Lecture 5
Supspace Tranformations
Eigendecompositions, kernel PCA and CCA

Pavel Laskov[1] Blaine Nelson[1]

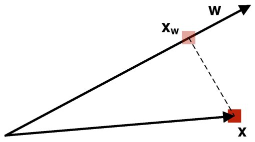
[1]Cognitive Systems Group
Wilhelm Schickard Institute for Computer Science
Universität Tübingen, Germany

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Advanced Topics in Machine Learning, 2012

Recall: Projections

- Projection of a point $\mathbf{x}$ onto a direction $\mathbf{w}$ is computed as:



$$\text{proj}_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}\frac{\mathbf{w}^\top \mathbf{x}}{\|\mathbf{w}\|^2}$$

- Directions in an RKHS expressed as linear combination of points:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i)$$

- The norm of the projection onto $\mathbf{w}$ thus can be expressed as

$$\|\text{proj}_{\mathbf{w}}(\mathbf{x})\| = \frac{w^\top x}{\|\mathbf{w}\|} = \frac{\sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})}{\sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j)}} = \sum_{i=1}^{N} \beta_i \kappa(\mathbf{x}_i, \mathbf{x})$$

Thus, the *size* of the projection onto $\mathbf{w}$ can be expressed as a linear combination of the kernel valuations with $\mathbf{x}$

Recall: Fisher/Linear Discriminant Analysis (LDA)

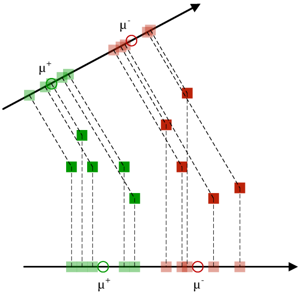
- In LDA, we chose a projection direction **w** to maximize the cost function

$$J(\mathbf{w}) = \frac{\left\|\mu_{\mathbf{w}}^{+} - \mu_{\mathbf{w}}^{-}\right\|^2}{(\sigma_{\mathbf{w}}^{+})^2 + (\sigma_{\mathbf{w}}^{-})^2} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T (S_W^{+} + S_W^{-})\mathbf{w}}$$

  where μ^+ & μ^- are the averages of the sets, σ^+ & σ^- are their standard deviations, $\mathbf{S}_B$ is the between scatter matrix & $\mathbf{S}_W^+$ and $\mathbf{S}_W^-$ are the within scatter matrices

- The optimal solution $\mathbf{w}^*$ is given by the first eigenvector of the matrix

$$(\mathbf{S}_W^+ + \mathbf{S}_W^-)^{-1}\mathbf{S}_B$$



Recall: Kernel LDA

- When the projection direction is in feature space, $\mathbf{w}_{\boldsymbol{\alpha}} = \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i)$
- From this, the LDA objective can be expressed as

$$\max_{\boldsymbol{\alpha}} \; J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \mathbf{N} \boldsymbol{\alpha}}$$

where

$$\mathbf{M} = (\mathbf{K}_+ - \mathbf{K}_-)\mathbf{1}_N \mathbf{1}_N^\top (\mathbf{K}_+ - \mathbf{K}_-)$$

$$\mathbf{N} = \mathbf{K}_+ \left(\mathbf{I}_{N^+} - \tfrac{1}{N^+} \mathbf{1}_{N^+} \mathbf{1}_{N^+}^\top \right) \mathbf{K}_+^\top + \mathbf{K}_- \left(\mathbf{I}_{N^-} - \tfrac{1}{N^-} \mathbf{1}_{N^-} \mathbf{1}_{N^-}^\top \right) \mathbf{K}_-^\top$$

- Solutions $\boldsymbol{\alpha}^*$ to the above generalized eigenvalue problem (as discussed later) allow us to project data onto this discriminant direction as

$$\|\mathrm{proj}_{\mathbf{w}}(\mathbf{x})\| = \sum_{i=1}^{N} \alpha_i^* \kappa(\mathbf{x}_i, \mathbf{x})$$

General Subspace Learning & Projections

- **Objective**: find a subspace that captures an important aspect of the training data…we find K axes that span this subspace
- **General Problem**: we will solve problems

$$\max_{g(\mathbf{w})=1} f(\mathbf{w})$$

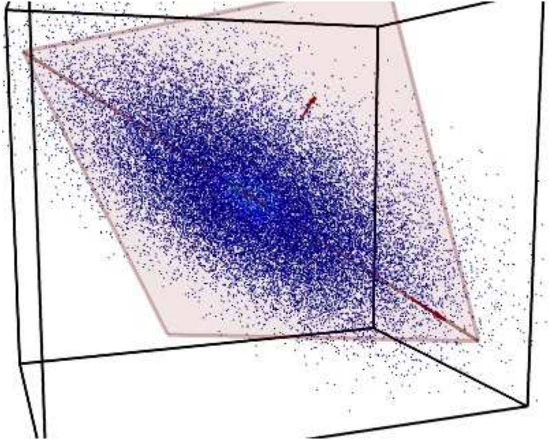
  for projection direction $\mathbf{w}$…iteratively solving these problems will yield a subspace defined by $\{\mathbf{w}_k\}_{k=1}^{K}$
- **General Approach**: find a center $\boldsymbol{\mu}$ and a set of K orthonormal directions $\{\mathbf{w}_k\}_{k=1}^{K}$ used to project data into the subspace:

$$\tilde{\mathbf{x}} \leftarrow \left(\mathbf{w}_k{}^{\top}(\mathbf{x}-\boldsymbol{\mu})\right)_{k=1}^{K}$$

