Arto Klami and Samuel Kaski. 2006. Generative models that discover dependencies between data sets. In: S. McLoone, T. Adali, J. Larsen, and M. Van Hulle (editors). Machine Learning for Signal Processing XVI. Piscataway, NJ, IEEE, pages 123-128.

GENERATIVE MODELS THAT DISCOVER DEPENDENCIES BETWEEN DATA SETS

Arto Klami, Samuel Kaski

Helsinki University of Technology
Adaptive Informatics Research Centre
P.O.Box 5400, FI-02015 HUT, Finland

ABSTRACT

We develop models for a kind of data fusion task: Combine multiple data sources under the assumption that data set specific variation is irrelevant and only between-data variation is relevant. We extend a recent generative modeling interpretation of Canonical Correlation Analysis (CCA), a traditional linear method applicable to this task, in a way which allows generalization to other types of models. The generative formulation makes all standard tools of Bayesian inference applicable. We finally introduce new dependency-seeking clustering models that outperform standard generative clustering models in their task.

1. INTRODUCTION

We study the task of modeling dependencies between two data sets of co-occurring or paired samples (x, y). In other words, the task is to find what is shared by, or statistically in common in x and y. The underlying assumption is that variation within either data set alone is more noisy, or at least less interesting than variation that is in common. Example tasks include translation where the x and y are sentences in different languages, or measurement data from two different kinds of noisy sensors such as gene expression arrays, that measure the same system.

This task has been classically solved by Canonical Correlation Analysis (CCA) [1], or more recently by other methods that maximize mutual information such as the Information Bottleneck [2]. Mutual information measures deviation from independence and is hence arguably a very good objective function for finding dependencies. Unfortunately it is defined for distributions and not data sets, and hence cannot handle well uncertainties caused by the finiteness of the data sets. Alternative Bayes factor-based dependency measures have been proposed for the task [3], but even they do not take all uncertainties into account.

Bayesian generative modeling of joint distributions, in this case of $p(x, y)$, is a traditional well-justified framework for modeling finite data sets. Complexity control of models can be formulated rigorously, which makes it possible to use flexible model structures and constrain their complexity according to the data.

Standard flexible models will, by default, try to model all variation within the data, and hence they would be even too flexible for modeling of dependency between data sets. In more focused modeling tasks it is customary to constrain the solution space using explicit prior knowledge to make independence assumptions between the variables, resulting in models structured according to the specific system being modeled.

Recently [4] it was suggested that a simple independence assumption would be sufficient for turning a generative joint distribution model into CCA. This is striking since dependency modeling and generative modeling of joint densities had earlier been considered very different tasks. The new finding raises the immediate questions of how general the relationship is, and under what conditions it holds. More generally, it would be very interesting to better understand the the relationship between the two tasks. We will start exploring these questions in this paper.

In summary, the main finding is that if a generative model has a very flexible model for both of the marginals (x and y) separately, then a very constrained model for the relationships will specialize in capturing dependencies between the data sets. We will explain and justify this in more detail, and introduce some new models which utilize this insight.

2. MODEL STRUCTURE FOR DEPENDENCY EXPLORATION

A Bayes network or graphical model can be used to represent independence assumptions in a model: if two nodes are not connected with an edge, there is no direct relation between the two variables. Traditionally, the model structure is learned from data to represent the real dependencies, but we can also use the same framework to *force* the model to use certain parameters for describing the dependencies. This can be done by imposing independence assumptions to suitable locations. The main goal of this paper is to assess whether this kind of structural focusing can help in finding dependencies.

1-4244-0657-9/06/$20.00 ©2006 IEEE.

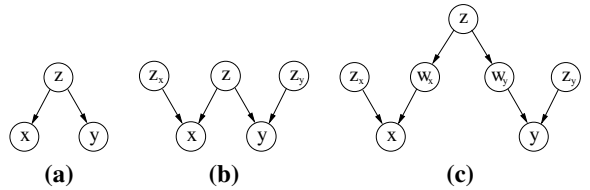


Fig. 1. Three model structures for dependency modeling, all sharing the property that the observed variables x and y have some latent variables in common.

