General latent variable models of PLS enhanced with spectral techniques and representations in Hilbert spaces

Marianna Bolla
Institute of Mathematics
Budapest University of Technology and Economics
marib@math.bme.hu
The research was supported by the
TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project.

Paris
27 May, 2014

Outline

- General latent variable models in function spaces
- solved via representation of joint distributions with SD or SVD of the conditional expectation operator between the two margins:
  - the role of the two spaces is symmetric (correspondence analysis, maximal correlation, canonical correlation analysis);
  - the role of the two spaces is asymmetric (generalized regression, ACE algorithm);
  - symmetric joint distribution (normalized modularity and discrepancy of edge-weighted graphs).
- Compromise factors of independent samples with a novel matrix decomposition algorithm and application.

Outline

- General latent variable models in function spaces
- solved via representation of joint distributions with SD or SVD of the conditional expectation operator between the two margins:
  - the role of the two spaces is symmetric (correspondence analysis, maximal correlation, canonical correlation analysis);
  - the role of the two spaces is asymmetric (generalized regression, ACE algorithm);
  - symmetric joint distribution (normalized modularity and discrepancy of edge-weighted graphs).
- Compromise factors of independent samples with a novel matrix decomposition algorithm and application.

Outline

- General latent variable models in function spaces
- solved via representation of joint distributions with SD or SVD of the conditional expectation operator between the two margins:
  - the role of the two spaces is symmetric (correspondence analysis, maximal correlation, canonical correlation analysis);
  - the role of the two spaces is asymmetric (generalized regression, ACE algorithm);
  - symmetric joint distribution (normalized modularity and discrepancy of edge-weighted graphs).
- Compromise factors of independent samples with a novel matrix decomposition algorithm and application.

Outline

- General latent variable models in function spaces
- solved via representation of joint distributions with SD or SVD of the conditional expectation operator between the two margins:
  - the role of the two spaces is symmetric (correspondence analysis, maximal correlation, canonical correlation analysis);
  - the role of the two spaces is asymmetric (generalized regression, ACE algorithm);
  - symmetric joint distribution (normalized modularity and discrepancy of edge-weighted graphs).
- Compromise factors of independent samples with a novel matrix decomposition algorithm and application.

Outline

- General latent variable models in function spaces
- solved via representation of joint distributions with SD or SVD of the conditional expectation operator between the two margins:
  - the role of the two spaces is symmetric (correspondence analysis, maximal correlation, canonical correlation analysis);
  - the role of the two spaces is asymmetric (generalized regression, ACE algorithm);
  - symmetric joint distribution (normalized modularity and discrepancy of edge-weighted graphs).
- Compromise factors of independent samples with a novel matrix decomposition algorithm and application.

Outline

- General latent variable models in function spaces
- solved via representation of joint distributions with SD or SVD of the conditional expectation operator between the two margins:
  - the role of the two spaces is symmetric (correspondence analysis, maximal correlation, canonical correlation analysis);
  - the role of the two spaces is asymmetric (generalized regression, ACE algorithm);
  - symmetric joint distribution (normalized modularity and discrepancy of edge-weighted graphs).
- Compromise factors of independent samples with a novel matrix decomposition algorithm and application.

Hilbert spaces of random variables

Let (ξ, η) be a pair of real-valued random variables over the product space $\mathcal{X} \times \mathcal{Y}$.
Their joint distribution is $\mathbb{W}$ with margins $\mathbb{P}$ and $\mathbb{Q}$.
Assume that the dependence between ξ and η is regular, i.e., $\mathbb{W}$ is absolutely continuous with respect to $\mathbb{P} \times \mathbb{Q}$, and let w denote the Radon–Nikodym derivative (Rényi, A., On measures of dependence, Acta Math. Acad. Sci. Hung., 1959).
$H = L^2(\xi)$ and $H' = L^2(\eta)$ are sets of random variables which are measurable functions of ξ and η, and have 0 expectation and finite variance with respect to $\mathbb{P}$ and $\mathbb{Q}$.
H and H' are Hilbert-spaces with the covariance as inner product; they are naturally embedded as subspaces into the L^2-space defined likewise by the (ξ, η) pair over the product space endowed with $\mathbb{W}$ (Breiman, L. and Friedman, J.H., Estimating optimal transformations for multiple regression and correlation, JASA, 1985).

