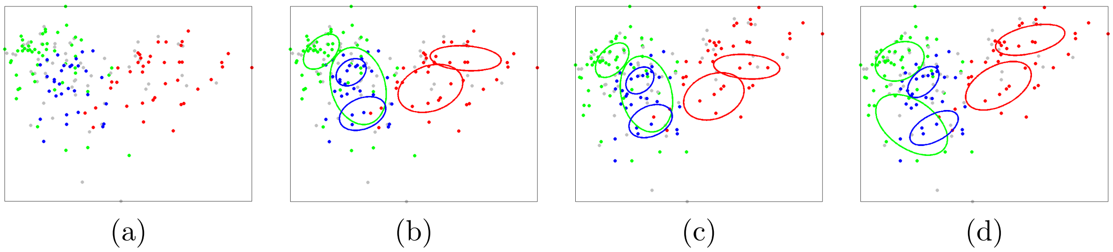

Figure 4.2: Sample iterations of optimizing the discriminative subspace. Dots show data in the subspace; ellipses show the shape of mixture model components used to model the distribution in the subspace. There are three classes (red, green, blue) and unlabeled samples (gray dots). **(a):** Initial transformation. **(b):** The mixture model is optimized for the transformation. **(c):** The transformation is optimized for the mixture model. **(d):** The mixture model is optimized for the new transformation. The iteration continues in alternating steps.

[3] Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, accepted for publication, 2008.

[4] Jaakko Peltonen, Jacob Goldberger, and Samuel Kaski. Fast Semi-supervised Discriminative Component Analysis. In Konstantinos Diamantaras, Tülay Adali, Ioannis Pitas, Jan Larsen, Theophilos Papadimitriou, and Scott Douglas, editors, *Machine Learning for Signal Processing XVII*, pages 312–317. IEEE, 2007.

[5] Jaakko Peltonen and Samuel Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16: 68–83, 2005.

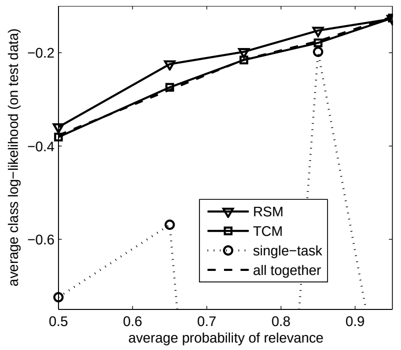

Figure 4.3: Relevant subtask learning model (RSM) outperforms a multi-task method that clusters tasks (TCM) and to two naive methods ("single-task" and "all together"), on news article data. The task was to predict relevance of news articles to a specific reader (the reader-of-interest), using articles rated by other readers as additional sources of information. Average results over 10 generated problems are shown, as a function of one experiment design parameter, the average probability that a sample is relevant to the reader-of-interest.

4.3 Relevant subtask learning

Having too little labeled training data is a common problem in classifier design. The problem is particularly hard for the high-dimensional data in genome-wide studies of modern bioinformatics, but appears also in image classification from few examples, finding of relevant texts, etc.

After realizing that the world is full of other data sets, the problem becomes how to simultaneously learn from a small data set and retrieve useful information from the other data sets. We have recently introduced a learning problem called *relevant subtask learning*, a variant of multi-task learning, which aims to solve the small-data problem by intelligently making use of other, potentially related "background" data sets.

Such potentially related "background" data sets are available for instance in bioinformatics, where there are databases full of data measured for different tasks, conditions or contexts; for texts there is the web. Such data sets are *partially relevant*: they do not come from the exact same distribution as future test data, but their distributions may still contain some useful part. Our research problem is, *can we use the partially relevant data sets to build a better classifier for the test data?*

Learning from one of the data sets is called a "task". Our scenario is then a special kind of *multi-task learning* problem. However, in contrast to typical multi-task learning, our problem is fundamentally asymmetric and more structured; test data fits one task, the "*task-of-interest*," and other tasks may contain *subtasks* relevant for the task-of-interest, but no other task needs to be wholly relevant.

