Extracting Coactivated Features from Multiple Data Sets

Michael U. Gutmann
University of Helsinki
michael.gutmann@helsinki.fi

Aapo Hyvärinen
University of Helsinki
aapo.hyvarinen@helsinki.fi

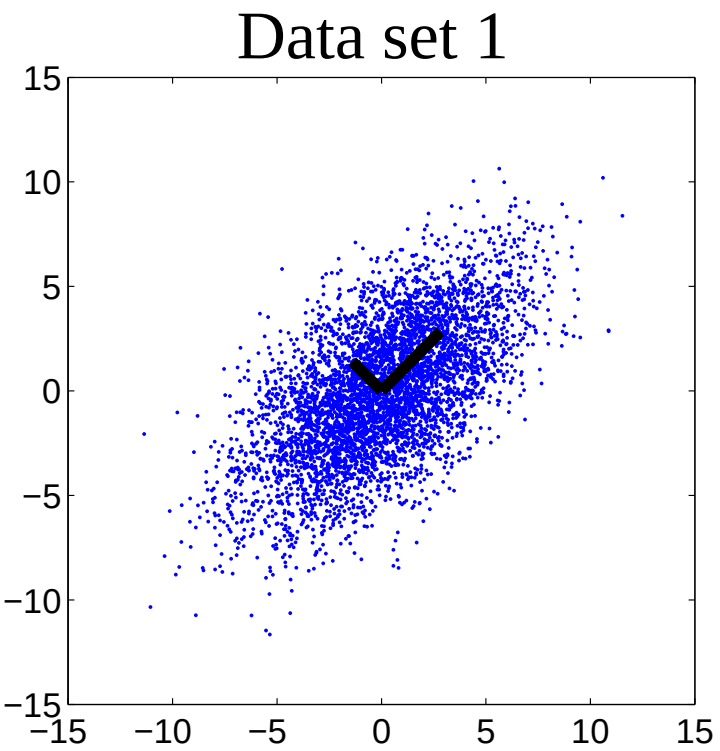
Contents

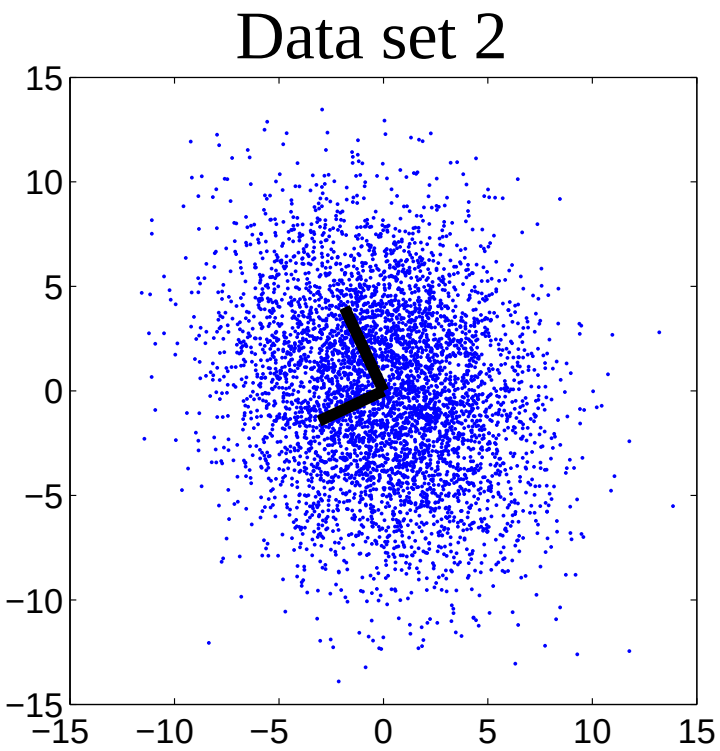
This talk is about a new method to find related features (structure) in multiple data sets.

