Associative Clustering by Maximizing a Bayes Factor

Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
{janne.sinkkonen,janne.nikkila,leo.lahti,samuel.kaski}@hut.fi

Abstract

Clustering by maximizing the dependency between (margin) groupings or partitionings of co-occurring data pairs is studied. We suggest a probabilistic criterion that generalizes discriminative clustering (DC), an extension of the information bottleneck (IB) principle to labeled continuous data. The criterion is the Bayes factor between models assuming dependence and independence of the two cluster sets, and it can be used as a well-founded criterion for IB for small data sets. With suitable prior assumptions the Bayes factor is equivalent to the hypergeometric probability of a contingency table with the optimized clusters at the margins, and for large data it becomes the standard mutual information. An algorithm for two-margin clustering of paired continuous data, associative clustering (AC), is introduced. Genes are clustered to find dependencies between gene expression and transcription factor binding, and dependencies between expression in different organisms.

1 Introduction

Distributional clustering by the information bottleneck (IB) principle [20] groups nominal values x of a random variable X by maximizing the dependency of the groups with another, co-occurring discrete variable Y. Clustering documents x by the occurrences of words y in them is an example. For a continuous X, the analogue of IB is to *partition* the space of possible values $\mathbf{x} \in \mathbb{R}^{d_x}$ by discriminative clustering (DC); then the dependency of the partitions and y is maximized [18]. IB can be used for grouping values at both the X and Y margins [7], but for continuous data no equivalent two-margin partition method has been presented so far.

Both DC and IB maximize dependency between representations of random variables. Their dependency measures are asymptotically equivalent to mutual information (MI);[1] the empirical mutual information used by IB and some forms of DC, is problematic for finite data sets, however. A likelihood interpretation of empirical MI [18, 21] opens a way to probabilistic dependency measures that are asymptotically equivalent to MI but perform better for finite data sets [19]. The current likelihood formulation, however, breaks down in the case of two-margin clustering.

[1] Yet another example of dependency maximization is canonical correlation analysis, which uses a second-moment criterion equivalent to mutual information assuming normally distributed data [13].

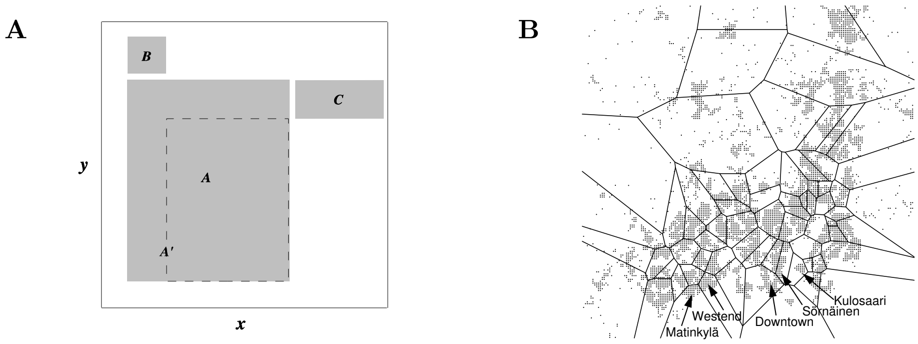

Figure 1: **A** Demonstration of the difference between dependency and joint density modeling. The hypothetical joint domain of two one-dimensional continuous margin variables x and y is shown by the outer square, and joint pdf is shown by shades of gray. A generative model of the joint distribution would focus on modeling the main probability mass in the region A, while AC that optimizes the Bayes factor (2) would focus on the areas not explainable as products of the marginal distributions. These would include B, C, and to a lesser extent the part of A denoted by A' and the dashed line. **B** Partitioning of Helsinki region into demographically homogeneous regions with AC. Here $\mathbf{x}$ contains geographic coordinates of buildings and demographic information about inhabitants indicating social status, family structure, etc. For instance downtown and close-by relatively rich (Kulosaari, Westend) areas become separated from less well-off areas.

In this paper we have two goals: (i) to present a general probabilistic criterion for IB and DC-like dependency modeling that is easily generalizable to the two-margin and other interesting cases [5], and (ii) to apply the criterion to two-way clustering of co-occurring continuous data.

