Extracting Coactivated Features from Multiple Data Sets

Michael U. Gutmann and Aapo Hyvärinen

Dept. of Computer Science and HIIT
Dept. of Mathematics and Statistics
P.O. Box 68, FIN-00014 University of Helsinki, Finland
{michael.gutmann,aapo.hyvarinen}@helsinki.fi

Abstract. We present a nonlinear generalization of Canonical Correlation Analysis (CCA) to find related structure in multiple data sets. The new method allows to analyze an arbitrary number of data sets, and the extracted features capture higher-order statistical dependencies. The features are independent components that are coupled across the data sets. The coupling takes the form of coactivation (dependencies of variances). We validate the new method on artificial data, and apply it to natural images and brain imaging data.

Keywords: Data fusion, coactivated features, generalization of CCA.

1 Introduction

This paper is about data fusion – the joint analysis of multiple data sets. We propose methods to identify for each data set features which are related to the identified features of the other data sets.

Canonical Correlation Analysis (CCA) is a classical method to find in two data sets features that are related. In CCA, "related" means correlated. CCA can be considered to consist of individual whitening of the data sets, followed by their rotation such that the corresponding coordinates are maximally correlated. CCA extracts features which capture both the correlation structure within and between the two data sets.

CCA has seen various extensions: More robust versions were formulated [2], sparsity priors on the features were imposed [1], it was combined with Independent Component Analysis (ICA) to postprocess the independent components of two data sets [7], and it was extended to find in two data sets related clusters [8].

Here, we propose a new method which generalizes CCA in three aspects:

1. Multiple data sets can be analyzed.
2. The features for each data set are maximally statistically independent.
3. The features across the data sets have statistically dependent variances; the features tend to be jointly activated.

In Section 2, we present our method to find coactivated features. In Section 3, we test its performance on artificial data. Applications to natural image and brain imaging data are given in Section 4. Section 5 concludes the paper.

T. Honkela et al. (Eds.): ICANN 2011, Part I, LNCS 6791, pp. 323–330, 2011.

2 Extraction of Coactivated Features

In Subsection 2.1, we present the general statistical model which underlies our data analysis method. In Subsection 2.2, we show that in some special case our method boils down to CCA. Subsection 2.3 focuses on the analysis of multiple data sets.

2.1 Modeling the Coupling between the Data Sets

As in CCA, we assume that each data set has been whitened. Denote by $\mathbf{z}^i$ the random vector whose i.i.d. observations form data set i. We assume that the total number of data sets is n. We use ICA to find, for each data set, features that are maximally statistically independent. That is, we model the $\mathbf{z}^i$ as

$$\mathbf{z}^i = \mathbf{Q}^i \mathbf{s}^i \quad (i = 1, \dots n), \tag{1}$$

where $\mathbf{z}^i \in \mathbb{R}^d$ and the $\mathbf{Q}^i$ are orthonormal matrices of size $d \times d$. Each vector $\mathbf{s}^i$ contains d independent random variables $s^i_k, k = 1, \dots d$ of variance one which follow possibly different distributions. The unknown features that we wish to identify are the columns of the $\mathbf{Q}^i$. We denote them by $\mathbf{q}^i_k$, $k = 1, \dots, d$.

We have assumed that the $s^i_k, k = 1, \dots, d$ are statistically independent in order to extract, for each data set i, meaningful features. In order to find features that are related across the data sets, we assume, in contrast, that across the index i, the s^i_k are statistically dependent. The joint density $p_{s^1_1, \dots, s^1_d, \dots, s^n_1, \dots, s^n_d}$ factorizes thus into d factors $p_{s^1_1, s^2_1, \dots, s^n_1}$ to $p_{s^1_d, s^2_d, \dots, s^n_d}$. To model coactivation, we assume that the dependent variables have a common variance component, that is

$$s^1_k = \sigma_k \tilde{s}^1_k \qquad s^2_k = \sigma_k \tilde{s}^2_k \qquad s^3_k = \sigma_k \tilde{s}^3_k \qquad \dots \qquad s^n_k = \sigma_k \tilde{s}^n_k, \tag{2}$$

where the random variable $\sigma_k > 0$ sets the variance, and the $\tilde{s}^i_k$ are Gaussian random variables. Treating the general case where the $\tilde{s}^i_k$ may be correlated becomes quickly complex. We are treating here two special cases: For correlated sources, we consider only the case of $n = 2$. This is done in the next subsection. For larger numbers of data sets, we are additionally assuming that the $\tilde{s}^i_k$ are independent random variables. This is the topic of Subsection 2.3.

