# Chapter 4

# Multi-source machine learning

**Samuel Kaski, Arto Klami, Gayle Leen, Jaakko Peltonen, Kai Puolamäki, Janne Sinkkonen, Zhirong Yang, Helena Aidos, Ilkka Huopaniemi, Kristian Nybo, Juuso Parkkinen, Eerika Savia, Tommi Suvitaival, Abhishek Tripathi**

## 4.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. In practical computational data analysis tasks a common problem is lack of sufficient amount of representative data. Modeling requires either data or prior knowledge which by definition does not exist in knowledge discovery or data mining tasks. If there was enough data, modern statistical machine learning toolboxes would contain powerful approaches to building flexible models that do not make strong assumptions about data, but flexible models are naturally weak given little data.

In many applications, for instance in molecular biology and neuroinformatics, there is data available in public or special-purpose databanks, but the problem is that not everything is relevant. We are developing new machine learning methods capable of learning from *multiple data sources* containing only *partially relevant* data, and generalizing to new contexts. The methods extend and generalize the current approaches called multi-view, multi-way and multi-task learning, on structured and unstructured domains.

Moreover, we have developed new principles and methods for the task of *visualizing* high-dimensional data; this task is central in any knowledge discovery process.

## 4.2 Multi-view learning

Multi-view learning tells how several data sources, or views, can be combined to extract more relevant information. We focus on unsupervised settings, where the relevance comes from statistical dependencies between multiple views of the same objects. For example, a collection of images with captions can be represented with two views, one capturing the contents of the image while the other describes the caption. Dependencies between these representations reveal more information on the intended content, or semantics, of the images than either view alone.

We have developed new theory for decomposing variation in multiple views into source-specific and shared components [1], building on Bayesian latent-variable models that capture the dependencies by assigning flexible source-specific models for describing the noise in each of the views. The same basic formulation extends to various practical models. A prime example is [2] that applies hierarchical non-parametric Bayesian models for making the source-specific parts extremely flexible, and builds a hierarchical grouping of human genes based on both mRNA and protein expression. The model, illustrated in Figure 4.1, reveals processes that could not be found by looking at either view alone.

Besides advanced Bayesian solutions, we have also developed novel multi-view algorithms for application purposes. [4] aimed at creating an easy-to-use data integration tool for bioinformatics applications and was accompanied by an open-source software package, while [3] introduces a fast algorithm for maximizing mutual information of linear projections, applied to brain imaging data.

Going beyond standard multi-view learning, we have also developed novel solutions for applications without co-occurring data. Traditional multi-view learning can only be applied for cases with clear one-to-one co-occurrence between the views. We showed in [5] that the co-occurrence itself can be learned by maximizing statistical dependency between two views with no known co-occurrence. In brief, the idea is to order the samples of one of the views so that the dependency between the views is maximal. This, in turn, requires efficient means for measuring the dependency, provided by classical data integration tools like canonical correlation analysis (CCA). A simple iterative algorithm alternating between optimizing the ordering (solved through a linear assignment problem) and finding a representation that maximally captures the dependency (solved through CCA) finds the co-occurrences with high accuracy. We have applied the algorithm for aligning probeset of various microarray brands, matching metabolite identities of different species or measurement batches, and aligning sentences of bi-lingual corpora.

## References

[1] Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72:39–46, 2008.

[2] Simon Rogers, Arto Klami, Janne Sinkkonen, Mark Girolami, and Samuel Kaski. Infinite Factorization of Multiple Non-parametric Views. *Machine Learning*, 79(1–2):201–226, 2010.

[3] Eerika Savia, Arto Klami, and Samuel Kaski. Fast dependent components for fMRI analysis. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1737–1740, 2009.