- Background information on the extraction of related features from multiple data sets
- Explanation of the statistical model underlying our method
- Testing our method on artificial data
- Application to real data
  (here: natural images, in the paper: also brain imaging data)

Introduction

An example





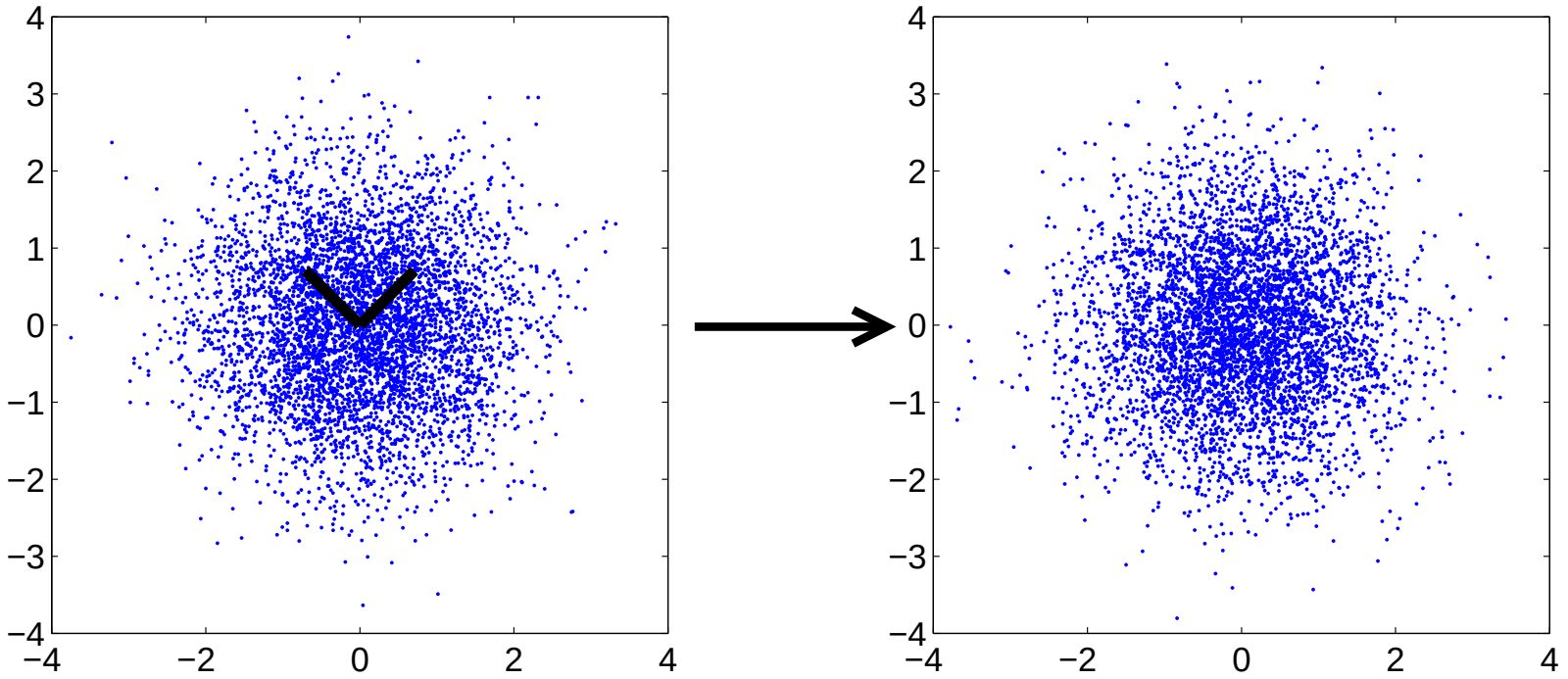
Goals:

1. Characterize each data set separately
   $\rightarrow$ Eigenvalues and eigenvectors of covariance matrices
2. Find relations between the two data sets

Correlation based method (1/3)

