Shared Space Component Analysis

Mihalis A. Nicolaou

Deparment of Computing,
Imperial College London, UK
iBug Talk
mihalis@imperial.ac.uk

February 2015

Intro

Component Analysis (CA): Collection of statistical methods aiming at the factorisation of a signal into components pertaining to a particular task (e.g., predictive analysis, clustering), usually of reduced dimensionality.

- CA utilised in the vast majority of ML and CV systems (e.g., PCA).
- Closely related to the process of dimensionality reduction, specifically wrt. feature extraction.

Intro

Component Analysis (CA): Collection of statistical methods aiming at the factorisation of a signal into components pertaining to a particular task (e.g., predictive analysis, clustering), usually of reduced dimensionality.

- CA utilised in the vast majority of ML and CV systems (e.g., PCA).
- Closely related to the process of dimensionality reduction, specifically wrt. feature extraction.

Roots of CA lie in Principal CA (PCA)

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane.

On lines and planes of closest fit to systems of points in space (Pearson, 1901)

Figure from Pearson's paper points out that the line of best fit is the direction with maximum variance. Also note, line of worst fit is orthogonal to the line of best fit.

Roots of CA lie in Principal CA (PCA)

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane.

On lines and planes of closest fit to systems of points in space (Pearson, 1901)

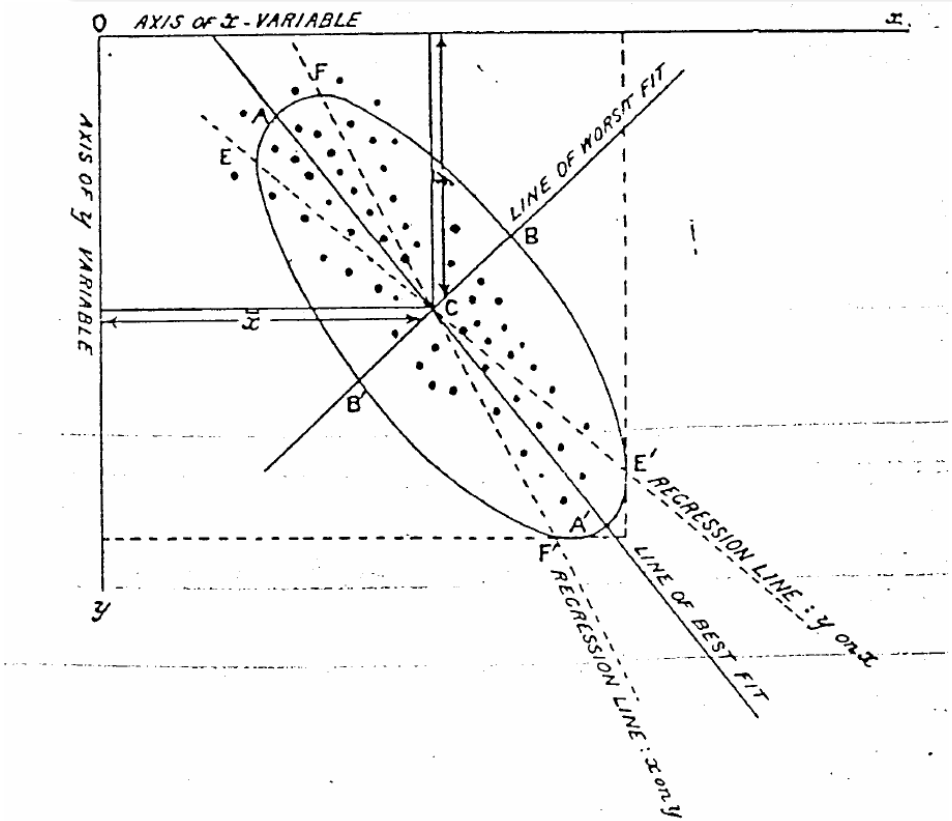


Figure from Pearson's paper points out that the line of best fit is the direction with maximum variance. Also note, line of worst fit is orthogonal to the line of best fit.

Roots of CA lie in Principal CA (PCA)

- PCA independently developed in 1933 by Hotelling, coining the term "Principal Components".
- In PCA and linear CA in general, we aim to learn a linear transformation $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where $\mathbf{X}$ is the observation matrix.
- $\mathbf{Y}$ is the latent, unobserved space which satisfies certain properties. In case of PCA, $\mathbf{Y}$ needs to maximally preserve variance (Hotelling), or equivalently, minimise the reconstruction error (Pearson).
- Similarly to PCA, many other well known CA techniques deal with just one set of observations.
- In order to define a notion of a "shared space" though, we need more than one set of observations.

Roots of CA lie in Principal CA (PCA)

- PCA independently developed in 1933 by Hotelling, coining the term "Principal Components".
- In PCA and linear CA in general, we aim to learn a linear transformation $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where $\mathbf{X}$ is the observation matrix.
- $\mathbf{Y}$ is the latent, unobserved space which satisfies certain properties. In case of PCA, $\mathbf{Y}$ needs to maximally preserve variance (Hotelling), or equivalently, minimise the reconstruction error (Pearson).
- Similarly to PCA, many other well known CA techniques deal with just one set of observations.
- In order to define a notion of a "shared space" though, we need more than one set of observations.

Roots of CA lie in Principal CA (PCA)

- PCA independently developed in 1933 by Hotelling, coining the term "Principal Components".
- In PCA and linear CA in general, we aim to learn a linear transformation $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where $\mathbf{X}$ is the observation matrix.
- $\mathbf{Y}$ is the latent, unobserved space which satisfies certain properties. In case of PCA, $\mathbf{Y}$ needs to maximally preserve variance (Hotelling), or equivalently, minimise the reconstruction error (Pearson).
- Similarly to PCA, many other well known CA techniques deal with just one set of observations.
- In order to define a notion of a "shared space" though, we need more than one set of observations.

Roots of CA lie in Principal CA (PCA)

- PCA independently developed in 1933 by Hotelling, coining the term "Principal Components".
- In PCA and linear CA in general, we aim to learn a linear transformation $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where $\mathbf{X}$ is the observation matrix.
- $\mathbf{Y}$ is the latent, unobserved space which satisfies certain properties. In case of PCA, $\mathbf{Y}$ needs to maximally preserve variance (Hotelling), or equivalently, minimise the reconstruction error (Pearson).
- Similarly to PCA, many other well known CA techniques deal with just one set of observations.
- In order to define a notion of a "shared space" though, we need more than one set of observations.

Roots of CA lie in Principal CA (PCA)

- PCA independently developed in 1933 by Hotelling, coining the term "Principal Components".
- In PCA and linear CA in general, we aim to learn a linear transformation $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where $\mathbf{X}$ is the observation matrix.
- $\mathbf{Y}$ is the latent, unobserved space which satisfies certain properties. In case of PCA, $\mathbf{Y}$ needs to maximally preserve variance (Hotelling), or equivalently, minimise the reconstruction error (Pearson).
- Similarly to PCA, many other well known CA techniques deal with just one set of observations.
- In order to define a notion of a "shared space" though, we need more than one set of observations.

Canonical Correlation Analysis

- Most well known method belonging in the so-called Shared-Space Component Analysis family is Canonical Correlation Analysis (CCA).
- Introduced by Hotelling in 1936 (Relations Between Two Sets of Variates, Biometrika), three years after he proposed PCA.
- CCA is actually a natural extension of PCA to two datasets, only instead of maximally preserving variance, we maximise the correlation (covariance) between the datasets.

$$\textbf{PCA} \rightarrow \mathbf{W}^T \mathbf{\Sigma}_{XX} \mathbf{W} = \mathbf{W}^T \mathbf{X}\mathbf{X}^T \mathbf{W}, \text{ where } \mathbf{Y} = \mathbf{W}^T \mathbf{X}$$

$$\textbf{CCA} \rightarrow \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2 = \mathbf{W}_1^T \mathbf{X_1} \mathbf{X_2} \mathbf{W}_2, \text{ where } \mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$$

where $Cov(\mathbf{X}, \mathbf{X}) = Var(\mathbf{X})$

Canonical Correlation Analysis

- Most well known method belonging in the so-called Shared-Space Component Analysis family is Canonical Correlation Analysis (CCA).
- Introduced by Hotelling in 1936 (Relations Between Two Sets of Variates, Biometrika), three years after he proposed PCA.
- CCA is actually a natural extension of PCA to two datasets, only instead of maximally preserving variance, we maximise the correlation (covariance) between the datasets.

$$\textbf{PCA} \rightarrow \mathbf{W}^T \mathbf{\Sigma}_{XX} \mathbf{W} = \mathbf{W}^T \mathbf{X}\mathbf{X}^T \mathbf{W}, \text{ where } \mathbf{Y} = \mathbf{W}^T \mathbf{X}$$

$$\textbf{CCA} \rightarrow \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2 = \mathbf{W}_1^T \mathbf{X}_1 \mathbf{X}_2 \mathbf{W}_2, \text{ where } \mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$$

where $Cov(\mathbf{X}, \mathbf{X}) = Var(\mathbf{X})$

Canonical Correlation Analysis

- Most well known method belonging in the so-called Shared-Space Component Analysis family is Canonical Correlation Analysis (CCA).
- Introduced by Hotelling in 1936 (Relations Between Two Sets of Variates, Biometrika), three years after he proposed PCA.
- CCA is actually a natural extension of PCA to two datasets, only instead of maximally preserving variance, we maximise the correlation (covariance) between the datasets.

  PCA $\rightarrow \mathbf{W}^T \mathbf{\Sigma}_{XX} \mathbf{W} = \mathbf{W}^T \mathbf{X}\mathbf{X}^T \mathbf{W}$, where $\mathbf{Y} = \mathbf{W}^T \mathbf{X}$

  CCA $\rightarrow \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2 = \mathbf{W}_1^T \boldsymbol{X_1 X_2} \mathbf{W}_2$, where $\mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$

  where $Cov(\mathbf{X}, \mathbf{X}) = Var(\mathbf{X})$

CCA is a natural extension of PCA (More Formally)

CCA: Given observations $\mathbf{X}_1 \in \Re^{F_1 x T}$, $\mathbf{X}_2 \in \Re^{F_2 x T}$ infer maximally correlated latent spaces $\mathbf{Y}_1 \in \Re^{N x T}$, $\mathbf{Y}_2 \in \Re^{N x T}$.

- Assuming linear projections, $\mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$, we can maximise the standard, pearson correlation coefficient, i.e.:

$$\frac{Cov(\mathbf{Y}_1, \mathbf{Y}_2)}{\sigma_{\mathbf{Y}_1} \sigma_{\mathbf{Y}_2}} = \frac{\mathbb{E}[\mathbf{Y}_1 \mathbf{Y}_2]}{\mathbb{E}[\mathbf{Y}_1^2]\mathbb{E}[\mathbf{Y}_2^2]} = \frac{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_1} \mathbf{W}_1} \sqrt{\mathbf{W}_2^T \mathbf{\Sigma}_{X_2 X_2} \mathbf{W}_2}}$$

- Due to scale invariance of correlation wrt. loadings, we have

$$\max \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2, \ s.t. \ \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

- An equivalent least-squares formulation,

$$\min ||\mathbf{W}_1^T \mathbf{X}_1 - \mathbf{W}_2^T \mathbf{X}_2||_F^2, \ s.t. \ \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

CCA is a natural extension of PCA (More Formally)

CCA: Given observations $\mathbf{X}_1 \in \Re^{F_1 x T}$, $\mathbf{X}_2 \in \Re^{F_2 x T}$ infer maximally correlated latent spaces $\mathbf{Y}_1 \in \Re^{N x T}$, $\mathbf{Y}_2 \in \Re^{N x T}$.