Here we will consider three simple structures (Figure 1) for modeling two (usually multivariate) variables (x and y). All models share the property that the variables interact only through a common latent variable (or group of variables). Some general-purpose methods can be derived already from these simple structures, and thus they serve as a good basis for studying the properties of modeling dependencies with generative models.

3. CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis (CCA) is a classical linear model for finding dependencies [1]. It is formulated as finding the linear transformations W_x and W_y such that each dimension of $W_x x$ correlates maximally with the corresponding dimension of $W_y y$. CCA thus finds what the two data sets have in common, and does that explicitly by maximizing the correlation. It can be effectively computed by solving a certain generalized eigenvalue problem, and the solution is a unique global optimum.

It was recently shown that, rather surprisingly, we can find the CCA solution also as the maximum likelihood solution of a certain probabilistic model [4]. This is somewhat counterintuitive, considering how traditional CCA optimizes a completely different criterion. We will here start by summarizing the interpretation given in [4], and then proceed to an extended model structure that helps to explain why the probabilistic version of CCA could be derived.

The model structure used by [4] corresponds to that in Figure 1 (a), and the actual model is given by

$$\begin{aligned} z &\sim N(0, I), \\ x|z &\sim N(W_x z, \Psi_x), \\ y|z &\sim N(W_y z, \Psi_y), \end{aligned}$$

where we have assumed zero mean data for simplicity, and do not explicitly mention the dimensionalities to keep the notation more compact. The technical derivations revealing the connection to CCA can be found in [4], but the main point is that the maximum likelihood estimates for the projections W_x and W_y have a connection to the CCA projections, and, particularly, that the posterior expectations of z given x and y lie on the CCA projection space.

An important observation is that the marginal covariance matrix of the pair (x, y) in the above model is

$$\begin{pmatrix} W_x W_x^T + \Psi_x & W_x W_y^T \\ W_y W_x^T & W_y W_y^T + \Psi_y \end{pmatrix}, \tag{1}$$

and that the connection to classical CCA is retained as long as we have a model that has identical marginal covariance. It is worth noticing how the different parameters affect different parts of the covariance. Most notably, the parameters Ψ only affect the part related to one data set. The key to why the above model leads to canonical directions must lie here, but [4] offers no explanations.

We now proceed to explain the interpretation by extending the probabilistic model. We adopt the model structure in Figure 1 (b) with separate latent variables for the marginals, and define the following model

$$\begin{aligned} z, z_x, z_y &\sim N(0, I), \\ x|z, z_x &\sim N(W_x z + B_x z_x, \sigma_x^2 I), \\ y|z, z_y &\sim N(W_y z + B_y z_y, \sigma_y^2 I). \end{aligned}$$

Given the latent variables each sample is thus generated as a signal fusion with fixed diagonal noise. Alternatively we can think of the model for each data set as probabilistic PCA [5, 6], because $W_x z + B_x z_x$ could be written simply as $\tilde{W}\tilde{z}$ for $\tilde{W} = [W_x, B_x]$ and $\tilde{z} = [z^T, z_x^T]^T$. The novelty is in sharing part of the latent variables between two PCAs.

The marginal covariance matrix of the model is

$$\begin{pmatrix} W_x W_x^T + B_x B_x^T + \sigma_x^2 I & W_x W_y^T \\ W_y W_x^T & W_y W_y^T + B_y B_y^T + \sigma_y^2 I \end{pmatrix},$$

which is structurally identical to that in (1). We can think of $B_x B_x^T + \sigma_x^2 I$ as an approximation to Ψ_x, and if the dimensionality of z_x is high enough (equals the dimensionality of the data x) we can directly factorize Ψ_x as $B_x B_x^T$, leading to $\sigma_x^2 = 0$. From this it follows that given complex enough latent variables the second model equals the first, and thus also finds the solution of classical CCA.

While the original CCA can be solved as an eigenvalue problem, the extended model that allows varying marginal model complexities needs to be optimized with an iterative method. An expectation maximization (EM) algorithm for optimization is given in Figure 2.