[4] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

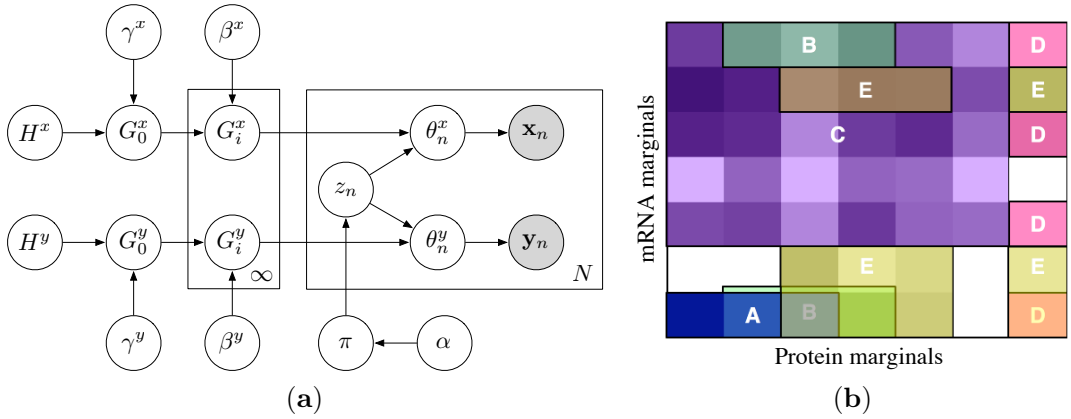


Figure 4.1: (**a**): Illustration of the hierarchical Dirichlet process model for cluster analysis of coupled data sources. (**b**): Application of the model on coupled analysis of mRNA and protein concentrations. Both marginals correspond to clusters of genes, automatically detected by the model, and the color-codes and letters indicate higher-level processes obtained by simultaneous clustering of the contingency table of cluster assigments.

[5] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Using dependencies to pair samples for multi-view learning. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1561–1564, 2009.

## 4.3 Multi-task learning

We have introduced two new multi-task learning setups, suitable for different scenarios, and solutions for them: *relevant subtask learning* and *paired multi-task learning*.

### Relevant subtask learning

It is all too common in classification tasks that there is too little training data to estimate sufficiently powerful models. 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.

We have 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. To fit the model, we have used both simple maximum likelihood fitting [1] and more advanced variational Bayesian inference [3]. 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. 4.2). The relevant subtask learner outperforms a comparable standard multi-task learning model [4].

The generalization error of relevant subtask learning has been analyzed theoretically in [5] in a slightly different setting, where the task is density estimation and supplementary tasks are assumed to be mixtures of a shared interesting density and a non-interesting task-specific density. Relevant subtask learning has smaller generalization error than learning from the task-of-interest alone or from a supplementary task alone.

### Paired Multi-task Learning

When faced with an abundance of tasks containing potentially relevant information to a desired learning task, we ask: how can we decide which tasks are relevant? And what is the relationship between the different tasks? Knowledge about the task relationships and problem structure can then be exploited in jointly learning multiple tasks. By sharing statistical strength between different tasks, this *multitask learning* set-up can overcome potential problems when there is little data for a single task.

