Publication 5

Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. 2004. Associative clustering. In: Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi (editors). Proceedings of the 15th European Conference on Machine Learning (ECML 2004). Pisa, Italy. 20-24 September 2004. Berlin, Heidelberg, Germany. Springer. Lecture Notes in Computer Science, volume 3201, pages 396-406. ISBN 3-540-23105-6.

Associative Clustering

Janne Sinkkonen[1], Janne Nikkilä[1], Leo Lahti[1], and Samuel Kaski[2,1]

[1]Helsinki University of Technology, Neural Networks Research Centre,
P.O. Box 5400, FIN-02015 HUT, Finland
[2]Department of Computer Science,
P.O. Box 68, FIN-00014 University of Helsinki, Finland
{Janne.Sinkkonen,Janne.Nikkila,Leo.Lahti,Samuel.Kaski}@hut.fi
http://www.cis.hut.fi/projects/mi

Abstract. Clustering by maximizing the dependency between two paired, continuous-valued multivariate data sets is studied. The new method, *associative clustering (AC)*, maximizes a Bayes factor between two clustering models differing only in one respect: whether the clusterings of the two data sets are dependent or independent. The model both extends Information Bottleneck (IB)-type dependency modeling to continuous-valued data and offers it a well-founded and asymptotically well-behaving criterion for small data sets: With suitable prior assumptions the Bayes factor becomes equivalent to the hypergeometric probability of a contingency table, while for large data sets it becomes the standard mutual information. An optimization algorithm is introduced, with empirical comparisons to a combination of IB and K-means, and to plain K-means. Two case studies cluster genes 1) to find dependencies between gene expression and transcription factor binding, and 2) to find dependencies between expression in different organisms.

1 Introduction

Distributional clustering by the information bottleneck (IB) principle [21] 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 continuous-valued X, the analogue of IB is to *partition* the space of possible values $\boldsymbol{x} \in \mathbb{R}^{d_x}$ by discriminative clustering (DC); then the dependency of the partitions and y is maximized [16].

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 variants of DC is problematic for finite data sets, however. A likelihood interpretation of empirical MI (see [16]) opens a way to probabilistic dependency measures that

[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 [11].

are asymptotically equivalent to MI but perform better for finite data sets [17]. The current likelihood formulation, however, breaks down when both margins are clustered simultaneously.

In this paper we introduce a novel method, *associative clustering (AC)*, for clustering of paired continuous-valued data by maximizing the dependency between the clusters of X and Y, later called *margin clusters*. A sample application is search for different types of city districts, by partitioning a city into demographically homogeneous regions (Fig. 1**B**). Here the paired data are the coordinates and demographics of the buildings of the city.

As a measure of dependency between the cluster sets, we suggest using a Bayes factor, extended from an optimization criterion for DC [17]. The criterion compares evidence for two models, one assuming independent margin clusters (clusters for $\boldsymbol{x}$ and $\boldsymbol{y}$), and the other allowing more general dependency of the margin clusters in generating data. With suitable prior assumptions the Bayes factor is equivalent to a hypergeometric probability commonly used as a dependency measure for contingency tables. It is well justified for finite data sets, avoiding the problems of empirical mutual information due to sampling uncertainty. Yet it is asymptotically equivalent to mutual information for large data sets. The Bayes factor is also usable as the cost function of IB [14].

AC will be applied for finding dependencies in gene expression data. It will be compared with standard K-means, computed independently for the two margins, which provides a baseline result. The comparison relevals how much is gained by explicit dependency modeling.

AC will additionally be compared with a new variant of IB. IB operates on discrete data, and therefore the continuous multivariates need first to be discretized into atomic regions, for example with K-means. The symmetric IB [5] can then compose discrete representations for the margins as combinations of the atomic regions. Again dependence of the representations is the criterion for clustering, and a dependency-maximized contingency table spanned by the margin clusters results. K-means discretization was chosen because its parameterization is similar to AC and, more importantly, because it is perhaps the most obvious alternative for multidimensional discretization.[2]

In IB, dependency has classically been measured by the (empirical) mutual information. As margin clusters are here combinations of very small Voronoi regions, IB finds dependencies between the data sets well, but on the other hand produces clusters that are potentially less local than those obtained by AC or standard K-means. We will evaluate the average dispersion of the clusters in the empirical tests of Section 4.