- Assuming linear projections, $\mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$, we can maximise the standard, pearson correlation coefficient, i.e.:

$$\frac{Cov(\mathbf{Y}_1, \mathbf{Y}_2)}{\sigma_{\mathbf{Y}_1} \sigma_{\mathbf{Y}_2}} = \frac{\mathbb{E}[\mathbf{Y}_1 \mathbf{Y}_2]}{\mathbb{E}[\mathbf{Y}_1^2]\mathbb{E}[\mathbf{Y}_2^2]} = \frac{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_1} \mathbf{W}_1} \sqrt{\mathbf{W}_2^T \mathbf{\Sigma}_{X_2 X_2} \mathbf{W}_2}}$$

- Due to scale invariance of correlation wrt. loadings, we have

$$\max \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2, \ s.t. \ \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

- An equivalent least-squares formulation,

$$\min ||\mathbf{W}_1^T \mathbf{X}_1 - \mathbf{W}_2^T \mathbf{X}_2||_F^2, \ s.t. \ \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

CCA is a natural extension of PCA (More Formally)

CCA: Given observations $\mathbf{X}_1 \in \Re^{F_1 x T}$, $\mathbf{X}_2 \in \Re^{F_2 x T}$ infer maximally correlated latent spaces $\mathbf{Y}_1 \in \Re^{N x T}$, $\mathbf{Y}_2 \in \Re^{N x T}$.

- Assuming linear projections, $\mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$, we can maximise the standard, pearson correlation coefficient, i.e.:

$$\frac{Cov(\mathbf{Y}_1, \mathbf{Y}_2)}{\sigma_{\mathbf{Y}_1} \sigma_{\mathbf{Y}_2}} = \frac{\mathbb{E}[\mathbf{Y}_1 \mathbf{Y}_2]}{\mathbb{E}[\mathbf{Y}_1^2]\mathbb{E}[\mathbf{Y}_2^2]} = \frac{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_1} \mathbf{W}_1} \sqrt{\mathbf{W}_2^T \mathbf{\Sigma}_{X_2 X_2} \mathbf{W}_2}}$$

- Due to scale invariance of correlation wrt. loadings, we have

$$\max \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2, \; s.t. \; \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

- An equivalent least-squares formulation,

$$\min ||\mathbf{W}_1^T \mathbf{X}_1 - \mathbf{W}_2^T \mathbf{X}_2||_F^2, \; s.t. \; \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

CCA is a natural extension of PCA (More Formally)

CCA: Given observations $\mathbf{X}_1 \in \Re^{F_1 x T}$, $\mathbf{X}_2 \in \Re^{F_2 x T}$ infer maximally correlated latent spaces $\mathbf{Y}_1 \in \Re^{N x T}$, $\mathbf{Y}_2 \in \Re^{N x T}$.

- Assuming linear projections, $\mathbf{Y}_i = \mathbf{W}_i^T \mathbf{X}_i$, we can maximise the standard, pearson correlation coefficient, i.e.:

$$\frac{Cov(\mathbf{Y}_1, \mathbf{Y}_2)}{\sigma_{\mathbf{Y}_1} \sigma_{\mathbf{Y}_2}} = \frac{\mathbb{E}[\mathbf{Y}_1 \mathbf{Y}_2]}{\mathbb{E}[\mathbf{Y}_1^2]\mathbb{E}[\mathbf{Y}_2^2]} = \frac{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_1} \mathbf{W}_1} \sqrt{\mathbf{W}_2^T \mathbf{\Sigma}_{X_2 X_2} \mathbf{W}_2}}$$

- Due to scale invariance of correlation wrt. loadings, we have

$$\max \mathbf{W}_1^T \mathbf{\Sigma}_{X_1 X_2} \mathbf{W}_2, \; s.t. \; \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

- An equivalent least-squares formulation,

$$\min ||\mathbf{W}_1^T \mathbf{X}_1 - \mathbf{W}_2^T \mathbf{X}_2||_F^2, \; s.t. \; \mathbf{W}_i^T \mathbf{\Sigma}_{X_i X_i} \mathbf{W}_i = \mathbf{I}, i = \{1, 2\}.$$

Time Warping is closely related to CA

Dynamic Time Warping

Given $\mathbf{X}_1 \in \mathbb{R}^{D \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D \times T_2}$ (equal D required)

$$\arg\min_{\mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

where $\mathbf{\Delta}_1 \in \{0,1\}^{T_1 \times T_\Delta}$, $\mathbf{\Delta}_2 \in \{0,1\}^{T_2 \times T_\Delta}$ are binary selection matrices, effectively re-mapping the samples in $\mathbf{X}_1$, $\mathbf{X}_2$. Optimal path inferred by dynamic programming in $\mathcal{O}(T_x T_y)$.

Least-Squares CCA

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$ (equal T required),

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2} ||\mathbf{W}_1^T\mathbf{X}_1 - \mathbf{W}_2^T\mathbf{X}_2||_F^2, \ s.t. \ldots.$$

Canonical Time Warping (CTW)

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T_2}$,

Time Warping is closely related to CA

Dynamic Time Warping

Given $\mathbf{X}_1 \in \mathbb{R}^{D \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D \times T_2}$ (equal D required)

$$\arg\min_{\mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Least-Squares CCA

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$ (equal T required),

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2} ||\mathbf{W}_1^T\mathbf{X}_1 - \mathbf{W}_2^T\mathbf{X}_2||_F^2, \ s.t. \ldots.$$

Canonical Time Warping (CTW)

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T_2}$,

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2, \mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{W}_1^T\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{W}_2^T\mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Time Warping is closely related to CA

Dynamic Time Warping

Given $\mathbf{X}_1 \in \mathbb{R}^{D \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D \times T_2}$ (equal D required)

$$\arg\min_{\mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Least-Squares CCA

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$ (equal T required),

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2} ||\mathbf{W}_1^T\mathbf{X}_1 - \mathbf{W}_2^T\mathbf{X}_2||_F^2, \ s.t. \ldots.$$

Canonical Time Warping (CTW) - Zhou and Torre, 2009

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T_2}$ (D_1 may be $\neq D_2$, T_1 may be $\neq T_2$),

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2, \mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{W}_1^T\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{W}_2^T\mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Time Warping is closely related to CA

Dynamic Time Warping

Given $\mathbf{X}_1 \in \mathbb{R}^{D \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D \times T_2}$ (equal D required)

$$\arg\min_{\mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Least-Squares CCA

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$ (equal T required),

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2} ||\mathbf{W}_1^T\mathbf{X}_1 - \mathbf{W}_2^T\mathbf{X}_2||_F^2, \ s.t. \ldots.$$

Canonical Time Warping (CTW) - Alt. Opt. Step 1

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T_2}$,

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2, \mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{W}_1^T\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{W}_2^T\mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Time Warping is closely related to CA

Dynamic Time Warping

Given $\mathbf{X}_1 \in \mathbb{R}^{D \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D \times T_2}$ (equal D required)

$$\arg\min_{\mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

Least-Squares CCA

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$ (equal T required),

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2} ||\mathbf{W}_1^T\mathbf{X}_1 - \mathbf{W}_2^T\mathbf{X}_2||_F^2, \ s.t. \ldots.$$

Canonical Time Warping (CTW) - Alt. Opt. Step 2

Given $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T_1}$, $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times T_2}$,

$$\arg\min_{\mathbf{W}_1, \mathbf{W}_2, \mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{W}_1^T\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{W}_2^T\mathbf{X}_2\mathbf{\Delta}_2||_F^2, \ s.t. \ldots.$$

So far...

- So far we have talked about PCA, and how CCA is a shared space component analysis method which can be considered as a generalisation of PCA to multiple datasets.
- We have also seen how the least-squares CCA can be elegantly combined with Dynamic Time Warping (DTW).
- We will now look into probabilistic interpretations of CCA which are of particular interest.

So far...

- So far we have talked about PCA, and how CCA is a shared space component analysis method which can be considered as a generalisation of PCA to multiple datasets.
- We have also seen how the least-squares CCA can be elegantly combined with Dynamic Time Warping (DTW).
- We will now look into probabilistic interpretations of CCA which are of particular interest.

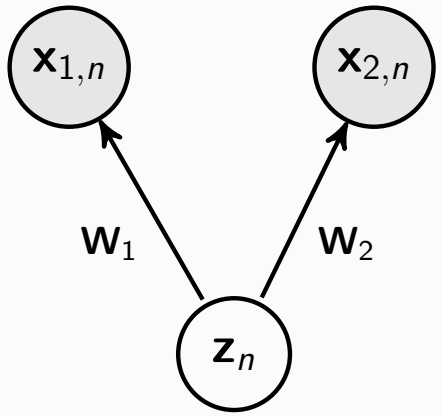
So far...

- So far we have talked about PCA, and how CCA is a shared space component analysis method which can be considered as a generalisation of PCA to multiple datasets.
- We have also seen how the least-squares CCA can be elegantly combined with Dynamic Time Warping (DTW).
- We will now look into probabilistic interpretations of CCA which are of particular interest.

A Probabilistic Interpretation of CCA

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{X}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T}\}$$
$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \epsilon_2$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i \mathbf{I}),\ \mathbf{z_n} \sim \mathcal{N}(0, \mathbf{I})$$



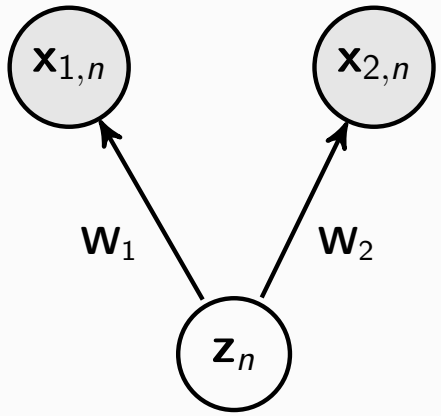
- The Maximum Likelihood (ML) parameter estimates of this model have been shown to be equivalent to deterministic CCA (Bach and Jordan, 2005).
- Note that in this generative model, the "shared-space" $\mathbf{Z}$ is now defined explicitly.

This is actually a special case of a model introduced in 1958, the Inter-battery Factor Analysis (IBFA).

A Probabilistic Interpretation of CCA

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{X}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T}\}$$
$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \epsilon_2$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i \mathbf{I}),\ \mathbf{z_n} \sim \mathcal{N}(0, \mathbf{I})$$



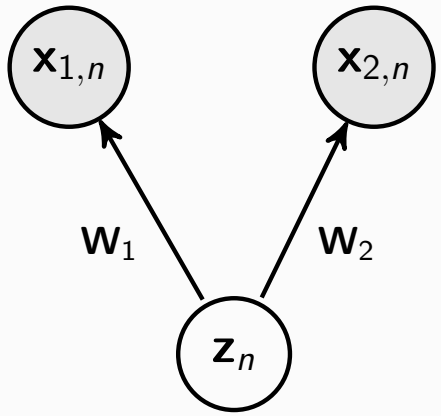
- The Maximum Likelihood (ML) parameter estimates of this model have been shown to be equivalent to deterministic CCA (Bach and Jordan, 2005).
- Note that in this generative model, the "shared-space" $\mathbf{Z}$ is now defined explicitly.

This is actually a special case of a model introduced in 1958, the Inter-battery Factor Analysis (IBFA).

A Probabilistic Interpretation of CCA

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{X}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T}\}$$
$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \epsilon_2$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i \mathbf{I}),\ \mathbf{z_n} \sim \mathcal{N}(0, \mathbf{I})$$



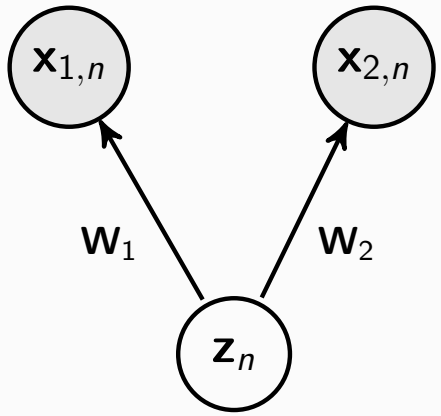
- The Maximum Likelihood (ML) parameter estimates of this model have been shown to be equivalent to deterministic CCA (Bach and Jordan, 2005).
- Note that in this generative model, the "shared-space" $\mathbf{Z}$ is now defined explicitly.

This is actually a special case of a model introduced in 1958, the Inter-battery Factor Analysis (IBFA).