We suggest using a Bayes factor, extended from a DC optimization criterion [19], as a measure of dependency for groupings or partitions formed of co-occurrence data. It compares the evidence for two models, one assuming dependent generation of data in the margin clusters, and the other assuming independent clusters for x and y. With suitable prior assumptions, the Bayes factor is equivalent to a hypergeometric probability commonly used as a dependency measure of contingency tables. It is well justified for finite data sets, avoiding the problems of empirical mutual information due to sampling uncertainty, yet asymptotically equivalent to mutual information for large data sets. The Bayes factor is usable as the cost function of DC and IB, as well as in novel setups such as fully continuous co-occurrence data, here called *associative clustering* (AC).

Another variant of AC, clustering exchangeable pairs of co-occurrence data from the same vector space, is related to "clustering with side data" of similarity [24]. The main difference is that instead of enhancing the similarity of the pairs AC maximizes their dependency, which is more flexible in the sense that pairs are not necessarily required to end up in the same cluster. Another difference in the current version of AC is that it cannot use additional non-paired data, analogous to non-labeled data in classification tasks, although it can map unpaired data into existing partitions. Extension to accommodate unpaired data along the lines of [12] is probably possible.

Both variants of AC will be applied in Section 3 for finding dependencies in gene expression data.

Another line of work on generative modeling of joint density of co-occurring data becomes close to our approach. Mixture and component models exist for discrete data [3, 4, 10], and the joint Gaussian mixture models of the MDA type [9, 15] could be extended to model two continuous margins. The difference is that the Bayes factor introduced here focuses only on the *dependencies* between the variables, skipping the parts of the joint distribution representable as a product of margins. Both goals are rigorous but different, as illustrated in Figure 1**A**.

2 Bayes factor for maximizing dependency between two sets of clusters

The dependency between two clusterings, indexed by i and j, can be measured by mutual information if their joint distribution p_{ji} is known. If only a *contingency table* of co-occurrence frequencies n_{ji} computed from a finite data set is available, the mutual information computed from the empirical distribution would be a biased estimate. A Bayesian finite-data alternative is the *Bayes factor* between models that assume dependent and independent margins. Bayes factors have been classically used as dependency measures for contingency tables (see, e.g., [8]). We will use the classical results as building blocks to derive the Bayes factor to be optimized; the novelty here is that we suggest maximizing the Bayes factor instead of only measuring fixed tables with it.

In general, frequencies over the cells of a contingency table are multinomially distributed. The model M_i of *independent margins* assumes that the multinomial parameters over cells are outer products of posterior parameters at the margins: $\theta_{ij} = \theta_i \theta_j$. The model M_d of *dependent margins* ignores the structure of the cells as a two-dimensional table and samples cell-wise frequencies directly from a table-wide multinomial distribution θ_{ij}. Dirichlet priors are assumed for both the margin and the table-wide multinomials.

Maximization of the Bayes factor

$$\mathrm{BF} = \frac{p(M_d|\{n_{ji}\})}{p(M_i|\{n_{ji}\})} \tag{1}$$

with respect to the margin clusters then gives a contingency table where the margins are maximally dependent, that is, which cannot be explained as a product of independent margins. In the two-way associative clustering introduced in this paper, margins are defined by the learning set and parameterized as partitionings of the continuous data spaces. Then BF is maximized with respect to the parameters. If applied to two-way IB, the margins would be determined as groupings of nominal values of the discrete margin variables, and BF would be maximized with respect to different groupings.

After marginalization over the multinomial parameters, the Bayes factor can be shown to take the form

$$\mathrm{BF} = \frac{\sum_{ji} \Gamma(n_{ji} + n^{(d)})}{\sum_i \Gamma(n_{\cdot i} + n^{(x)}) \sum_j \Gamma(n_{j\cdot} + n^{(y)})} \;, \tag{2}$$

with $n_{\cdot i} = \sum_j n_{ji}$ and $n_{j\cdot} = \sum_i n_{ji}$ expressing the margins and the parameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arising from the Dirichlet priors. We have set all three to unity, which makes BF equivalent with the hypergeometric probability classically used as a dependency measure of contingency tables. For large data sets, (2) is approximated by mutual information of the margins; [19] outlines the proof for the case of one fixed and one parameterized margin.

2.1 Associative clustering for partitioning continuous margins

For paired data $\{(\mathbf{x}_k, _k)\}$ of real vectors $(\mathbf{x},) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, we search for partitionings $\{V_j^{(x)}\}$ for $\mathbf{x}$ and $\{V_i^{(y)}\}$) for . The partitions can be interpreted as clusters in the same way as in K-means; they are Voronoi regions parameterized by their centroids $\mathbf{m}_j$: $\mathbf{x} \in V_j^{(x)}$ if $\|\mathbf{x} - \mathbf{m}_j\| \leq \|\mathbf{x} - \mathbf{m}_k\|$ for all k, and correspondingly for . The Bayes factor (2) will be maximized with respect to the Voronoi centroids.