  - This is a K-dimensional representation of the data *regardless* of the original space's dimensionality—the coordinates in the space spanned by $\{\mathbf{w}_k\}_{k=1}^{K}$
  - This projection will be centered at $\mathbf{0}$ (in feature space)

Subspace Learning

We want to find subspace that captures important aspects of our data



Overview

- LDA found 1 direction for discriminating between 2 classes
- In this lecture, we will see 3 subspace projection objectives / techniques:
  - Find directions that maximize variance in X (PCA)
  - Find directions that maximize covariance between X & Y (MCA)
  - Find directions that maximize correlation X & Y (CCA)
- These techniques extract underlying structure from the data allowing us to...
  - Capture fundamental structure of the data
  - Represent the data in low dimensions
- Each of these techniques can be kernelized to operate in a feature space yielding kernelized projections onto $\mathbf{w}$:

$$\|\mathrm{proj}_{\mathbf{w}}(\phi(\mathbf{x}))\| = \mathbf{w}^\top \phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) \quad (1)$$

where α is the vector of dual values defining $\mathbf{w}$

Part I

Principal Component Analysis

Motivation: Directions of Variance

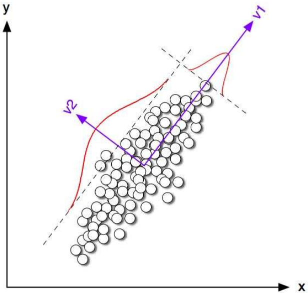
- We want to find a direction $\mathbf{w}$ that maximizes the data's variance
- Consider a random variable $\mathbf{x} \sim P_{\mathcal{X}}$ (Assume $\mathbf{0}$-mean). The variance of its projection onto (normalized) $\mathbf{w}$ is

$$\mathrm{E}_{\mathbf{x}\sim\mathcal{X}}\left[\mathrm{proj}_{\mathbf{w}}(\mathbf{x})^2\right] = \mathrm{E}\left[\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}\right] = \mathbf{w}^\top \underbrace{\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]}_{\mathbf{C}_{xx}} \mathbf{w} = \mathbf{w}^\top\mathbf{C}_{xx}\mathbf{w}$$

- In input space $\mathcal{X}$, the empirical covariance matrix (of centered data) is

$$\hat{\mathbf{C}}_{\mathbf{x},\mathbf{x}} = \frac{1}{N}\mathbf{X}^\top\mathbf{X} \ ;$$

  an $D \times D$ matrix
- How can we find directions that maximize $\mathbf{w}^\top\mathbf{C}_{xx}\mathbf{w}$? How can we kernelize it?



Recall: Eigenvalues & Eigenvectors

- Given an $N \times N$ matrix $\mathbf{A}$, an eigenvector of $\mathbf{A}$ is a *non-trivial* vector $\mathbf{v}$ that satisfies $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$; the corresponding value λ is an eigenvalue
- Eigen-values/vector pairs satisfy Rayleigh quotients:

$$\lambda = \frac{\mathbf{v}^\top \mathbf{A}\mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \qquad \lambda_1 = \max_{\|\mathbf{x}\|=1} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

- Eigen-vectors/values form orthonormal matrix $\mathbf{V}$ & diagonal matrix $\mathbf{\Lambda}$

$$\mathbf{V} = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_N \\ | & | & & | \end{bmatrix} \qquad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1(\mathbf{A}) & 0 & \ldots & 0 \\ 0 & \lambda_2(\mathbf{A}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_N(\mathbf{A}) \end{bmatrix}$$

  which form the eigen-decomposition of $\mathbf{A}$: $\quad \mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$
- Deflation: for any eigen-value/vector pair $(\lambda, \mathbf{v})$ of $\mathbf{A}$, the transform

$$\tilde{\mathbf{A}} \leftarrow \mathbf{A} - \lambda\mathbf{v}\mathbf{v}^\top$$

  deflates the matrix; *i.e.*, $\mathbf{v}$ is an eigenvector of $\tilde{\mathbf{A}}$ but has eigenvalue 0

Principle Components Analysis (PCA)

- Principle Components Analysis (PCA) - algorithm for finding the principle axes of a dataset
- PCA finds subspace spanned by $\{\mathbf{u}_i\}$ that maximizes the data's variance:

$$\mathbf{u}_1 = \operatorname*{argmax}_{\|\mathbf{w}\|=1} \mathbf{w}^\top \mathbf{C}_{xx} \mathbf{w} \qquad\qquad \mathbf{C}_{xx} = \frac{1}{N}\mathbf{X}^\top \mathbf{X}$$

- This is achieved by computing $\mathbf{C}_{xx}$'s eigenvectors
  1. Compute the data's mean: $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = \frac{1}{N}\mathbf{X}^\top \mathbf{1}_N$
  2. Compute the data's covariance: $\mathbf{C}_{xx} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$
  3. Find its principle axes: $[\mathbf{U}, \boldsymbol{\Lambda}] = eig(\mathbf{C}_{xx})$
  4. Project data $\{\mathbf{x}_i\}$ onto the first K eigenvectors: $\tilde{\mathbf{x}}_i \leftarrow \mathbf{U}_{1:K}^\top (\mathbf{x}_i - \boldsymbol{\mu})$