A Probabilistic Interpretation of CCA

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{X}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T}\}$$
$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \epsilon_2$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i \mathbf{I}),\ \mathbf{z_n} \sim \mathcal{N}(0, \mathbf{I})$$



- The Maximum Likelihood (ML) parameter estimates of this model have been shown to be equivalent to deterministic CCA (Bach and Jordan, 2005).
- Note that in this generative model, the "shared-space" $\mathbf{Z}$ is now defined explicitly.

This is actually a special case of a model introduced in 1958, the Inter-battery Factor Analysis (IBFA).

Inter-Battery Factor Analysis (IBFA)

Battery (tests) refers to a series of psychological, behaviour or cognitive assessment tests. This term was often used in statistics since data from multiple batteries were essentially the one of the first datasets which consisted of multiple modalities.

- Many seminal works have been published in journals such as Psychometrika (devoted to the advancement of theory and methodology for behavioural data).
- A prominent example lies in Tucker's Inter-Battery Factor Analysis (IBFA), (1958, Psychometrika).
- CCA was introduced out of the increasing need for analysing multiple sets of data. It is considered the first "shared-space" model.
- Similarly, IBFA is the first model which extends shared-space models to the private-shared space paradigm.

Inter-Battery Factor Analysis (IBFA)

Battery (tests) refers to a series of psychological, behaviour or cognitive assessment tests. This term was often used in statistics since data from multiple batteries were essentially the one of the first datasets which consisted of multiple modalities.

- Many seminal works have been published in journals such as Psychometrika (devoted to the advancement of theory and methodology for behavioural data).
- A prominent example lies in Tucker's Inter-Battery Factor Analysis (IBFA), (1958, Psychometrika).
- CCA was introduced out of the increasing need for analysing multiple sets of data. It is considered the first "shared-space" model.
- Similarly, IBFA is the first model which extends shared-space models to the private-shared space paradigm.

Inter-Battery Factor Analysis (IBFA)

Battery (tests) refers to a series of psychological, behaviour or cognitive assessment tests. This term was often used in statistics since data from multiple batteries were essentially the one of the first datasets which consisted of multiple modalities.

- Many seminal works have been published in journals such as Psychometrika (devoted to the advancement of theory and methodology for behavioural data).
- A prominent example lies in Tucker's Inter-Battery Factor Analysis (IBFA), (1958, Psychometrika).
- CCA was introduced out of the increasing need for analysing multiple sets of data. It is considered the first "shared-space" model.
- Similarly, IBFA is the first model which extends shared-space models to the private-shared space paradigm.

Inter-Battery Factor Analysis (IBFA)

Battery (tests) refers to a series of psychological, behaviour or cognitive assessment tests. This term was often used in statistics since data from multiple batteries were essentially the one of the first datasets which consisted of multiple modalities.

- Many seminal works have been published in journals such as Psychometrika (devoted to the advancement of theory and methodology for behavioural data).
- A prominent example lies in Tucker's Inter-Battery Factor Analysis (IBFA), (1958, Psychometrika).
- CCA was introduced out of the increasing need for analysing multiple sets of data. It is considered the first "shared-space" model.
- Similarly, IBFA is the first model which extends shared-space models to the private-shared space paradigm.

Inter-Battery Factor Analysis (IBFA)

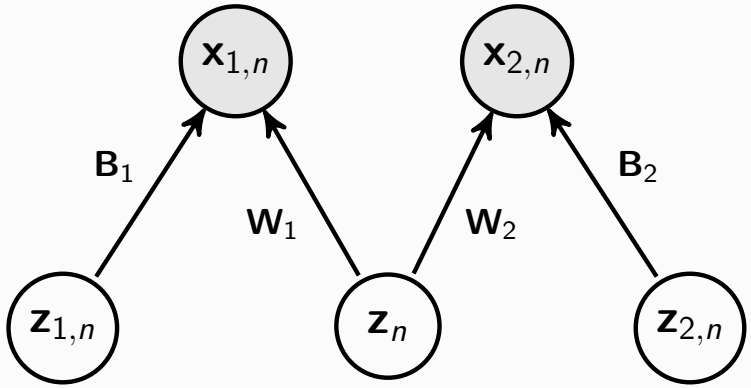
Battery (tests) refers to a series of psychological, behaviour or cognitive assessment tests. This term was often used in statistics since data from multiple batteries were essentially the one of the first datasets which consisted of multiple modalities.

- Many seminal works have been published in journals such as Psychometrika (devoted to the advancement of theory and methodology for behavioural data).
- A prominent example lies in Tucker's Inter-Battery Factor Analysis (IBFA), (1958, Psychometrika).
- CCA was introduced out of the increasing need for analysing multiple sets of data. It is considered the first "shared-space" model.
- Similarly, IBFA is the first model which extends shared-space models to the private-shared space paradigm.

Private-Shared Space Models

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \mathbf{B}_1 \mathbf{z}_{1,n} + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \mathbf{B}_2 \mathbf{z}_{2,n} + \epsilon_2$$
$$\mathbf{z_n}, \mathbf{z_1}, \mathbf{z_2} \sim \mathcal{N}(0, \mathbf{I})$$

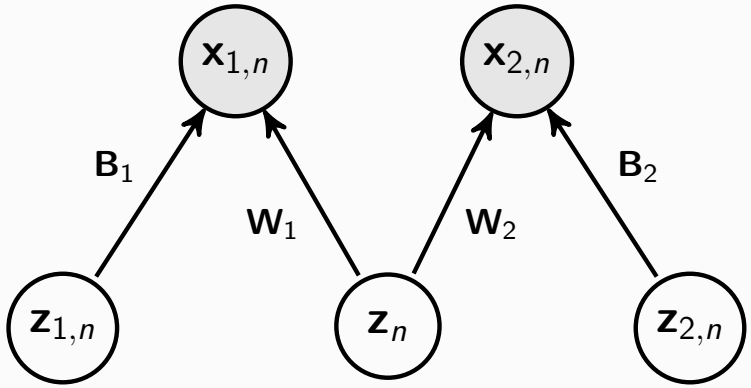


- Generative interpretation of CCA (Bach and Jordan, 2005) is essentially equivalent to a special case of the probabilistic interpretation of IBFA (Browne, 1979).
- Kaski and Klami (2008) re-introduce IBFA as Probabilistic CCA (PCCA), providing an EM algorithm. Terms have been used interchangeably (Kaski and Klami in JMLR, 2013).
- IBFA actually complements CCA by providing a description of the variation not captured by the correlating components.

Private-Shared Space Models

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \mathbf{B}_1 \mathbf{z}_{1,n} + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \mathbf{B}_2 \mathbf{z}_{2,n} + \epsilon_2$$
$$\mathbf{z_n}, \mathbf{z_1}, \mathbf{z_2} \sim \mathcal{N}(0, \mathbf{I})$$

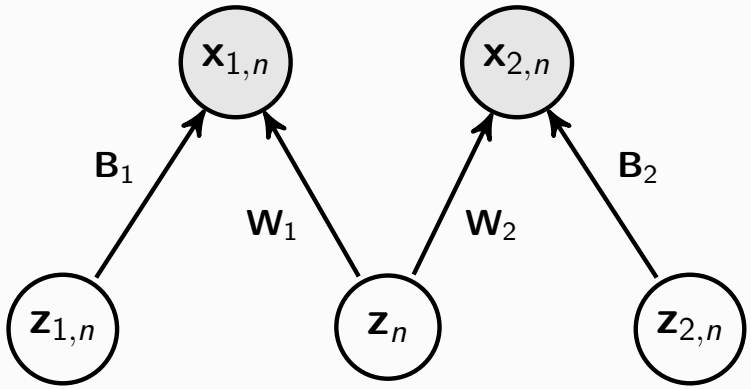


- Generative interpretation of CCA (Bach and Jordan, 2005) is essentially equivalent to a special case of the probabilistic interpretation of IBFA (Browne, 1979).
- Kaski and Klami (2008) re-introduce IBFA as Probabilistic CCA (PCCA), providing an EM algorithm. Terms have been used interchangeably (Kaski and Klami in JMLR, 2013).
- IBFA actually complements CCA by providing a description of the variation not captured by the correlating components.

Private-Shared Space Models

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \mathbf{B}_1 \mathbf{z}_{1,n} + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \mathbf{B}_2 \mathbf{z}_{2,n} + \epsilon_2$$
$$\mathbf{z_n}, \mathbf{z_1}, \mathbf{z_2} \sim \mathcal{N}(0, \mathbf{I})$$

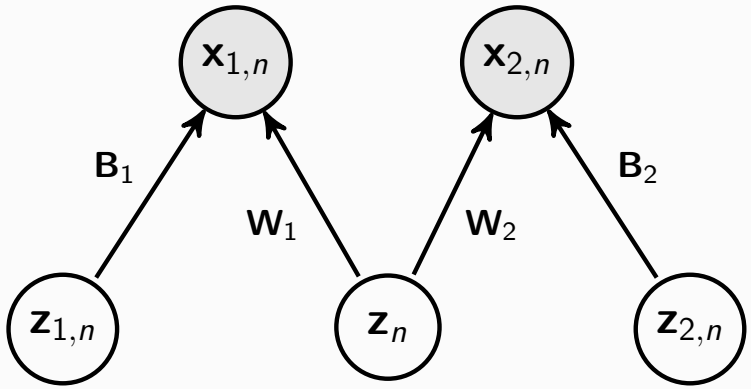


- Generative interpretation of CCA (Bach and Jordan, 2005) is essentially equivalent to a special case of the probabilistic interpretation of IBFA (Browne, 1979).
- Kaski and Klami (2008) re-introduce IBFA as Probabilistic CCA (PCCA), providing an EM algorithm. Terms have been used interchangeably (Kaski and Klami in JMLR, 2013).
- IBFA actually complements CCA by providing a description of the variation not captured by the correlating components.

Private-Shared Space Models

Given datasets $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times T}, \mathbf{X}_2 \in \mathbb{R}^{D_2 \times T}$, the gen. model is defined as:

$$\mathbf{x}_{1,n} = \mathbf{W}_1 \mathbf{z}_n + \mathbf{B}_1 \mathbf{z}_{1,n} + \epsilon_1$$
$$\mathbf{x}_{2,n} = \mathbf{W}_2 \mathbf{z}_n + \mathbf{B}_2 \mathbf{z}_{2,n} + \epsilon_2$$
$$\mathbf{z_n}, \mathbf{z_1}, \mathbf{z_2} \sim \mathcal{N}(0, \mathbf{I})$$



- Generative interpretation of CCA (Bach and Jordan, 2005) is essentially equivalent to a special case of the probabilistic interpretation of IBFA (Browne, 1979).
- Kaski and Klami (2008) re-introduce IBFA as Probabilistic CCA (PCCA), providing an EM algorithm. Terms have been used interchangeably (Kaski and Klami in JMLR, 2013).
- IBFA actually complements CCA by providing a description of the variation not captured by the correlating components.

An interesting observation...

Browne (1979) clarifies.

- The generative formulation maintains a single latent variable $\mathbf{z}$ that captures the shared variation, whereas classical CCA results in two separate but correlating variables obtained by projecting.

$\rightarrow$ i.e. $P(\mathbf{Z}|\mathbf{X}_1, \mathbf{X}_2)$, $P(\mathbf{Z}|\mathbf{X}_1)$ vs. $\mathbf{W}^T\mathbf{X}_1$.

Based on this, Kaski and Klami (2013) note.

- "the extended model provides novel application opportunities not immediately apparent in the more restricted CCA model"

An interesting observation...

Browne (1979) clarifies.

- The generative formulation maintains a single latent variable $\mathbf{z}$ that captures the shared variation, whereas classical CCA results in two separate but correlating variables obtained by projecting.

$\rightarrow$ i.e. $P(\mathbf{Z}|\mathbf{X}_1, \mathbf{X}_2)$, $P(\mathbf{Z}|\mathbf{X}_1)$ vs. $\mathbf{W}^T\mathbf{X}_1$.

Based on this, Kaski and Klami (2013) note.

- "the extended model provides novel application opportunities not immediately apparent in the more restricted CCA model"

So far...