Both mixture models for discrete data [2,3,8] and Mixture Discriminant Analysis (MDA)-like [7,13] models for continuous data have common elements with our approach, and can readily be extended for the double-margin case. An im-

[2] Note that discretizing the dimensions independently of each other and using the Cartesian product as the multidimensional partitioning would fail badly for high-dimensional $\boldsymbol{x}$ or $\boldsymbol{y}$. As far as we know, better discretization methods or other comparable methods for co-clustering of continuous data have not been published.

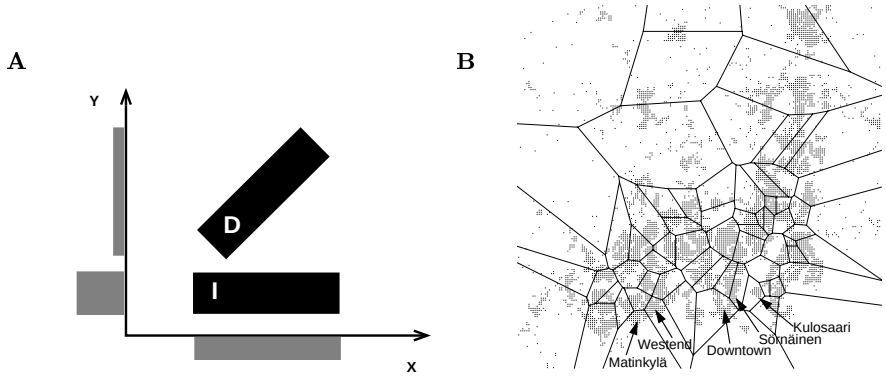

Fig. 1. A Demonstration of the difference between dependency modeling and joint density modeling. The hypothetical joint density of two one-dimensional variables x and y is plotted with black, and the respective marginal densities are depicted as histograms (*grey*). The marginals, here for simplicity univariate, correspond to the paired data of the AC setting. The visualized joint distribution consists of two equally-sized parts: a block in which x and y are independent (denoted by I) and another block (D) where x and y are dependent. Models for the joint distribution would focus equally on both blocks, whereas AC and IB focus on the dependent block D not explainable as products of the marginals, and neglect the independent block I. **B** Partitioning of Helsinki region into demographically homogeneous regions with AC. Here $\boldsymbol{x}$ contains geographic coordinates of buildings and $\boldsymbol{y}$ demographic information about inhabitants indicating social status, family structure, etc. Spatially relatively compact yet demographically homogeneous clusters emerge. For instance downtown and close-by relatively rich (Kulosaari, Westend) areas become separated from less well-off areas

portant difference is that our optimization criterion (as well as that of the Information Bottleneck) 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 Associative Clustering

2.1 Bayes Factor for Maximizing Dependency between Two Sets of Clusters

The dependency between two clusterings, indexed by i and j, for the same set of objects can be measured by mutual information if their joint distribution p_{ij} is known. If only a *contingency table* of co-occurrence frequencies n_{ij} computed from a finite data set is available, the mutual information computed from the empirical distribution would be a biased estimate. A *Bayes factor*, to be introduced below, then has the advantage of properly taking into account the finiteness of the data while still being asymptotically equivalent to mutual information. Bayes factors have been classically used as dependency measures for contingency tables (see, e.g., [6]) by comparing a model of dependent margins to another one

for independent margins. We will use the classical results as building blocks to derive an optimizable criterion for associative clustering; the novelty here is that the Bayes factor is optimized instead of only using it to measure dependency in a fixed table.

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

$$BF = \frac{p(\{n_{ij}\}|M_d)}{p(\{n_{ij}\}|M_i)}$$

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 associative clustering introduced in this paper, the data counts are defined by the training data set and the parameters that determine how the continuous data spaces are partitioned into margin clusters. Then *BF* is maximized with respect to the parameters. If this principle were applied to two-way IB, the margins would be determined as groupings of nominal values of the discrete margin variables, and the *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

$$BF = \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_i \Gamma(n_{i\cdot} + n^{(x)}) \prod_j \Gamma(n_{\cdot j} + n^{(y)})} \,, \tag{1}$$

with $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$ expressing the margins. The parameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arise from Dirichlet priors. We have set all three parameters to unity, which makes *BF* equivalent to the hypergeometric probability classically used as a dependency measure of contingency tables. In the limit of large data sets, (1) becomes mutual information of the margins; [17] outlines the proof for the case of one fixed and one parameterized margin.