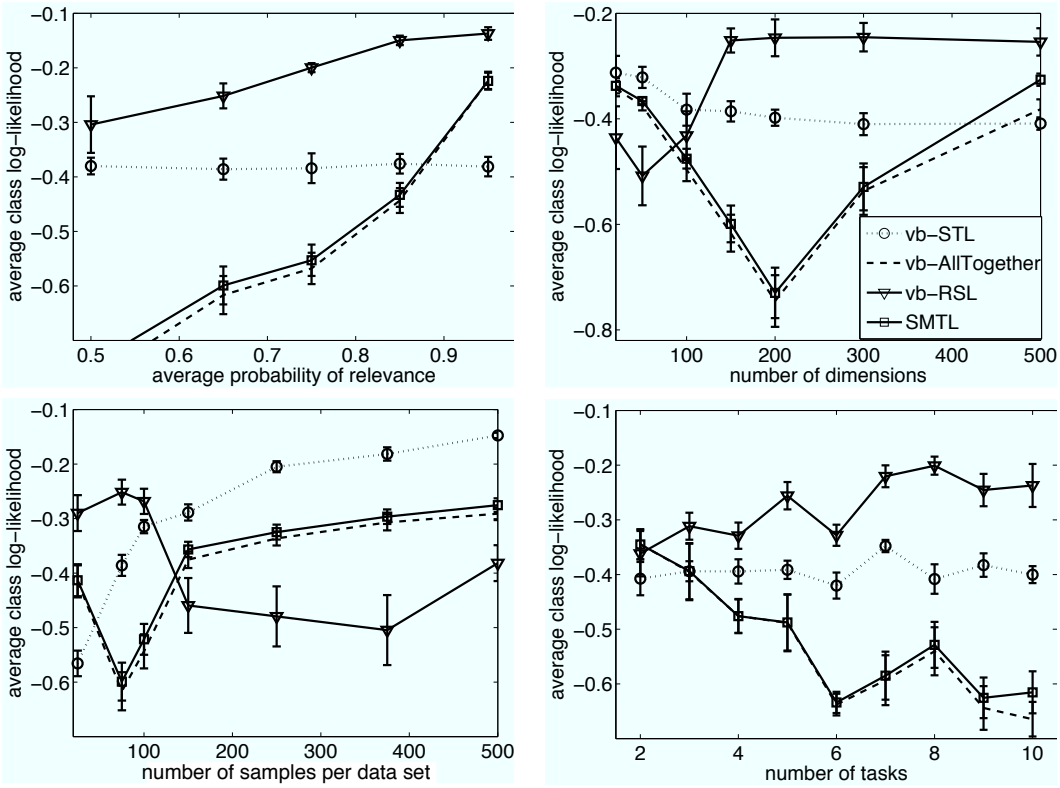


Figure 4.2: Comparison of multi-task learning approaches on news article data. The task was to predict relevance of news articles to a specific reader (the reader-of-interest), using articles rated by other readers as additional sources of information. Results are shown as a function of several design parameters: the proportion of relevant samples (**t**op left), data dimensionality (**t**op right), the number of samples per data set (**b**ottom left) and the number of tasks (data sets; **b**ottom right). Relevant subtask learning (vb-RSL) outperforms a multi-task method that clusters tasks (SMTL; [4]) and to two naive methods ("vb-STL" and "vb-AllTogether") when there are many dimensions but few samples per data set (less than 100), which is a realistic scenario.

We address a specific problem in bioinformatics: learning to choose control samples for use in a differential gene expression experiment in cancer (case vs control). Gene expression measurements are likely to contain bias due to factors such as patient-specific and laboratory-specific effects, and typically there are only a small number of samples available for each experimental condition. These factors make it problematic to select a set of pairs of control and tumor tissue (case) samples, such that the differential gene expression of the case samples is solely due to cancer-specific variation. However, there is potentially a huge amount of useful information about cancer, and the relationship to control tissue contained in publicly available gene expression databases. If two cancer types are similar, then it is likely that they will use similar control samples.

The suitable controls for each experiment form a group of controls. These groups are considered as classes / control tissues, and the task is to classify each case sample to one of these classes. We formulate this as a multi-task learning problem in [2] where we have *a set of primary tasks* (choosing the control class for each sample for a cancer type) which we want to learn, and *a set of auxiliary tasks* (choosing the control class for each control sample). This follows a paired structure, such that each primary task is paired