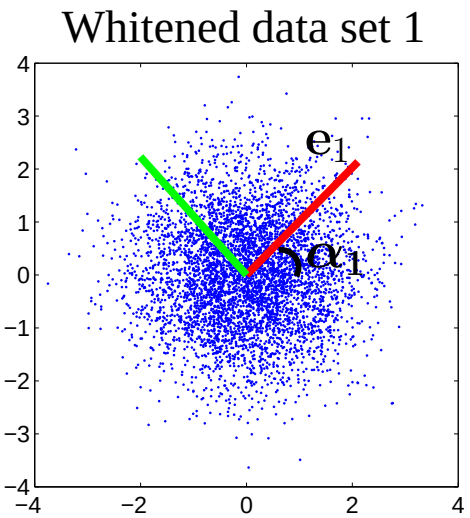
Whitened data set 1
(normalized representation)

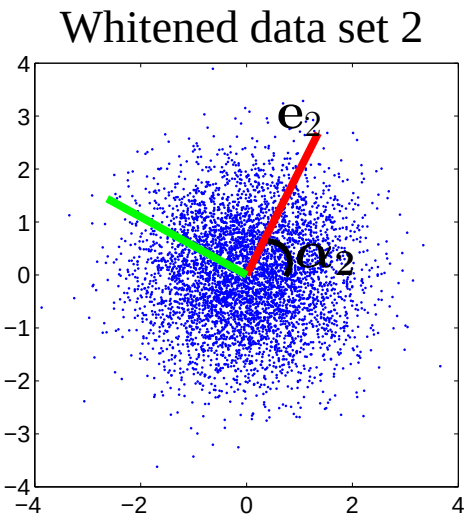
After change of basis:
data is still white

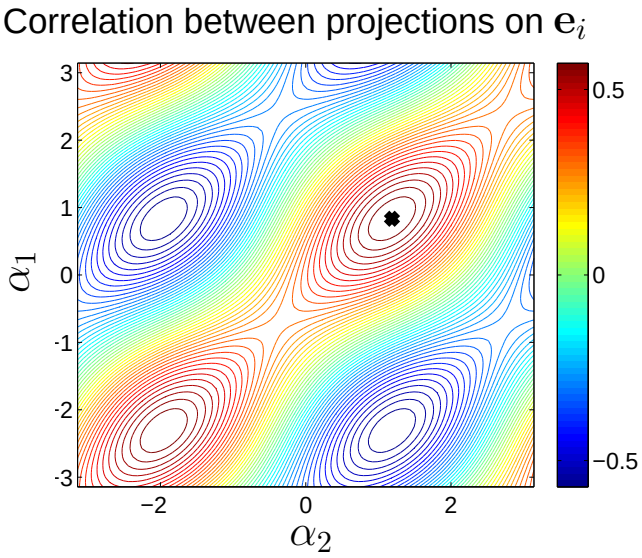


- Whitening is defined up to a rotation
- Choose coordinate systems which best describe the relation between the two data sets 1 and 2

Correlation based method (2/3)