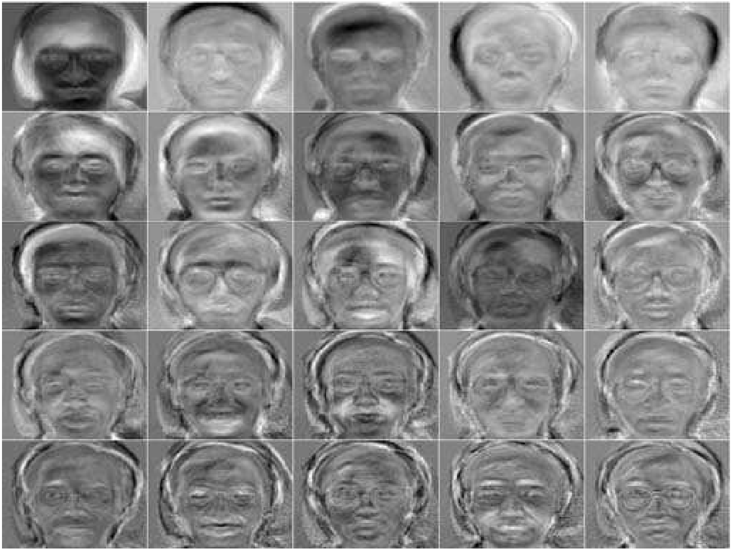
Properties of PCA

- Directions found by PCA are orthonormal: $\mathbf{u}_i{}^\top \mathbf{u}_j = \delta_{i,j}$
- When projected onto the space spanned by $\{\mathbf{u}_i\}$, resulting data has diagonal covariance matrix
- The eigenvalues λ_i are the amount of variance captured by the direction $\mathbf{u}_i$
- Variance captured by 1^{st} K directions is $\sum_{i=1}^{K} \lambda_i \left(\mathbf{C}_{xx}\right)$
- Using all directions, we can completely reconstruct the data in an alternative basis.
- Directions with low eigenvalues $\lambda_i \ll \lambda_1$ correspond to irrelevant aspects of data. . . often we use top K directions to re-represent the data.

Applications of PCA

- **Denoising/Compression**: PCA removes the $(D-K)$-dimensional subspace with the least information. The PCA transform thus retains the most salient information about the data.
- **Correction**: Reconstruction of data that has been damaged or has missing elements
- **Visualization**: The PCA transform produces a small dimensional projection of data which is convenient for visualizing high dimensional datasets
- **Document Analysis**: PCA can be used to find common themes in a set of documents

Application: Eigenfaces for Face Recognition [1]



Application: Eigenfaces for Face Recognition [1]

Part II

Kernel PCA

Kernelizing PCA

- PCA works in the primal space, but not all data structure is well-captured by these linear projections
- How can we kernelize PCA?

Singular Value Decomposition I

- Suppose $\mathbf{X}$ is any $N \times D$ matrix
- The eigen-decomposition of PSD matrices $\mathbf{C}_{xx} = \mathbf{X}^\top \mathbf{X}$ & $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ are

$$\mathbf{C}_{xx} = \mathbf{U}\mathbf{\Lambda}_D\mathbf{U}^\top \qquad\qquad \mathbf{K} = \mathbf{V}\mathbf{\Lambda}_N\mathbf{V}^\top$$

  where $\mathbf{U}$ & $\mathbf{V}$ are orthogonal and $\mathbf{\Lambda}_D$ & $\mathbf{\Lambda}_N$ have the eigenvalues
- Consider any eigen-pair $(\lambda, \mathbf{v})$ of $\mathbf{K}$... then $\mathbf{X}^\top\mathbf{v}$ is an eigenvector of $\mathbf{C}_{xx}$:

$$\mathbf{C}_{xx}\mathbf{X}^\top\mathbf{v} = \mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{v} = \mathbf{X}^\top\mathbf{K}\mathbf{v} = \lambda\mathbf{X}^\top\mathbf{v}$$

  and $\left\|\mathbf{X}^\top\mathbf{v}\right\| = \sqrt{\lambda}$. Thus there is an eigenvector of $\mathbf{C}_{xx}$ such that $\mathbf{u} = \frac{1}{\sqrt{\lambda}}\mathbf{X}^\top\mathbf{v}$
- In fact, we have the following correspondences:

$$\mathbf{u} = \lambda^{-1/2}\mathbf{X}^\top\mathbf{v} \qquad\qquad \mathbf{v} = \lambda^{-1/2}\mathbf{X}\mathbf{v}$$

Singular Value Decomposition II

- Further, let $t = rank\,(\mathbf{X}) \leq \min\,[D, N]$. It can be shown that

$$rank\,(\mathbf{C}_{xx}) = rank\,(\mathbf{K}) = t$$

- The singular value decomposition (SVD) of non-square $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top$$

  where $\mathbf{U}$ is $D \times D$ & orthogonal, $\mathbf{V}$ is $N \times N$ & orthogonal, and $\boldsymbol{\Sigma}$ is $N \times D$ with diagonal given by values $\sigma_i = \sqrt{\lambda_i}$
- The SVD is an analog of eigen-decomposition for non-square matrices.
  - $\mathbf{X}$ is non-singular iff all its singular values are non-zero
  - It yields a spectral decomposition:

$$\mathbf{X} = \sum_i \sigma_i \mathbf{v}_i \mathbf{u}_i^\top$$

  - Matrix-vector multiply $\mathbf{Xw}$ can be viewed as first projecting $\mathbf{w}$ into eigen-space $\{\mathbf{u}_i\}$ of $\mathbf{X}$, deforming according to its singular values σ_i and reprojecting into N-space using $\{\mathbf{v}_i\}$