3.1. Properties of the model

The above formulation for CCA has a few interesting consequences, which will be discussed here. First, being a probabilistic derivation it allows the use of standard techniques of generative modeling, such as model complexity selection and utilizing prior information, to be used in canonical correlation based analysis. For this the original derivation by [4] is already sufficient.

1. Marginalize over z_x and z_y to get model with covariance matrix of the form (1) where we have $\Psi_x = B_x B_x^T + \sigma_x^2 I$ and $\Psi_y = B_y B_y^T + \sigma_y^2 I$. Update the parameters $W = [W_x, W_y]$ using the EM step
$$W = \Sigma A^T \left(M + A\Sigma A^T\right)^{-1}.$$
Here $M = \left(I + W^T \Psi^{-1} W\right)^{-1}$, $A = MW^T\Psi^{-1}$, and Ψ is a block-diagonal matrix that consists of Ψ_x and Ψ_y. Σ is the joint sample covariance matrix.

2. Marginalize over z, and optimize the parameters related to x. The update rule for B_x is identical to the above one, but $\Psi = W_x W_x^T + \sigma_x^2 I$, $M = \left(I + B_x^T \Psi^{-1} B_x\right)^{-1}$, $A = MB_x^T\Psi^{-1}$, and Σ is the sample covariance of x. For σ_x^2 we get
$$\sigma_x^2 = \frac{1}{d_x}\text{trace}\left(\Sigma - \Sigma A^T B_x^T - W_x W_x^T\right),$$
where d_x is the dimensionality of x, and B_x is the new value just updated. Do exactly the same for parameters related to y.

Fig. 2. EM algorithm for optimizing the extended probabilistic CCA repeats the two steps until convergence. The second step can be repeated a few times in a row to improve the convergence of the marginal models, avoiding unnecessary use of parameters W to model the marginals.

The model also allows generalizing CCA-based analysis from normal distribution to other (exponential family) distributions, already suggested in [4]. For this purpose, however, our derivation from a more complete model structure is an important step. It illustrates how the canonical correlation solution is found only when the marginal models (that were implicit in the original derivation) are capable of modeling any possible variation within each data set alone. A similar observation was made in the context of discrete variables by [7], who showed that maximizing the likelihood of a co-clustering is equivalent to maximizing the mutual information between the clusters if we assume that the marginal densities given clusters are known exactly.

If marginal latent variables of lower dimensionality are used, we lose some of the capacity required for modeling the within-data variation, and the optimal solution uses W for modeling individual data sets as well. While the model may still be a reasonably good generative model, it does not capture the dependencies correctly. This will be demonstrated empirically in Section 5.

Another interesting observation is that modeling the marginal distributions is related to the whitening operation used for preprocessing data. It has previously been shown that classical CCA can be though of as whitening both data sets separately, followed by principal component analysis (PCA) of the concatenated whitened variables (see e.g. [8]). Here the role of the whitening step is played by the marginal models, which suggests a probabilistic interpretation of whitening in this context, as well as a generalization of similar preprocessing step to model families other than those consisting of linear projections.

4. DEPENDENCY-SEEKING CLUSTERING

An interesting generalization of the above formulation is to use clustering models, as several real-world data sets have inherent cluster structure instead of linear relations. Unfortunately, the requirement of being able to model all possible variation using the marginal models is a lot more difficult to satisfy when the data no longer comes from a single exponential family distribution, such as a Gaussian. Still, we can derive practical dependency-seeking clustering algorithms by making simplifying assumptions.

4.1. Simple model

In the simplest case we assume that only the common effects have cluster structure, but the variation within each cluster is still linear. The actual model with assumption of normality is then given by

$$\begin{aligned} z \sim \text{Mult}(\theta), \quad z_x, z_y \sim N(0, I), \\ x|z, z_x \sim N(\mu_x^z + B_x z_x, \sigma_x^2 I), \\ y|z, z_y \sim N(\mu_y^z + B_y z_y, \sigma_y^2 I), \end{aligned}$$

where μ_x^z denotes the mean vector for the x-space corresponding to the cluster z. In principle we could have an uninformative model for z as in the CCA case, but allowing different weights makes more sense in most clustering tasks.

