Arto Klami and Samuel Kaski. 2005. Non-parametric dependent components. In: Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005). Philadelphia, PA, USA. 18-23 March 2005. Piscataway, NJ, IEEE, pages V-209 - V-212.

NON-PARAMETRIC DEPENDENT COMPONENTS

Arto Klami, Samuel Kaski

University of Helsinki
Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland

ABSTRACT

Canonical correlation analysis (CCA) is equivalent to finding mutual information-maximizing projections for normally distributed data. We remove the restriction of normality by non-parametric estimation, and formulate the problem of finding dependent components with a connection to Bayes factors. The method is applied for characterizing yeast stress by finding what is in common in several different stress conditions.

1. INTRODUCTION

Given two data sets, CCA [1] finds a component from both of them, such that the components correlate maximally. The second pair of components is sought in the same way, with the additional constraint of being uncorrelated with the first, and so on. CCA can be formulated as a generalized linear eigenvalue problem, which generalizes nicely to multiple data sets. Several other generalizations exist (some listed in Section 3).

It can be shown [2] that the CCA-components maximize mutual information between the data sets, assuming multinormal data (pursued further for normal data in [3]). The idea of maximizing mutual information between representations of data sets has earlier been applied to nonlinear projections [4] and clustering (where the projection is on multinomial cluster indexes) [5, 6, 7] as well. The key idea in these works is that it is not necessary to define a (generative) model for the dependency between the data sets; a semi-parametric representation will learn the necessary dependencies.

Mutual information is a natural measure of statistical dependency for generalizing CCA to non-Gaussian distributions. In fact, a closely related method, discriminant analysis, has been generalized by maximizing variants of or approximations to mutual information, computed of non-parametric estimates of the data distribution (after the projection) [8].

We have shown earlier for discriminant analysis [9] that a simpler approach of finding the projection that maximizes likelihood of predictions outperforms the earlier approximations, while the cost function asymptotically converges to mutual information. The approach can be considered as a rigorous finite-data method for non-parametric estimation of discriminant components.

The same formulation does not work for CCA since the cost function is symmetric, and cannot be expressed as a likelihood. For clustering we have, however, been able to formulate a rigorous finite-data cost function that asymptotically converges to mutual information [10, 11]. It is a Bayes factor comparing two models; in the one the data sets are assumed dependent and in the other not.

In this paper we take a step towards a rigorous generalization of CCA to non-Gaussian data. We present a computationally feasible method for finding dependency-maximizing projections, and discuss how to take the finite-data uncertainty into account.

2. METHOD

2.1. Measuring dependency

Denote by $D = \{D_1, ..., D_N\}$ a collection of N data sets, with M vectorial samples each. The data come in tuples of co-occurring samples, one from each set. Denote by $f = \{f_1, ..., f_N\}$ a set of deterministic mappings from D to respective latent variables $S = \{S_1, ..., S_N\}$. The mappings are parameterized by $\theta = \{\theta_1, ..., \theta_N\}$. Consider two hypotheses, H_d and H_i, where H_d assumes that there is some kind of dependency between the data sets which can be captured by the latent variables. H_i assumes there is no such dependency.

The likelihood ratio

$$\frac{P(f(D;\theta)|H_d)}{P(f(D;\theta)|H_i)} \tag{1}$$

measures the dependency captured by the latent variables of a model with parameters θ. Maximizing (1) is asymptotically equivalent to maximizing the mutual information of the latent variables. This can be shown easily by taking the log of (1) to the limit of an infinite amount of data.

Maximizing (1) gives the model that captures dependencies optimally, but the parameters may overfit a given finite data set. We will next discuss a theoretical connection which results in new ways of handling the uncertainty.