Covariance & Kernel Matrix Duality

- The SVD decomposition of $\mathbf{X}$ showed a duality in eigenvectors of $\mathbf{C}_{xx}$ and $\mathbf{K}$ that allows us to *kernelize* it
- If $\mathbf{u}_j$ is the j^{th} eigenvector of $\mathbf{C}_{xx}$, then

$$\mathbf{u}_j = \lambda_j^{-1/2} \mathbf{X}^\top \mathbf{v}_j = \lambda_j^{-1/2} \sum_{i=1}^{N} \mathbf{X}_{i,\bullet} v_{j,i}$$

  i.e., a linear combination of the data points
- Replacing $\mathbf{X}_{i,\bullet}$ with $\phi(\mathbf{x}_i)$, the eigenvector $\mathbf{u}_j$ in feature space is

$$\mathbf{u}_j = \lambda_j^{-1/2} \sum_{i=1}^{N} v_{j,i} \phi(\mathbf{x}_i) = \sum_{i=1}^{N} \alpha_{j,i} \phi(\mathbf{x}_i)$$
$$\boldsymbol{\alpha}_j = \lambda_j^{-1/2} \mathbf{v}_j$$

  with $\boldsymbol{\alpha}_j$ acting as a *dual vector* defined by eigen-vector $\mathbf{v}_j$ of the *kernel matrix* $\mathbf{K}$

Projections into Feature Space

- Suppose $\mathbf{u}_j = \sum_{i=1}^{N} \alpha_{j,i}\phi(\mathbf{x}_i)$ is a normalized direction in the feature space
- For any data point $\mathbf{x}$, the projection of $\phi(\mathbf{x})$ onto $\mathbf{u}_j$ is

$$\|\mathrm{proj}_{\mathbf{u}_j}(\phi(\mathbf{x}))\| = {\mathbf{u}_j}^\top \phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_{j,i}\kappa(\mathbf{x}_i, \mathbf{x})$$

  which represents the *value* of $\phi(\mathbf{x})$ in terms of the j^{th} axis
- Thus, if we have a set of K orthonormal basis vectors $\{\mathbf{u}_j\}_{j=1}^{K}$, the projection of $\phi(\mathbf{x})$ onto each would produce a new K-vector—

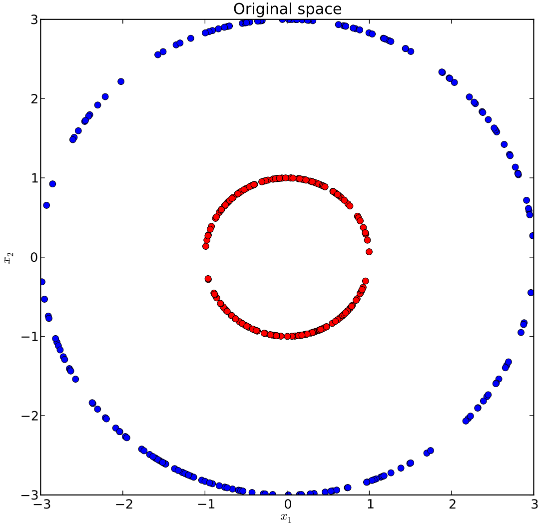
$$\tilde{\mathbf{x}} = \begin{bmatrix} \|\mathrm{proj}_{\mathbf{u}_1}(\phi(\mathbf{x}))\| \\ \|\mathrm{proj}_{\mathbf{u}_2}(\phi(\mathbf{x}))\| \\ \vdots \\ \|\mathrm{proj}_{\mathbf{u}_K}(\phi(\mathbf{x}))\| \end{bmatrix}$$

  the representation of $\phi(\mathbf{x})$ in that basis
- Thus, we can perform the PCA transform *in feature space*

Kernel PCA

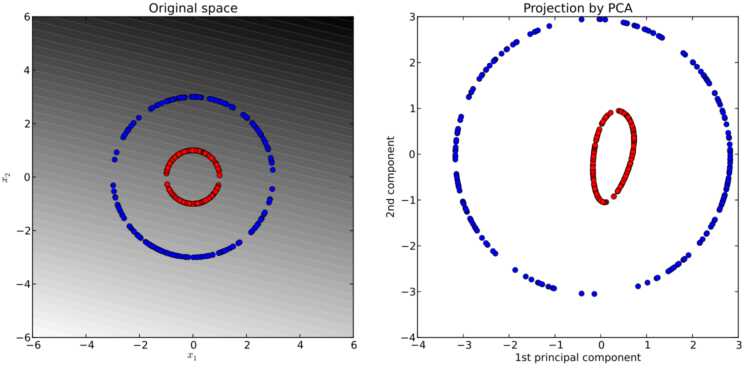
- Performing PCA directly in feature space is not feasible since the covariance matrix is $D \times D$
- However, duality between $\mathbf{C}_{xx}$ & $\mathbf{K}$ allows us to perform PCA indirectly
- Projecting data onto 1st K directions yields a K-dimensional representation
- The algorithm is thus
  1. Center kernel matrix: $\hat{\mathbf{K}} = \mathbf{K} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{K} - \frac{1}{N}\mathbf{K}\mathbf{1}\mathbf{1}^\top + \frac{\mathbf{1}^\top\mathbf{K}\mathbf{1}}{N^2}\mathbf{1}\mathbf{1}^\top$
  2. Find its eigenvectors: $[\mathbf{V}, \mathbf{\Lambda}] = eig\left(\hat{\mathbf{K}}\right)$
  3. Find dual vectors: $\boldsymbol{\alpha}_j = \lambda_j^{-1/2}\mathbf{v}_j$
  4. Project data onto subspace: $\tilde{\mathbf{x}} \leftarrow \left(\sum_{i=1}^{N} \alpha_{j,i}\kappa\left(\mathbf{x}_i, \mathbf{x}\right)\right)_{j=1}^{K}$