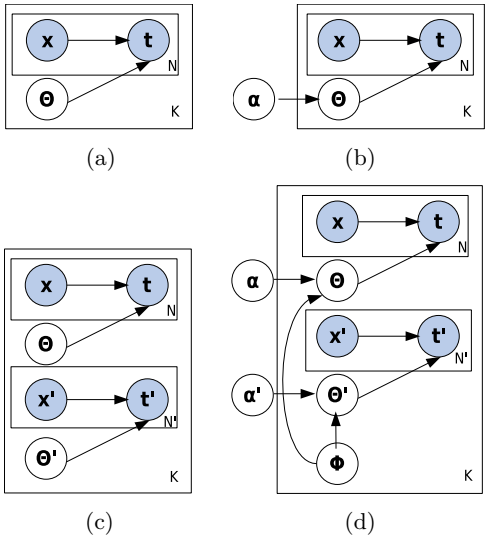


Figure 4.3: Schematic illustration of statistical strength sharing in multitask learning scenarios. Learning a set of $K$ tasks as in (a) amounts to finding different parameterisations $\theta_i, i = 1, ..., K$ for the tasks. If the tasks are assumed to be related, multitask learning approaches assume some shared structure across all $K$ tasks through a common parameterisation via $\alpha$ (b). We consider the situation where there are $K$ pairs of tasks (c), and propose the structure in (d) to share information between the tasks. There is shared structure within each task set's parameterisation $\theta, \theta'$ through $\alpha, \alpha'$ and across each of the $K$ pairs through $\phi$.

with a corresponding auxiliary task. We transfer information about the *relatedness of the auxiliary set of tasks* to the set of primary tasks (see Figure 4.3 and its caption for more details). We formulate the model using the Gaussian process framework; the task functions are given Gaussian process priors and the task structure is modeled through the parameterisation of the covariance functions. For each set of tasks, the task functions are assumed to come from a linear combination of an underlying set of latent functions. This linear combination, which models the inter-task similarity in each set, is constrained to be the same for both the primary and auxiliary task set.

In learning the classification, we use knowledge about the relationships between the case and the control samples. This pairing is transferred to new pairs, such that our model can infer a suitable control sample for a new case sample (see Figure 4.4).

## 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] Gayle Leen, David R. Hardoon and Samuel Kaski. Automatic Choice of Control Measurements, In *Advances in Machine Learning (Proc. ACML'09, The 1st Asian Conference on Machine Learning)*,5828:206–219. Springer, 2009.

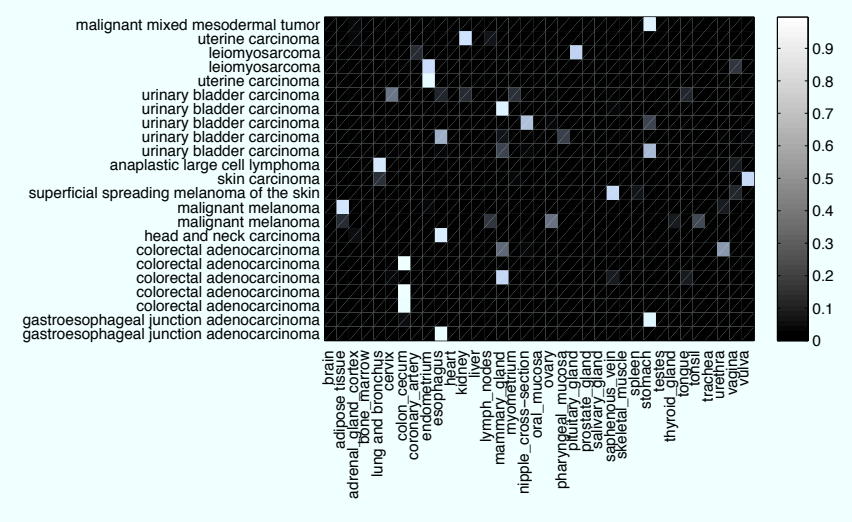


Figure 4.4: Visualization of the probability distribution over the control classes (x axis) for some tumor samples (y axis) with unknown control classes

[3] Jaakko Peltonen, Yusuf Yaslan, and Samuel Kaski. Relevant subtask learning by constrained mixture models. *Intelligent Data Analysis*, to appear.

[4] 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.

[5] Keisuke Yamazaki and Samuel Kaski. An Analysis of Generalization Error in Relevant Subtask Learning. In Mario Köppen, Nikola Kasabov, and George Cooghill, editors, *Advances in Neuro-Information Processing, 15th International Conference, ICONIP 2008*, pages 629–637. Springer-Verlag, Berlin Heidelberg, 2009.