In [1] we introduced a method that uses logistic regression classifiers. The key is to assume that each data set is a mixture of relevant and irrelevant samples. By fitting this model to all data sets, the common model for relevant samples learns from all tasks. We model the irrelevant part with a sufficiently flexible model such that irrelevant samples cannot distort the model for relevant data. A sample application is a news recommender for one user, where classifications from other users are available (Fig. 4.3). The relevant subtask learner outperforms a comparable standard multi-task learning model (related to [2]).

References

[1] Samuel Kaski and Jaakko Peltonen. Learning from relevant tasks only. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 608–615. Springer-Verlag, Berlin, Germany, 2007.

[2] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, 8: 35–63, 2007.

4.4 Discriminative generative modeling

The more traditional counterpart to supervised mining is *discriminative learning* where the data set is the same but the task is different. Given paired data $(\mathbf{x}, c)$, the task is to predict c for a test set where only the values of $\mathbf{x}$ are known.

There exist two traditional modeling approaches for predicting c, discriminative and generative. Discriminative models optimize the conditional probability $p(c|\mathbf{x})$ (or some other discriminative criterion) directly. The models are good classifiers since they do not waste resources on modeling those properties of the data that do not affect the value of c, that is, the marginal distribution of $\mathbf{x}$. The alternative approach is generative modeling of the joint distribution $p(c, \mathbf{x})$. Generative models add prior knowledge of the distribution of $\mathbf{x}$ into the task. This facilitates for example inferring missing values, since the model is assumed to generate also the covariates $\mathbf{x}$. The generative models are often additionally simpler to construct, and their parameters offer simple explanations in terms of expected sufficient statistics.

Discriminative Joint Density Models. In discriminative generative modeling we study discriminative inference given a generative model family $p(c, \mathbf{x}, \theta)$. The model family is assumed to be as good as possible but still known to be incorrect, and the objective is to obtain a distribution or point estimate that is optimal for predicting the values of c given $\mathbf{x}$. The Bayesian approach of using the posterior of the generative model family $p(c, \mathbf{x}, \theta)$ is not particularly well justified in this case, and it is known that it does not always generalize well to new data [1].

One way of learning discriminative classifiers is to take a joint density model, and then change the objective function from joint likelihood $\prod_i p(c_i, \mathbf{x}_i|\theta)$ to conditional likelihood $\prod_i p(c_i|\mathbf{x}_i, \theta)$. Earlier, we have presented an EM algorithm for obtaining discriminative point estimates [2]. The point estimate is (asymptotically) consistent for discrimination, given the model family. In [3] we proved that this applies for distributions as well; we derived an axiomatic proof that a *discriminative posterior* is consistent for conditional inference; using the discriminative posterior is standard practice in Bayesian regression, but we show that it is rigorous for model families of joint densities as well.

Compared to pure discriminative models, the benefit of the approach is that prior knowledge about $\mathbf{x}$ is brought in. The models operate in the same parameter space as ordinary discriminative models, but the generative formulation constrains the model manifold. Additionally, the density estimate for $\mathbf{x}$ from the model can be used for inferring missing values in the data [3].

References

[1] Peter D. Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2–3):119–149, 2007.

[2] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Expectation maximization algorithms for conditional likelihoods. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pages 753–760, New York, USA, 2005. ACM press.

[3] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, and Samuel Kaski. Discriminative MCMC. Report E1, Publications in Computer and Information Science, Helsinki University of Technology, 2006.

4.5 Visualization methods

Visualization of mutual similarities of entries in large high-dimensional data sets is a central subproblem in exploratory analysis and mining. It is makes sense to "look at the data" in all stages of data analysis, and reducing the dimensionality to two or three gives a scatterplot visualization.

When the intrinsic dimensionality of the data is higher than the dimensionality of the visualization, as is often the case, the visualization cannot represent the data flawlessly; some properties are necessarily lost or misrepresented. A compromise is unavoidable, but which compromise is the best for visualization? Many existing nonlinear dimensionality reduction methods practically ignore this question altogether, because they are not designed to reduce the dimensionality of the data set lower than is possible without losing information. Some methods choose the compromise implicitly in that they produce the lower-dimensional representation by minimizing a cost function, but the cost function has not been motivated from the point of view of visualization, that is, it is not obvious why a projection that minimizes the cost function should be a good visualization. We have filled this gap by introducing rigorously motivated measures for the quality of a visualization, as well as a nonlinear dimensionality reduction method that optimizes these measures and is therefore specifically designed for optimal visualization.