The likelihood ratio (1) is related to a Bayesian measure of dependency where the parameters have been integrated out, namely the Bayes factor

$$BF = \frac{P(D|H_d)}{P(D|H_i)} \ . \tag{2}$$

Theorem 1 *The Bayes factor* (2) *can equivalently be written as*

$$BF = \int \frac{P(f(D;\theta)|H_d)}{P(f(D;\theta)|H_i)} P(\theta|D, H_i) d\theta \ , \tag{3}$$

under the following three assumptions:

1. *The prior probabilities for the mappings do not depend on the hypothesis, that is,* $P(\theta|H_d) = P(\theta_1|H_i)...P(\theta_N|H_i)$.
2. *The probabilities of observed data, given the latent variables, do not depend on the hypothesis.*
3. *The mappings* f *are deterministic.*

The proof is given in Appendix A.

The second assumption is the crucial modeling assumption. If the models are not allowed to model dependencies between the observed data sets given the latent variables, then the possible dependencies must be captured already in the latent space.

According to Theorem 1, the Bayes factor (3) is an integral over different mappings. The first part of the integrand is exactly (1), and it would be tempting to search for dependencies by maximizing the whole integrand. However, the result would be identical to MAP estimation of the joint model $P(D|H_d)$, and our goal is not joint density modeling; while it captures dependencies it models data set-specific variation as well.

Nevertheless, the formula (3) suggests that we could form a continuum between the two extremes of joint density modeling and maximization of the (empirical) mutual information by using a tunable "regularization" term in place of $P(\theta|D, H_i)$. In this paper we take the limit of uninformative (flat) term, and take care of the regularization by a simpler practical method: Leave-one-out during density estimation (details below).

2.2. Estimation of the dependency-maximizing projections

In this work the model family for both hypotheses is the set of all linear projections of a specific dimensionality. It is parameterized by a separate projection matrix for each data space. The vectors in θ_n are restricted to unit length for convenience, and in (1) we have $f_n(D_n; \theta_n) = \theta_n^T D_n$ for all $1 \leq n \leq N$.

Computing (1) is impossible without some further assumptions, because $P(\theta^T D|H)$ is unknown for both hypotheses. We estimate it for both hypotheses with non-parametric Parzen density estimators in the latent space. The approach is feasible, because the projections are assumed to be low-dimensional, and thus the estimates do not suffer from the curse of dimensionality. A main advantage is that the method is non-parametric and makes no assumptions about the distribution of the latent variables.

As a further simplification we seek for only one-dimensional projection for each data space. If more dimensions are needed, one can remove the already projected dimension by transforming the data by $\mathbf{I} - \theta\theta^T$, and then repeat the process. This is an approximation to the solution that would be found by directly estimating a projection to some chosen dimensionality.

The densities are estimated on a leave-one-out basis. If needed, we can speed up the computation by picking a subset of K samples as kernels, and use only them for density estimation. The number of kernels needed for sufficient accuracy depends mainly on the dimensionality of the latent space (here the number of data sets).

We use isotropic Gaussian kernels for estimating the densities. For one-dimensional projections (used for computing $P(\theta_n^T D_n|H_i)$ for each n) the kernel has a single parameter, the width of the Gaussian. In theory we could optimize the widths during learning, but it could possibly lead to serious overfitting. Thus the widths are here estimated individually for each data set by a simple heuristic: select the kernel width that gives maximal likelihood on the first principal component of the data.

The joint probability $P(\theta^T D|H_d)$ is estimated using Gaussian kernels with diagonal covariance matrix Σ. The diagonal elements are given by

$$\Sigma_{n,n} = \sigma_n K^{\frac{1}{5} - \frac{1}{N+4}} ,$$

where σ_n is the width of the kernel for the nth data set. Thus the width for the joint kernel is slightly larger than the width of one-dimensional kernels. The scaling is based on the rule [12] of using width proportional to $K^{-\frac{1}{N+4}}$ for K-kernel estimate in N-dimensional space when the underlying distribution is normal.