- We have talked about CCA (shared space) and IBFA (private-shared space).
- We referred to their probabilistic interpretations.
- We have also seen how DTW can be elegantly combined with component analysis.
- In what follows, we present a case-study of a private-shared space model applied to the problem of fusing multiple subjective annotations.

So far...

- We have talked about CCA (shared space) and IBFA (private-shared space).
- We referred to their probabilistic interpretations.
- We have also seen how DTW can be elegantly combined with component analysis.
- In what follows, we present a case-study of a private-shared space model applied to the problem of fusing multiple subjective annotations.

So far...

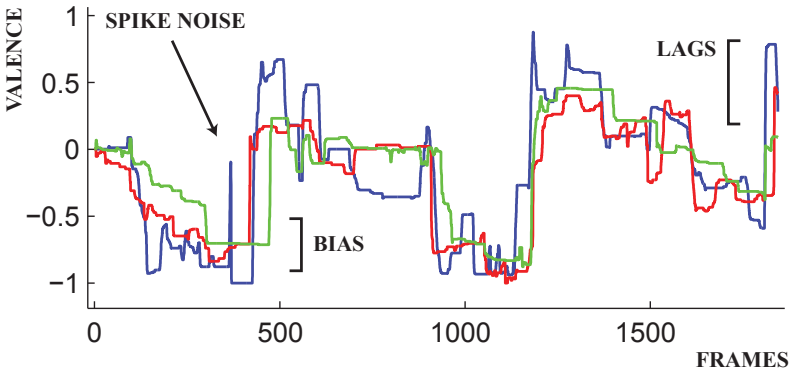
- We have talked about CCA (shared space) and IBFA (private-shared space).
- We referred to their probabilistic interpretations.
- We have also seen how DTW can be elegantly combined with component analysis.
- In what follows, we present a case-study of a private-shared space model applied to the problem of fusing multiple subjective annotations.

So far...

- We have talked about CCA (shared space) and IBFA (private-shared space).
- We referred to their probabilistic interpretations.
- We have also seen how DTW can be elegantly combined with component analysis.
- In what follows, we present a case-study of a private-shared space model applied to the problem of fusing multiple subjective annotations.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").

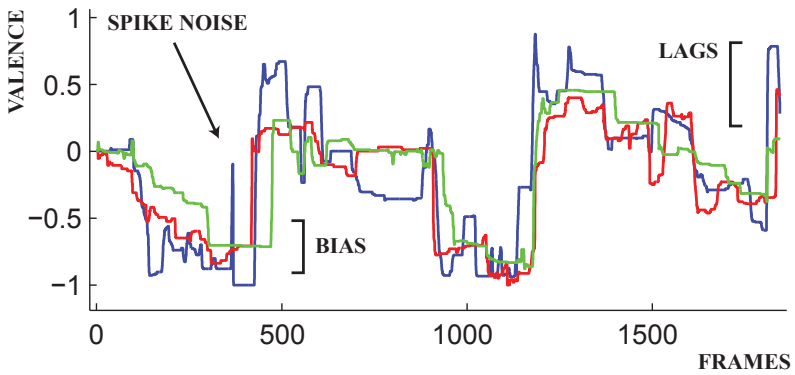


Challenges.

- Account for annotator-specific bias and noise
- Model the common signal shared by all annotators
- Account for temporal discrepancies amongst annot.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").

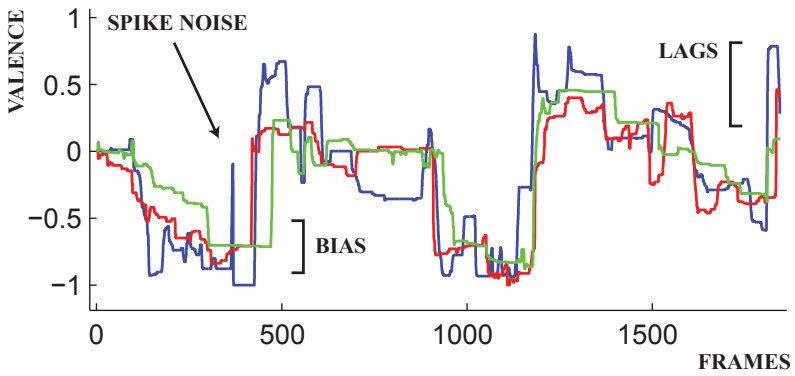


Challenges.

- Account for annotator-specific bias and noise
- Model the common signal shared by all annotators
- Account for temporal discrepancies amongst annot.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").

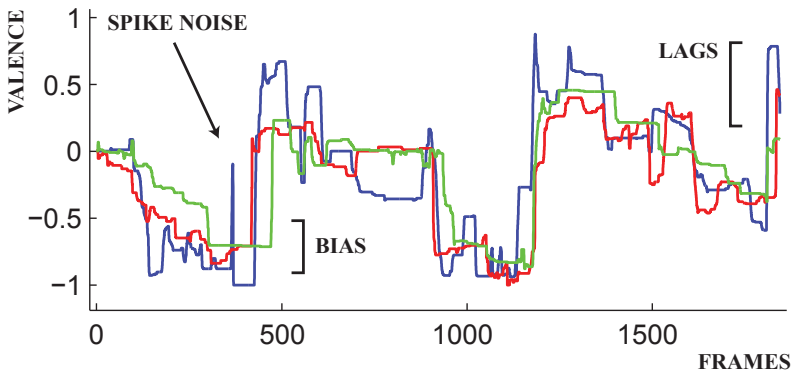


Challenges.

- Account for annotator-specific bias and noise
- Model the common signal shared by all annotators
- Account for temporal discrepancies amongst annot.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").

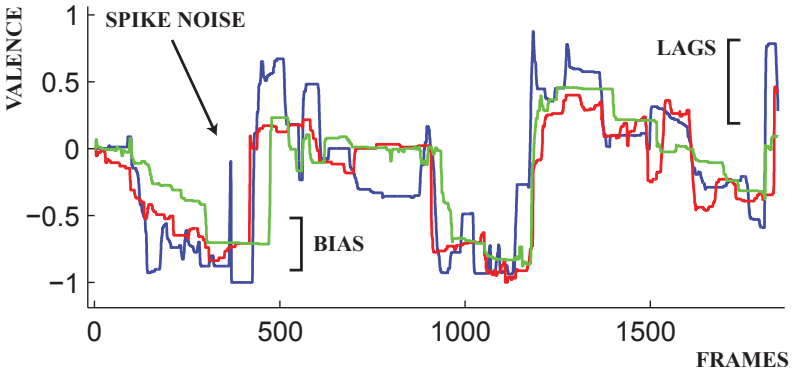


Challenges.

- Account for annotator-specific bias and noise
- Model the common signal shared by all annotators
- Account for temporal discrepancies amongst annot.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").

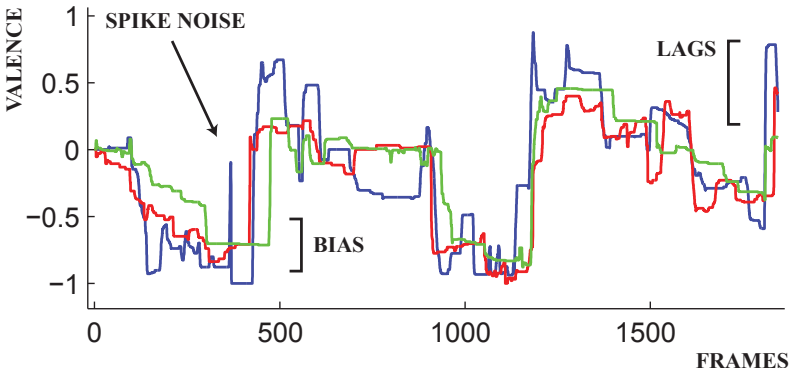


Challenges and Proposed Methodology (P/S).

- Account for annotator-specific bias and noise → private space
- Model the common signal shared by all annotators
- Account for temporal discrepancies amongst annot.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").

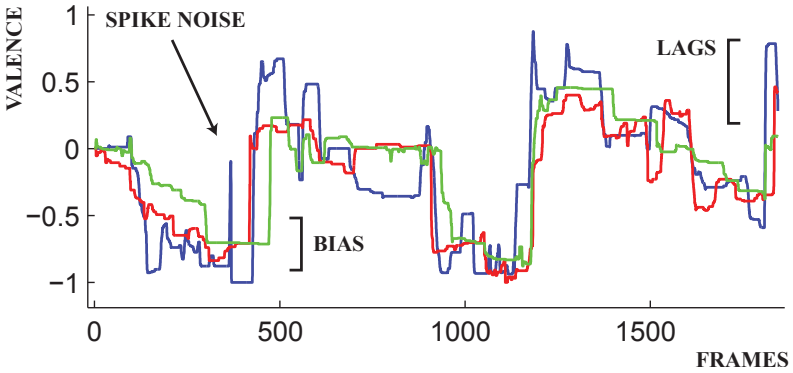


Challenges and Proposed Methodology (P/S).

- Account for annotator-specific bias and noise → private space
- Model the common signal shared by all annotators → shared space
- Account for temporal discrepancies amongst annot.

Continuous Subjective Annotations

Problem: Given multiple annotations (time-series) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, infer the common, consensus annotation (i.e., the "ground truth").



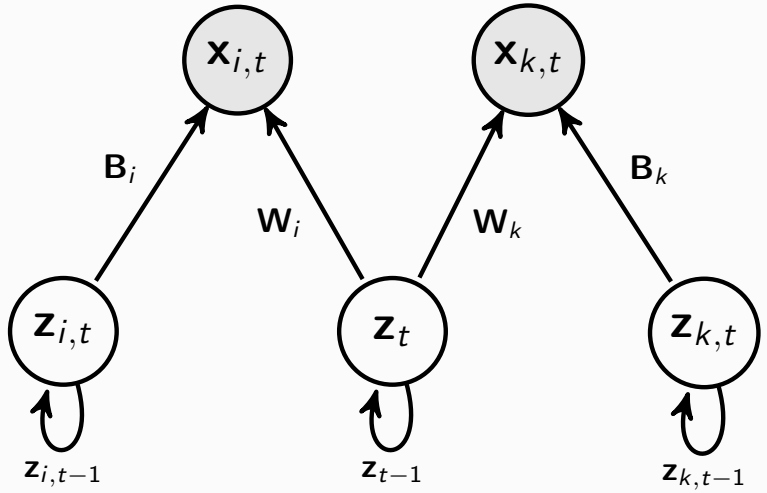
Challenges and Proposed Methodology (P/S).

- Account for annotator-specific bias and noise → private space
- Model the common signal shared by all annotators → shared space
- Account for temporal discrepancies amongst annot. → time warping

Dynamic Probabilistic CCA (DPCCA)

The generative model of DPCCA is as follows[1].

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I})$$
$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1}, \mathbf{V}_Z))$$
$$P(\mathbf{z}_{i,t}|\mathbf{z}_{i,t-1}) \sim \mathcal{N}(\mathbf{A}_{z_i}\mathbf{z}_{i,t-1}, \mathbf{V}_{Z_i})$$



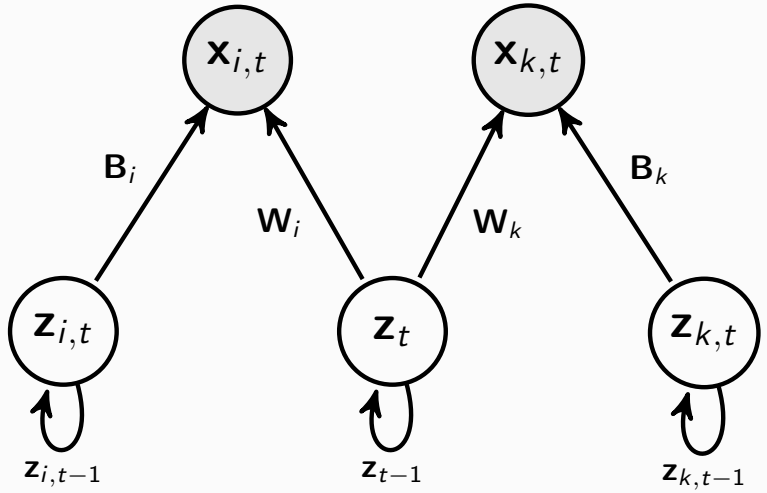
- DPCCA models temporal dynamics in both private and shared latent spaces by incorporating a linear dynamical system prior.
- First order moments attained by applying smoothing (RTS).
- Inherent model flexibility. Application-wise, translates to e.g., being able to condition the shared space (i.e. "ground-truth") on any subset of the available annotators, i.e. $P(\mathbf{Z}|\mathbf{X}_1)$, $P(\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N)$.