The optimization problem is combinatorial for hard clusters, but gradient methods are applicable after the cluster borders are smoothed. Gradients for the simpler fixed-margin problem have been derived in [19], and are analogous here. An extra trick has been found to improve the optimization in the fixed-margin case [12], and applied here as well: The denominator of the Bayes factor is given extra weight to effectively smooth the distribution of data over the margin clusters. This is asymptotically equivalent to an additional term that rewards for high margin cluster entropies.

Clustering by AC is demonstrated in Figure 1**B**.

2.2 Clustering discrete margins

For discrete $\{(x_k, y_k)\}$, the clustering means grouping the nominal margin values. Dependencies in this kind of data sets have been classically analyzed by the information bottleneck, using empirical mutual information as the cost function [20].

The Bayes factor (2) is an alternative criterion, particularly suitable for small data sets. For very large data sets the two criteria would be equivalent, for both (2) and empirical mutual information would approach the real mutual information.

In the sequential information bottleneck algorithm [20], a randomly chosen sample is iteratively re-assigned into the cluster which minimizes a Jensen-Shannon divergence-based criterion. Plugging in the Bayes factor (2) would result in a Bayesian re-assignment criterion that is asymptotically equivalent to Jensen-Shannon.

2.3 Associative clustering of exchangeable pairs

If all data are in the same space, additional information about known dependencies between data pairs can be incorporated into the clustering with variants of basic AC. A special case introduced here is that of "swappable" (internally exchangeable) pairs $(\mathbf{x}^1, \mathbf{x}^2)$, that would just describe an undirected probabilistic association between $\mathbf{x}^1$ and $\mathbf{x}^2$. The one set of clusters will be optimized to maximize the Bayes factor

$$\mathrm{BF} = \frac{\sum_{j \leq i} \Gamma(n_{ji} + n^{(d)})}{\sum_k \Gamma(n_k + n^{(i)})} \tag{3}$$

between dependent assignment of pair members to clusters, and assignment of the members by ignoring the information of pairings available in the data. Now $n_{ji} = n_{ij}$ denotes the number of pairs with one member falling to cluster i and the other to j, and n_k is the number of pair members in cluster V_i.

3 Clustering of genes

The DNA of an activated gene is transcribed into mRNA molecules that are transported in the cell as blueprints for protein manufacturing. With microarray techniques, mRNA concentration or *gene expression* of thousands of genes can be measured simultaneously. Expression is regulated by a complex network, in which proteins called *transcription factors* (TFs) play a major part by binding into a regulatory portion (promoter) of the DNA of the gene. Expression regulation directs the internal states of the cell and is therefore a fundamental determinant of all biological function.

Expression data is noisy due to measurement errors and irrelevant biological variation. Before building more complex models [17], it is therefore commonly explored by clustering [6, 2]. AC is then optimal if dependencies of gene expression patterns with other data are sought, as is done in this paper with respect to TF binding intensities and expression in another organism. AC as a nonparametric dependency model will reveal the presence or lack of dependencies between data sets—the latter being not so uncommon in the case of expression data. If the margin data sets are dependent, AC will find clusters with potentially interesting interactions.

3.1 Clusters with coherent expression and TF binding patterns

AC was applied to 6185 genes of common yeast, *Saccharomyces cerevisiae*. The first margin data was 300-dimensional, consisting of expressions after 300 knock-out mutations[2] [11]. The second margin data consisted of 113-dimensional patterns of binding intensities of putative regulatory TFs [14]. Margin clusters, being Voronoi regions, would then be sets of mutually relatively similar expressions and TFs, selected to produce interactions (or cells of a contingency table) with unexpectedly high or low numbers of co-occurrences. Richly populated cells of the table indicate a large number of genes regulated in the same way by a set of TFs, which is interesting from the viewpoint of understanding gene activity regulation of cells.