Kernel PCA - Application



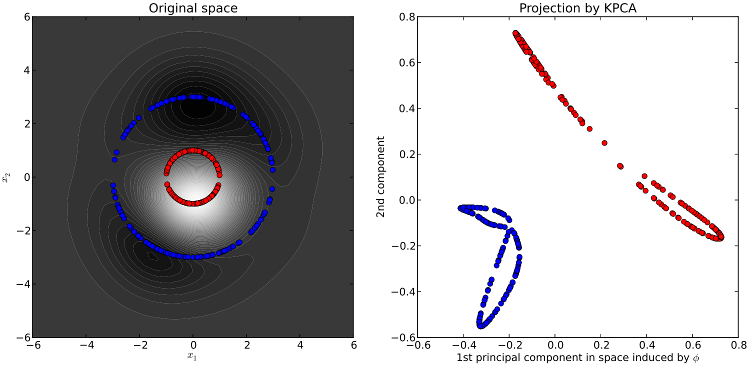
Kernel PCA - Application

Usual PCA fails to capture the data's two ring structure—the rings are not separated in the first two components.



Kernel PCA - Application

Kernel PCA (RBF) does capture the data's two ring structure & the resulting projections separate the two rings



Part III

Maximum Covariance Analysis

Motivation: Directions that Capture Covariance

- Suppose we have a pair of related variables: input variable $\mathbf{x} \sim P_{\mathcal{X}}$ and output variable $\mathbf{y} \sim P_{\mathcal{Y}}$—paired data
- We'd like to find directions of high covariance in spaces $\mathbf{w}_x \in \mathcal{X}$ and $\mathbf{w}_y \in \mathcal{Y}$ such that changes in direction $\mathbf{w}_x$ yield changes in $\mathbf{w}_y$
- Assuming mean-centered variables, we again have that the covariance of its projection onto (normalized) $\mathbf{w}_x$ & $\mathbf{w}_y$ is

$$\mathrm{E}_{\mathbf{x}\sim\mathcal{X},\mathbf{y}\sim\mathcal{Y}}\left[{\mathbf{w}_x}^\top \mathbf{x} {\mathbf{w}_y}^\top \mathbf{y}\right] = \mathbf{w}_x^\top \underbrace{\mathrm{E}\left[\mathbf{x}\mathbf{y}^\top\right]}_{\mathbf{C}_{xy}} \mathbf{w}_y = \mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{w}_y$$

- The empirical covariance matrix (of centered data) is

$$\hat{\mathbf{C}}_{\mathbf{x},\mathbf{y}} = \frac{1}{N} \mathbf{X}^\top \mathbf{Y} \quad ;$$

  an $D_{\mathcal{X}} \times D_{\mathcal{Y}}$ matrix
- How can we find directions that maximize $\mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{w}_y$ for non-square, non-symmetric matrix? How can we kernelize it in space $\mathcal{X}$?

Maximum Covariance Analysis (MCA)

- PCA captures structure in data $\mathbf{X}$, but what data is paired $(\mathbf{x}, y)$? We would like to find correlated directions in X and Y
- Suppose we project $\mathbf{x}$ onto direction $\mathbf{w}_x$ and y onto direction $\mathbf{w}_y$... the covariance of these random variables is

$$\mathrm{E}\left[{\mathbf{w}_x}^\top \mathbf{x} {\mathbf{w}_y}^\top \mathbf{y}\right] = {\mathbf{w}_x}^\top \mathrm{E}\left[\mathbf{x}\mathbf{y}^\top\right] \mathbf{w}_y = {\mathbf{w}_x}^\top \mathbf{C}_{xy} \mathbf{w}_y$$

- The problem we want to solve can again be cast as

$$\max_{\|\mathbf{w}_x\|=1, \|\mathbf{w}_y\|=1} \frac{1}{N} {\mathbf{w}_x}^\top \mathbf{X}^\top \mathbf{Y} \mathbf{w}_y$$

  that is, finding a pair of directions to maximize the covariance
- The solution is simply the first singular vectors $\mathbf{w}_x = \mathbf{u}_1$ & $\mathbf{w}_y = \mathbf{v}_1$ of the SVD $\mathbf{C}_{xy} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Naturally, singular vectors $(\mathbf{u}_2, \mathbf{v}_2), (\mathbf{u}_3, \mathbf{v}_3), \ldots$ capture additional covariance

Kernelized MCA

- As with PCA, MCA can also be kernelized by projecting $\mathbf{x} \rightarrow \phi(\mathbf{x})$
- Consider that eigen-analysis of $\mathbf{C}_{xy}\mathbf{C}_{xy}^\top$ gives us $\mathbf{U}$ & of $\mathbf{C}_{xy}^\top\mathbf{C}_{xy}$ gives us $\mathbf{V}$ of the SVD of $\mathbf{C}_{xy}$...in fact

$$\mathbf{C}_{xy}^\top\mathbf{C}_{xy} = \frac{1}{N^2}\mathbf{Y}^\top\mathbf{K}_{xx}\mathbf{Y}$$

  which has dimension $D_y \times D_y$ & eigen-analysis of this matrix yields (kernelized) directions $\mathbf{v}_k$
- Then, in decomposing $\mathbf{C}_{xy}\mathbf{C}_{xy}^\top$, we have again a relationship between $\mathbf{u}_k$ & $\mathbf{v}_k$: $\mathbf{u}_k = \frac{1}{\sigma_k}\mathbf{C}_{xy}\mathbf{v}_k$, allowing us to project onto $\mathbf{u}_k$ when X is kernelized:

$$\|\mathrm{proj}_{\mathbf{u}_k}(\phi(\mathbf{x}))\| = \sum_{i=1}^{N} \alpha_{k,i}\kappa(\mathbf{x}_i, \mathbf{x}) \qquad \boldsymbol{\alpha}_k = \frac{1}{N\sigma_k}\mathbf{Y}\mathbf{v}_k$$