[1] Nicolaou et al. @ ECCV '12 & TPAMI '14

Dynamic Probabilistic CCA (DPCCA)

The generative model of DPCCA is as follows[1].

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I})$$
$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1}, \mathbf{V}_Z))$$
$$P(\mathbf{z}_{i,t}|\mathbf{z}_{i,t-1}) \sim \mathcal{N}(\mathbf{A}_{z_i}\mathbf{z}_{i,t-1}, \mathbf{V}_{Z_i})$$



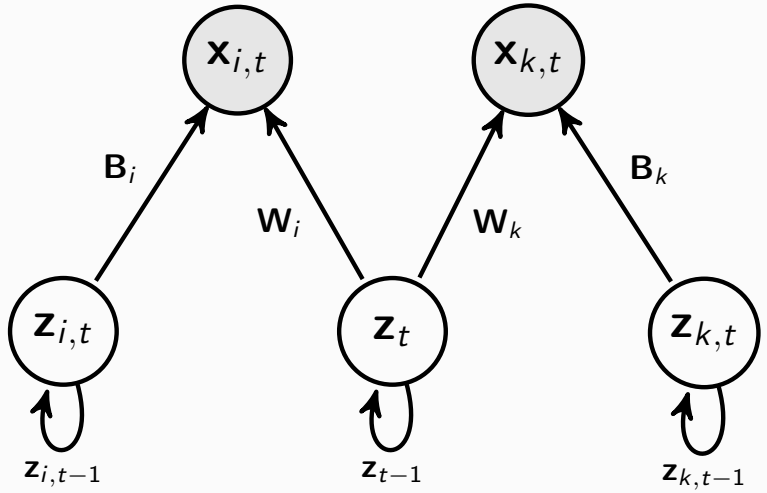
- DPCCA models temporal dynamics in both private and shared latent spaces by incorporating a linear dynamical system prior.
- First order moments attained by applying smoothing (RTS).
- Inherent model flexibility. Application-wise, translates to e.g., being able to condition the shared space (i.e. "ground-truth") on any subset of the available annotators, i.e. $P(\mathbf{Z}|\mathbf{X}_1)$, $P(\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N)$.

[1] Nicolaou et al. @ ECCV '12 & TPAMI '14

Dynamic Probabilistic CCA (DPCCA)

The generative model of DPCCA is as follows[1].

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I})$$
$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1}, \mathbf{V}_Z))$$
$$P(\mathbf{z}_{i,t}|\mathbf{z}_{i,t-1}) \sim \mathcal{N}(\mathbf{A}_{z_i}\mathbf{z}_{i,t-1}, \mathbf{V}_{Z_i})$$



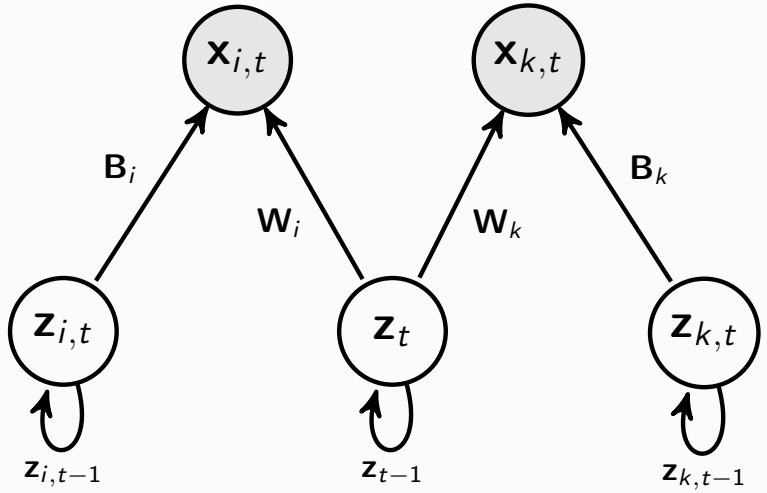
- DPCCA models temporal dynamics in both private and shared latent spaces by incorporating a linear dynamical system prior.
- First order moments attained by applying smoothing (RTS).
- Inherent model flexibility. Application-wise, translates to e.g., being able to condition the shared space (i.e. "ground-truth") on any subset of the available annotators, i.e. $P(\mathbf{Z}|\mathbf{X}_1)$, $P(\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N)$.

[1] Nicolaou et al. @ ECCV '12 & TPAMI '14

Dynamic Probabilistic CCA (DPCCA)

The generative model of DPCCA is as follows[1].

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I})$$
$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1}, \mathbf{V}_Z))$$
$$P(\mathbf{z}_{i,t}|\mathbf{z}_{i,t-1}) \sim \mathcal{N}(\mathbf{A}_{z_i}\mathbf{z}_{i,t-1}, \mathbf{V}_{Z_i})$$



- DPCCA models temporal dynamics in both private and shared latent spaces by incorporating a linear dynamical system prior.
- First order moments attained by applying smoothing (RTS).
- Inherent model flexibility. Application-wise, translates to e.g., being able to condition the shared space (i.e. "ground-truth") on any subset of the available annotators, i.e. $P(\mathbf{Z}|\mathbf{X}_1)$, $P(\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N)$.

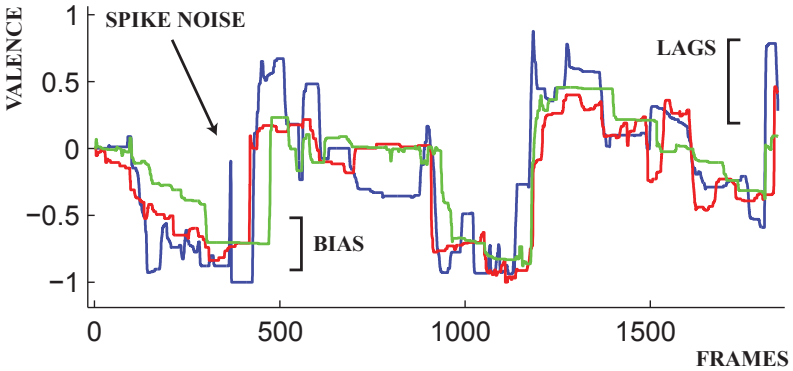
[1] Nicolaou et al. @ ECCV '12 & TPAMI '14

Dynamic Probabilistic CCA (DPCCA)

- DPCCA enables the inference of the individual, annotator-specific bias, model the annotation variance and discover the underlying, shared by all annotators signal, i.e. $\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N$ while modelling temporal dynamics.
- Nevertheless, temporal discrepancies in the annotations introduce noise in the derived spaces (e.g., lagging peaks will remain misaligned).

Dynamic Probabilistic CCA (DPCCA)

- DPCCA enables the inference of the individual, annotator-specific bias, model the annotation variance and discover the underlying, shared by all annotators signal, i.e. $\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N$ while modelling temporal dynamics.
- Nevertheless, temporal discrepancies in the annotations introduce noise in the derived spaces (e.g., lagging peaks will remain misaligned).

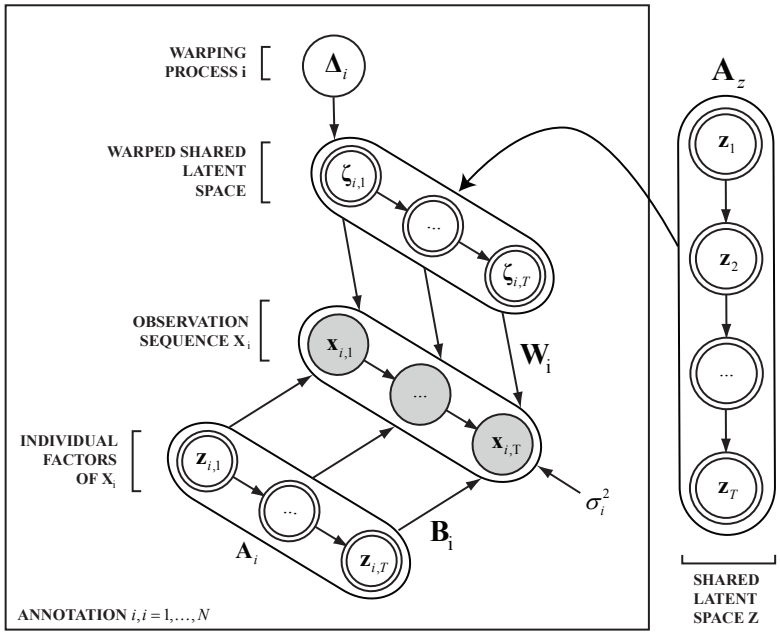


DPCCA with Time Warpings

Solution

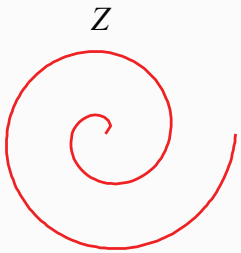
Incorporate warpings into DPCCA in order to eliminate temporal discrepancies in the "clean", shared space.

$$\mathcal{L}_{\text{(D)PCTW}} = \sum_{i}^{N} \sum_{j,j\neq i}^{N} \frac{||\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]\mathbf{\Delta}_i - \mathbb{E}[\mathbf{Z}|\mathbf{X}_j]\mathbf{\Delta}_j||_F^2}{N(N-1)}$$



DPCCA with Time Warpings: Experiment I

(a) 2D spiral.



(a)

(b)

(c)

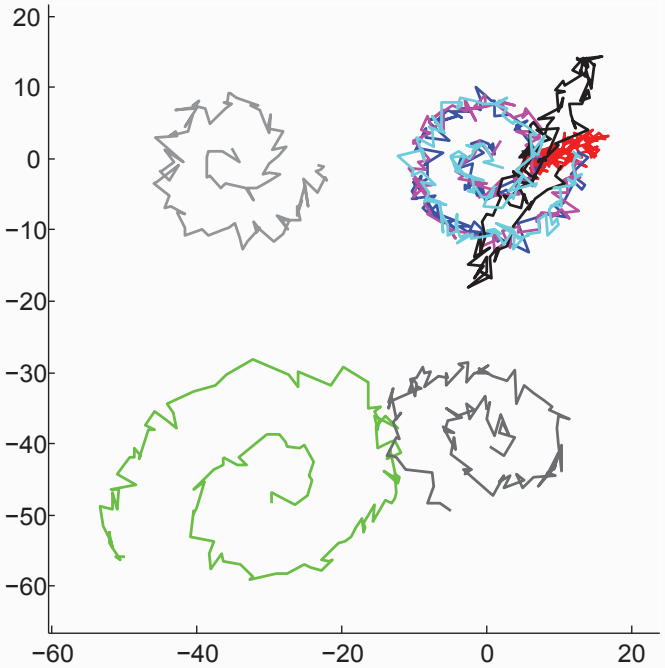
(d)

(e)

DPCCA with Time Warpings: Experiment I

(b) Noisy spirals generated from (a).