The numbers of margin clusters were chosen to produce table cells with ten data points on average. AC margin clusters were initialized by K-means, and K-means by choosing the best of three runs. Optimization parameters were chosen with a validation set. There was a significant interaction between gene expression and TF binding, as evaluated by comparing contingency tables produced by AC and independent margin K-means (tables evaluated by (2), 10-fold cross-validation, paired t-test, $p<0.0001$). The latent dependency structure revealed by AC is also visually apparent from the contingency tables produced by the two methods: In the AC table, more cells differ from the null hypothesis of independent margins (left-out data of one cross-validation fold shown in Fig. 2).

Finally, AC of all data was computed to find and interpret interesting clusters (Fig. 3**A**). A set of cells with most unexpectedly high data counts was chosen. One of these cells contained genes related to the cell cycle: secretion activity, cytokinesis, and mitosis. It is plausible that proteins for this natural sequence of processes are produced simultaneously and hence regulated by the same set of TFs, as suggested by the clustering. Other genes of the same cluster are potentially novel contributors to the cell cycle. We also found a group contributing to building cell walls and cytoskeleton organization, and possibly subsequently building mitochondria.

3.2 Of mice and men

After the human genome has been sequenced, annotation of genes by their functions is the major on-going effort of genomics. Functions are easier to study in model

[2]Knocking out means elimination of single genes. In all the data sets, missing values were imputed by gene-wise averages, and variances of dimensions were each separately normalized to unity.

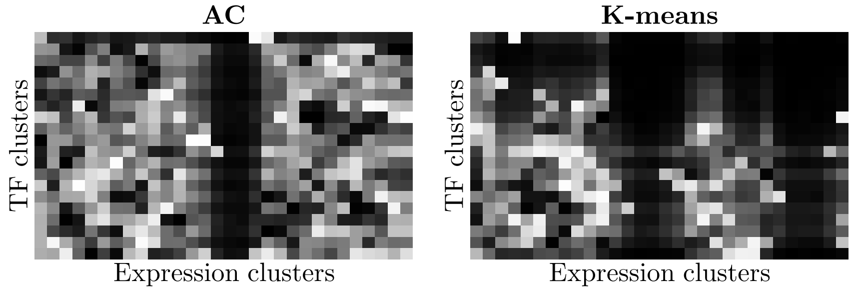

Figure 2: Deviation of contingency table frequencies from the null hypothesis of independent margins (for left-out data). Shades of gray denote p-values for the observed counts (light: p=0, count is abnormally high or low; black: p=1, the count has the expected value). The p-values are from a simulation of 1000 tables from the Dirichlet posterior representing the null hypothesis. Rows and columns are ordered with simulated annealing to maximize similarity of neighbor cells. Note that margins or cell locations of the two tables are not comparable.

organisms such as mouse, since expression can be compared in various treatments and in mutants. Knowledge on gene functions can be generalized from one organism to another on the basis of DNA sequences: their similarity suggests common evolutionary origin of the genes and hence similar function, or *orthology* of genes. The problems are that (i) functions may have diverged, (ii) the putative orthology may be erroneous, and (iii) orthologies may be unknown or nonexistent.

We clustered human and mouse expression profiles by AC, to maximize dependency between putatively orthologous gene pairs. Clusters will then be Voronoi regions of expression, that is, in a sense functionally definitely similar instead of being only putatively similar. After the AC solution is obtained, genes with no known orthologs can be mapped to the margin clusters and the contingency table suggests dependencies in their function.

Gene expression from 46 and 45 cell-lines (tissues) of human and mouse were available, respectively [22]. After removing non-expressed genes (Affymetrix AD<200), 4499 putative orthologs from the the Homologene [16] data base were available. After experiments analogous to those of Section 3.1, we found the orthologs in human and mouse to be significantly dependent (p<0.001).

The contingency table obtained by AC for all data (Fig. 3**B**) helps in interpreting the relationships between gene function of the two organisms. Each human cluster (column) is related to only a few mouse clusters (has only a few light squares), and vice versa. Many of the contingency table cells contained genes that were expressed only in a certain tissue in both mouse and man. Sample literature searches confirmed known expression of a joint cluster in the cardiac muscle and another in the testis. The mouse clusters that become divided into differently expressed genes in humans are particularly important, and their study is under way.