The operators taking conditional expectation

Integral operators with kernel w:

$$P_{\mathcal{X}} : H' \to H, \quad \psi = P_{\mathcal{X}}\phi = \mathbb{E}(\phi \,|\, \xi), \quad \psi(x) = \int_{\mathcal{Y}} w(x,y)\phi(y)\,\mathbb{Q}(dy)$$

$$P_{\mathcal{Y}} : H \to H', \quad \phi = P_{\mathcal{Y}}\psi = \mathbb{E}(\psi \,|\, \eta), \quad \phi(y) = \int_{\mathcal{X}} w(x,y)\psi(x)\,\mathbb{P}(dx)$$

$P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are projections and $P^*_{\mathcal{X}} = P_{\mathcal{Y}}$, $P^*_{\mathcal{Y}} = P_{\mathcal{X}}$, since

$$\langle P_{\mathcal{X}}\phi, \psi\rangle_H = \langle P_{\mathcal{Y}}\psi, \phi\rangle_{H'} = \operatorname{cov}_{\mathbb{W}}(\psi, \phi)$$

where $\operatorname{cov}_{\mathbb{W}}$ is the covariance-function:

$$\begin{aligned}\operatorname{cov}_{\mathbb{W}}(\psi, \phi) &= \int_{\mathcal{X}\times\mathcal{Y}} \psi(x)\phi(y)\mathbb{W}(dx, dy) \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \psi(x)\phi(y)w(x,y)\mathbb{Q}(dy)\mathbb{P}(dx)\end{aligned}$$

Riesz, F. and Sz.-Nagy, B., Leçons d'analyse fonctionnelle, Academic Publishing House, Budapest, 1952.

Assume that

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} w^2(x,y)\mathbb{Q}(dy)\mathbb{P}(dx) < \infty.$$

With discrete joint distribution $\{w_{ij}\}$ and margins $\{p_i\}$ ($p_i = \sum_j w_{ij}$), $\{q_j\}$ ($q_j = \sum_i w_{ij}$):

$$\sum_{i\in\mathcal{X}} \sum_{j\in\mathcal{Y}} \left(\frac{w_{ij}}{p_i q_j}\right)^2 p_i q_j = \sum_{i\in\mathcal{X}} \sum_{j\in\mathcal{Y}} \frac{w_{ij}^2}{p_i q_j} < \infty.$$

With absolutely continuous joint distribution $f(x,y)$ and margins $f_1(x)$ ($f_1(x) = \int f(x,y)\,dy$), $f_2(y)$ ($f_2(y) = \int f(x,y)\,dx$):

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} \left(\frac{f(x,y)}{f_1(x)f_2(y)}\right)^2 f_1(x)f_2(y)\,dx\,dy = \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{f^2(x,y)}{f_1(x)f_2(y)}\,dx\,dy < \infty.$$

Spectral and Singular Value Decompositions (SD and SVD)

Under this assumption, $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are Hilbert–Schmidt operators $\Longrightarrow$ compact (absolutely continuous) with discrete spectra:

$$P_{\mathcal{X}} = \sum_{i=1}^{\infty} s_i \langle ., \phi_i \rangle_{H'} \psi_i \quad \text{és} \quad P_{\mathcal{Y}} = \sum_{i=1}^{\infty} s_i \langle ., \psi_i \rangle_{H} \phi_i \quad \text{SVD},$$

where $1 > s_1 \geq s_2 \geq \cdots \geq 0$ ($\lim_{i\to\infty} s_i = 0$).
When $\mathbb{W}$ is symmetric (H and H' are isomorphic), then $P_{\mathcal{X}} = P_{\mathcal{Y}}$ is self-adjoint and the SD of $P_{\mathcal{X}} : H' \to H$ is

$$P_{\mathcal{X}} = \sum_{i=1}^{\infty} \lambda_i \langle ., \psi_i' \rangle_{H'} \psi_i$$

where $|\lambda_i| \leq 1$ and

$$P_{\mathcal{X}} \psi_i' = \lambda_i \psi_i.$$

ψ_i and ψ_i' are identically distributed (i.d.) with joint distribution $\mathbb{W}$.