The optimal set of projections is here sought by maximizing the logarithm of (1) with a conjugate gradient algorithm. The projections can be initialized randomly, or by using the first principal component. The final cost function is

$$\sum_{i=1}^{M} \left[\log \sum_{j \neq i} G_j(i|\theta, \Sigma) - \sum_{n=1}^{N} \log \sum_{j \neq i} g_j(i|\theta_n, \sigma_n) \right] , \quad (4)$$

where $g_j(i|\theta_n, \sigma)$ denotes the value of jth kernel at ith sample, computed in the projection space of nth data set. The $G_j(i|\theta, \Sigma)$ is a similar term, but the samples are in the joint space of all projections. If a regularizing distribution were used in (3), it would appear here as an additive term.

3. RELATED CCA EXTENSIONS

CCA can be extended to more than two data sets in several ways [13]. We compare our method with the following extension: search for the minimal generalized eigenvalues of the problem $A\xi = \lambda B\xi$, where A is the covariance matrix of D and B is the block-diagonal matrix of covariances of each data set D_n. Generalized eigenvectors ξ are the projection vectors, and the method is here called generalized CCA (gCCA).

Of the several non-linear extensions we consider KernelCCA [14], which leads to a similar eigenvalue problem, but the data covariances are replaced with Gram matrices of paired data sets. The method can be interpreted as a non-linear mapping to a high-dimensional feature space, where the correlations are maximized. In practice the feature mappings are not explicitly computed, because the kernel trick allows directly working with the Gram matrices. The kernelization has one drawback: KernelCCA does not have an explicit representation for the projection matrices. Such representations are useful when interpreting the results.

A generalization similar to ours was introduced very recently [15]. The task was to test the hypothesis of two data sources being dependent, and the proposed solution was to search for projections that have maximal mutual information (used as a lower bound for the test statistic). The result can be interpreted as a generalization of CCA that is similar to ours in the sense of making no distributional assumptions. The technical difference is in the approximations made for finite data: We regularized likelihood ratios, whereas in [15] simple approximations to (continuous) mutual information were made.

4. EXPERIMENTS

4.1. Toy demonstration

We wish to demonstrate that the proposed method finds the same solution as gCCA when the data is jointly normal and hence dependency means correlation. In addition, we show that the proposed method works even if the assumption of normality does not hold, while gCCA may fail.

For these purposes we created three two-dimensional data sets that have simple linear dependency, that is, there are projections to one-dimensional spaces that capture the dependency. Each data set is generated as a $M \times 2$ matrix. The first column of each data is sorted to increasing order, which gives a perfect dependency.

Dependent component	gCCA	Proposed method
Normal	1.5	0.9
Exponential	6.0	1.1
Gamma	3.1	1.2

Table 1. Both gCCA and the proposed method find the correct projections if all dependent components are normal (first line, the figures are average angles in degrees between the true and the found projection). gCCA errors increase noticeably when one of the dependent components is not normal (second and third line).

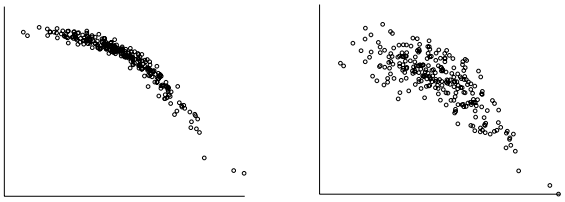

Fig. 1. The proposed method (left) finds a strong non-linear dependency between a normal (horizontal axis) and an exponential (vertical axis) variable. gCCA (right) has here found roughly similar, but clearly noisier, dependency.

After that the data sets are rotated to random orientation, and some additive Gaussian noise (standard deviation 0.1) is added.

For the first task all data sets are sampled from the normal distribution having unit variance. For the second task the dependent component of one data set is sampled from another distribution, but everything else is still normal. We used exponential and gamma distributions.