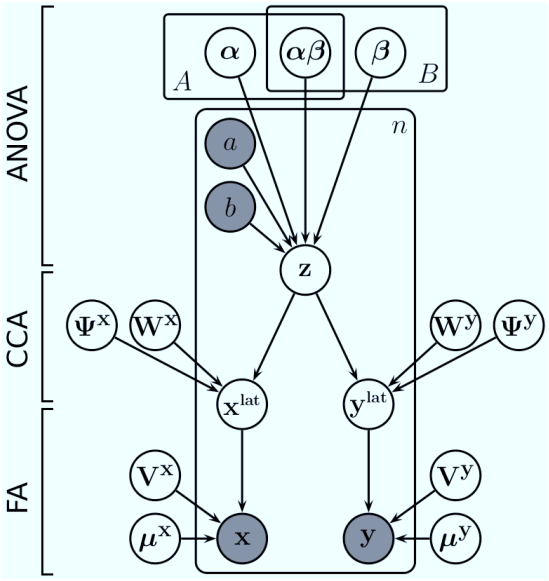


Figure 4.5: Plate diagram of the graphical model for Multi-Way, Multi-View Learning

## 4.4 Multi-way learning

Finding effects of one or multiple known covariates from the data is one of the most common statistical problems, commonly solved by tranditional Analysis of Variance (ANOVA), its multivariate generalization (MANOVA), or in general by linear models. The traditional methods are not applicable and very limited alternative methods exist to currently increasingly important problems in molecular biology where the dimensionality of the problem p is very large and the number of observations n is (relatively) small. The same "large p, small n" problem recurs also in other fields.

In biological experiments typical covariates are disease, drug treatment groups, gender or time-series, resulting in a multi-way experimental setup. The main challenge in biology is that the number of samples (for instance mice or human patients) is small due to economical and ethical cost, whereas the number of variables (such as genes or metabolites) is huge. Due to this, the traditional multivariate methods cannot be used, and on the other hand little research of multi-way analysis has been presented in the machine learning literature.

We have recently introduced a Bayesian method for solving this burning problem of multi-way analysis of small sample-size, high-dimensional datasets [1]. Moreover, the multi-way data-analysis problem becomes even more complicated when heterogeneous data with multiple covariates are integrated from multiple sources. Different data sources usually have distinct, unmatched variable-spaces with different dimensionalities. We have generalized ANOVA-type analysis to the case of multiple sources by considering the source ("view") as an additional covariate in the ANOVA-type analysis. The problem is impossible for traditional methods due to the different variable-spaces, but by utilizing dependencies between the sources the problem can be solved. We introduced a model (Figure 4.5; [2]) which is able to find the multi-way covariate-effects and to partion them into shared and source-specific effects. The method is applicable to any small sample-size, multi-source experiments, currently very popular in biological research.

## References

[1] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Samuel Kaski, and Matej Orešič. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261–276, 2009.

[2] Ilkka Huopaniemi and Tommi Suvitaival and Janne Nikkilä and Matej Orešič and Samuel Kaski. Multi-Way, Multi-View Learning. In *NIPS 2009 workshop on Learning from Multiple Sources with Applications to Robotics*, 2009

## 4.5 Information visualization

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.

It is generally not possible to show all the similarity relationships within a high-dimensional data set perfectly on a low-dimensional display; some properties are necessarily lost or misrepresented. All linear or nonlinear dimensionality reduction methods must make a compromise about which kinds of similarity relationships they aim to show, but which compromise is best for visualization? Many methods practically ignore this question because they are not designed to reduce the dimensionality of the data set lower than is possible without losing information; several such methods have difficulties when producing low-dimensional displays. 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.

It has been difficult to assess the quality of visualizations since the task of visualization has not been well-defined. We have addressed this problem and introduced 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. An analyst looking at a scatterplot can choose any point (data item) and find its neighbors (similar other items) in the visualization. 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 $\lambda$ 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) [7, 8]. NeRV outperforms several recent nonlinear dimensionality reduction methods both by the new measures and by traditional measures.