When the role of the two spaces is symmetric

Definition

We say that the pair $(\mathbf{X}, \mathbf{Y})$ of k-dimensional random vectors with components in H and H', respectively, form a *k-dimensional representation* of the product space endowed with the measure $\mathbb{W}$ if $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T = \mathbf{I}_k$ and $\mathbb{E}_{\mathbb{Q}}\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_k$ (i.e., the components of $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated with zero expectation and unit variance, respectively); further, X_i and Y_j are uncorrelated for $i \neq j$, whereas the joint distribution of X_i and Y_i is $\mathbb{W}$ $(i = 1, \ldots, k)$. The cost of this representation is defined as

$$Q_k(\mathbf{X}, \mathbf{Y}) = \mathbb{E}_{\mathbb{W}}\|\mathbf{X} - \mathbf{Y}\|^2.$$

The couple $(\mathbf{X}^*, \mathbf{Y}^*)$ is an *optimal representation* if it minimizes the above cost.

Representation theorem for joint distributions

Theorem

Let $\mathbb{W}$ be a joint distribution with margins $\mathbb{P}$ and $\mathbb{Q}$. Assume that among the singular values of the conditional expectation operator, $P_{\mathcal{X}}$, there are at least k positive ones, and denote by $1 > s_1 \geq s_2 \geq \cdots \geq s_k > 0$ the largest ones. The minimum cost of a k-dimensional representation is $2\sum_{i=1}^{k}(1 - s_i)$, and it is attained with $\mathbf{X}^* = (\psi_1, \ldots, \psi_k)$ and $\mathbf{Y}^* = (\phi_1, \ldots, \phi_k)$, where ψ_i, ϕ_i is the singular function pair corresponding to the singular value s_i, $i = 1, \ldots, k$.

The optimal representation resembles the PLS objective: we are looking for uncorrelated components in function spaces, which are not necessarily linear transformations of the original ξ and η.

Maximal correlation (Gebelein, Rényi, and Sarmanov)

We are looking for the maximally correlated functions of ξ and η:

$$\max_{\psi\in H,\,\phi\in H'} \mathrm{cor}_{\mathbb{W}}(\psi,\phi) = \max_{\|\psi\|=\|\phi\|=1} \mathrm{cov}_{\mathbb{W}}(\psi,\phi) = s_1$$

and the maximum is attained at the ψ_1, ϕ_1 pair. We used the extremal properties of the SVD; further, $\|\psi\| = \sqrt{\mathrm{var}_{\mathbb{P}}\psi}$ and $\|\phi\| = \sqrt{\mathrm{var}_{\mathbb{Q}}\phi}$.

It also follows and will be illustrated in the forthcoming finite case that the maximal correlation, s_1, is 0 if and only if ξ and η are independent, or equivalently, $\mathbb{W} = \mathbb{P} \times \mathbb{Q}$.

The optimal 1-dimensional representation is also attained at the ψ_1, ϕ_1 pair, and its cost is

$$\min_{\|\psi\|=\|\phi\|=1} \|\psi-\phi\|^2 = \min_{\|\psi\|=\|\phi\|=1} (\|\psi\|^2+\|\phi\|^2-2\mathrm{cov}_{\mathbb{W}}(\psi,\phi)) = 2(1-s_1).$$

Correspondence analysis

Product space: contingency table with entries $w_{ij} \geq 0$ $(\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} = 1)$.
$\mathcal{X} = \{1, \ldots, m\}$: rows, $\mathcal{Y} = \{1, \ldots, n\}$: columns.
Margins: $p_1, \ldots, p_m$ and $q_1, \ldots, q_n$ in the diagonals of $\mathbf{P}$ and $\mathbf{Q}$.
The effect of $P_{\mathcal{X}} : H' \to H$, $P_{\mathcal{X}}\phi = \psi$:

$$\psi(i) = \frac{1}{p_i} \sum_{j=1}^{n} w_{ij}\phi(j) = \sum_{j=1}^{n} \frac{w_{ij}}{p_i q_j}\phi(j)q_j, \quad i = 1, \ldots, m.$$