2.2 Optimization of AC

For paired data $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}$ of real vectors $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, we search for partitionings $\{V_i^{(x)}\}$ for $\boldsymbol{x}$ and $\{V_j^{(y)}\}$) for $\boldsymbol{y}$. The partitions can be interpreted as clusters in the same way as in K-means; they are Voronoi regions parameterized by their centroids $\boldsymbol{m}_i$: $\boldsymbol{x} \in V_i^{(x)}$ if $\|\boldsymbol{x} - \boldsymbol{m}_i\| \leq \|\boldsymbol{x} - \boldsymbol{m}_k\|$ for all k, and correspondingly for $\boldsymbol{y}$. The Bayes factor (1) will be maximized with respect to the Voronoi centroids.

The optimization problem is combinatorial for hard clusters, but gradient methods are applicable after the clusters are smoothed. Gradients for the simpler one-margin problem have been derived in [17], and are analogous here. An extra trick, found to improve the optimization in the fixed-margin case [10], is applied here as well: The denominator of the Bayes factor is given extra weight. A choice of $\lambda^{(\cdot)} > 1$ introduces a regularizing term to the cost function that for large sample sizes approaches margin cluster entropy, and thereby in general favors solutions with uniform margin distributions.

The smoothed *BF*, here called *BF'*, is then optimized with respect to the $\{\boldsymbol{m}\}$ by a conjugate-gradient algorithm (see, for example [1]). We have

$$\log BF' = \sum_{ij} \log\Gamma\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) g_j^{(y)}(\boldsymbol{y}_k) + n^{(d)}\right)$$
$$- \lambda^{(x)} \sum_i \log\Gamma\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) + n^{(x)}\right) - \lambda^{(y)} \sum_j \log\Gamma\left(\sum_k g_j^{(y)}(\boldsymbol{y}_k) + n^{(y)}\right) ,$$
$$g_i^{(x)}(\boldsymbol{x}) \equiv Z^{(x)}(\boldsymbol{x})^{-1} \exp\left(-\|\boldsymbol{x} - \boldsymbol{m}_i^{(x)}\|^2 / \sigma_{(x)}^2\right) ,$$

and similarly for $g^{(y)}$. The $g(\cdot)$ are the smoothed Voronoi regions at the margins. The $Z(\cdot)$ is set to normalize $\sum_i g_i^{(x)}(\boldsymbol{x}) = \sum_j g_j^{(y)}(\boldsymbol{y}) = 1$. The parameters σ control the degree of smoothing of the Voronoi regions.

The gradient of $\log BF'$ with respect to an X-prototype $\boldsymbol{m}_i^{(x)}$ is

$$\nabla_{\boldsymbol{m}_i^{(x)}} \log BF' = \frac{1}{\sigma_{(x)}^2} \sum_{k,i'} (\boldsymbol{x}_k - \boldsymbol{m}_i^{(x)}) g_i^{(x)}(\boldsymbol{x}_k) g_{i'}^{(x)}(\boldsymbol{x}_k) \left(L_i^{(x)}(\boldsymbol{y}_k) - L_{i'}^{(x)}(\boldsymbol{y}_k)\right) ,$$

where

$$L_i^{(x)}(\boldsymbol{y}) \equiv \sum_j \Psi\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) g_j^{(y)}(\boldsymbol{y}_k) + n^{(d)}\right) g_j^{(y)}(\boldsymbol{y}) - \lambda^{(x)} \Psi\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) + n^{(x)}\right) ,$$

and for y accordingly. In the gradient, $\Psi(\cdot)$ is the digamma function.

Note that the smoothing is used during optimization only. Results are evaluated with hard clusters and the original *BF*.

3 Reference Methods

3.1 Information Bottleneck with K-means (K-IB)

For discrete X and Y, AC-type of clustering translates to grouping the nominal margin values to two sets of clusters that are maximally dependent. The setup is that of the information bottleneck [18,21].

