FAST DEPENDENT COMPONENTS FOR FMRI ANALYSIS

Eerika Savia, Arto Klami, Samuel Kaski

Helsinki University of Technology
Department of Information and Computer Science
P.O. Box 5400, FI-02015 TKK, Finland

ABSTRACT

Canonical correlation analysis (CCA) can be used to find correlating projections of two data sets with co-occurring samples. Instead of correlation, we would typically want to find more general dependencies, measured by mutual information. Variants of CCA based on non-parametric estimation of mutual information have been proposed previously; they outperform traditional CCA for non-Gaussian data but require infeasible amounts of computation for already quite modest sample sizes. We introduce a novel variant that uses a semiparametric estimate leading to a considerably faster algorithm, and apply the method on searching for statistical dependencies between multi-sensory stimuli and functional magnetic resonance imaging (fMRI) of brain activity.

Index Terms— Canonical correlation, component models, fMRI, mixture model, mutual information

1. INTRODUCTION

In fMRI analysis of brain signals related to natural stimulation, both the brain signals and the stimuli are very complex and difficult to analyze. Both contain variation that is not interesting to the analyst. The interesting aspects are the dependencies between the two data sets, and hence the analysis should be focused on those. The same setup of searching for statistical dependencies between data sets of co-occurring samples recurs in many applications.

A classical approach to searching for dependencies is to project the data sets onto lower-dimensional subspaces, in which it is easier to estimate dependencies than in the original high-dimensional spaces. When a projection is optimized to maximize dependency, it discards variation that is not present in the other data set, while keeping the shared variation.

The authors belong to Adaptive Informatics Research Centre, a CoE of the Academy of Finland, and Helsinki Institute for Information Technology HIIT. We thank Riitta Hari, Sanna Malinen and Yevhen Hlushchuk (Advanced Magnetic Imaging Centre and and Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology) for collaboration in the original fMRI study and providing us invaluable help in interpreting the findings from the neuroscientific point of view in the earlier study involving this data. This work was supported by the PASCAL2 NoE of the EC.

The task of finding maximally correlating projections between two data sets can be solved by a classical method called canonical correlation analysis (CCA) [1]. The method is fast and robust, but for many applications correlation is too simple a measure; it only measures linear dependency.

Replacing correlation with mutual information makes discovery of more general types of dependency possible. However, mutual information cannot be computed as easily as correlation, and we need to resort to approximations. One option is to empirically estimate the probability density in the projection space, and estimate mutual information based on the density estimate. In [2, 3], methods based on non-parametric Parzen-kernel estimates have been used. We call these methods dependent component analysis (DeCA).

The Parzen estimates are consistent and accurate, but computationally demanding for large data sets. Replacing the non-parametric density estimates with semiparametric estimates should give comparable results while being scalable. We introduce a novel algorithm that uses a mixture of Gaussians to estimate the density in the projection space. An analogous method has been shown to improve efficiency in a related task of discriminant analysis [4].

We applied the novel DeCA-variant to the task of finding dependencies between measured brain activity and multisensory stimuli. Following an earlier application of CCA to the same task [5], we used a two-step approach: First, spatially independent patterns of brain activity were extracted from fMRI data with independent component analysis (ICA) [6]. In the second step the new DeCA-variant was used to find dependencies between the brain patterns and the stimuli.

2. METHOD

The task in DeCA is to find linear projections of two data sets, $\mathbf{X}$ and $\mathbf{Y}$, so that the mutual information between the projections $\mathbf{s}_x = \mathbf{w}_x^T\mathbf{X}$ and $\mathbf{s}_y = \mathbf{w}_y^T\mathbf{Y}$ is maximized. The objective is thus to maximize

$$I(s_x, s_y) = \int\int p(s_x, s_y) \log \frac{p(s_x, s_y)}{p(s_x)\, p(s_y)} ds_x ds_y \qquad (1)$$

with respect to linear transformations $\mathbf{w}_x$ and $\mathbf{w}_y$. Here s_x and s_y are the random variables $s_x = \mathbf{w}_x^T\mathbf{x}$ and $s_y = \mathbf{w}_y^T\mathbf{y}$,

and $\mathbf{x}$ and $\mathbf{y}$ are random vectors corresponding to data sets $\mathbf{X}$ and $\mathbf{Y}$. Mutual information has, however, two severe difficulties as a cost function: It requires knowledge of the joint probability density $p(s_x, s_y)$, and it involves an integral over the whole projection space. As a practical solution, we used a mixture of Gaussians-based density estimate $\hat{p}(s_x, s_y)$, and estimated the integral as an average over the observations. That is, we maximized the objective function

$$\hat{I}(s_x, s_y) = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{p}(s_x^i, s_y^i)}{\hat{p}(s_x^i)\,\hat{p}(s_y^i)} \ , \qquad (2)$$

where N is the number of observations.