$P_{\mathcal{X}}$ is integral operator with kernel $\frac{w_{ij}}{p_i q_j}$ and its SVD is obtained via the transformation

$$\sqrt{p_i}\psi(i) = \sum_{j=1}^{n} \frac{w_{ij}}{\sqrt{p_i}\sqrt{q_j}}(\sqrt{q_j}\phi(j)), \quad i = 1, \ldots, m$$

with the SVD of the normalized contingency table
$\mathbf{W}_{\text{nor}} = \mathbf{P}^{-1/2}\mathbf{W}\mathbf{Q}^{-1/2}$.

If $r = \operatorname{rank} \mathbf{W}_{\mathrm{nor}}$, this SVD is

$$\mathbf{W}_{\mathrm{nor}} = \sum_{k=0}^{r-1} s_k \mathbf{v}_k \mathbf{u}_k^T$$

with singular values $1 = s_0 \geq s_1 \geq \cdots \geq s_{r-1} > 0$ together with orthonormal singular vector pairs $(\mathbf{v}_0, \mathbf{u}_0), (\mathbf{v}_1, \mathbf{u}_1), \ldots, (\mathbf{v}_{r-1}, \mathbf{u}_{r-1})$. If the matrix $\mathbf{W}\mathbf{W}^T$ is irreducible (i.e., the contingency table cannot be divided into blocks with permuting its rows or columns), then 1 is a single singular value with vector pair $\mathbf{v}_0 = (\sqrt{p_1}, \ldots, \sqrt{p_m})^T$, $\mathbf{u}_0 = (\sqrt{q_1}, \ldots, \sqrt{q_n})^T$.

The SVD of the operator $P_{\mathcal{X}}$ is obtained by back transformation: the singular values are the same, and the values taken on by the $\mathbb{P}$-distributed categorical variable ψ_k are the coordinates of the vector $\mathbf{P}^{-1/2}\mathbf{v}_k$, while those taken on by the $\mathbb{Q}$-distributed categorical variable ϕ_k are the coordinates of $\mathbf{Q}^{-1/2}\mathbf{u}_k$, $k = 1, \ldots, r-1$.

Maximal correlations sequentially

The pair ψ_0 and ϕ_0 take on values constantly 1 (they are not in H and H'), and is called trivial correspondence factor pair.
The non-trivial pairs solve the following sequential maximization task, when $r > 1$:

$$\max_{\substack{\mathbb{E}_{\mathbb{P}}\psi=0,\,\mathrm{var}_{\mathbb{P}}\psi=1\\ \mathbb{E}_{\mathbb{Q}}\phi=0,\,\mathrm{var}_{\mathbb{Q}}\phi=1\\ \mathrm{cov}_{\mathbb{P}}\psi\psi_i=0\,(i=1,\ldots,k-1)\\ \mathrm{cov}_{\mathbb{Q}}\phi\phi_i=0\,(i=1,\ldots,k-1)}} \mathrm{cov}_{\mathbb{W}}(\psi,\phi) = \mathrm{cov}_{\mathbb{W}}(\psi_k,\phi_k) = s_k, \quad k = 1,\ldots,r-1.$$

The $r = 1$ case is equivalent to $s_1 = 0$, i.e., to zero maximal correlation; this case corresponds to an independent table: $w_{ij} = p_i q_j$, $i = 1,\ldots,m$, $j = 1,\ldots,n$.
Representation Theorem $\Longrightarrow$ the first k non-trivial correspondence factor pairs solve a minimum placement problem with due regard to the categories (Benzécri, J. P. et al., L'Analyse des correspondances, Dunod, Paris, 1980).