Our continuous data must be discretized before IB can be applied. One approach is to first quantize the vectorial margins $\boldsymbol{x}$ and $\boldsymbol{y}$ separately by, for instance., K-means, without paying attention to possible dependencies between the two margins. This results in two sets of margin partitions which span a large, sparse contingency table that can be filled with frequencies of training data pairs $(\boldsymbol{x}_k, \boldsymbol{y}_k)$. The number of elementary

Voronoi regions is chosen by a validation set, as detailed in Section 4. In the second phase, the large table is compressed by standard IB to the desired size by aggregating the atomic margin clusters. At this stage, joins at the margins are made to explicitly maximize the dependency of margins in the resulting smaller contingency table.

IB algorithms are well described in the literature. We have used the symmetric sequential information bottleneck, described fully in [18]. The algorithm measures dependency of the margins by empirical mutual information, and it is optimized by re-assigning of individual samples (here atomic margin partitions) to clusters until a differential, local version of the cost function does not decrease. Optimization is robust and fast.

The final partitions obtained by the combination of K-means and IB are of a very flexible form, and therefore the method is expected to model the dependencies of the margin variables well—as long as one does not overfit to the data with too many K-means clusters. As a drawback, the final margin clusters will consist of many atomic Voronoi regions, and they are therefore not guaranteed to be especially homogeneous with respect to the original continuous variables ($\boldsymbol{x}$ or $\boldsymbol{y}$). Interpretation of the clusters may then be difficult. Our empirical results support both the good performance of K-IB and the non-localness of the resulting clusters.

3.2 K-means

The data sets will also be clustered by independent K-means clusterings in both data spaces. Results will represent a kind of a baseline, with no attempt to model dependency.

4 Experiments

4.1 Dependencies between Gene Expression Patterns and TF Binding Patterns

We sought gene regulation patterns by exploring dependencies between gene expression on the one hand, and measurement data about potential regulatory interactions on the other. The latter was measurements of binding patterns of putative regulatory proteins, transcription factors (TFs), in the promoter regions of the same genes. Associative clustering, K-IB, and K-means were applied to 6185 genes of the common yeast, *Saccharomyces cerevisiae*. The first margin data ($\boldsymbol{x}$) was 300-dimensional, consisting of expressions after 300 knock-out mutations[3] [9]. The second margin data ($\boldsymbol{y}$) consisted of 113-dimensional patterns of binding intensities of TFs [12]. Margin clusters would then ideally be internally homogeneous sets of expressions and TFs, selected to produce combinations (contingency table cells) with unexpectedly high or low numbers of genes.

For AC, the numbers of margin clusters were chosen to produce cross clusters (contingency table cells) with ten data samples on average. During the cross-validation runs margin clusters were initialized by K-means, and in each fold the best of three AC runs was chosen as the final AC clustering. The parameters $\sigma_{(\cdot)}$ were chosen with

[3] 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.

a validation set (half of the data as a training set, and half of the data as validation set), and based on the previous experiments $\lambda^{(\cdot)}=1.2$.

Essentially the same test was conducted for the combination of K-means and information bottleneck (K-IB). Now the number of atomic K-means clusters was chosen with a validation set, resulting in 400 clusters for the expression space and 300 clusters for the transcription factor binding space. In the cross-validation runs, the atomic clusters were computed by K-means from three different random initializations, and for each of these a symmetric IB was sequentially optimized [18]. Of the three runs the best clustering (in the sense of IB cost) was chosen.

K-IB and AC tables were compared to each other and to tables obtained by bare margin K-means (10-fold cross validation, tables evaluated by equation 1, paired t-test). For this data, AC outperformed K-IB ($p<0.01$) and found more dependent clusters. Not surprisingly, significant differences to K-means were found ($p<0.01$) for both AC and K-IB.

The internal dispersion of the margin clusters was measured for all methods by the sum of intra-cluster component-wise variances. As expected, K-IB clusters are more scattered (Figure 2**A**) in both data spaces. Significant difference was found between AC and K-IB, but not between AC and K-means, nor between the random partitioning and K-IB.