Visualization as information retrieval

We view visualization as an information retrieval task. Consider an analyst studying a scatterplot of countries, organized according to their welfare indicators. Being interested in Finland, she wants to know which other countries are similar. The visualization helps in this task of retrieving similar items, and quality of retrieval can be measured with standard information retrieval measures *precision* and *recall.* Any information retrieval method needs to make a compromise between these measures, parameterized by the relative cost of false positives and misses. Since a visualizer is an information retrieval device as well, it needs to make the same compromise.

We have adapted the information retrieval measures to visualization by smoothing them and representing them as differences between distributions of points being neighbors. It turns out that the traditional measures are limiting cases of these more general measures. Once the relative cost λ of false positives and misses has been fixed, we can directly optimize the visualization to minimize the retrieval cost. We call the resulting visualization method the Neighborhood Retrieval Visualizer (NeRV) [1].

The NeRV is a further development of our earlier method *local multidimensional scaling* [2], a faster method where the trade-off between precision and recall was heuristic and hence the results were less accurate.

Later we added the Self-Organizing Map to the comparison [3]. The SOM was very good in terms of (smoothed) precision, even producing a slightly better result than NeRV in some cases. In terms of recall the SOM performed poorly.

References

[1] Jarkko Venna and Samuel Kaski. Nonlinear dimensionality reduction as information retrieval. In Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS*07), San Juan, Puerto Rico, March 21-24, 2007.

[2] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–899, 2006.

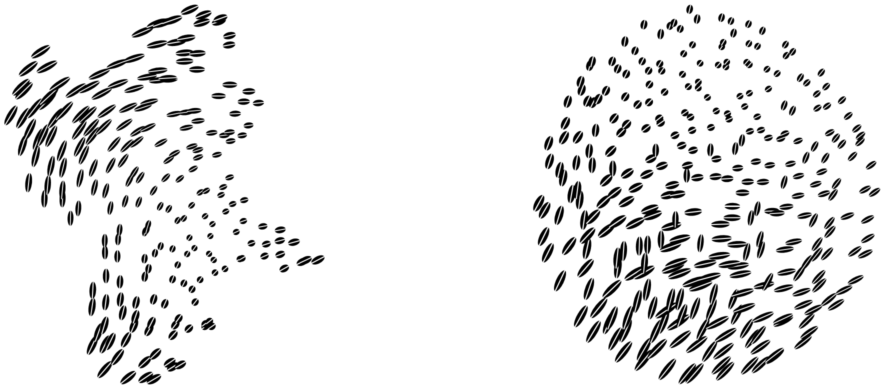

Figure 4.4: Two nonlinear projections of data that lies on the surface of a three-dimensional sphere. One of the input coordinates governs the rotation of the glyphs, the second their scale, and the third their degree of elongation. Hence, points having similar glyphs are close to each other in the input space. On the *left*, precision has been maximized; the sphere has become split open and the glyphs change smoothly, but on the opposite ends of the projection there are similar glyphs that are projected far from each other. On the *right*, recall has been maximized and the sphere has become squashed flat. There are areas where the different kinds of glyphs are close to each other, but there are no areas where similar glyphs are very far from each other.

[3] Kristian Nybo, Jarkko Venna and Samuel Kaski. The Self-Organizing Map as a Visual Neighbor Retrieval Method. In *Proceedings of 6th Int. Workshop on Self-Organizing Maps (WSOM '07).* Bielefeld University, Bielefeld, Germany, 2007.

4.6 Networks

Machine Learning is in the midst of a "structural data revolution". After many decades of focusing on independent and identically-distributed examples, many researchers are now modelling inter-related entities that are linked together into complex graphs. A major driving force is the explosive growth of heterogeneous data collected on diverse sectors of the society. Example domains include bioinformatics, communication networks, and social network analysis.

Networks are a special case of structural data. Inferring properties of the network nodes, or vertices, from the links, or edges, has become a common data mining problem. Network data are typically not a complete description of reality but come with errors, omissions and uncertainties. Some links may be spurious, for instance due to measurement noise in biological networks, and some potential links may be missing, for instance friendship links of newcomers in social networks. Probabilistic generative models are a tool for modeling and inference under such uncertainty. They treat the links as random events, and give an explicit structure for the observed data and its uncertainty. Compared to non-stochastic methods, they are therefore likely to perform well as long as their assumptions are valid; they may reveal properties of networks that are difficult to observe with non-statistical techniques from the noisy and incomplete data, and they also offer a groundwork for new conceptual developments.