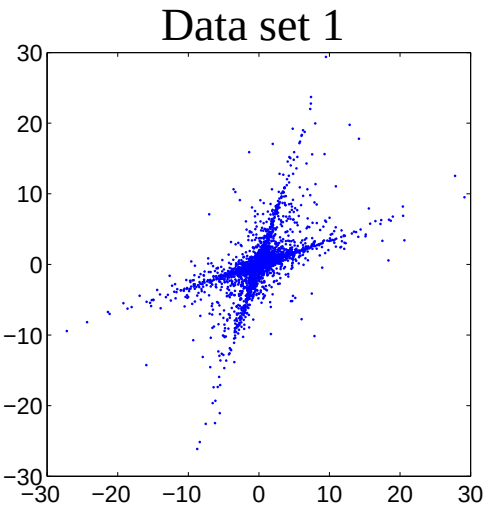


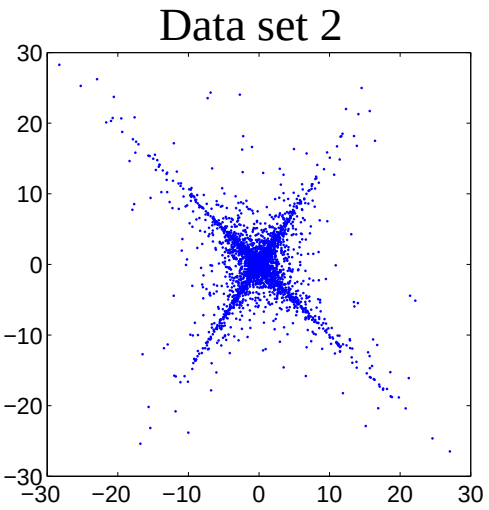


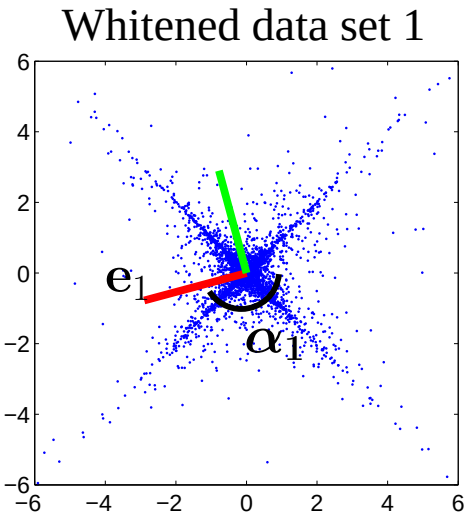


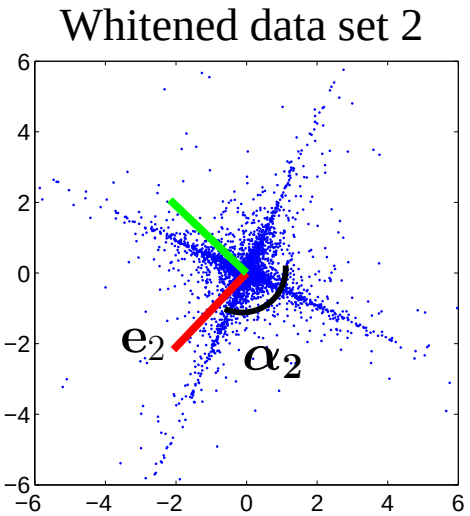
- The x-coordinate of a data point in the coordinate system defined by $\mathbf{e}_i$ is given by its projection on $\mathbf{e}_i$.
- Compute the correlation between the x-coordinates for different coordinate systems.
- Choose the coordinate systems for which the x-coordinates are most strongly correlated (here: correlation coefficient of ≈ 0.6).
- Described method is called Canonical Correlation Analysis (CCA).

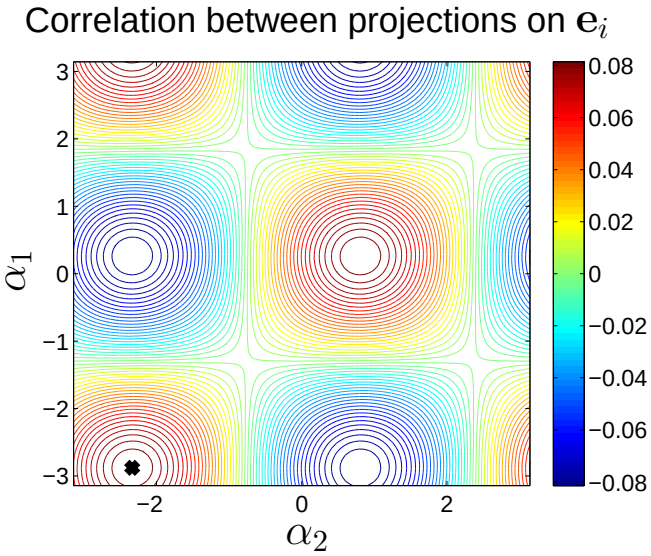
Correlation based method (3/3)











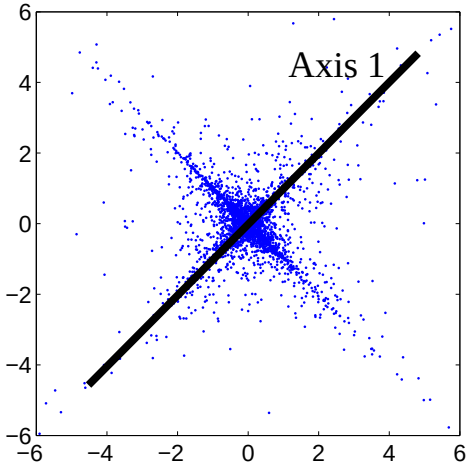
$\rightarrow$ Method does not seem to work here.