We can again marginalize over z_x and z_y, and end up with a model where $x \sim N(\mu_x^z, B_x B_x^T + \sigma_x^2 I)$. Using z_x of full dimensionality gives equivalent parameterization in the form $x \sim N(\mu_x^z, \Psi_x)$, and we can directly write the final clustering model as

$$\begin{aligned} z \sim \text{Mult}(\theta), \\ (x, y)|z \sim N(\mu^z, \Psi), \end{aligned} \tag{2}$$

where

$$\Psi = \begin{pmatrix} \Psi_x & 0 \\ 0 & \Psi_y \end{pmatrix}.$$

In summary, the model is a normal mixture model for data where the two feature vectors have been concatenated, with the restriction that the covariance of the clusters is

shared and has a block-diagonal structure. The intuitive approach to clustering such data would be to use the full covariance matrix. It would in this case lead to individual clusters modeling also some of the dependencies, and even though it might be better in terms of the likelihood it would still be worse for making inference on the dependencies. In the other extreme where the covariance matrix would be restricted to be completely diagonal, the model would use cluster structure to model also within-data variation, again losing some of the dependencies.

Note that this suggests that the covariance matrix should be restricted also in cases where the variables in both data sets are expected to have correlation (for example, due to being measurements of the same actual property conducted with different measurement techniques). This is because we specifically want to capture the real link between the two data sets into the cluster structure, instead of the within-cluster covariance. This is in contradiction to the traditional approach, where all prior information naturally should be included in the model structure as well as possible.

4.2. More structured variant

The clustering model (2) requires quite strict assumptions for the data. Even though it will still lead to improved performance in many dependency exploration tasks, it is worthwhile to study if we can do something better.

In principle we would like to build a model where z acts as a cluster index, and the marginal models are complex enough to represent any meaningful structure in each data set within one cluster. Without restricting to any special cases, the best we could do would be to allow some general mixture model for the marginals as well, but marginalizing over the latent variables of such a model would be practically impossible. We thus take an alternative approach, where we accept that the marginal models are going to be insufficient, and try to focus on dependencies by specifying a more complex model for the joint effects. With suitable structure we can avoid at least part of the cases where the clusters are used to describe marginal variation.

We replace the middle part of the model with a hierarchy that also contains an independence assumption between two new latent variables w_x and w_y (Figure 1 (c)). The actual model is given by

$$\begin{aligned} z &\sim \text{Mult}(\theta_z), \\ w_x|z \sim \text{Mult}(\theta^z_{w_x}), &\quad w_y|z \sim \text{Mult}(\theta^z_{w_y}), \\ x|w_x \sim N(\mu^{w_x}_x, \Psi_x), &\quad y|w_y \sim N(\mu^{w_y}_y, \Psi_y), \end{aligned} \tag{3}$$

where we have already marginalized over z_x and z_y. The model can be optimized using an EM algorithm (details omitted), which reminds closely the EM algorithm for the classical Gaussian mixture.

The improvement compared to the model (2) is that the hierarchy can be used to detect and model correctly cases where either of the marginal distributions within a single cluster is multimodal. The algorithm solves this kind of situation by choosing the θ^z_w for that cluster to have two or more active components.

The model is formulated so that first a higher-level cluster z is selected, and based on that we independently select lower-level clusters w_x and w_y. Alternatively we can think of this as a process where a pair of lower-level clusters is chosen from a joint distribution of w_x and w_y. In this interpretation the prior $p(w_x, w_y)$ is not independent, but not completely free either. The model corresponds to representing the prior as a sum of independent priors.

The latter interpretation links the model to the associative clustering (AC) [3]. In AC the task is to find marginal clusterings for two data sets so that the contingency table formed by the sample counts is as dependent as possible. Here the samples could be assigned to the (w_x, w_y) clusters in a probabilistic way, leading to a table of "counts" with similar interpretation. From this perspective the proposed clustering method is a probabilistic alternative to AC that directly maximizes a measure of dependency. Comparing the two alternatives more thoroughly is, however, beyond the scope of this paper.

5. EXPERIMENTS

Here we verify empirically some of the properties claimed in the previous sections. These experiments are not a comprehensive study on the performance of the methods, but aim to demonstrate the kind of effects one should anticipate, and know how to deal with, in dependency exploration tasks performed by probabilistic modeling.