2.2 Two Data Sets: A Generalization of Canonical Correlation Analysis

We consider here the case $n = 2$. Let $\mathbf{s}_k = (s^1_k,\ s^2_k)^T$ contain the k-th component of the vectors $\mathbf{s}^1$ and $\mathbf{s}^2$. If $(\sigma_k)^2$ follows the inverse Gamma distribution with parameter ν_k, the variance variable σ_k can analytically be integrated out.[1] The factors $p_{\mathbf{s}_k} = p_{s^1_k, s^2_k}$, $k = 1, \dots, d$, follow a student's t-distribution,

$$p_{\mathbf{s}_k}(\mathbf{s}_k; \nu_k; \mathbf{\Lambda}_k) = \frac{\Gamma\left(\frac{\nu_k+2}{2}\right)}{(\pi(\nu_k - 2))\Gamma\left(\frac{\nu_k}{2}\right)} |\mathbf{\Lambda}_k|^{\frac{1}{2}} \left(1 + \frac{1}{(\nu_k - 2)} \mathbf{s}_k^T \mathbf{\Lambda}_k \mathbf{s}_k\right)^{-\frac{\nu_k+2}{2}}. \tag{3}$$

[1] Proofs are omitted due to a lack of space. Supplementary material is available from the first author.

Here, $\Gamma()$ is the gamma function and $\mathbf{\Lambda}_k$ is the inverse covariance matrix of $\mathbf{s}_k$,

$$\mathbf{\Lambda}_k = \frac{1}{1-\rho_k^2}\begin{pmatrix} 1 & -\rho_k \\ -\rho_k & 1 \end{pmatrix}. \tag{4}$$

The parameter ρ_k is the correlation coefficient between s_k^1 and s_k^2. As ν_k becomes larger, the distribution $p_{\mathbf{s}_k}$ approaches a Gaussian.

Together with Eq. (1), the density $p_{\mathbf{s}_k}$ leads to the log-likelihood ℓ,

$$\ell(\mathbf{q}_1^1, \mathbf{q}_2^1, \ldots, \mathbf{q}_d^1, \mathbf{q}_d^2, \rho_1, \ldots, \rho_d, \nu_1, \ldots, \nu_d) = \sum_{t=1}^{T}\sum_{k=1}^{d} \log p_{\mathbf{s}_k}(\mathbf{y}_k(t)), \tag{5}$$

where $\mathbf{y}_k(t) = (\mathbf{q}_k^{1T}\mathbf{z}^1(t),\ \mathbf{q}_k^{2T}\mathbf{z}^2(t))^T$ contains the two inner products between the feature vectors $\mathbf{q}_k^i$ and the t-th observation of the white random vector $\mathbf{z}^i$. As denoted in the equation, maximization of the log-likelihood ℓ can be used to find the features $\mathbf{q}_k^i$ (the columns of the orthonormal matrices $\mathbf{Q}^i$), the correlation coefficients ρ_k, as well as the parameters ν_k. If the learned ν_k have small values there are higher-order statistical dependencies between the features; large values mean that the correlation coefficient ρ_k captures already most of the dependency.

We show now that maximization of Eq. (5) generalizes CCA. More specifically, we show that for large values of ν_k, the vectors $\mathbf{q}_k^i$ which maximize ℓ are those found by CCA: The objective ℓ considered as function of the $\mathbf{q}_k^i$ is

$$\ell(\mathbf{q}_1^1, \ldots, \mathbf{q}_d^2) = \text{const} - \sum_{t=1}^{T}\sum_{k=1}^{d} \frac{\nu_k+2}{2} \log\left(1 + \frac{1}{\nu_k - 2}\mathbf{y}_k(t)^T\mathbf{\Lambda}_k\mathbf{y}_k(t)\right). \tag{6}$$