Part IV

Generalized Eigenvalues & CCA

Motivation: Directions of Correlation

- Suppose that instead of input & output variables, we have 2 variables that are different representations of the same data $\mathbf{x}$:

$$\mathbf{x}_a \leftarrow \psi_a(\mathbf{x}) \qquad\qquad \mathbf{x}_b \leftarrow \psi_b(\mathbf{x})$$

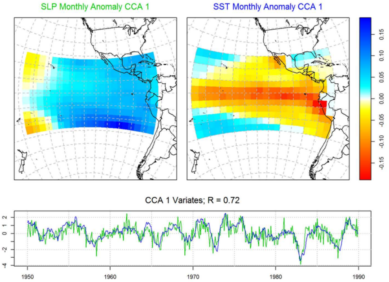
- We'd like to find directions of high correlation in these spaces $\mathbf{w}_a \in \mathcal{X}_a$ and $\mathbf{w}_b \in \mathcal{X}_b$ such that changes in direction $\mathbf{w}_a$ yield changes in $\mathbf{w}_b$
- Assuming mean-centered variables, we have that the correlation of its projection onto (normalized) $\mathbf{w}_a$ & $\mathbf{w}_b$ is

$$\rho_{ab} = \frac{\mathrm{E}_{\mathbf{x}_a \sim \mathcal{X}, \mathbf{x}_b \sim \mathcal{X}_b}\left[{\mathbf{w}_a}^\top \mathbf{x}_a {\mathbf{w}_b}^\top \mathbf{x}_b\right]}{\sqrt{\mathrm{E}\left[{\mathbf{w}_a}^\top \mathbf{x}_a {\mathbf{w}_a}^\top \mathbf{x}_a\right] \mathrm{E}\left[{\mathbf{w}_b}^\top \mathbf{x}_b {\mathbf{w}_b}^\top \mathbf{x}_b\right]}} = \frac{\mathbf{w}_a^\top \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a^\top \mathbf{C}_{aa} \mathbf{w}_a \cdot \mathbf{w}_b^\top \mathbf{C}_{bb} \mathbf{w}_b}}$$

  where $\mathbf{C}_{ab}$, $\mathbf{C}_{aa}$ & $\mathbf{C}_{bb}$ are the covariance matrices between $\mathbf{x}_a$ & $\mathbf{x}_b$ (with usual empirical versions)
- How can we find directions that maximize ρ_{ab}? How can we kernelize it in spaces $\mathcal{X}_a$ & $\mathcal{X}_b$?

Applications of CCA

- Climate Prediction: Researchers have used CCA techniques to find correlations in sea level pressure & sea surface temperature:



- CCA is used with bilingual corpora (same text in two languages) aiding in translation tasks.

Canonical Correlation Analysis (CCA) I

- Our objective is to find directions of maximal correlation:

$$\max_{\mathbf{w}_a, \mathbf{w}_b} \rho_{ab}(\mathbf{w}_a, \mathbf{w}_b) = \frac{\mathbf{w}_a^\top \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a^\top \mathbf{C}_{aa} \mathbf{w}_a \cdot \mathbf{w}_b^\top \mathbf{C}_{bb} \mathbf{w}_b}} \tag{2}$$

  a problem we call canonical correlation analysis (CCA)
- As with previous problems this can be expressed as

$$\max_{\mathbf{w}_a, \mathbf{w}_b} \quad \mathbf{w}_a^\top \mathbf{C}_{ab} \mathbf{w}_b \tag{3}$$

$$\text{such that} \quad \mathbf{w}_a^\top \mathbf{C}_{aa} \mathbf{w}_a = 1 \text{ and } \mathbf{w}_b^\top \mathbf{C}_{bb} \mathbf{w}_b = 1$$

Canonical Correlation Analysis (CCA) II

- The Lagrangian function for this optimization is

$$\mathcal{L}(\mathbf{w}_a, \mathbf{w}_b, \lambda_a, \lambda_b) = \mathbf{w}_a^\top \mathbf{C}_{ab} \mathbf{w}_b - \frac{\lambda_a}{2}(\mathbf{w}_a^\top \mathbf{C}_{aa} \mathbf{w}_a - 1) - \frac{\lambda_b}{2}(\mathbf{w}_b^\top \mathbf{C}_{bb} \mathbf{w}_b - 1)$$

- Differentiating it w.r.t. $\mathbf{w}_a$ & $\mathbf{w}_b$ & setting equal to 0 gives

$$\mathbf{C}_{ab}\mathbf{w}_b - \lambda_a \mathbf{C}_{aa}\mathbf{w}_a = 0 \qquad \mathbf{C}_{ba}\mathbf{w}_a - \lambda_b \mathbf{C}_{bb}\mathbf{w}_b = 0$$

$$\lambda_a \mathbf{w}_a^\top \mathbf{C}_{aa}\mathbf{w}_a = \lambda_b \mathbf{w}_b^\top \mathbf{C}_{bb}\mathbf{w}_b$$

which implies that $\lambda_a = \lambda_b = \lambda$

- The constraints on $\mathbf{w}_a$ & $\mathbf{w}_b$ can be written in matrix form as

$$\begin{bmatrix} \mathbf{0} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{bmatrix} \tag{4}$$

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w} \ ;$$

a generalized eigenvalue problem for the primal problem

Generalized Eigenvectors I

- Suppose $\mathbf{A}$ & $\mathbf{B}$ are symmetric & $\mathbf{B} \succ 0$, then the generalized eigenvalue problem (GEP) is to find $(\lambda, \mathbf{w})$ s.t.