In earlier works, $\hat{p}(s_x, s_y)$ has typically been a Parzen estimate, which is non-parametric. Hence, optimizing the cost of Eq. (2) has been straightforward; derive the gradient of the cost and use any standard optimization method to find a local optimum. Here, we consider parametric estimates of the form

$$\hat{p}(s_x, s_y) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}([s_x; s_y] \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \ , \qquad (3)$$

where $\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at $\mathbf{x}$. The π_k represent the probabilities of the K mixture components. This estimate has a set of parameters $\boldsymbol{\Theta}$ that need to be learned, $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$. Hence, straightforward optimization with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$ is not possible.

We propose an alternating algorithm following the work in [4]. Starting with some initial projections, we learn a mixture of Gaussians in the projection space using the expectation maximization (EM) algorithm. After learning the density estimate, we optimize the projections $\mathbf{w}_x$ and $\mathbf{w}_y$. The algorithm then proceeds by alternating these two steps.

With a mixture of Gaussians density estimate, the objective function Eq. (2) can be easily differentiated, and we can use gradient-based methods to learn the projections. We used a conjugate gradient method, with the number of iterations equal to the dimensionality of the parameter space. The density estimate was always optimized until convergence of the EM algorithm.

We optimized the components one at a time because density estimation in high-dimensional spaces is very difficult. By using one-dimensional projections we can estimate the joint density $p(s_x, s_y)$ in a two-dimensional space, which can be done accurately enough already with a reasonably small data sets.

After finding the first component, we can proceed to search for the next maximally dependent component with the following constraint: The projections on the consecutive components should be independent of the projections on the previous components, in both the $\mathbf{X}$- and $\mathbf{Y}$-spaces. In practice, searching for a component that maximizes dependency with the other data set while minimizing dependency with the earlier component(s) is difficult. We used an approximation, where instead of full independency, we required the components to be uncorrelated with the earlier projections, analogously to how successive CCA components are defined. This can be satisfied with a simple deflation procedure, as follows. After extracting a component $\mathbf{w}_x$, we transformed the data by

$$\bar{\mathbf{X}} = \mathbf{X} \left(\mathbf{I} - \frac{1}{\mathbf{w}_x^T \mathbf{X} \mathbf{X}^T \mathbf{w}_x} \mathbf{X}^T \mathbf{w}_x \mathbf{w}_x^T \mathbf{X} \right) \qquad (4)$$

(and analogously for $\mathbf{Y}$), and searched for the next component by applying the algorithm to $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$. The procedure can be continued up to the minimum of the $\mathbf{X}$- and $\mathbf{Y}$-space dimensionalities, or until there are no significant dependencies left between $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$.

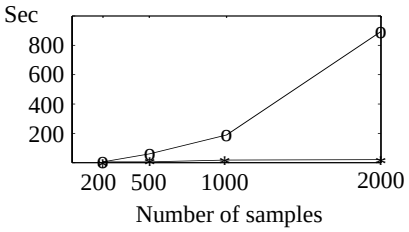


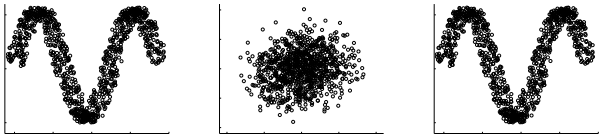
Fig. 1. Running times of non-parametric DeCA (o) and mixture of 20 Gaussians DeCA (*) in seconds as a function of the number of data samples.

3. TECHNICAL VALIDATION

The main advantage of the proposed method, compared to earlier DeCA methods, is in the computational speed. The Parzen estimate used in the earlier works has a computational complexity of $\mathcal{O}(N^2)$, and each iteration of a gradient-based optimization algorithm requires evaluating the densities again. A mixture model with K mixture densities has, however, only a complexity of $\mathcal{O}(NK)$ for evaluating the density or the gradient with respect to the projections. We show in Fig. 1 the computation times as a function of data size, N. We also demonstrate the functionality of the DeCA variants compared to CCA on an artificial data set. The data consisted of 10-dimensional data $\mathbf{X}$ and 7-dimensional data $\mathbf{Y}$ of 1000 samples, each dimension being uniform random noise except in one of the dimensions, where there was a clear dependency between $\mathbf{X}$ and $\mathbf{Y}$, as shown in panel (a) of Fig. 2. CCA could not find such non-linear dependency, whereas both DeCA variants worked correctly.