What does “not work” mean?

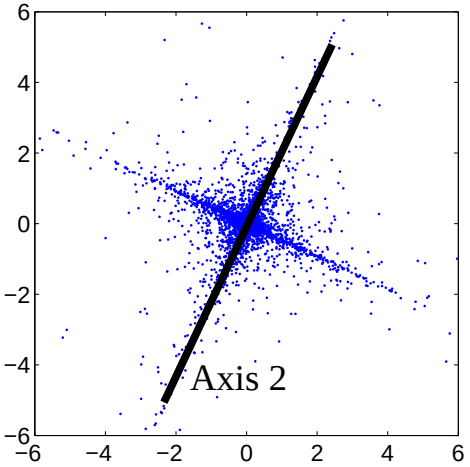
It means

1. that we did not find meaningful features within each data set
2. that we did not find any relation between the two data sets

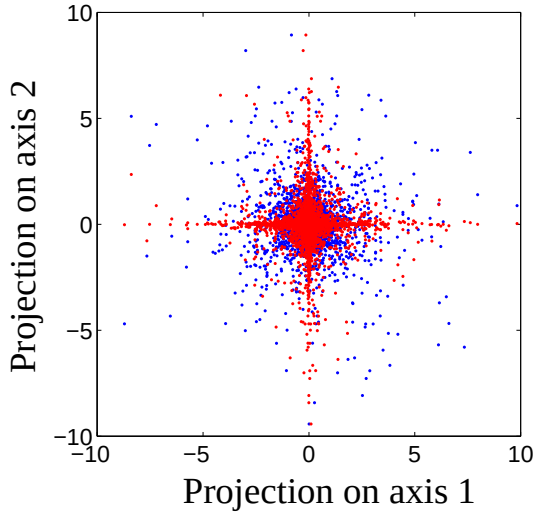
Whitened data set 1



Whitened data set 2



Blue: original data
Red: data with same marginal distr.



⇓ Identify independent components

⇓ Identify coactivation

In this talk . . .

I will present a method where

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

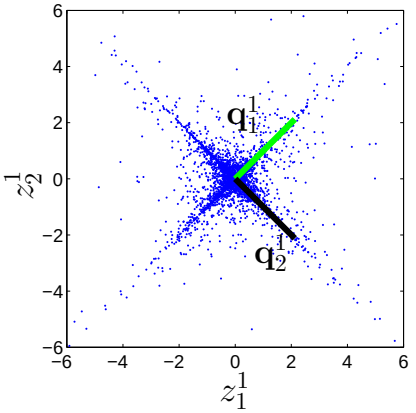
Extraction of coactivated features

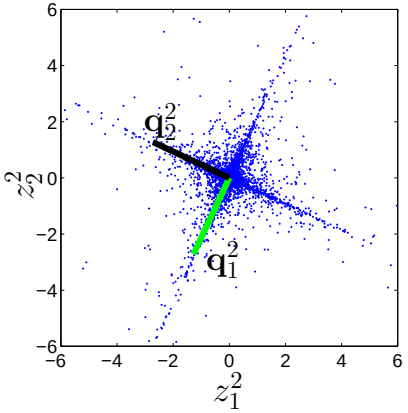
Statistical model underlying our method (1/2)

- Given n data sets, we assume that each set is formed by iid. observations of a random vector $\mathbf{z}^i \in \mathbb{R}^d$.
- To model structure within a data set, we assume that