Finally, data from the AC cross clusters was studied more closely to find potential biologically interesting gene concentrations, focusing on contingency table cells with the most unexpectedly high data counts. In two of the cells, for example, genes showed a clear and significant bias towards an over-representation of ribosomal protein coding genes. In the one cell, most of the genes coding for constituent proteins of the cellular ribosomal complex are present. In the other cell several genes coding for the mitochondrial ribosomal subunits are present, and also another set of genes coding for cellular ribosomal protein subunits.

4.2 Of Mice and Men

As a second test, we clustered human-mouse expression profiles of putative orthologs, that is, gene pairs sequence-wise similar enough to be suspected to have the same evolutionary origin (see Figure 3). Ideal margin clusters would be internally homogeneous by expression in at least one species. Cross clusters (cells of the contingency table formed from margin clusters) would then be cross-species clusters and will be optimized to detect cross-species regularities in gene expression.

Gene expression from 46 and 45 cell-lines (tissues) of human and mouse were available, respectively [19]. After removing non-expressed genes (Affymetrix AD<200), 4499 putative orthologs from the the HomoloGene [15] data base were available. After experiments analogous to those of Section 4.1, we found the human-mouse orthologs of left-out data to be significantly dependent according to both K-IB and AC (10-fold cross validation against K-means, paired t-test, $p<0.001$). Differences between AC and K-IB were not very clear. AC clusters, however, were probably more condensed ($p<0.05$; Fig. 2**B**) while tables obtained by K-IB were more dependent ($p<0.02$).

To illustrate the use of AC for finding interesting relationships, that is, groups of genes with functional similarity, we picked some cross clusters with significant deviation from the null hypothesis of independent margin clusters (see also Figure 3).

In the first example AC found a gene pair with a rare and potentially interesting functional relationship. This cell had unexpectedly few genes ($p<0.01$), in fact only

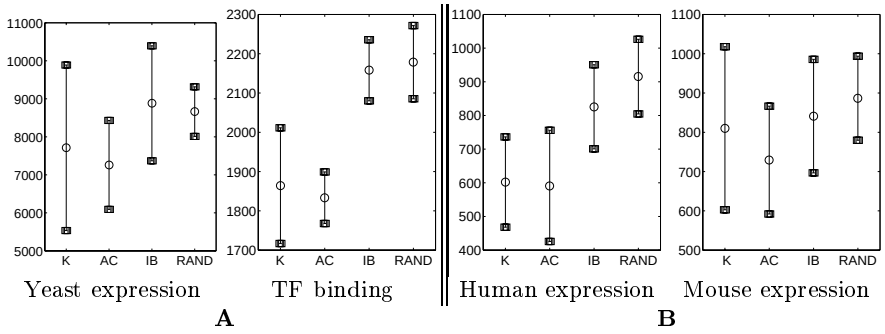

Fig. 2. Average internal dispersion of expression and TF margin clusters obtained by four methods. 'K' denotes independent K-means for the margins, and is supposed to produce very compact clusters. Clusters in 'RAND' are produced by random assignment and therefore represent an upper limit of dispersion. The AC and K-means clusters are more condensed in the expression and in the TF binding space than the IB clusters. Circles denote the average component-wise intra-cluster variances in left-out data of cross-validation folds (n=10), and squares show the approximate 99 percent confidence interval for the means over the cross validation folds. **A**: Yeast expression and TF binding. The differences between neither AC and K-means, nor between IB and RAND are statistically significant ($p>0.1$, 10-fold cross validation, paired t-test), but the difference between IB and AC is significant for both expression and TF binding ($p<0.01$). **B**: Homologous genes of human and mouse. The AC and K-means clusters seem to be more condensed in human and in mouse expression space than the IB or RAND clusters ($p<0.05$; paired t-test). Differences between AC and K-means or between IB and RAND are not statistically significant ($p>0.1$).

a single gene pair (LocusIDs 1808 and 12934). Average margin profiles of the cluster suggested activity in the human brain co-occurring with no activity in the mouse at all. Combining the margin profile information and the fact that only one this kind of gene exists in the contingency table, we may deduce that homologues which are active in human brain but totally silent in the mouse are very rare. Examples of such gene pairs may highlight interesting functional differences between the species. Indeed, the function of the gene was found to be related to embryo-stage brains *and* later brain activity only in humans (see Figure 3B).