4. APPLICATION TO FMRI ANALYSIS

Natural stimuli are increasingly used in fMRI studies to imitate real-life situations. Consequently, it is no longer feasible to assume single features of the experimental design



(a) **(b)** **(c)**

Fig. 2. **(a)** Shows a dependent but uncorrelated subspace of the toy data. The CCA projection **(b)** completely missed the dependency, whereas DeCA found the true projections. **(c)** shows the result of the new mixture-based variant, but the non-parametric variant produced an indistinguishable solution. In both figures x-axis represents the projection of data set $\mathbf{X}$ and y-axis the projection of data set $\mathbf{Y}$.

alone to account for the measured brain activity. Instead, relevant combinations of stimulus features could explain the more complex brain activation patterns.

We have earlier proposed a two-step approach [5, 7], where ICA was first used to identify spatially independent brain patterns. As the second step, temporal dependencies between stimuli and the brain patterns were detected using CCA or some CCA-variant, like DeCA. The idea is to look for dependencies between combinations of stimulus features and the corresponding combinations of brain patterns.

This framework has been studied in two earlier papers, [5] and [7], one of which used non-parametric version of DeCA [2] and the other CCA for the dependency exploration step. We applied the new DeCA-variant to the data from study [5] where the fMRI measurements originally come from [8]. The preprocessed stimulus data and the result of ICA for the fMRI data (1932 data points) were taken from the earlier study without any modifications.

The original stimulation sequence consisted of 7 different stimuli: one tactile stimulus, 3 visual stimuli containing hands, faces or buildings, and 3 audio stimuli consisting of either tone pips, a voice reading about history, or a voice giving guitar fingering instructions. In [5], the original features of the stimuli were augmented with six features, extracted from the spectrogram of the actual auditory stimuli. The 13 stimulus time courses were quite far from being normally distributed, justifying the need for DeCA that does not make an implicit global normality assumption like CCA. Performing these experiments using non-parametric DeCA, on the other hand, would have been computationally infeasible.

5. EXPERIMENTS

We initialized the projection vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ to the first PCA-components of the data sets $\mathbf{X}$ and $\mathbf{Y}$, and initialized the mixture estimate in the projection space as follows. The initial values of means $\boldsymbol{\mu}_k$ and mixture probabilities π_k were determined by k-means separately in each projection space, by using the centroids as initial means of the mixture components and the proportions of cluster sizes as π_k. The initial covariance matrix $\boldsymbol{\Sigma}_k$ of each mixture component was set to a diagonal matrix diag $\left(\sigma_x^2, \sigma_y^2\right)$ containing the variances of the initial projections s_x and s_y. Finally, the number of mixture components was initially set to 5, but if a mixture component was left empty during an EM-iteration, it was removed. The full algorithm was run for 10 iterations.

A critical issue in analyzing small data sets is to avoid overfitting. Already classical CCA suffers severely from overfitting with small enough data (see e.g. [10] for a demonstration and solution for classical CCA using Bayesian learning), which shows as spurious canonical correlations. Here we assessed the quality of the results by bootstrapping. DeCA (like CCA) identifies the components only up to a sign, and to average over a set of components obtained from the bootstrap replicates requires choosing the signs to match each other. Here, we solved the unidentifiability issue of the components by matching the signs of the bootstrapped projections with an algorithm resembling a one-component k-means (details omitted due to lack of space).

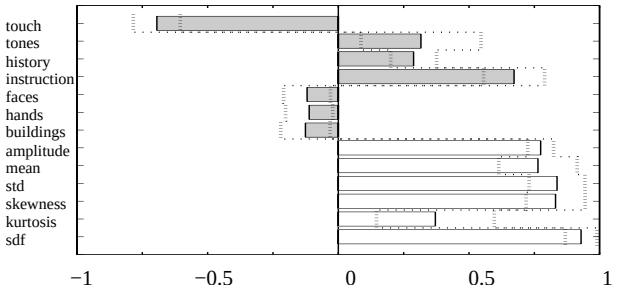
The statistical significance was tested with Wilcoxon signed rank test compared to zero factor loadings. We used the p-value threshold 0.01, corrected for multiple testing with Bonferroni correction.

6. RESULTS

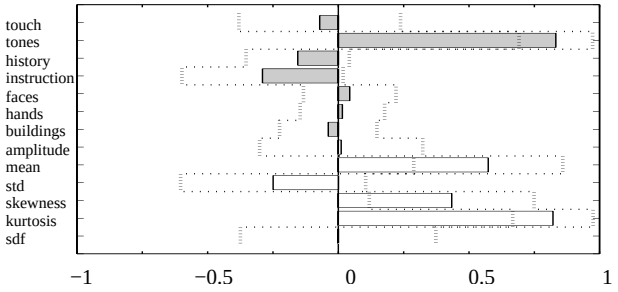
We were able to distinguish 8 individual significant DeCA components, which show clear discriminative patterns, and they can be roughly described as follows: Component 1 discriminates sound related features from other senses, especially from tactile stimulus. Component 2 discriminates tone pips and the related auditory features from speech. Component 3 discriminates between the two different speech features, *history* and *instruction*. Components 4–6 each represent one of the visual features *buildings*, *faces* and *hands*. Component 7 discriminates between two auditory features, *skewness* and *mean*. Component 8 represents the auditory feature *amplitude*. In Fig. 3 , we show the factor loadings[1] of components 1–3, *i*.e., the correlations between the DeCA component and each of the original feature dimensions.