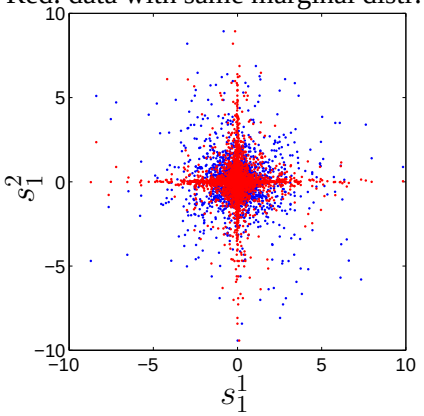
$$\mathbf{z}^i = \sum_{k=1}^{d} \mathbf{q}_k^i s_k^i \quad (i = 1, \ldots n)$$

  $\mathbf{q}_k^i$: orthonormal, $s_1^i, \ldots, s_d^i$: statistically independent
- To model structure across the data sets, we assume that $s_k^1, \ldots, s_k^n$ are statistically *dependent*.





Blue: original data
Red: data with same marginal distr.

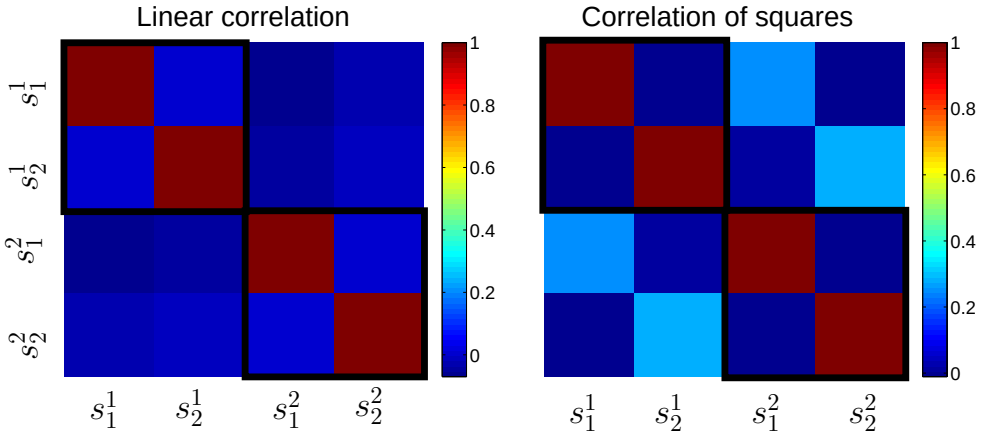


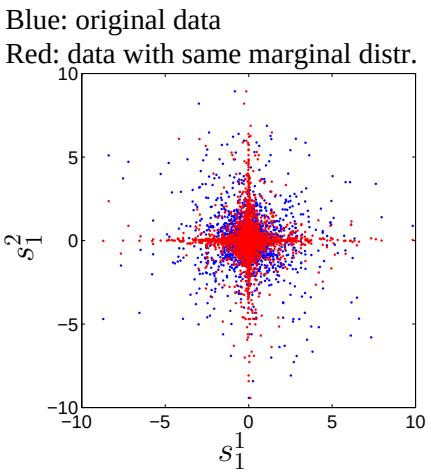
Statistical model underlying our method (2/2)

- Dependency assumptions: The k-th sources from all the data sets share a common (latent) variance variable σ_k:

$$s_k^1 = \sigma_k \tilde{s}_k^1 \qquad s_k^2 = \sigma_k \tilde{s}_k^2 \qquad \ldots \qquad s_k^n = \sigma_k \tilde{s}_k^n$$

- The $\tilde{s}_k^1, \ldots, \tilde{s}_k^n$ are Gaussian (zero mean, possibly correlated).
- Choosing a prior for σ_k completes the model specification.