We generated 50 replications of both types of data sets with $M = 500$ samples. The estimation error is measured as the average angle (over all data sets and all replications) between the estimated and true projections. The results are represented in Table 1, and we see that both gCCA and the proposed method always find a solution close to the real one if all distributions are normal.

If the dependent component of one data set is non-normal, gCCA often fails to find the true projections. This is seen as a larger average angle, even though gCCA still occasionally finds the correct solution. Note that the failures are not because of optimization; gCCA always finds the global optimum.

The difference is illustrated in Figure 1, which shows a scatterplot of two one-dimensional projections for both methods. Here the dependent component in the first data set is normally distributed, and in the second data set it is exponential. The dependency found is not linear, but still clear.

4.2. Yeast stress data

We applied the method on gene expression data from [16]. The data consists of expressions of 5998 yeast genes in different stressful conditions. The expressions are measured a few times in each condition, giving a short time series for each. We chose five conditions (two heat shocks, DTT exposure, diamide treatment, and menadione exposure), that is, five data sets. The main common thing in the conditions should be stress, and hence by searching for the dependencies we hope to be able to characterize stress.

The proposed method is compared with gCCA and KernelCCA. The samples were divided into five parts, and two-fold testing was performed in each part. This gives 10 independent training and test sets, each having 600 genes.

For KernelCCA we used Gaussian kernels with unit variance, and regularization value 0.001 for all pairwise kernels. These parameters were chosen based on preliminary tests, and they are close to those used in [14] in KernelICA (an ICA algorithm using KernelCCA as a contrast function). In more systematical tests these values could naturally be chosen using a validation set.

To demonstrate the speedup, we also performed ten-fold cross validation, where each learning set had 5400 genes and randomly picked $K = 270$ kernels were used. KernelCCA was left out from this test because of computational reasons; some non-trivial tricks would have been needed to solve the eigenproblem with matrices of size 27.000×27.000.

Measuring the performance of the methods is non-trivial, because the true projections are not known. Multi-information of the projections could be used as a quality measure in principle, but it is difficult to estimate it if the dimensionality of the joint projection space is high. Binned estimates do not work because the number of bins would clearly exceed the number of samples, and simple plug-in estimates like (4) were not considered because of possible bias in favor of the proposed method.

While we cannot come up with a general performance measure, we have a way to validate the results externally. Gasch [16] has classified some of the genes as environmental stress response (ESR) genes, that is, genes that react to various stress conditions. We can then measure how well the ESR genes are separated from the rest after projection. For that we used a leave-one-out k-nearest neighbor classifier (with k arbitrarily fixed to 9) in the joint projection space, and report the average errors in Table 2.

Besides the classification error, we measured the strength of the found dependencies by computing the empirical mutual information between all pairs of data sets, averaged over the ten folds, and summarized by their average.

The classification accuracy of the proposed method is roughly equal to that obtained in the original data, without projections, and hence the dependent components capture most of the interesting dependencies. The method outperforms both gCCA and KernelCCA significantly (paired t-test, $p < 0.01$).

KernelCCA has roughly equal performance to gCCA; the classification error is a bit worse, but the pairwise mutual information is better on the average. A reason for the relatively poor performance of KernelCCA could be that there seems to be a strong linear dependency between two of the studied data sets (the heat shocks), and Gaussian kernels might not be able to find linear dependencies optimally. We should also remember that the list of ESR genes is not final but only the current draft.

5. DISCUSSION

We have presented a method that generalizes canonical correlation analysis to non-normal data. The algorithm is based on maximizing dependency, and distributional assumptions are circumvented by non-parametric density estimation in the projection spaces.

Using simple generated data, we showed that the proposed method finds the canonical components if the data is jointly normal. If it is not, the method still finds the true dependent components, while CCA may fail. We also applied the method on a gene expression data set, and again it outperformed both generalized CCA and KernelCCA, a non-linear variant of CCA.