We have extended NeRV to supervised visualization [4], to linear visualization [2], and to visualization with ontological annotation [3].

In addition to NeRV, we have introduced methods for the specific application of visualizing convergence of Markov chain Monte Carlo (MCMC) sampling methods commonly used in Bayesian inference [5].

One of the popular nonlinear dimensionality reduction methods, Stochastic Neighbor Embedding (SNE; [1]) is a special case of NeRV, corresponding to optimizing recall only. We have additionally introduced other generalizations of SNE and efficient algorithms for computing it, as described in the following.

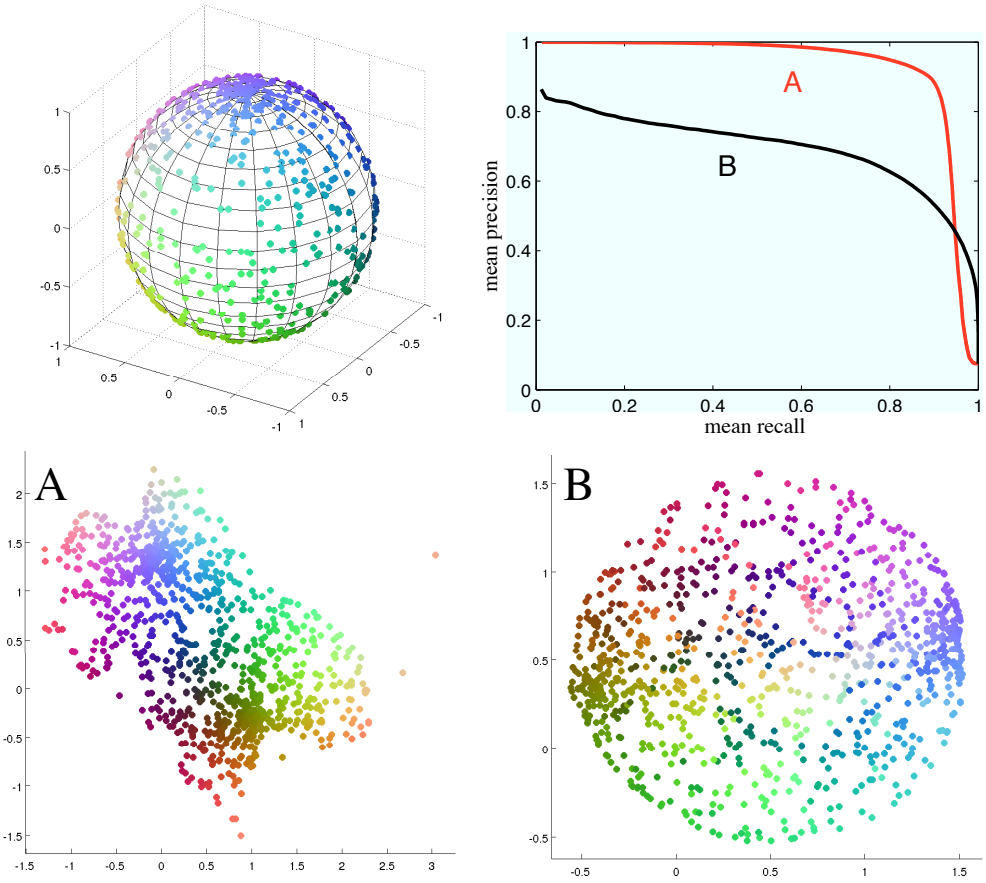


Figure 4.6: Demonstrating the precision-recall tradeoff in visualization. The task is to retrieve neighbors of points in the original space, based on their locations in the visualization. **Top left:** A three-dimensional dataset sampled from the surface of a sphere. **Bottom:** Two embeddings of the dataset. In **A**, the sphere has been cut open and folded out. This eliminates *false positives* (false neighbors), but there are some *misses* (missed neighbors) because points on different sides of the tear end up far away from each other. In contrast, **B** minimizes the number of misses by simply squashing the sphere flat; this yields many false positives because points on opposite sides of the sphere are mapped close to each other. **Top right:** mean precision–mean recall curves for the two projections. **A** has better precision (yielding higher values at the left end of the curve) **B** has better recall (yielding higher values at the right end of the curve).