Canonical correlation analysis

Any or both of the starting random variables ξ, η can as well be a random vector (with real components).
For example, if they have m- and n-dimensional Gaussian distribution, respectively, then their maximum correlation is the largest canonical correlation between them, and it is realized by appropriate linear combinations of the components of ξ and η, respectively.
Akin to the way above, we can find canonical correlations one after the other with corresponding function pairs (under some orthogonality constraints), as many as the rank of the cross-covariance matrix of ξ and η. The procedure relies on the SVD of the $m \times n$ matrix $\mathbf{C}_{11}^{-1/2}\mathbf{C}_{12}\mathbf{C}_{22}^{-1/2}$, where $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ are the covariance matrices of ξ and η, while $\mathbf{C}_{12}$ is their cross-covariance matrix, usually estimated from a sample.

When the role of the two spaces is asymmetric

ψ: response, ϕ: predictor.

Only $\|\psi\| = 1$ is assumed when $\mathbb{E}_{\mathbb{W}}\|\psi - \phi\|^2$ is minimized.

With the notation $\hat{\phi} = \frac{\phi}{\|\phi\|}$ ($\|\hat{\phi}\| = 1$):

$$\mathbb{E}_{\mathbb{W}}\|\psi - \phi\|^2 = 1 - 2\mathrm{cov}_{\mathbb{W}}(\psi, \hat{\phi})\|\phi\| + \|\phi\|^2 \geq 1 - 2s_1\|\phi\| + \|\phi\|^2.$$

$$\mathrm{cov}_{\mathbb{W}}(\psi, \hat{\phi}) \leq s_1 \Longrightarrow \min \mathbb{E}_{\mathbb{W}}\|\psi - \phi\|^2 = 1 - s_1^2$$

attained at the $\psi_1, s_1\phi_1$ pair.

The ACE algorithm (imitating conditional expectation by smoothing) converges to the solution, and is applicable to non-parametric regression (Breiman, L. and Friedman, J.H., Estimating optimal transformations for multiple regression and correlation, JASA, 1985).

The procedure can be extended to random vectors with components in H and H', when the predictor and response have the same dimension.

We consider k-dimensional representation of the product space endowed with the measure $\mathbb{W}$ such that only $\mathbb{E}_{\mathbb{Q}}\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_k$ is assumed, while the components of $\mathbf{X}$ are also uncorrelated with zero expectation (but usually not of unit variance); further, X_i and Y_j are uncorrelated for $i \neq j$, whereas the joint distribution of X_i and Y_i is $\mathbb{W}$ $(i = 1, \ldots, k)$.

By a componentwise minimization,

$$\mathbb{E}_{\mathbb{W}}\|\mathbf{Y} - \mathbf{X}\|^2 = \sum_{i=1}^{k} \mathbb{E}_{\mathbb{W}}\|Y_i - X_i\|^2 \geq \sum_{i=1}^{k}(1 - s_i^2)$$

attained with the choice $Y_i = \phi_i$ and $X_i = s_i\psi_i$, $i = 1, \ldots, k$.

The case of a symmetric joint distribution

Definition

We say that the k-dimensional random vector $\mathbf{X}$ with components in H forms a *k-dimensional representation* of the product space $H \times H'$ (H and H' are isomorphic) endowed with the *symmetric measure* $\mathbb{W}$ (and margin $\mathbb{P}$) if $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T = \mathbf{I}_k$. Further, the cost of this representation is defined as

$$Q_k(\mathbf{X}) = \mathbb{E}_{\mathbb{W}}\|\mathbf{X} - \mathbf{X}'\|^2,$$

where $\mathbf{X}$ and $\mathbf{X}'$ are identically distributed and the joint distribution of their coordinates X_i and X_i' is $\mathbb{W}$ $(i = 1, \ldots, k)$, while X_i and X_j' are uncorrelated if $i \neq j$.
The k-dimensional random vector $\mathbf{X}^*$ is an *optimal representation* if it minimizes the above cost.