Applying the method – learning the parameters

- The most interesting parameters of the model are the features $\mathbf{q}_k^i$.
- They can be learned by maximizing the log-likelihood $\ell(\mathbf{q}_1^1, \ldots, \mathbf{q}_d^n)$.
- For the special case of uncorrelated sources,

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

where

$$G_k(u) = \log \int \frac{p_{\sigma_k}(\sigma_k)}{(2\pi\sigma_k^2)^{\frac{n}{2}}} \exp\left(-\frac{u^2}{2\sigma_k^2} \right) \mathrm{d}\sigma_k. \qquad (2)$$

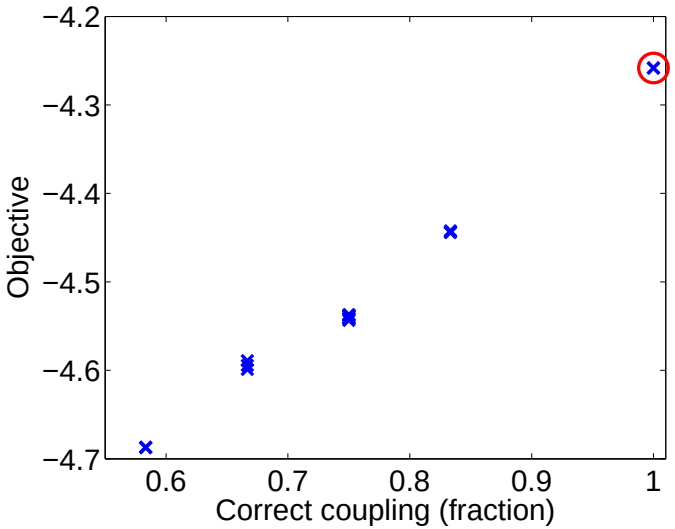
Equations show that the presented method is related to Independent Subspace Analysis (see paper)

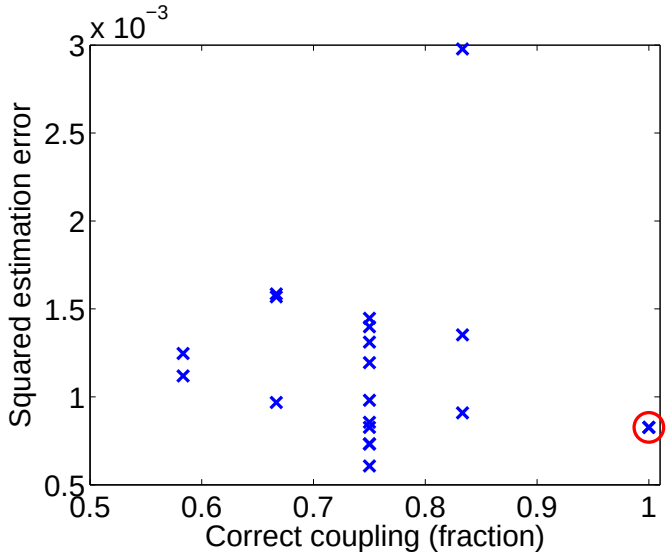
Testing on artificial data

Simulation setup and goals

- Setup
  - Three data sets ($n = 3$) of dimension four ($d = 4$)
  - No linear correlation in the sources s_k^i
  - Randomly chosen orthonormal mixing matrices $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3$
  - 10000 observations
  - Learning of the parameters by maximization of log-likelihood, with ad hoc nonlinearity $G(u) = -\sqrt{0.1 + u}$
- Quantities of interest
  - Error in the mixing matrices
  - Identification of the coupling

Results





Results for *one* estimation problem. Optimization performed for 20 different initializations.

- Correct coupling at maximum of the log-likelihood
- Presence of local maxima (not nice! but no catastrophe, see following slides)
- Learning the right mixing matrices without learning the right coupling seems possible.