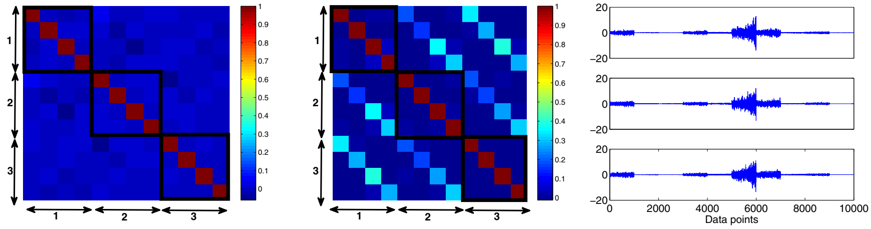
For large ν_k the term $1/(\nu_k - 2)\mathbf{y}_k(t)^T\mathbf{\Lambda}_k\mathbf{y}_k(t)$ is small so that we can use the first-order Taylor expansion $\log(1+x) = x + O(x^2)$. Taking further into account that the $\mathbf{z}^i$ are white and that the $\mathbf{q}_k^i$ have unit norm, we obtain with Eq. (4)

$$\ell(\mathbf{q}_1^1, \mathbf{q}_1^2, \ldots, \mathbf{q}_d^1, \mathbf{q}_d^2) \approx \text{const} + T\sum_{k=1}^{d} \frac{1}{1-\rho_k^2}\left(\rho_k \mathbf{q}_k^{1T}\widehat{\mathbf{\Sigma}}_{12}\mathbf{q}_k^2\right), \tag{7}$$

where $\widehat{\mathbf{\Sigma}}_{12}$ is the sample cross-correlation matrix between $\mathbf{z}^1$ and $\mathbf{z}^2$. Since $1-\rho_k^2$ is positive, ℓ is maximized when $|\mathbf{q}_k^{1T}\widehat{\mathbf{\Sigma}}_{12}\mathbf{q}_k^2|$ is maximized for all k under the orthonormality constraint for the matrices $\mathbf{Q}^i = (\mathbf{q}_1^i \ldots \mathbf{q}_d^i)$. We need here the absolute value since ρ_k can be positive or negative. This set of optimization problems is solved by CCA, see for example [3, ch. 3]. Normally, CCA maximizes $\mathbf{q}_k^{1T}\widehat{\mathbf{\Sigma}}_{12}\mathbf{q}_k^2$ so that for negative ρ_k, one of the $\mathbf{q}_k^i$ obtained via maximization of ℓ would have switched signs compared to the one obtained with CCA.

2.3 Analysis of Multiple Data Sets

We return now to Eq. (2), and consider the case where the $\tilde{s}_k^i$ are independent random variables which follow a standard normal distribution. The random variables $s_k^1, \ldots, s_k^n$ are then linearly uncorrelated but have higher order



(a) Linear correlation (b) Correlation of squares (c) Coactivation

Fig. 1. We illustrate with artificial data the coactivation of the features across the data sets. (a) Correlation coefficients between the s_k^i. (b) Correlation coefficients between the squared s_k^i. The black rectangles indicate each data set. In this example, there are three data sets ($n = 3$), each has four dimensions ($d = 4$). (c) Illustration of the dependencies between the s_1^i. Row i shows s_1^i, $i \in \{1, 2, 3\}$. Correlation of squares means that the sources tend to be concurrently activated. Note that the data points $s_k^i(t)$, $t = 1, \ldots, 10000$ do not have an order. To visualize coactivation, we chose the order in the figure.

dependencies. The dependencies can be described by the terms "coactivation" or "variance-coupling": whenever one variable is strongly nonzero the others are likely to be nonzero as well. Figure 1 illustrates this for the case of three coupled data sets ($n = 3$) with dimensionality four ($d = 4$).

Under the assumption of uncorrelated Gaussian $\tilde{s}_k^i$, the log-likelihood ℓ to estimate the features $\mathbf{q}_k^i$ is

$$\ell(\mathbf{q}_1^1, \ldots, \mathbf{q}_d^n) = \sum_{t=1}^{T} \sum_{k=1}^{d} G_k \left(\sum_{i=1}^{n} (\mathbf{q}_k^{i\,T} \mathbf{z}^i(t))^2 \right), \tag{8}$$

where $\mathbf{z}^i(t)$ is the t-th data point in data set $i = 1, \ldots, n$, and G_k is a nonlinearity which depends on the distribution of the variance variable σ_k.