Theorem

Let $\mathbb{W}$ be a symmetric joint distribution with margin $\mathbb{P}$. Assume that among the eigenvalues of the conditional expectation operator $P_{\mathcal{X}} : H' \to H$ (H and H' are isomorphic) there are at least k positive ones and denote by $1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$ the largest ones. Then the minimum cost of a k-dimensional representation is $2\sum_{i=1}^{k}(1-\lambda_i)$, and it is attained by $\mathbf{X}^* = (\psi_1, \ldots, \psi_k)$ where ψ_i is the eigenfunction corresponding to the eigenvalue λ_i $(i = 1, \ldots k)$.

Guess for the optimal choice of k.

Finite $\mathcal{X}$: edge-weighted graph ($\mathcal{X} = V$)

$G = (V, \mathbf{W})$ edge-weighted graph, $|V| = n$,
$w_{ij} = w_{ji} \geq 0$ $(i \neq j)$ pairwise similarities, $w_{ii} = 0$ (i=1,. . . ,n),
$\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$.
$d_i := \sum_{j=1}^{n} w_{ij}$ $(i = 1, \ldots, n)$: generalized degrees
$\mathbf{d} := (d_1, \ldots, d_n)^T$: degree-vector, $\sqrt{\mathbf{d}} := (\sqrt{d_1}, \ldots, \sqrt{d_n})^T$
$\mathbf{D} := \operatorname{diag}(d_1, \ldots, d_n)$: degree-matrix
$\operatorname{vol}(X) = \sum_{i \in X} d_i$: volume of $X \subset V$
$\mathbf{M} = \mathbf{W} - \mathbf{d}\mathbf{d}^T$: modularity matrix (Newman–Girvan)
The SD of $P_{\mathcal{X}}$ is obtained by the SD of the normalized modularity matrix

$$\mathbf{M}_{\mathrm{nor}} = \mathbf{D}^{-1/2}\mathbf{M}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$$

with eigenvalues $1 > \lambda_1 \geq \cdots \geq \lambda_n \geq -1$ (1 cannot be an eigenvalue if the underlying graph is connected, i.e., $\mathbf{W}$ is irreducible).

Representation of the vertices

$Q_{k-1} = \sum_{i<j} w_{ij}\|\mathbf{r}_i - \mathbf{r}_j\|^2$ with representatives $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^{k-1}$ of the vertices. Then run the k-means algorithm with them.

- $k-1$ positive structural eigenvalues: community structure;

$$\min_{\substack{\sum_{i=1}^n d_i \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_{k-1} \\ \sum_{i=1}^n d_i \mathbf{r}_i = \mathbf{0}}} Q_{k-1} = \sum_{i=1}^{k-1} (1 - \lambda_i)$$

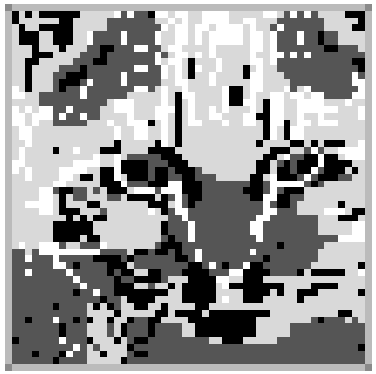
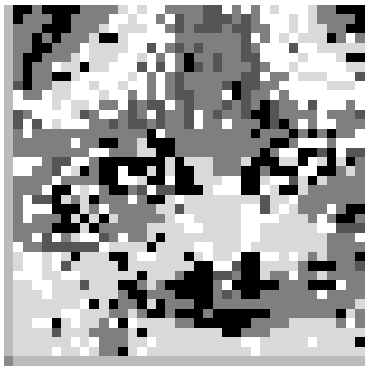
- $k-1$ negative structural eigenvalues: anticommunity structure;

$$\max_{\substack{\sum_{i=1}^n d_i \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_{k-1} \\ \sum_{i=1}^n d_i \mathbf{r}_i = \mathbf{0}}} Q_{k-1} = \sum_{i=1}^{k-1} (1 - \lambda_{n-i+1})$$

- both positive and negative structural eigenvalues: regular structure (cluster pairs with small discrepancy, Szemerédi). Expander Mixing Lemma (discrepancy): for every $X, Y \subset V$, $|w(X,Y) - \mathrm{vol}(X)\mathrm{vol}(Y)| \leq \|\mathbf{M}_{\mathrm{nor}}\| \sqrt{\mathrm{vol}(X)\mathrm{vol}(Y)}$.