The main difference between results of DeCA and CCA was that DeCA was able to find more components. Some of the found DeCA components had counterparts in the CCA results (DeCA components 1,2 and 4). But there were also new ones; *e*.g. DeCA components 3, 7 and 8 discriminate fine differences between auditory signals, which could not be consistently distinguished with CCA.

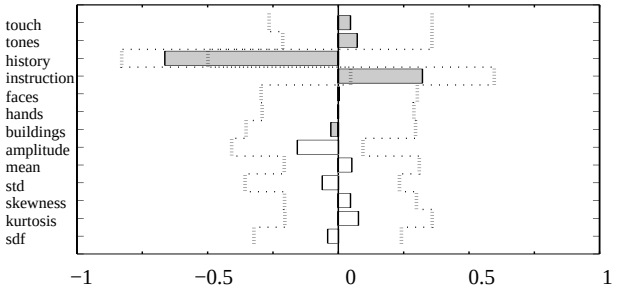
[1] An alternative would have been to study the projection vectors directly, but when there are within-dataset-correlations between the dimensions, the projection vectors might be misleading.



(a) DeCA component 1. All 13 features were significant.



(b) DeCA component 2. Significant features were tones, history, instruction, mean, std, skewness and kurtosis.



(c) DeCA component 3. Significant features were history, instruction and amplitude.

Fig. 3. Factor loadings of DeCA components 1–3 for the stimulus features. In each caption we list the features that both deviated from zero statistically significantly (P-value < 0.01) and have factor loadings above 0.1 The gray bars illustrate the experimental stimulus features and the white bars the extracted stimulus features. The dotted lines indicate the level of one standard deviation between the bootstrap results.

7. DISCUSSION

We introduced an algorithm for finding linear projections of two data sets, so that the projections have high mutual information. The proposed method improves earlier solutions by replacing non-parametric density estimators by mixtures of Gaussians, resulting in an optimization algorithm that is linear instead of quadratic in the number of data points.

We applied the method to a time-paired data set from an earlier study. One of the data sets consisted of 13-dimensional time-series description of the stimuli presented to test subjects, and the other data set of spatially independent brain patterns produced from fMRI measurements by so-called reliable ICA approach [9]. We found 8 different components that seem reasonable in terms of the stimulus features. Some of the found combinations of brain patterns were already familiar from the earlier study, and have there been interpreted to be meaningful [5]. In that study, CCA found only 5 components consistently from this data. DeCA was able to find more dependencies that were not strong enough in terms of the correlation, but still manifested themselves in mutual information.

We also showed with artificial data set that the results of the novel semiparametric version and non-parametric version of DeCA are comparable and, indeed, can find dependency structures when correlation is zero and CCA fails. The novel variant was also verified to be considerably faster.

8. REFERENCES

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

[2] A. Klami and S. Kaski, "Non-parametric dependent components," *ICASSP'05*, pp. V–209–V–212, 2005.

[3] X. Yin, "Canonical correlation analysis based on information theory," *Journal of Multivariate Analysis*, vol. 91, pp. 161–176, 2004.

[4] J. Peltonen, J. Goldberger, and S. Kaski, "Fast semi-supervised discriminative component analysis," *MLSP'07*, pp. 312–317, 2007.

[5] J. Ylipaavalniemi, E. Savia, S. Malinen, R. Hari, R. Vigário, and S. Kaski, "Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli," *NeuroImage*, under revision.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, New York, NY, 2001.

[7] J. Ylipaavalniemi, E. Savia, R. Vigário, and S. Kaski, "Functional elements and networks in fMRI," *ESANN'07*, pp. 561–566, 2007.

[8] S. Malinen, Y. Hlushchuk, and R. Hari, "Towards natural stimulation in fMRI – issues of data analysis," *NeuroImage*, vol. 35, no. 1, pp. 131–139, 2007.

[9] J. Ylipaavalniemi and R. Vigário, "Analyzing consistency of independent components: An fMRI illustration," *NeuroImage*, vol. 39, no. 1, pp. 169–180, 2008.

[10] A. Klami and S. Kaski, "Local dependent components," *ICML 2007*, pp. 425–432. Omnipress, 2007.