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w} \tag{5}$$

  which are equivalent to

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathbf{w}^\top \mathbf{B}\mathbf{w}} \qquad\qquad \max_{\mathbf{w}^\top \mathbf{B}\mathbf{w}=1} \mathbf{w}^\top \mathbf{A}\mathbf{w}$$

  Note, eigenvalues are special case with $\mathbf{B} = \mathbf{I}$

- Since $\mathbf{B} \succ 0$, any GEP can be converted to an Eigenvalue problem by inverting $\mathbf{B}$:

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{w} = \lambda \mathbf{w}$$

Generalized Eigenvectors II

- However, to ensure symmetry, we can instead use $\mathbf{B} \succ 0$ to decompose $\mathbf{B} = \mathbf{B}^{-1/2}\mathbf{B}^{-1/2}$ where $\mathbf{B}^{-1/2} = \sqrt{\mathbf{B}}^{-1}$ is a symmetric real matrix—taking $\mathbf{w} = \mathbf{B}^{-1/2}\mathbf{v}$ for some $\mathbf{v}$ we obtain (symmetric)

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{v} = \lambda\mathbf{v}$$

an eigenvalue problem for $\mathbf{C} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ providing solutions to Eq. (5)

$$\mathbf{w}_i = \mathbf{B}^{-1/2}\mathbf{v}_i$$

Generalized Eigenvectors III

Proposition 1

Solutions to GEP of Eq. (5) *have following properties: if eigenvalues are distinct, then*

$$\mathbf{w}_i^\top \mathbf{B} \mathbf{w}_j = \delta_{i,j}$$
$$\mathbf{w}_i^\top \mathbf{A} \mathbf{w}_j = \lambda_i \delta_{i,j}$$

that is, the vectors $\mathbf{w}_i$ *are orthonormal after applying transformation* $\mathbf{B}^{1/2}$*—that is, they are conjugate with respect to* $\mathbf{B}$*.*

Generalized Eigenvectors IV

Theorem 2

If $(\lambda_i, \mathbf{w}_i)$ *are eigen-solutions to GEP of Eq.* (5)*, then* $\mathbf{A}$ *can be decomposed as*

$$\mathbf{A} = \sum_{i=1}^{N} \lambda_i \mathbf{B}\mathbf{w}_i (\mathbf{B}\mathbf{w}_i)^\top$$

This yields the generalized deflation of $\mathbf{A}$*:*

$$\tilde{\mathbf{A}} \leftarrow \mathbf{A} - \lambda_i \mathbf{B}\mathbf{w}_i \mathbf{w}_i^\top \mathbf{B}^\top$$

while $\mathbf{B}$ *is unchanged.*

Solving CCA as a GEP

- As shown in Eq. (4), CCA is a GEP $\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}$ where

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{0} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{bmatrix}$$

- Since this is a solution to Eq. (2), the eigenvalues will be correlations $\Rightarrow$ $\lambda \in [-1, +1]$. Further, the eigensolutions will pair: for each $\lambda_i > 0$ with eigenvector $\begin{bmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{bmatrix}$, there is a $\lambda_j = -\lambda_i$ with eigenvector $\begin{bmatrix} \mathbf{w}_a \\ -\mathbf{w}_b \end{bmatrix}$. *Hence, we only need to consider the positive spectrum*.
- Larger eigenvalues correspond to the strongest correlations.
- Finally, the solutions are conjugate w.r.t. matrix $\mathbf{B}$ which reveals that for $i \neq j$

$$\mathbf{w}_{a,j}^{\top}\mathbf{C}_{aa}\mathbf{w}_{a,i} = 0 \qquad \mathbf{w}_{b,j}^{\top}\mathbf{C}_{bb}\mathbf{w}_{b,i} = 0$$

However, the directions will not be orthogonal in the original input space.

Dual Form of CCA I

- Let's take the directions to be linear combinations of data:

$$\mathbf{w}_a = \mathbf{X}_a^\top \boldsymbol{\alpha}_a \qquad\qquad \mathbf{w}_b = \mathbf{X}_b^\top \boldsymbol{\alpha}_b$$

- Substituting these directions into Eq. (3) gives

$$\max_{\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_b} \quad \boldsymbol{\alpha}_a^\top \mathbf{K}_a \mathbf{K}_b \boldsymbol{\alpha}_b$$
$$\text{such that} \quad \boldsymbol{\alpha}_a^\top \mathbf{K}_a^2 \boldsymbol{\alpha}_a = 1 \text{ and } \boldsymbol{\alpha}_b^\top \mathbf{K}_b^2 \boldsymbol{\alpha}_b = 1$$

where $\mathbf{K}_a = \mathbf{X}_a \mathbf{X}_a^\top$ and $\mathbf{K}_b = \mathbf{X}_b \mathbf{X}_b^\top$.