5.1. CCA and marginal model complexity

In Section 3 we claimed that the probabilistic formulation of CCA only holds when sufficiently complex marginal models are used. Here we demonstrate that the EM algorithm for extended CCA indeed converges to the classical CCA solution given full complexity, and show that this does not hold with lower complexity.

For this purpose we use a simple generated data set that has a subspace (three dimensions) with significant correlation and the rest of the dimensions (three) in both data spaces are more or less independent noise. 1000 samples are drawn from the distribution, and the solutions are computed using the EM algorithm (Figure 2) for various marginal model complexities. The results are computed as averages over 100 different data sets from the same distribution.

We compare the variants by measuring squared correlation (sum of squared canonical correlations when there is

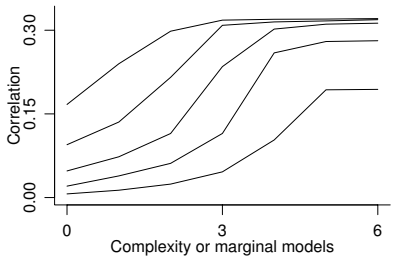


Fig. 3. The correlation of the posterior expectations increases as a function of the marginal model complexity, and moves from the PCA solution (no marginal models) to CCA solution (full complexity). The lines present different subspace dimensionalities, increasing from 1 (bottom line) to 5 (topmost line).

more than one dimension) between the posterior expectations $E[z_x|x]$ and $E[z_y|y]$. CCA is known to find the maximal value, whereas PCA for the concatenated variables has no particular reason to find correlating projections. In Figure 3 the correlation within the subspace is illustrated for varying marginal model complexities and dimensionalities of the subspace, and it is evident that too low a complexity leads to decreased performance in finding the canonical directions. What is sufficient depends on the number of dimensions sought, and ultimately on the data in question.

5.2. Dependency-seeking clustering

Evaluating the performance of a clustering algorithm is often quite difficult task, as no definitive measures for assessing the quality exist. Here even the validation data likelihood cannot be used, as it is not necessarily optimal criterion for dependency exploration task. A more direct measure of dependency is required, and as the clustering algorithms are here extended from canonical correlation analysis, we decided to study how closely they can mimic the solution of CCA. The algorithms are compared to the alternative of using an unrestricted joint model, here a normal Gaussian mixture with full covariance matrix, to see whether they can find the dependencies better than the naive approach of modeling all variation.

We use the same data set used in the previous demonstration. Even though the data comes from a single Gaussian, the independence assumption means that clusters are required for modeling the dependent parts. As a practical measure for the similarity of the results, we measure the variance of the cluster centroids in the CCA subspace: the solutions that focus on modeling directions perpendicular to the subspace will have small variation within the subspace.

We averaged the variances of a normal mixture model with full covariance matrix and the model (2) over 100 runs. The resulting average variance for normal mixture model was 0.09 (standard deviation 0.04), and 0.26 (deviation 0.08) for the block-diagonal variant, showing that the latter reveals clearly more (98 out of 100 runs) structure in the canonical subspace. The values are for 6 clusters, but similar results are obtained with other complexities as well.

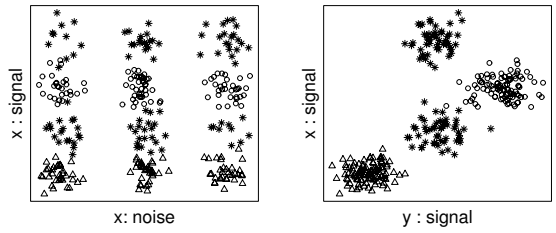
As a more real example, we cluster yeast genes based on two expression measurements of different stressful treatments (time-series of heat shock and diamide treatment). The measurement data was obtained from [9], and preprocessed like in [8]. The common thing between the measurements should be general stress, and thus we compare the clustering results to a list of environmental stress response (ESR) genes by [9] (two-class problem, each gene either is or is not an ESR). The average "classification accuracy" (percentage of training samples from the same class in a cluster; random assignment gives 75.7%) over 10 random data splits (half of the samples for training and half for validation) was 81.1% for normal mixture model, and 86.0% for model (2). The latter is significantly better (Wilcoxon signed rank test, $p < 0.002$). Again the number of clusters was arbitrarily fixed to 6. It is worth noticing that the normal mixture model had significantly higher likelihood on both training and validation data due to modeling more within-data variation, but it still tells less about yeast stress.