Finally, AC for exchangeable pairs was used to collapse all mouse and human genes into *one* set of clusters. Here expression vectors consisted of measurements from 21 tissues common to both organisms. As preprocessing, the variance of each tissue in each organism was normalized to unity. Again, the ortholog pairs were significantly dependent (p<0.001). In the contingency table (Fig. 3**C**) most data occurs near the diagonal, suggesting that the model is capable of constructing clusters for which pairings imply similarity, that is, the members of a human–mouse expression pair

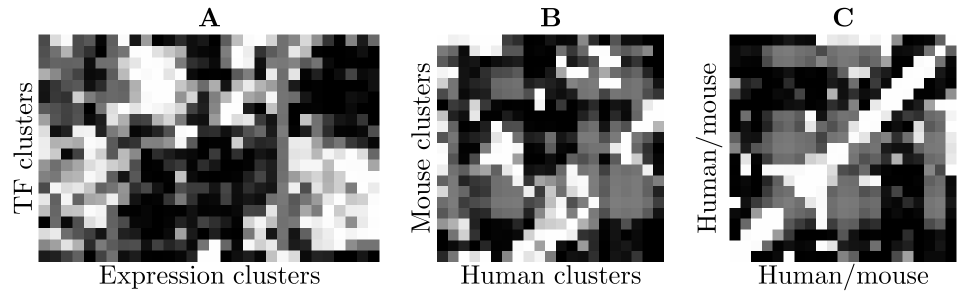

Figure 3: Contingency tables for (**A**) transcription factor clusters vs. gene expression clusters in yeast, (**B**) human gene expression clusters vs. mouse gene expression clusters, and (**C**) human/mouse AC with assumption of exchangeable pairs. For interpretation see Figure 2, except that the gray shades are different here. Light: count of the cell is unexpectedly high; black: low; gray: expected.

fall into the same cluster. The next step is to use the result to interpret new genes without known orthologs.

4 Discussion

We have presented a probabilistic measure of dependency for clustering, applicable to the information bottleneck and partitioning of continuous co-occurrence data. The latter method, associative clustering (AC), was found capable of extracting interesting structure from paired gene expression data sets.

Maximization of the suggested Bayes factor is asymptotically equivalent to maximization of mutual information, and could therefore be seen as a dependency criterion alternative to empirical mutual information. It additionally gives information-bottleneck type dependency modeling a new justification that is clearly different from joint distribution models but still rigorously probabilistic.

The work could possibly be extended towards a compromise between strict dependency modeling and a model of the joint density (as has been done for one-sided clustering, [12]). Then the margins could be estimated in part from non-paired data. This is analogous to "semisupervised learning" from partially labeled data (see e.g. [23]), the labels having been replaced by samples of co-occurring paired data.

Acknowledgments

This work has been supported by the Academy of Finland, grants 50061 and 52123. We thank Eero Castrén and Christophe Roos for biological interpretations of the results.

References

[1] O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *PNAS*, pages 3351–3356, March 2003.

[2] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–97, 1999.

[3] D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] W. Buntine. Variational extensions to EM and multinomial PCA. In *Proc ECML'02*, Lecture Notes in Artificial Intelligence 2430. Springer, Berlin, 2002.

[5] G. Chechik and N. Tishby. Extracting relevant structures with side information. In *NIPS, 2002*. MIT Press, 2002. To appear.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS, USA*, 95:14863–14868, 1998.

[7] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Proc. UAI 7*. Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[8] I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4(6):1159–1189, Nov. 1976.

[9] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *JRSS B: Methodological*, 58:155–176, 1996.

[10] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[11] T. R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[12] S. Kaski, J. Sinkkonen, and A. Klami. Regularized discriminative clustering. In *Proc. NNSP*. 2003. Accepted for publication.

[13] J. Kay. Feature discovery under contextual supervision using mutual information. In *Proc IJCNN'92*, pages 79–84. IEEE, 1992.

[14] T. Lee et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[15] D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *NIPS 9*, pages 571–577. MIT Press, Cambridge, MA, 1997.

[16] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29:137–141, 2001.

[17] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17(Suppl 1):243–252, 2003.

[18] J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[19] J. Sinkkonen, S. Kaski, and J. Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In *Proc. ECML'02*, Lecture Notes in Artificial Intelligence 2430, pages 418–430, Berlin, 2002. Springer.

[20] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proc. ACM SIGIR*, pages 129–136. ACM Press, 2002.

[21] N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck. In *NIPS 14*. MIT Press, Cambridge, MA, 2002.

[22] A. I. Su et al.. Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99:4465–4470, 2002.

[23] M. Szummer and T. Jaakkola. Kernel expansions with unlabeled examples. *NIPS 13*. MIT Press, Cambridge, MA, 2001.

[24] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side information. In *NIPS 14*. MIT Press, 2002. to appear.