Component models for large networks

Being among the easiest ways to find meaningful structure from discrete data, Latent Dirichlet Allocation (LDA) and related component models have been applied widely. They are simple, computationally fast and scalable, interpretable, and admit flexible nonparametric priors. In the currently popular field of network modeling, relatively little work has taken uncertainty of data seriously in the Bayesian sense, and component models have been introduced to the field only recently. We have developed a component model of networks that finds community-like structures like the earlier methods motivated by physics. With Dirichlet Process priors and an efficient implementation the models are highly scalable.

References

[1] Janne Sinkkonen, Janne Aukia, and Samuel Kaski. Inferring vertex properties from topology in large networks. In *The 5th International Workshop on Mining and Learning with Graphs (MLG'07)*, Florence, Italy, 2007. Universita Degli Studi di Firenze.

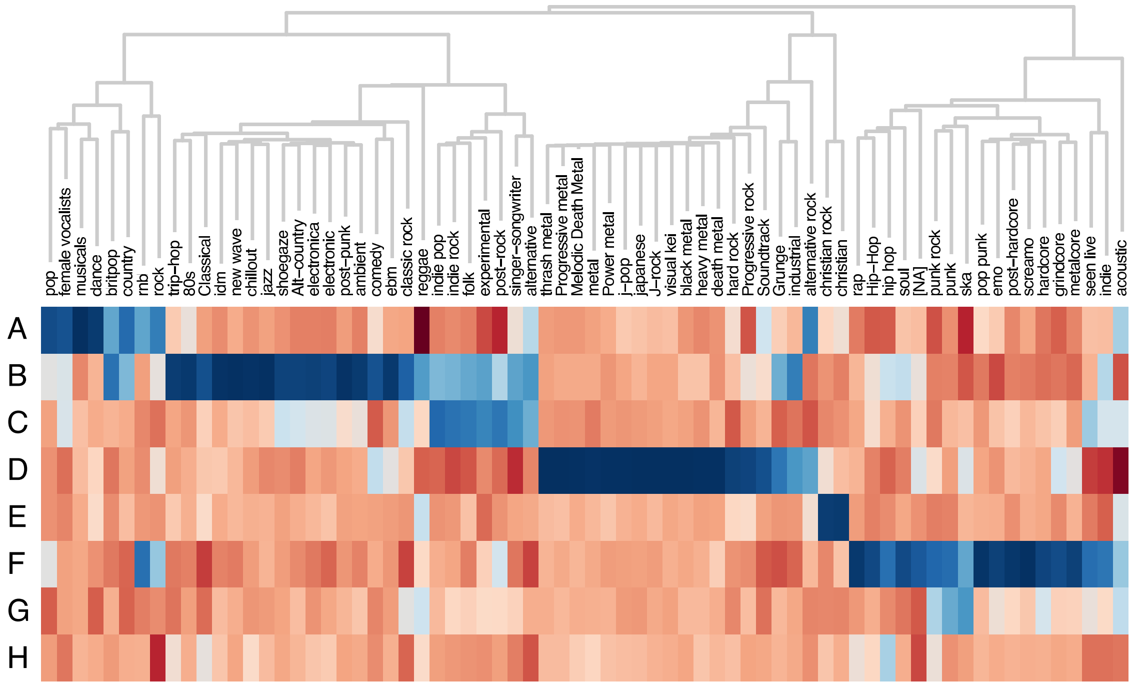

Figure 4.5: Last.fm is an Internet site that learns the musical taste of its members on the basis of examples, and then constructs a personalized, radio-like music feed. The web site also has a richer array of services, including a possibility to announce friendships with other users. The friendship network alone, when divided into components, reveals musical structures, because the music tastes of friends tend to be similar. Here the latent components found by our model were afterwards correlated with user's listening habits; songs are aggregated by tags given to them. Tags are intuitively grouped into genres. The network has 147,000 nodes and 353,000 links, but the running time with an efficient implementation by our collaborators at Xtract Ltd. was just 8.4 hours.