### Heavy-tailed Symmetric Stochastic Neighbor Embedding

Stochastic Neighbor Embedding (SNE) has been shown to be quite promising for data visualization. Currently, the most popular implementation, t-SNE [6], is restricted to a particular Student t-distribution as its embedding distribution. Moreover, it uses a gradient descent algorithm that may require users to tune parameters such as the learning step size, momentum, etc., in finding its optimum.

In [9], we have rigorously investigated the working mechanism of Heavy-tailed Symmetric Stochastic Neighbor Embedding (HSSNE). The several findings are: (1) we propose to use a negative score function to characterize and parameterize the heavy-tailed embedding similarity functions; (2) this finding has provided us with a power family of functions that

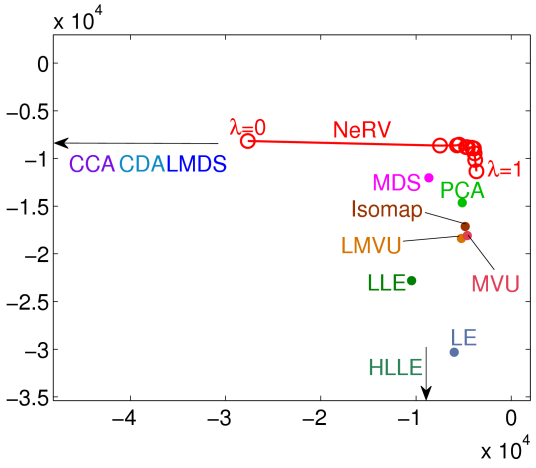


Figure 4.7: Comparison of visualization performance between several recent methods on a data set of mouse gene expression profiles, in terms of two novel measures: mean smoothed precision (vertical axis) and mean smoothed recall (horizontal axis). Our method NeRV performs best (best values near the upper right corner).

convert distances to embedding similarities; and (3) we have developed a fixed-point algorithm for optimizing SSNE, which greatly saves the effort in tuning program parameters and facilitates the extensions and applications of heavy-tailed SSNE. We have presented two empirical studies, one for unsupervised visualization showing that our optimization algorithm runs as fast and as good as the best known t-SNE implementation and the other for semi-supervised visualization showing quantitative superiority using the homogeneity measure as well as qualitative advantage in cluster separation over t-SNE. The latter results are shown in Figure 4.8.

## References

[1] Geoffrey Hinton and Sam T. Roweis. Stochastic Neighbor Embedding. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 833–840. MIT Press, Cambridge, MA, 2002.

[2] Jaakko Peltonen. Visualization by Linear Projections as Information Retrieval. In José Príncipe and Risto Miikkulainen, editors, *Advances in Self-Organizing Maps (proceedings of WSOM 2009)*, pages 237-245. Springer, Berlin Heidelberg, 2009.

[3] Jaakko Peltonen, Helena Aidos, Nils Gehlenborg, Alvis Brazma, and Samuel Kaski. An information retrieval perspective on visualization of gene expression data with ontological annotation. In *Proceedings of ICASSP 2010*, to appear.

[4] Jaakko Peltonen, Helena Aidos, and Samuel Kaski. Supervised Nonlinear Dimensionality Reduction by Neighbor Retrieval. In *Proceedings of ICASSP 2009, the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1809–1812. IEEE, 2009.

[5] Jaakko Peltonen, Jarkko Venna, and Samuel Kaski. Visualizations for Assessing Convergence and Mixing of Markov Chain Monte Carlo Simulations. *Computational Statistics and Data Analysis*, 53:4453–4470, 2009.