Application to real data

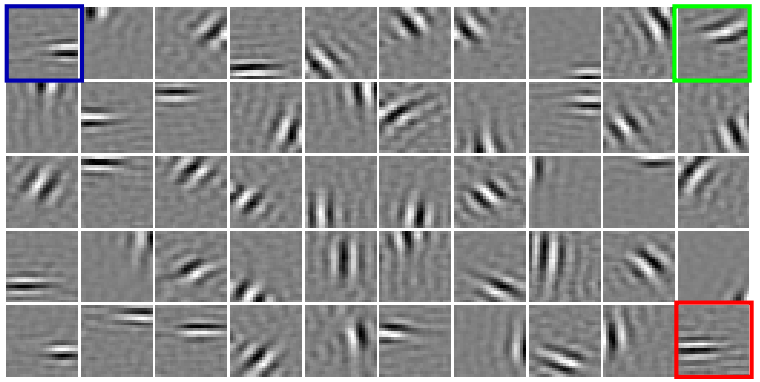
Setup

- Data set 1: 10000 image patches of size 25px $\times$ 25px, extracted at random locations from natural video data
- Data set 2: same image patches, 40ms later
- For each data set separately: whitening and dimension reduction (98% of variance retained)
- Learning of 50 features per data set by maximization of the log-likelihood of the model (same objective function as for the artificial data).

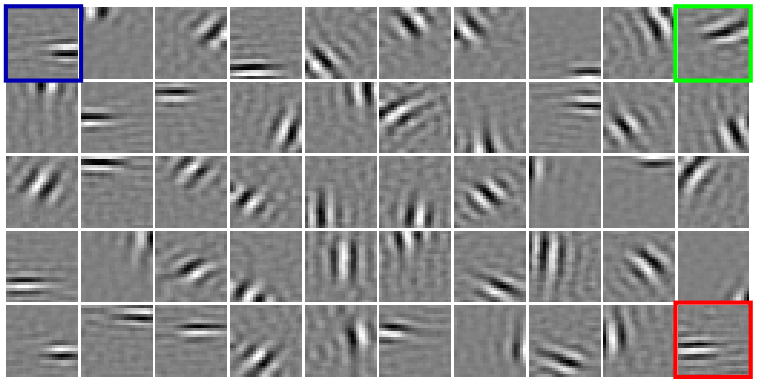
Learned features

Our method:

First data set

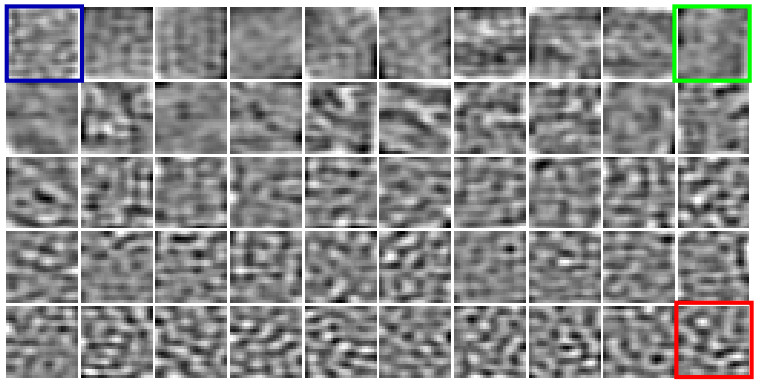


Second data set

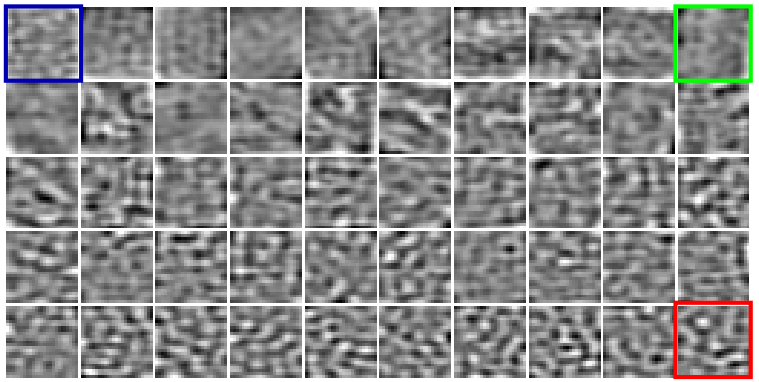


Canonical correlation analysis:

First data set



Second data set



Summary

Summary

- Presented a new method to find related features (structure) in multiple data sets:

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

- In the paper:
  - more theory (in particular, more on the relation to CCA)
  - more simulations with natural images
  - simulations with brain imaging data