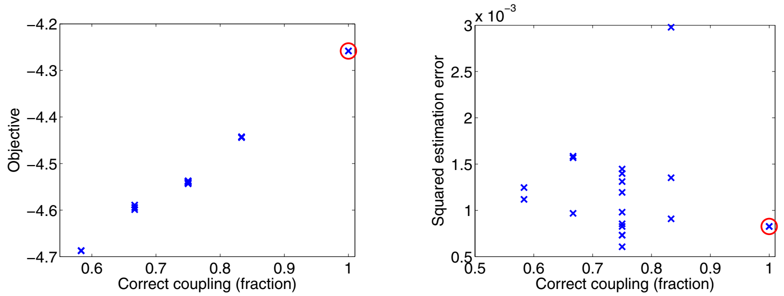
This model is closely related to Independent Subspace Analysis (ISA) [5, ch. 20]. ISA is a generalization of ICA; the sources are not assumed to be statistically independent but, like above, some groups of sources (subspaces) are dependent through a common variance variable. ISA was proposed for the analysis of a single data set but by imposing constraints on the feature vectors we can relate it to our model: Denote by $\mathbf{z}$ and $\mathbf{s}$ the vectors in $\mathbb{R}^{dn}$ which are obtained by stacking the $\mathbf{z}^i$ and $\mathbf{s}^i$ on each other. Eq. (1) can then be written as $\mathbf{z} = \mathbf{Q}\mathbf{s}$. The matrix $\mathbf{Q}$ is orthonormal and block-diagonal, with blocks given by the $\mathbf{Q}_i$. Our dependency assumptions for the sources s_k^i in this subsection correspond to the dependency assumptions in ISA. This means that our model corresponds to an ISA model with a block-diagonality constraint for the mixing matrix. This correspondence allows us to maximize the log-likelihood in Eq. (8) with an adapted version of the FastISA algorithm [6].

3 Simulations with Artificial Data

In this section, we use artificial data to both illustrate the theory and to test our methods. To save space, we only show results for the method in Subsection 2.3. We generated data which follows the model of Subsection 2.1 and 2.3; the dependencies for that kind of data were illustrated in Figure 1. As in the figure, we set the number of data sets to three ($n = 3$), and the dimension of each data set to four ($d = 4$). The variance variables σ_k in Eq. (2) were generated by squaring Gaussian random variables. The sources s_k^i were then normalized to unit variance. The three orthonormal mixing matrices $\mathbf{Q}_i$ were drawn at random. This defined the three random variables $\mathbf{z}_i$. For each, we drew $T = 10000$ observations, which gave the coupled data sets.

Given the data sets, we optimized the log-likelihood ℓ in Eq. (8) to estimate the coupled features (the columns $\mathbf{q}_i^k$ of the mixing matrices $\mathbf{Q}_i$). As nonlinearity, we chose $G_k(u) = G(u) = -\sqrt{0.1 + u}$, as in [6]. Comparison of the estimates with the true features allows to assess the method. In particular, we can assess whether the coupling is estimated correctly. The ICA model for each of the data sets, see Eq. (1), can only be estimated up to a permutation matrix. That is, the order of the sources is arbitrary. However, for the coupling between the features to be correct, the permutation matrix for each of the data sets must be the same. Comparison of the permutation matrices allows to assess the estimated coupling.

We tested the algorithm for ten toy data sets (each consisting of three coupled data sets of dimension four). In each case, we found the correct coupling at the maximum of the objective in Eq. (8). However, we observed that the objective has local maxima. Figure 2 shows that only the global maximum corresponds to the



(a) Correct coupling vs. objective (b) Correct coupling vs. estim. error

Fig. 2. Local maxima in the objective function in Eq. (8). The figures show simulation results where, for the same data, we started from 20 different random initializations of the $\mathbf{Q}_i$. The red circle indicates the trial with the largest objective. (a) We plot the value of the objective function versus the fraction of the correct learned coupling. The larger the value of the objective, the better the estimated coupling. (b) We plot the sum of the estimation errors in the $\mathbf{Q}_i$ versus the learned coupling. The estimation error can be very small but the estimated coupling can be wrong. This happens when the $\mathbf{Q}_i$ are individually well estimated but they do not have the same permutation matrix.

correct estimates of the coupling. We have used the adapted FastISA algorithm to maximize Eq. (8). It has been pointed out that FastISA converges to local maxima [6]. When we used a simple gradient ascent algorithm to maximize Eq. (8), we observed also the presence of local maxima – the results were as in Figure 2.