In another example, a cross cluster contained unexpectedly many genes ($p<0.01$), most of them testis-specific (see Figure 3D). Due to their tissue specificity and importance for reproduction, they may have sustained their function during evolution.

5 Discussion

We have presented a novel method, associative clustering (AC), for clustering continuous paired data. It maximizes a Bayes factor between two sets of clusters. AC was found to perform better or equally well than a combination of K-means and information bottleneck (IB), and better than standard K-means. AC was also capable of extracting biologically interesting structure from paired gene expression data sets.

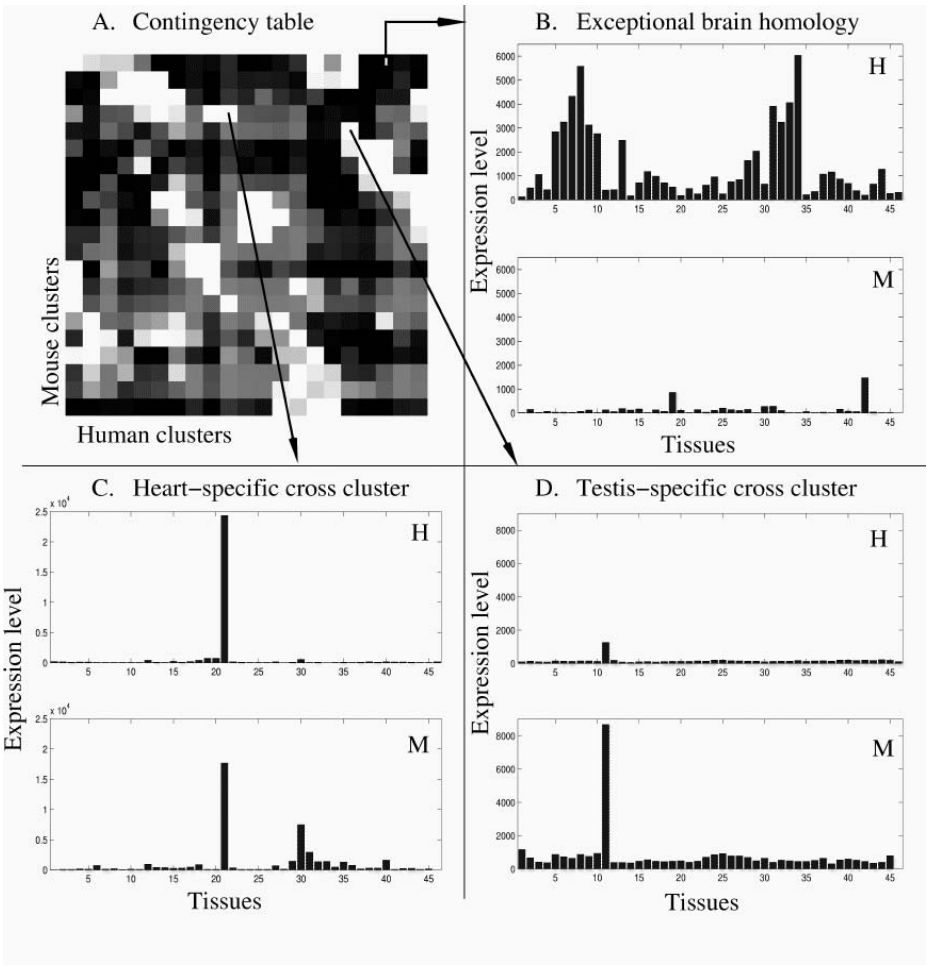

Fig. 3. **A** The contingency table from associative clustering of orthologous human-mouse gene pairs (orthologous genes are supposed or known to have a common evolutionary ancestor gene). *White cross clusters* contain an unexpectedly high number of genes compared to the margin-based expectation. *Black cross clusters* contain examples of exceptional gene pairs. **B** An example of an interesting outlier homology from a black cross cluster: the gene is highly active in most human tissues but is hardly expressed at all in mouse. The first 21 tissues are common for both species in B, C and D. **C** Cluster-wide average profiles reveal activity in heart tissue, and additional strong activity in mouse skeletal muscle. Measuring human skeletal muscle would reveal either a more complete homology or a species difference. **D** A densely populated cross cluster of testis-specific genes.