(a)



(b)

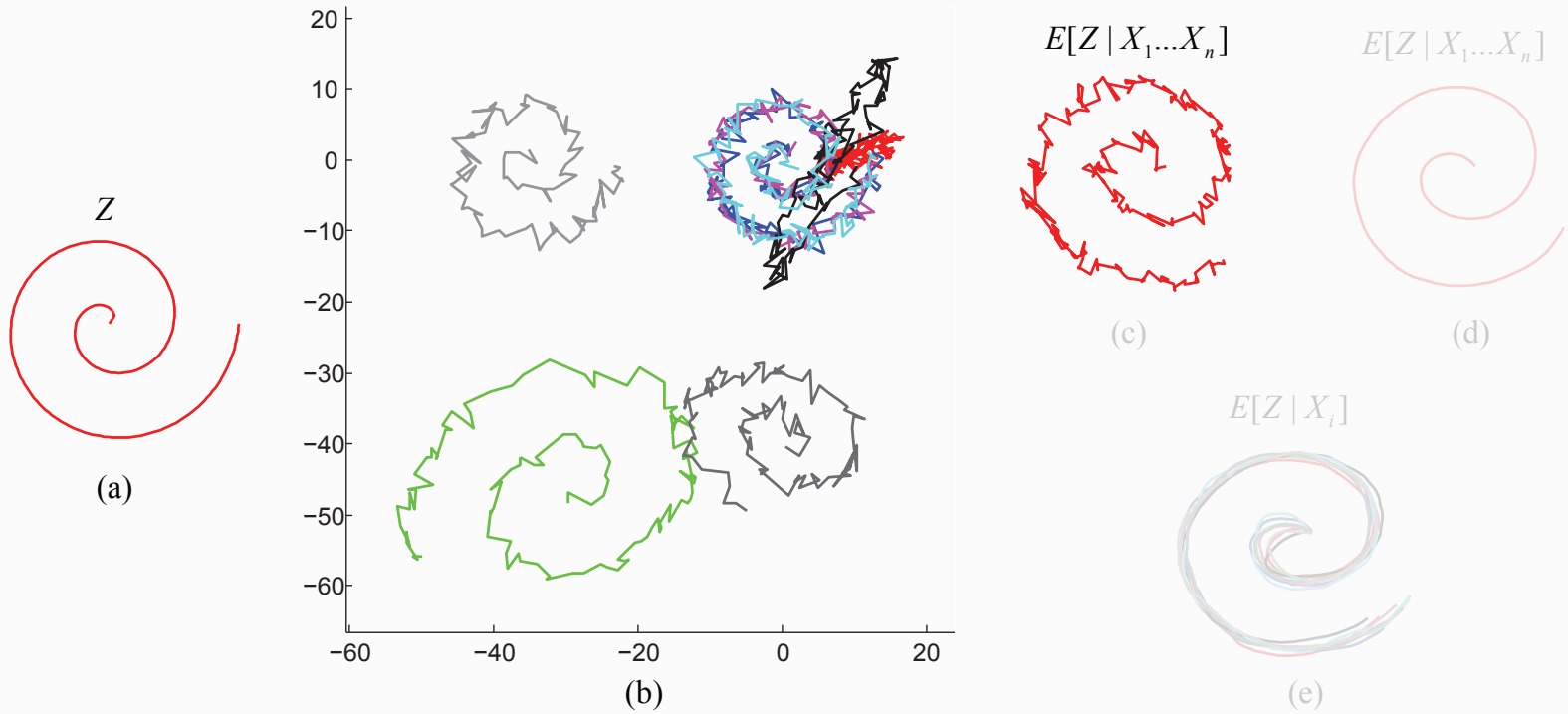
(c)

(d)

(e)

DPCCA with Time Warpings: Experiment I

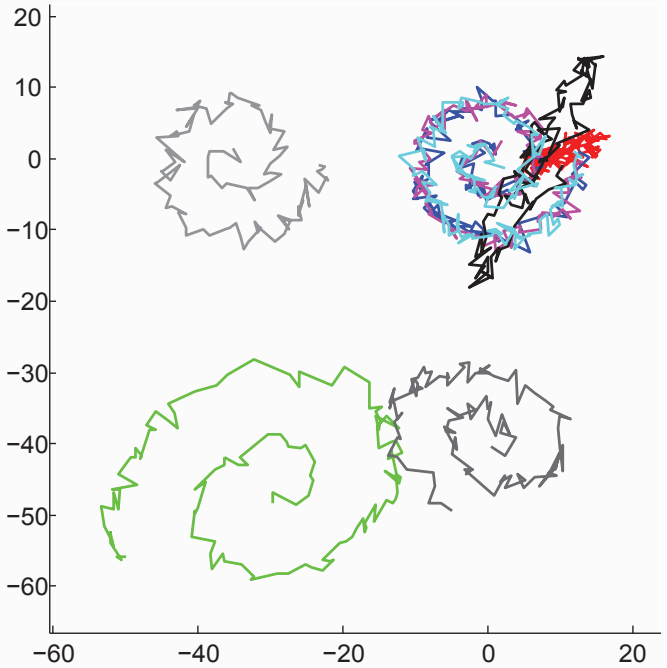
(c) Shared space given all annotations ("ground-truth", PCCA).



DPCCA with Time Warpings: Experiment I

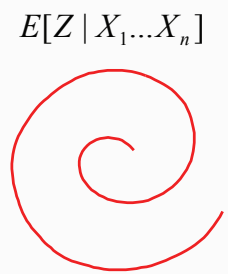
(d) Shared space given all annotations ("ground-truth", DPCCA).

(a)



(b)

(c)

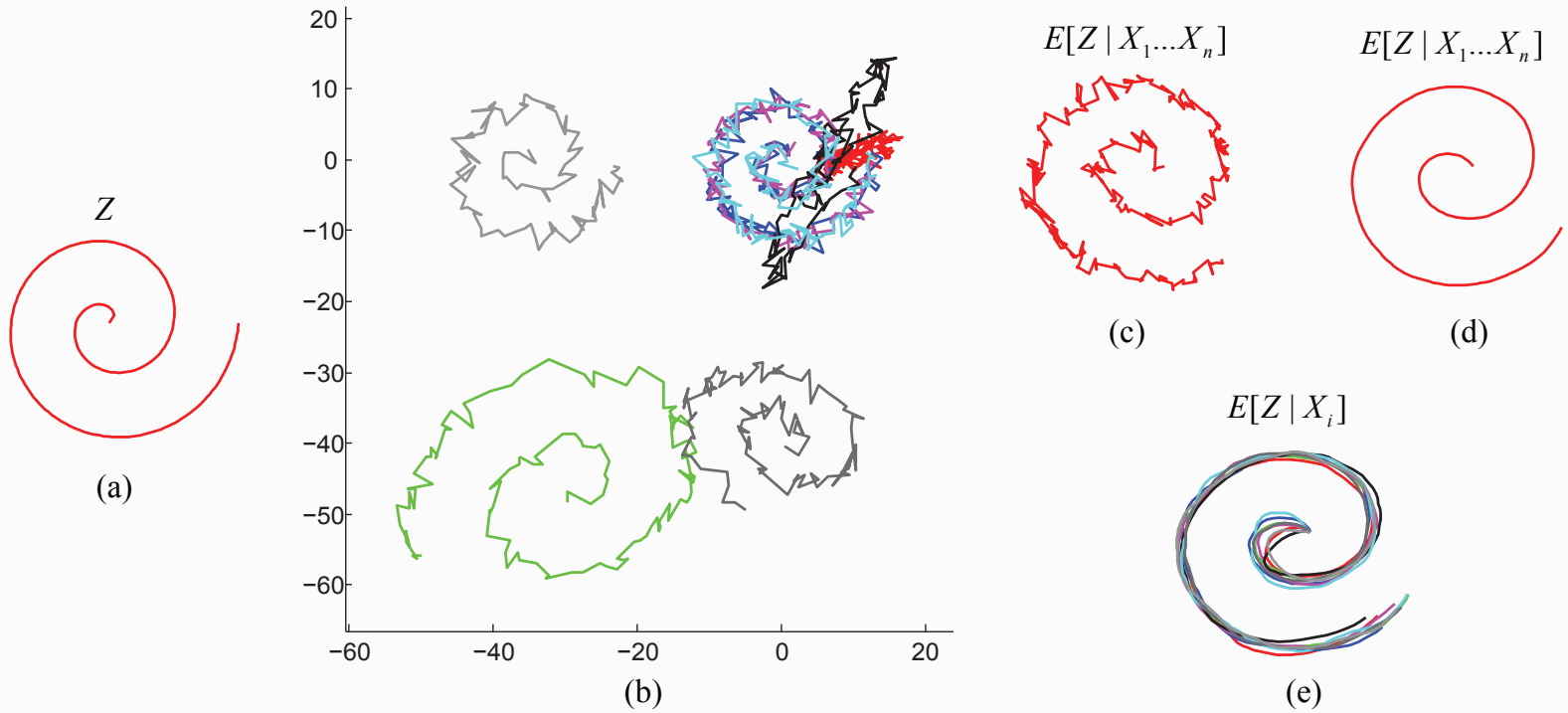


(d)

(e)

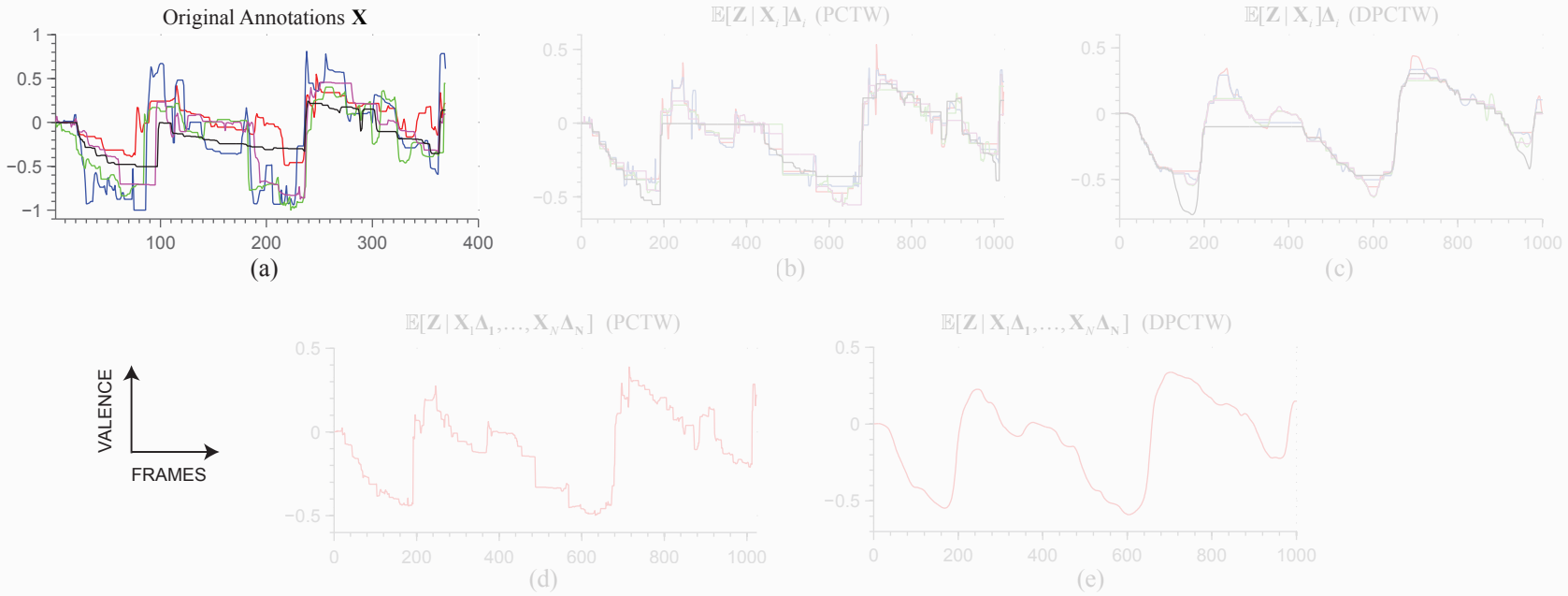
DPCCA with Time Warpings: Experiment I

(e) Aligned shared space given each annotation (DPCCA).