Simulations with the method for two data sets, outlined in Subsection 2.2, showed that local maxima also exist in that case (results not shown).

4 Simulations with Real Data

In Subsection 4.1, we apply our new method to the analysis of structure in natural images; we are learning from image sequences (video clips) features that are related over time. In Subsection 4.2, we apply the method to brain imaging data.

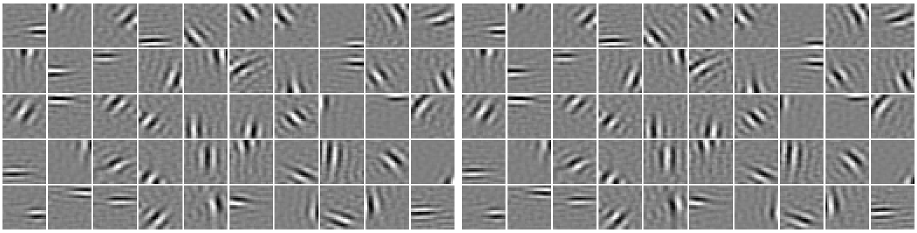
4.1 Simulations with Natural Images

We use here the method outlined in Subsection 2.3 for the analysis of $n = 2$ and $n = 5$ coupled data sets. First, we consider the case of two data sets, and compare our results with those obtained with CCA. The two data sets were constructed from natural image sequences. The database consisted of the 129 videos used in [4].[2] From this data, we extracted $T = 10000$ image patches of size 25px × 25px at random locations and at two time points. The first time points were also random; the resulting image patches formed the first data set. The second time points were 40ms after the first time points; these image patches formed the second data set. As preprocessing, we whitened each data set individually and retained in both cases 50 dimensions (98% of the variance). This gave our data $\mathbf{z}^i(t) \in \mathbb{R}^{50}, i \in \{1, 2\}$ and $t = 1, \ldots, 10000$, for the learning of the $\mathbf{q}_k^i, k = 1, \ldots, 50$. We run the algorithm five times, and picked the features giving the highest log-likelihood.

Figure 3 shows the learned features where we included the whitening matrices in the visualization: the features $(\mathbf{q}_k^{i\,T}\mathbf{V}^i)^T$ are shown, where $\mathbf{V}^i$ is the whitening matrix for the i-th data set. The learned features are Gabor-like. The features are arranged such that the k-th feature of the first data set is coupled with the k-th feature of the second data set. It can be clearly seen that the coupled features are very similar. This shows that, for natural video, the Gabor features produce temporally stable responses. This result is in line with previous research on natural images which explicitly learned temporally stable features from the same database [4]. This shows that the presence of local maxima in the objective ℓ is not really harmful; our learned features, which most likely correspond to a local maximum, also produced meaningful insight into the structure of the investigated coupled data sets.

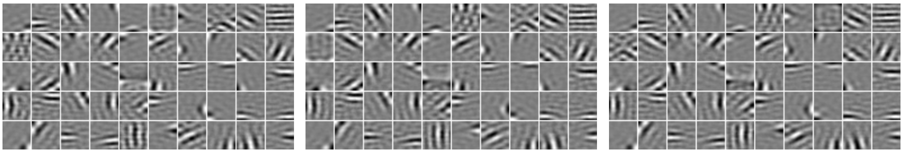
As a baseline for this simulation, we also applied CCA to the two coupled data sets. The extracted features were highly correlated but they did not identify meaningful structure in the data. The features were noise-like (results not shown). This shows the advantages of having a method at hand which takes both within and across the data sets higher-order statistics into account.

[2] For more details on the database, see [4], and references within.