Dual Form of CCA II

- Differentiating the Lagrangian again yields equations

$$\mathbf{K}_a\mathbf{K}_b\boldsymbol{\alpha}_b - \lambda\mathbf{K}_a^2\boldsymbol{\alpha}_a = \mathbf{0} \qquad \mathbf{K}_b\mathbf{K}_a\boldsymbol{\alpha}_a - \lambda\mathbf{K}_b^2\boldsymbol{\alpha}_b = \mathbf{0}$$

- However, these equations reveal a problem. When the dimension of the feature space is large compared number of data points ($D_a \gg N$), solutions will overfit the data.
- For the Gaussian kernel, data will always be independent in feature space & $\mathbf{K}_a$ will be invertible. Hence, we have

$$\boldsymbol{\alpha}_a = \frac{1}{\lambda}\mathbf{K}_a^{-1}\mathbf{K}_b\boldsymbol{\alpha}_b$$

$$\mathbf{K}_b^2\boldsymbol{\alpha}_b - \lambda^2\mathbf{K}_b^2\boldsymbol{\alpha}_b = \mathbf{0}$$

but the latter holds for all $\boldsymbol{\alpha}_b$ with perfect correlation $\lambda = 1$—*Solution is Overfit!!!*

Regularized CCA I

- To avoid overfitting, we can regularize the solutions $\mathbf{w}_a$ & $\mathbf{w}_b$ by controlling their norms. The Regularized CCA Problem is

$$\max_{\mathbf{w}_a, \mathbf{w}_b} \tilde{\rho}_{ab}(\mathbf{w}_a, \mathbf{w}_b) = \frac{\mathbf{w}_a^\top \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\left((1-\tau_a)\mathbf{w}_a^\top \mathbf{C}_{aa} \mathbf{w}_a + \tau_a \|\mathbf{w}_a\|^2\right) \cdot \left((1-\tau_b)\mathbf{w}_b^\top \mathbf{C}_{bb} \mathbf{w}_b + \tau_b \|\mathbf{w}_b\|^2\right)}}$$

  where $\tau_a \in [0,1]$ & $\tau_b \in [0,1]$ serve as regularization parameters
- Again this yields an optimization program for the dual variables

$$\begin{aligned} \max_{\mathbf{w}_a, \mathbf{w}_b} \quad & \boldsymbol{\alpha}_a^\top \mathbf{K}_a \mathbf{K}_b \boldsymbol{\alpha}_b \\ \text{such that} \quad & (1-\tau_a)\boldsymbol{\alpha}_a^\top \mathbf{K}_a^2 \boldsymbol{\alpha}_a + \tau_a \boldsymbol{\alpha}_a^\top \mathbf{K}_a \boldsymbol{\alpha}_a = 1 \\ \text{and} \quad & (1-\tau_b)\boldsymbol{\alpha}_b^\top \mathbf{K}_b^2 \boldsymbol{\alpha}_b + \tau_b \boldsymbol{\alpha}_b^\top \mathbf{K}_b \boldsymbol{\alpha}_b = 1 \end{aligned}$$

Regularized CCA II

- Using the Lagrangian technique, we again arrive at a GEP:

$$\begin{bmatrix} \mathbf{0} & \mathbf{K}_a\mathbf{K}_b \\ \mathbf{K}_b\mathbf{K}_a & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_a \\ \boldsymbol{\alpha}_b \end{bmatrix} = \lambda \begin{bmatrix} (1-\tau_a)\mathbf{K}_a^2 + \tau_a\mathbf{K}_a & \mathbf{0} \\ \mathbf{0} & (1-\tau_b)\mathbf{K}_b^2 + \tau_b\mathbf{K}_b \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_a \\ \boldsymbol{\alpha}_b \end{bmatrix}$$

- Solutions $(\boldsymbol{\alpha}_a^*, \boldsymbol{\alpha}_b^*)$ can now be used as usual projection directions of Eq. (1)
- Solving CCA using the above GEP is *impractical!* The matrices required are $2N \times 2N$. Instead, the usual approach is to make an incomplete Cholesky decomposition of the kernel matrices:

$$\mathbf{K}_a = \mathbf{R}_a^\top \mathbf{R}_a \qquad\qquad \mathbf{K}_b = \mathbf{R}_b^\top \mathbf{R}_b$$

  The resulting GEP can be solved more efficiently (see book for algorithms details)

Regularized CCA III

- Finally CCA can be extended to multiple representations of the data, which result in the following GEP:

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \dots & \mathbf{C}_{1k} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \dots & \mathbf{C}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{k1} & \mathbf{C}_{k2} & \dots & \mathbf{C}_{kk} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_k \end{bmatrix} = \rho \begin{bmatrix} \mathbf{C}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{C}_{kk} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_k \end{bmatrix}$$

LDA as a GEP

You should note, that the Fisher Discriminant Analysis problem can be expressed as

$$\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \mathbf{N} \boldsymbol{\alpha}}$$

which is a GEP. In fact, this is how solutions to LDA are obtained.

Summary

- In this lecture, we saw how different objectives for projection directions yield different subspaces... we saw 3 different algorithms:
  1. Principal Component Analysis
  2. Maximum Covariance Analysis
  3. Canonical Correlation Analysis
- We saw that each of these techniques can be solved using eigenvalue, singular value, and generalized eigenvector decompositions.
- We saw that each of these techniques yielded linear projections and thus could be kernelized.
- In the next lecture, we will explore the general technique of minimizing loss & how allows us to develop a wide range of kernel algorithms. In particular, we will see the Support Vector Machine for classification tasks.

Bibliography I

The Majority of the work from this talk can be found in the lecture's accompanying book, "Kernel Methods for Pattern Analysis."

[1] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.