5.3. Multimodal marginal clusters

The advantage of the structured variant (3) over the simpler clustering model is that it can cope with multimodalities in marginal clusters. The ability is demonstrated on toy data with two two-dimensional data sets. First dimensions of both data sets are dependent, whereas the other dimensions are independent but structured noise. A three-cluster solution by the model (3) is illustrated in Figure 4. Not only has the algorithm ignored the dimensions containing structured noise, but it has also assigned the two modes of one cluster in the dependent plane to the same higher-level cluster.

As another example, we also clustered the Multiple Features Database[1] that contains several feature sets for handwritten number recognition. We picked the Karhunen-Loeve coefficients and the Zernike moments, and reduced both feature sets to just two dimensions using PCA to make the problem more challenging. Furthermore, we added three dimensions with bimodal Gaussian distribution to both data sets to increase within-data variation. We used the same accuracy measure as above, this time averaging over 20 runs. With 10-cluster solutions we got 15.4% for normal mixture model, 17.5% for (2), and 27.4% for (3) (using 15 values for w_x and w_y). All differences are significant according to the Wilcoxon signed rank test (all p-values below 0.004), and low values are due to the high noise ratio.

[1] from http://www.ics.uci.edu/∼mlearn/MLRepository.html



(a) **(b)**

Fig. 4. Analysis of two data sets, x and y, where we are only interested in what they have in common. Both figures are scatter-plots between two dimensions, and the samples are marked according to their cluster index (three clusters). **(a)** The model learns to ignore a noise dimension, even though it has clear structure. **(b)** Multimodality in signal dimensions does not affect clusters if it appears in only one of the data sets.

6. DISCUSSION

In this paper we studied the use of generative models in finding dependencies between two data sets. Traditionally, dependencies have been sought by explicitly optimizing a criterion for dependency, using methods such as canonical correlation analysis (CCA) or various clustering methods optimizing the mutual information. Recently CCA was interpreted as a generative model, which lead us to study whether generative models could be used for dependency exploration tasks in other cases as well.

Based on a re-interpretation of the probabilistic CCA we were able to show that a necessary condition for a generative model to reveal dependencies is that the model contains flexible enough parts for both of the marginals. We then extended the principle to clustering, and derived two clustering models for seeking dependencies. Both were demonstrated to find dependencies better than an unrestricted joint model, and in particular the simpler model performed surprisingly well in a practical application of combining two gene expression data sets of yeast stress.

The exact relationship between explicit dependency optimization and generative dependency-seeking models remains to be studied. The latter allows rigorous treatment of finite data, but the first can be used also in cases where building a sufficiently good generative model would be impossible. It is also worth studying whether the two alternatives could be combined.

7. ACKNOWLEDGMENTS

This work was supported in part by the Academy of Finland, decision numbers 79017 and 207467, and in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. All right are reserved because of other commitments.

8. REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of The 37th Annual Allerton Conference on Communication, Control, and Computing*, Bruce Hajek and R. S. Sreenivas, Eds., pp. 368–377. University of Illinois, Urbana, Illinois, 1999.

[3] S. Kaski, J. Nikkilä, J. Sinkkonen, L. Lahti, J. Knuuttila, and C. Roos, "Associative clustering for exploring dependencies between functional genomics data sets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, vol. 2, no. 3, pp. 203–216, 2005.

[4] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[5] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, pp. 443–482, 1999.

[6] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, pp. 305–345, 1999.

[7] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of KDD'03, The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 89–98. ACM Press, New York, NY, USA, 2003.

[8] J. Nikkilä, C. Roos, E. Savia, and S. Kaski, "Explorative modeling of yeast stress response and its regulation with gCCA and associative clustering," *International Journal of Neural Systems*, vol. 15, no. 4, pp. 237–246, 2005.

[9] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, pp. 4241–4257, 2000.