(a) Features, first data set (b) Features, second data set

Fig. 3. Activity-coupled features in natural image sequences. The natural image patches in the second data set showed the same image sections as those in the first data set but 40ms later. The k-th feature of the first data set is coupled with the k-th feature of the second data set. The coupled features are very similar. This shows that Gabor features produce temporally stable responses [4].



(a) First data set (b) Third data set (c) Fifth data set

Fig. 4. Activity-coupled features in natural image sequences. The image patches in the five data sets showed the same image sections at different time points, each 40ms apart. The features for only three of five data sets are shown.

Next, we consider the case of $n = 5$ data sets. The image patches in the different data sets showed the same image sections at different time points, each 40ms apart. Figure 4 shows the results. The learned coupled features are again very similar, albeit less localized than those in Figure 3. The similarity of the features in the different data sets means that, for natural image sequences, the Gabor features tend to be active for a longer time period, see also [4].

4.2 Simulations with Brain Imaging Data

Finally, we apply the method of Subsection 2.2 to magnetoencephalography (MEG) data. [3] A subject received alternating visual, tactile and auditory stimulation interspersed with rest [9]. We estimated sources by a blind source separation method and chose for further analysis two sources which were located close to each other in the somatosensory or motor areas. We took at random time points windows of size 300ms for each source. This formed the two data sets which we analyzed with our method.

Figure 5 shows three selected pairs of the learned coupled features. The results indicate the presence of highly synchronized activity in the brain. The correlation

[3] We thank Pavan Ramkumar and Riitta Hari from the Brain Research Unit of Aalto University for the access to the data.

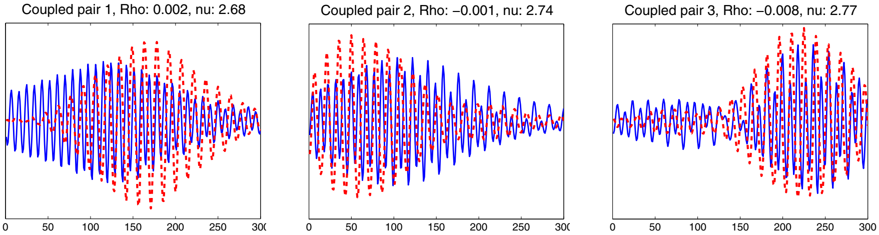


Fig. 5. Coupled features in MEG data. The feature outputs show no linear correlation ($\rho_k \approx 0$) but are nonlinearly correlated ($\nu_k \approx 2.7$).

coefficients ρ_k between the feature outputs are practically zero which shows that higher-order dependencies need to be detected in order to find this kind of synchronization.

5 Conclusions

We have presented a data analysis method which generalizes canonical correlation analysis to higher-order statistics and to multiple data sets. The method finds independent components which, across the data sets, tend to be jointly activated ("coactivated features"). The method was tested on artificial data, and its applicability to real data was demonstrated on natural images and brain imaging data.

References

1. Archambeau, C., Bach, F.: Sparse probabilistic projections. In: Advances in Neural Information Processing Systems (NIPS), vol. 21 (2009)
2. Archambeau, C., Delannay, N., Verleysen, M.: Mixtures of robust probabilistic principal component analyzers. Neurocomputing 71(7-9), 1274–1282 (2008)
3. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Heidelberg (2009)
4. Hurri, J., Hyvärinen, A.: Simple-cell-like receptive fields maximize temporal coherence in natural video. Neural Computation 15(3), 663–691 (2003)
5. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Chichester (2001)
6. Hyvärinen, A., Köster, U.: FastISA: A fast fixed-point algorithm for independent subspace analysis. In: 14th European Symposium on Artificial Neural Networks, ESANN (2006)
7. Karhunen, J., Ukkonen, T.: Extending ICA for finding jointly dependent components from two related data sets. Neurocomputing 70(16-18), 2969–2979 (2007)
8. Klami, A., Kaski, S.: Probabilistic approach to detecting dependencies between data sets. Neurocomputing 72(1-3), 39–46 (2008)
9. Ramkumar, P., Parkkonen, L., Hari, R., Hyvärinen, A.: Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. Human Brain Mapping (in press)