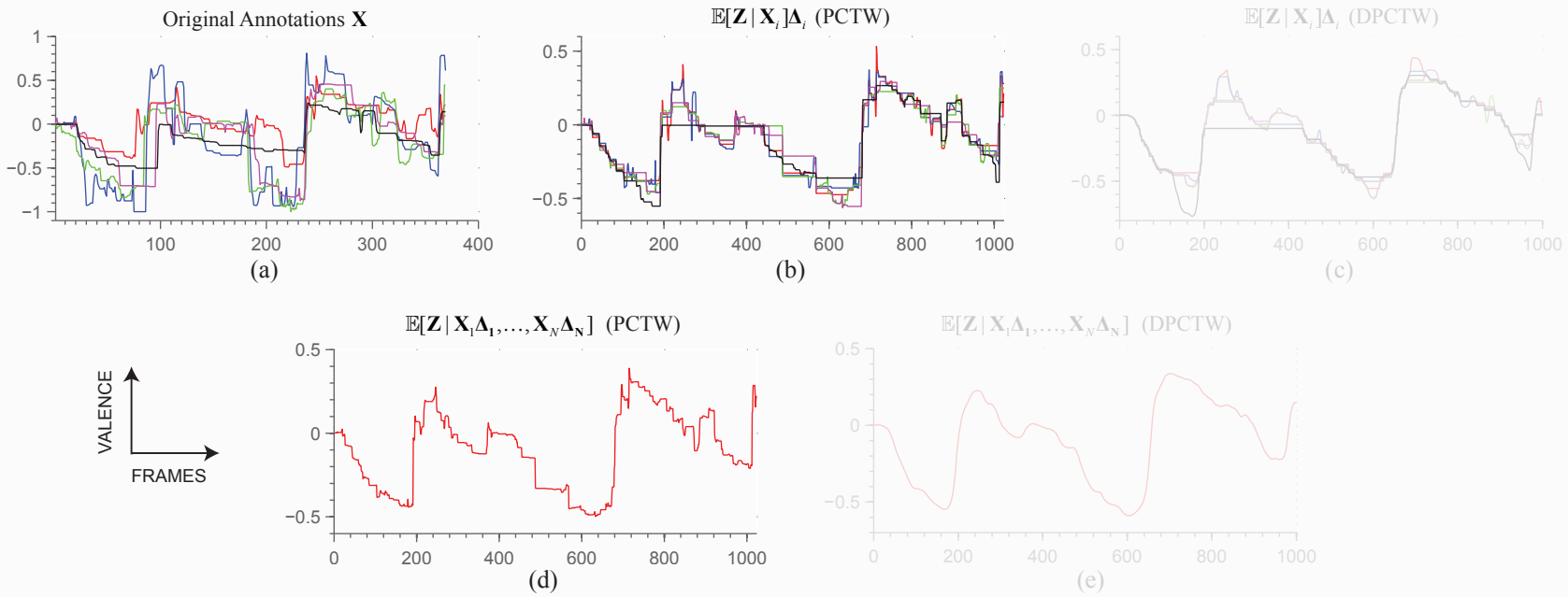
DPCCA with Time Warpings: Experiment II

DPCTW applied to continuous emotion annotations.



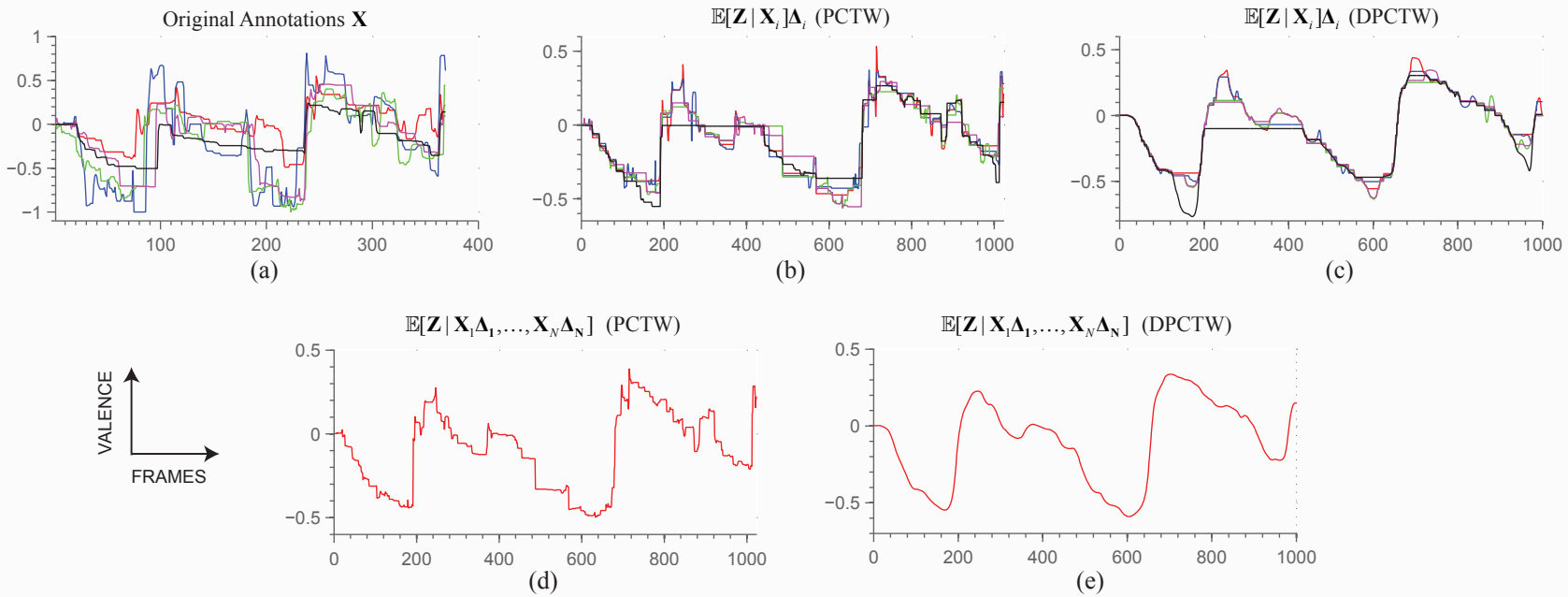
DPCCA with Time Warpings: Experiment II

DPCTW applied to continuous emotion annotations.



DPCCA with Time Warpings: Experiment II

DPCTW applied to continuous emotion annotations.



DPCCA and Annotator Ranking.

Aumann's agreement theorem

"If two people are Bayesian rationalists with common priors, and if they have common knowledge of their individual posteriors, then their posteriors must be equal".
- *Agreeing to disagree, RJ Aumann, Annals of Statistics 1976.*

- If annotators are rational in a Bayesian sense, the shared space posterior (common knowledge) conditioned on each annotator ($\mathbf{Z}|\mathbf{X}_i$) should be close to each other.
- We can rank the annotators based on the latent posterior by computing a probabilistic measure of difference (i.e. KL div.).
- Can detect "bad" annotators, e.g., adversarial or malicious, spammers etc.

DPCCA and Annotator Ranking.

Aumann's agreement theorem

"If two people are Bayesian rationalists with common priors, and if they have common knowledge of their individual posteriors, then their posteriors must be equal".
- *Agreeing to disagree, RJ Aumann, Annals of Statistics 1976.*

- If annotators are rational in a Bayesian sense, the shared space posterior (common knowledge) conditioned on each annotator $(\mathbf{Z}|\mathbf{X}_i)$ should be close to each other.
- We can rank the annotators based on the latent posterior by computing a probabilistic measure of difference (i.e. KL div.).
- Can detect "bad" annotators, e.g., adversarial or malicious, spammers etc.

DPCCA and Annotator Ranking.

Aumann's agreement theorem

"If two people are Bayesian rationalists with common priors, and if they have common knowledge of their individual posteriors, then their posteriors must be equal".
- *Agreeing to disagree, RJ Aumann, Annals of Statistics 1976.*

- If annotators are rational in a Bayesian sense, the shared space posterior (common knowledge) conditioned on each annotator $(\mathbf{Z}|\mathbf{X}_i)$ should be close to each other.
- We can rank the annotators based on the latent posterior by computing a probabilistic measure of difference (i.e. KL div.).
- Can detect "bad" annotators, e.g., adversarial or malicious, spammers etc.

DPCCA and Annotator Ranking.

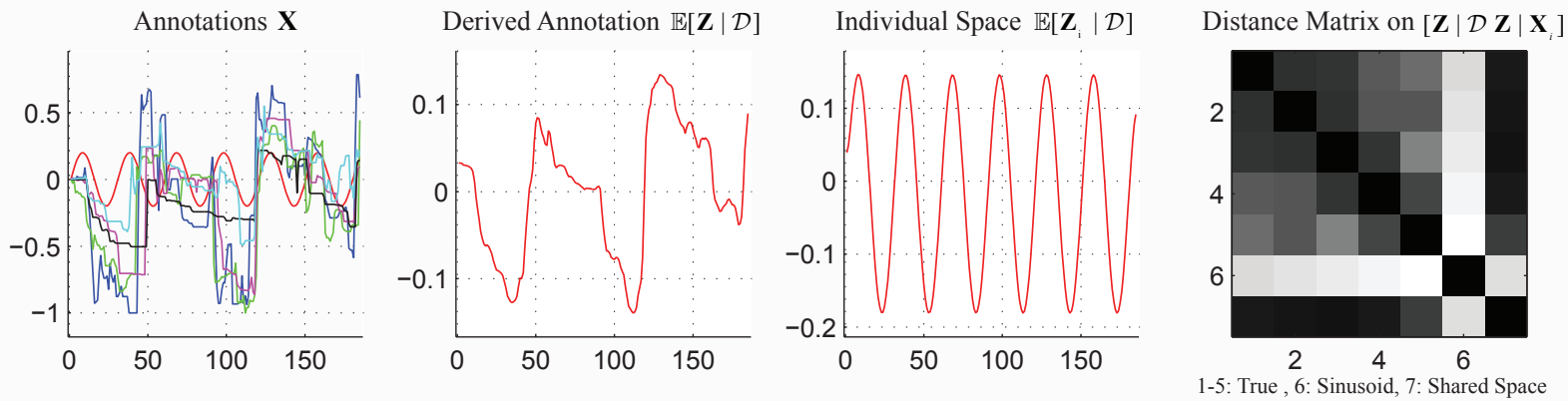
Aumann's agreement theorem

"If two people are Bayesian rationalists with common priors, and if they have common knowledge of their individual posteriors, then their posteriors must be equal".
- *Agreeing to disagree, RJ Aumann, Annals of Statistics 1976.*

- If annotators are rational in a Bayesian sense, the shared space posterior (common knowledge) conditioned on each annotator $(\mathbf{Z}|\mathbf{X}_i)$ should be close to each other.
- We can rank the annotators based on the latent posterior by computing a probabilistic measure of difference (i.e. KL div.).
- Can detect "bad" annotators, e.g., adversarial or malicious, spammers etc.

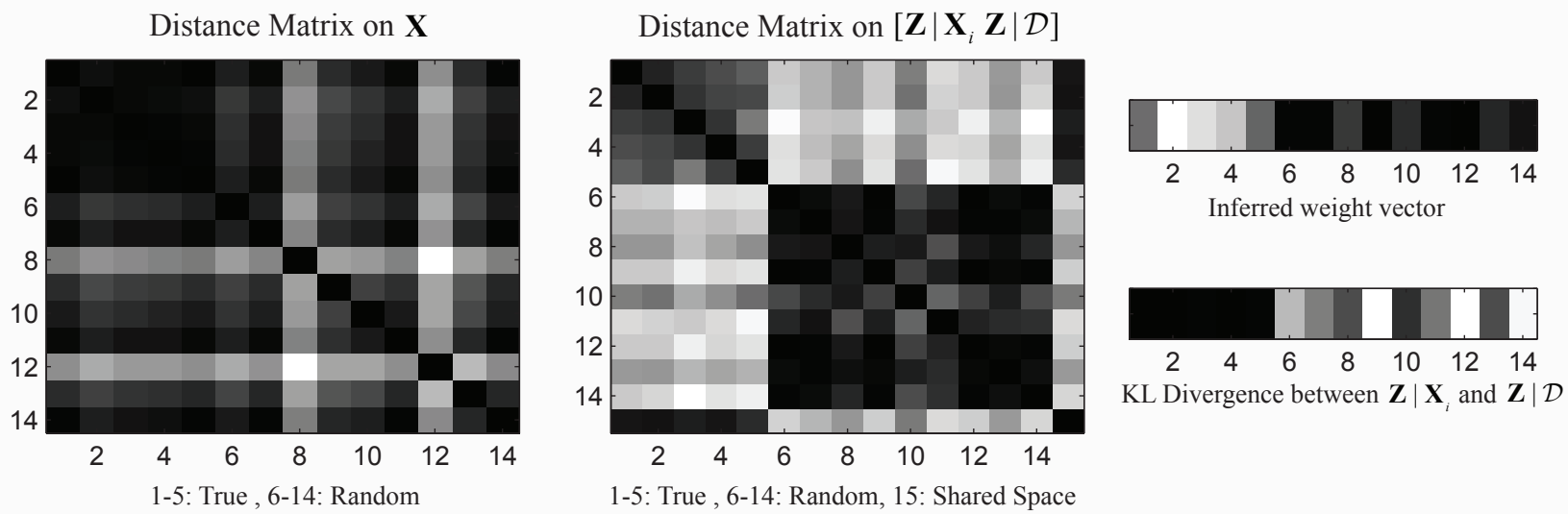
DPCCA and Annotator Ranking: Experiment

A random, structured annotation (sinusoid).