Small training set

Method	Classification error	Pairwise MI
Original data	6.6%	-
Proposed method	6.7%	0.41
Generalized CCA	8.3%	0.21
KernelCCA	8.7%	0.27

Large training set

Method	Classification error	Pairwise MI
Original data	6.6%	-
Proposed method	6.9%	0.40
Generalized CCA	8.3%	0.21

Table 2. The proposed method finds dependent components where classification of the ESR genes is as easy as in the original data space, and outperforms other CCA-variants, measured either with the classification error or the pairwise mutual information. Performances of gCCA and KernelCCA are roughly equal.

The current version of the method uses several computational approximations and assumptions. The main ones are: (i) Currently one component is sought at a time. Several possible ways to extend the method to overcome that limitation exist, e.g. parameterizing the whole projection matrix using Givens rotations. (ii) The method currently effectively optimizes the likelihood ratio, and needs further development. A compromise between dependency and joint modeling was suggested.

6. 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. The authors would like to thank Janne Nikkilä for insights on the data and Bayes factors, and acknowledge that access rights to the materials produced in this project are restricted due to other commitments. Arto Klami is additionally supported by graduate school HeCSE.

A. PROOF OF THEOREM 1

Multiplying (2) by $\iint P(\theta, S|D, H_d)d\theta dS = 1$ gives

$$BF = \frac{P(D|H_d)}{P(D|H_i)} \iint P(\theta, S|D, H_d)d\theta dS$$
$$= \iint \frac{P(D|H_d)P(\theta, S|D, H_d)}{P(D|H_i)P(\theta, S|D, H_i)} P(\theta, S|D, H_i)d\theta dS \ . \quad (5)$$

Notice that $P(D|H)P(\theta, S|D, H) = P(D, \theta, S|H)$ for both hypotheses.

We can also write (for both hypotheses) $P(D, \theta, S|H)$ as a product $P(D|\theta, S, H)P(S|\theta, H)P(\theta|H)$. The terms $P(\theta|H)$ and $P(D|\theta, S, H)$ cancel in (5) because of the first and second assumptions, respectively. The last term in (5) can equivalently be written as $P(S|\theta, D, H_i)P(\theta|D, H_i)$, and we end up with

$$BF = \iint \frac{P(S|\theta, H_d)}{P(S|\theta, H_i)} P(S|\theta, D, H_i)P(\theta|D, H_i)d\theta dS \ .$$

As a final step we use the third assumption, which states that $P(S|\theta, D, H_i)$ is a delta-distribution so that $P(S = f(D)) = 1$. Thus $P(S|\theta, H) = P(f(D)|H)$ and the integration over S can be dropped. This finally gives (3).

7. REFERENCES

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

[2] J. Kay, "Feature discovery under contextual supervision using mutual information," in *Proc. of Int. Joint Conference on Neural Networks*, pp. 79–84. IEEE, Piscataway, NJ, 1992.

[3] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," in *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA, 2003.

[4] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, pp. 161–163, 1992.

[5] S. Becker, "Mutual information maximization: models of cortical self-organization," *Network: Computation in Neural Systems*, vol. 7, pp. 7–31, 1996.

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

[7] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. of the 23rd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 208–215. ACM Press, New York, NY, USA, 2000.

[8] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[9] S. Kaski and J. Peltonen, "Informative discriminant analysis," in *Proc. of the 20th Int. Conference on Machine Learning*, pp. 329–336. AAAI Press, Menlo Park, CA, 2003.

[10] J. Sinkkonen, S. Kaski, and J. Nikkilä, "Discriminative clustering: Optimal contingency tables by learning metrics," in *Proc. of the 13th European Conference on Machine Learning*, pp. 418–430. Springer, Berlin, 2002.

[11] J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski, "Associative clustering," in *Proc. of the 15th European Conference on Machine Learning*, pp. 647–654. Springer, Berlin, 2004.

[12] B. W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hall, London, 1986.

[13] J.R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[14] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[15] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[16] 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.