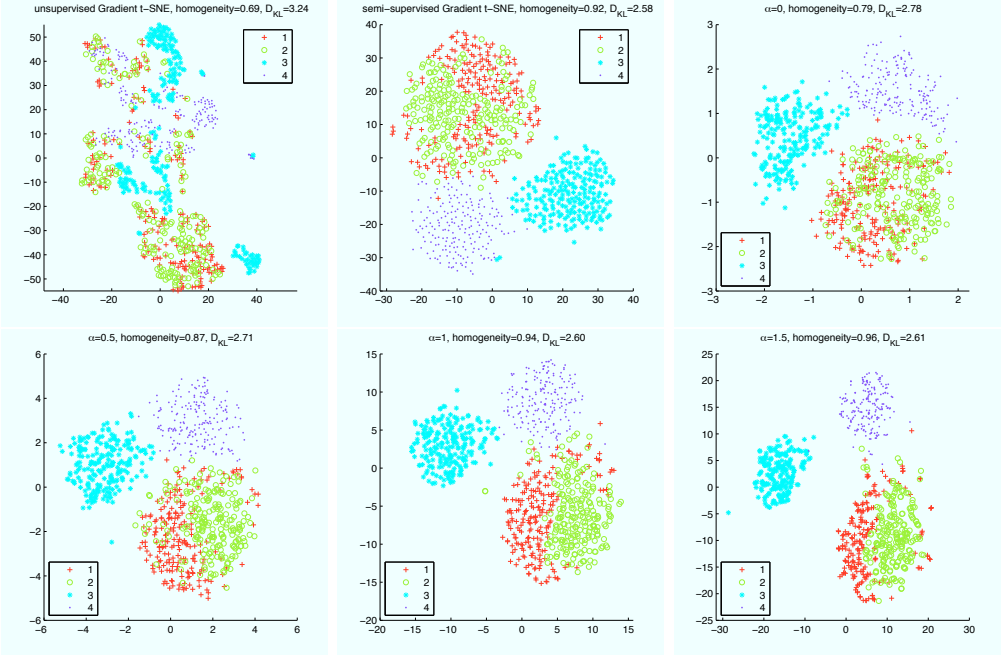


Figure 4.8: Semi-supervised visualization for the *vehicle* data set. The plots titled with $\alpha$ values are produced using the fixed-point algorithm of the power family of HSSNE.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[7] Jarkko Venna and Samuel Kaski. Nonlinear dimensionality reduction as information retrieval. In M. Meila and X. Shen, editors, *Proceedings of AISTATS*07, the 11th International Conference on Artificial Intelligence and Statistics (JMLR Workshop and Conference Proceedings Volume 2)*, pages 572–579, 2007.

[8] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[9] Zhirong Yang, Irwin King, Zenglin Xu, and Erkki Oja. Heavy-tailed symmetric stochastic neighbor embedding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2169–2177, 2009.

## 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 in 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.

We have earlier introduced a family of Bayesian probablistic component models for analyzing interactions or graphs, called Interaction Component Model (ICM). We applied ICM to the task of detecting dense subnetworks from noisy protein-protein interaction networks, and additionally from multiple views; protein-protein interactions and gene expression data [2]. Such subnetworks are interpretable as functional gene modules or protein complexes. Our methods outperformed other state-of-the-art methods in this task of discovering functional subnetworks.

We further extended the ICM framework to handle multi-relational data [3], and to detect block structures [1]. For example, protein complexes consist of tightly interacting proteins, and the complexes in turn interact with other complexes.

## References

[1] Juuso Parkkinen, Adam Gyenge, Janne Sinkkonen and Samuel Kaski. A block model suitable for sparse graphs. In *The 7th International Workshop on Mining and Learning with Graphs (MLG'09), Leuven, Belgium, July 2-4* (2009).

[2] Juuso Parkkinen and Samuel Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology* 2010 4:4.

[3] Janne Sinkkonen, Janne Aukia and Samuel Kaski. Infinite mixtures for multi-relational categorical data. In *Proceedings of the 6th International Workshop on Mining and Learning with Graphs (MLG 2008)*, Helsinki, Finland, 2008.