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 Bayes factor could probably replace mutual informaton in the Information-Theoretic Co-Clustering Algorithm [4] as well.

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, [10]). Then the margins could be estimated in part from non-paired data. This would be analogous to "semisupervised learning" from partially labeled data (see e.g. [20]), the labels having been replaced by samples of co-occurring paired data.

Acknowledgments This work has been supported by the Academy of Finland, decisions #79017 and #207467. We thank Jaakko Peltonen for the code for sequential IB, and Juha Knuuttila and Christophe Roos for help with biological interpretation of the results.

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms (1993). Wiley, New York
2. Blei, D., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Machine Learning Res. **3** (2003) 993–1022
3. Buntine, W.: Variational extensions to EM and multinomial PCA. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.): Proc. of the ECML'02, Lecture Notes in Artificial Intelligence, 2430 (2002). Springer, Berlin, pp. 23–34
4. Dhillon, I.S., Mallela, S., Kumar, R.: A divisive information-theoretic feature clustering algorithm for text classification. J. Machine Learning Res. **3** (2003) 1265–1287
5. Friedman, N., Mosenzon, O., Slonim, N., Tishby, N.: Multivariate information bottleneck. In: Proc. of UAI'01, The 17th Conference on Uncertainty in Artificial Intelligence (2001). Morgan Kaufmann Publishers, San Francisco, CA, pp. 152–161
6. Good, I.J.: On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. Annals of Statistics, **4** (1976) 1159–1189
7. Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixtures. J. of the R. Stat. Soc. B **58** (1996) 155–176
8. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **42** (2001) 177–196
9. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffrey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H.: Functional discovery via a compendium of expression profiles. Cell **102** (2000) 109–126
10. Kaski, S., Sinkkonen, J., and Klami, A.: Regularized discriminative clustering. In: Molina, C., Adali, T., Larsen, J., van Hulle, M., Douglas, S., Rouat, J. (eds.): Neural Networks for Signal Processing XIII (2003). IEEE, New York, NY, pp. 289–298

11. Kay, J.: Feature discovery under contextual supervision using mutual information. In: Proc. of IJCNN'92, International Joint Conference on Neural Networks (1992). IEEE, Piscataway, NJ, pp. 79–84
12. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Tomphson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.-B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science **298** (2002) 799–804
13. Miller, D.J., Uyar, H.S.: A mixture of experts classifier with learning based on both labelled and unlabelled data. In: Mozer. M, Jordan, M., Petsche, T. (eds.): Advances in Neural Information Processing Systems, 9 (1997). MIT Press, Cambridge, MA, pp. 571–577
14. Peltonen, J., Sinkkonen, J., and Kaski, S.. Sequential information bottleneck for finite data. In: Proc. of the International Conference on Machine Learning (to appear)
15. Pruitt, K.D., Maglott, D.R.: RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Research **29** (2001) 137–141
16. Sinkkonen, J. and Kaski, S.: Clustering based on conditional distributions in an auxiliary space. Neural Computation **14** (2002) 217–239
17. Sinkkonen, J., Kaski, S., and Nikkilä, J.: Discriminative clustering: Optimal contingency tables by learning metrics. In Elomaa, T., Mannila, H., Toivonen, H. (eds.): Proc. of the ECML'02, 13th European Conference on Machine Learning (2002). Springer, Berlin, pp. 418–430
18. Slonim, N.: The information bottleneck: theory and applications (2002). PhD thesis, Hebrew University, Jerusalem
19. Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz, P.G., and Hogenesch, J.B.: Large-scale analysis of the human and mouse transcriptomes. PNAS **99** (2002) 4465–4470
20. Szummer, M. and Jaakkola, T.: Kernel expansions with unlabeled examples. In: Leen, T, Dietterich, T., Tresp, V. (eds.): Advances in Neural Information Processing Systems, 13 (2001). MIT Press, Cambridge, MA, pp. 626–632
21. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: Hajek, B. and Sreenivas, R.S. (eds.): Proc. of The 37th Annual Allerton Conference on Communication, Control, and Computing (1999). University of Illinois, Urbana, Illinois, pp. 368–377