DPCCA and Annotator Ranking: Experiment

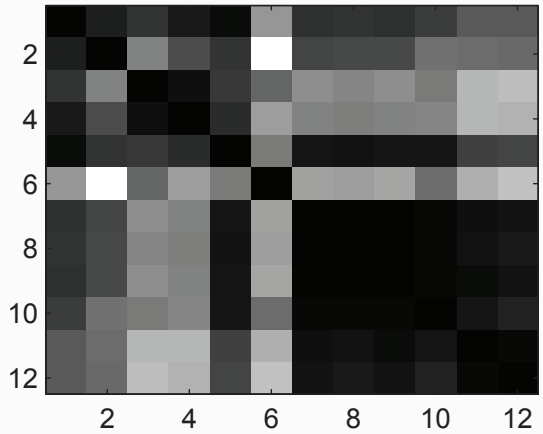
Set of random annotations.



DPCCA and Annotator Ranking: Experiment

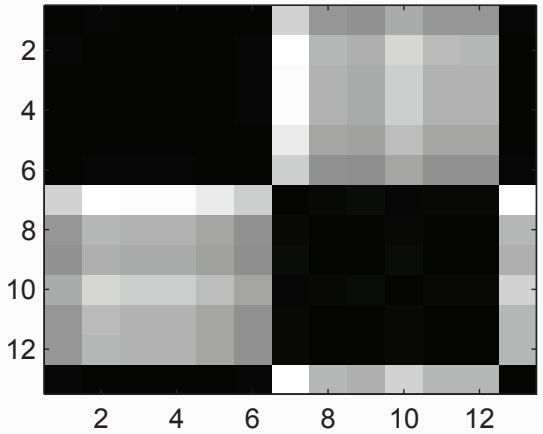
Two clusters of annotations with facial points as features.

Distance Matrix on $\mathbf{X}$



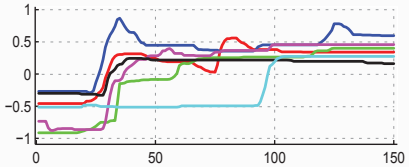
1-6: True, 7-12: Irrelevant

Distance Matrix on $[\mathbf{Z}|\mathbf{X}_i\ \mathbf{Z}|\mathbf{Y}]$

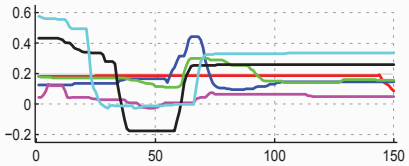


1-6: True, 7-12: Irrelevant, 13: Facial Points

Annotations $\mathbf{X}$



True Annotations (1-6)



Irrelevant Annotations (7-12)

Robust Canonical Correlation Analysis (RCCA)

RCCA is a robust-to-gross-noise variant of CCA. It is motivated by the wide presence of non-Gaussian noise in features extracted under unregulated, real-world conditions, e.g.,

→ in Visual Features: texture occlusions, tracking errors, spike noise.

→ in Audio Features: irrelevant (uncorrelated) sounds, equipment noise.

RCCA Formulation (1/2)

Given high-dimensional feature spaces $\mathbf{Z} \in \mathbb{R}^{dz \times T}$ and $\mathbf{A} \in \mathbb{R}^{da \times T}$ (representing e.g., video/audio cues), we pose the problem

$$\begin{aligned} &\underset{\mathbf{P}_z, \mathbf{P}_a, \mathbf{E}_z, \mathbf{E}_a}{\operatorname{argmin}} \quad \operatorname{rank}(\mathbf{P}_z) + \operatorname{rank}(\mathbf{P}_a) \\ &+\lambda_1 \|\mathbf{E}_z\|_0 + \lambda_2 \|\mathbf{E}_a\|_0 + \frac{\mu}{2} \|\mathbf{P}_z \mathbf{Z} - \mathbf{P}_a \mathbf{A}\|_F^2 \\ &\text{s.t.} \quad \mathbf{Z} = \mathbf{P}_z \mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a \mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{1}$$

$\mathbf{P}_z, \mathbf{P}_a$: Low-rank subspace spanning the correlated observations.

$\mathbf{E}_z, \mathbf{E}_a$: Capturing uncorrelated components, accounting for noise/outliers.

$\lambda_1, \lambda_2, \mu$: non-negative parameters.

$\rightarrow$ λ tunes the contribution of each signal to the clean space.

Problem (1) deemed difficult to solve due to the discrete nature of the rank function and the ℓ_0 norm (NP-Hard).

RCCA Formulation (1/2)

Given high-dimensional feature spaces $\mathbf{Z} \in \mathbb{R}^{dz \times T}$ and $\mathbf{A} \in \mathbb{R}^{da \times T}$ (representing e.g., video/audio cues), we pose the problem

$$\begin{aligned} \underset{\mathbf{P}_z, \mathbf{P}_a, \mathbf{E}_z, \mathbf{E}_a}{\operatorname{argmin}} \quad & \operatorname{rank}(\mathbf{P}_z) + \operatorname{rank}(\mathbf{P}_a) \\ & + \lambda_1 \|\mathbf{E}_z\|_0 + \lambda_2 \|\mathbf{E}_a\|_0 + \frac{\mu}{2} \overbrace{\|\mathbf{P}_z \mathbf{Z} - \mathbf{P}_a \mathbf{A}\|_F^2}^{\text{LS-CCA}} \\ \text{s.t.} \quad & \mathbf{Z} = \mathbf{P}_z \mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a \mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{1}$$

$\mathbf{P}_z, \mathbf{P}_a$: Low-rank subspace spanning the correlated observations.

$\mathbf{E}_z, \mathbf{E}_a$: Capturing uncorrelated components, accounting for noise/outliers.

$\lambda_1, \lambda_2, \mu$: non-negative parameters.

$\rightarrow$ λ tunes the contribution of each signal to the clean space.

Problem (1) deemed difficult to solve due to the discrete nature of the rank function and the ℓ_0 norm (NP-Hard).

RCCA Formulation (1/2)

Given high-dimensional feature spaces $\mathbf{Z} \in \mathbb{R}^{dz \times T}$ and $\mathbf{A} \in \mathbb{R}^{da \times T}$ (representing e.g., video/audio cues), we pose the problem

$$\begin{aligned} &\underset{\mathbf{P}_z, \mathbf{P}_a, \mathbf{E}_z, \mathbf{E}_a}{\operatorname{argmin}} \quad \operatorname{rank}(\mathbf{P}_z) + \operatorname{rank}(\mathbf{P}_a) \\ &+\lambda_1 \|\mathbf{E}_z\|_0 + \lambda_2 \|\mathbf{E}_a\|_0 + \frac{\mu}{2} \|\mathbf{P}_z \mathbf{Z} - \mathbf{P}_a \mathbf{A}\|_F^2 \\ &\text{s.t.} \;\; \mathbf{Z} = \mathbf{P}_z \mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a \mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{1}$$

$\mathbf{P}_z, \mathbf{P}_a$: Low-rank subspace spanning the correlated observations.

$\mathbf{E}_z, \mathbf{E}_a$: Capturing uncorrelated components, accounting for noise/outliers.

$\lambda_1, \lambda_2, \mu$: non-negative parameters.

$\rightarrow$ λ tunes the contribution of each signal to the clean space.

Problem (1) deemed difficult to solve due to the discrete nature of the rank function and the ℓ_0 norm (NP-Hard).

Part II: RCCA Formulation (2/2)

Solution: Adopt convex relaxations of (1) by replacing the ℓ_0 norm and rank function with their convex envelopes, the ℓ_1 and nuclear norm respectively, as follows

$$\begin{aligned} &\underset{\mathbf{P}_z,\mathbf{P}_a,\mathbf{E}_z,\mathbf{E}_a}{\operatorname{argmin}} \quad \|\mathbf{P}_z\|_* + \|\mathbf{P}_a\|_* \\ &+\lambda_1\|\mathbf{E}_z\|_1 + \lambda_2\|\mathbf{E}_a\|_1 + \frac{\mu}{2}\|\mathbf{P}_z\mathbf{Z} - \mathbf{P}_a\mathbf{A}\|_F^2 \\ &\text{s.t.} \;\; \mathbf{Z} = \mathbf{P}_z\mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a\mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{2}$$

$\rightarrow$ Problem (2) can be solved by employing the Linearized Alternating Directions Method (LADM), a variant of the Alternating Direction Augmented Lagrange Multiplier method.

Wrapping up

- There are many other private-shared space models, conceptually similar to IBFA, such as Common and Individual Features Analysis (COBE) and Joint and Individual Variation Explained (JIVE).
- The private (or individual) space can be extremely useful in many applications since it can be discriminative (e.g., face clustering). Prince, Elder et al. propose a PLDA (for inferences about identity, ICCV 2007, TPAMI 2012) based on a similar setting as IBFA.
- Partial CCA is another interesting variant: Eliminate the influence of a third variable, and subsequently project observations into maximally correlated space (c.f., Mukuta, ICML 2014).
- There are also several works on non-linear shared-space models based on GPs.

Wrapping up

- There are many other private-shared space models, conceptually similar to IBFA, such as Common and Individual Features Analysis (COBE) and Joint and Individual Variation Explained (JIVE).
- The private (or individual) space can be extremely useful in many applications since it can be discriminative (e.g., face clustering). Prince, Elder et al. propose a PLDA (for inferences about identity, ICCV 2007, TPAMI 2012) based on a similar setting as IBFA.
- Partial CCA is another interesting variant: Eliminate the influence of a third variable, and subsequently project observations into maximally correlated space (c.f., Mukuta, ICML 2014).
- There are also several works on non-linear shared-space models based on GPs.

Wrapping up

- There are many other private-shared space models, conceptually similar to IBFA, such as Common and Individual Features Analysis (COBE) and Joint and Individual Variation Explained (JIVE).
- The private (or individual) space can be extremely useful in many applications since it can be discriminative (e.g., face clustering). Prince, Elder et al. propose a PLDA (for inferences about identity, ICCV 2007, TPAMI 2012) based on a similar setting as IBFA.
- Partial CCA is another interesting variant: Eliminate the influence of a third variable, and subsequently project observations into maximally correlated space (c.f., Mukuta, ICML 2014).
- There are also several works on non-linear shared-space models based on GPs.

Wrapping up

- There are many other private-shared space models, conceptually similar to IBFA, such as Common and Individual Features Analysis (COBE) and Joint and Individual Variation Explained (JIVE).
- The private (or individual) space can be extremely useful in many applications since it can be discriminative (e.g., face clustering). Prince, Elder et al. propose a PLDA (for inferences about identity, ICCV 2007, TPAMI 2012) based on a similar setting as IBFA.
- Partial CCA is another interesting variant: Eliminate the influence of a third variable, and subsequently project observations into maximally correlated space (c.f., Mukuta, ICML 2014).
- There are also several works on non-linear shared-space models based on GPs.

Thank you.