Reproducing Kernel Hilbert Spaces



$(48 \times 48$ pixels$)$ $n = 48^2$

Structural eigenvalues of $\mathbf{M}_D$:

$0.137259, \quad 0.014255, \quad 0.000925,$

$-0.0006707, \quad -0.0006706, \ldots$

Gaussian kernel, image segmentation.

For treating non-linearities in the data, methods of regularization and kernelization also fit into this setup, as they merely work with the covariances. It is almost folklore that, due to the kernel-trick, the maps into the RKHS need not be performed, but how to select an appropriate kernel is the crucial question here (B, Spectral Clustering and Biclustering, Wiley, 2013).

Compromise factors of independent samples

Having k independent samples for the underlying n-dimensional random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_k$ $(k \leq n)$, we are looking for compromise factors, i.e., linear combinations $\mathbf{a}_i^T \mathbf{X}_i$ so that

$$\mathrm{var}(\sum_{i=1}^{k} \mathbf{a}_i^T \mathbf{X}_i) = \sum_{i=1}^{k} \mathbf{a}_i^T \mathbf{C}_i \mathbf{a}_i \to \max .$$

subject to $\mathbf{a}_i^T \mathbf{a}_j = \delta_{ij}$ $(i, j = 1, \ldots, k)$, where $\mathbf{C}_i$ is the covariance matrix of $\mathbf{X}_i$.

The problem is to find maxima of sums of heterogeneous quadratic forms (B, Michaletzky, Gy., Tusnády, G., Ziermann, M., Extrema of sums of heterogeneous quadratic forms, Lin. Alg. Appl., 1998).

Solution

- Theoretical solution (by Lagrange's multipliers): $\mathcal{C}(\mathbf{A}) = \mathbf{AS}$, where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$, $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$, $\mathcal{C}(\mathbf{A}) = (\mathbf{C}_1\mathbf{a}_1, \ldots, \mathbf{C}_k\mathbf{a}_k)$, and $\mathbf{S}$ is $k \times k$ symmetric matrix of the multipliers. A non-trivial solution exists if $\det(\mathbf{C} - \mathbf{I}_n \otimes \mathbf{S}) = 0$, characteristic polynomial of degree $k(k+1)/2$ of the diagonal and upper-diagonal entries of the compromise matrix $\mathbf{S}$, where $\mathbf{C} = \mathbf{C}_1 \oplus \cdots \oplus \mathbf{C}_k$.
- Algorithm (iteration converging to the solution): Starting with a suborthogonal matrix $\mathbf{A}^{(0)}$, the t-th step of the iteration: $\mathcal{C}(\mathbf{A}^{(t-1)}) := \mathbf{A}^{(t)}\mathbf{S}^{(t)}$ polar decomposition. $\operatorname{tr}\mathbf{S}^{(t)}$ converges to the maximum. An orthonormal system $\mathbf{a}_1^*, \ldots, \mathbf{a}_k^*$ giving the maximum is called a *compromise system* of the matrices $\mathbf{C}_1, \ldots, \mathbf{C}_k$.

In the possession of k independent samples, we use the empirical covariance matrices. The algorithm is also applicable to small samples with positive semidefinite matrices.

Application

The coordinates of the compromise vectors $\mathbf{a}_i$'s, called compromise factor loadings, help us to identify the variables that best characterize the samples in relation to the others, and hence, together with other sample scores they accomplish the best possible compromise between the samples, in contrast to the discriminant analysis, where we rather want to separate the samples. Sometimes there are no remarkable differences between the sample means, or even the sample covariances, to make discriminant analysis techniques applicable.

We applied the method for clinical measurements (protein, triglyceride and other organic matter levels in the urine) of 100 patients suffering from nephrosis. We distinguished between three stages of the illness; a no symptoms stage and two nephrotic stages: one in an intermediate stage, the other in a more seriously developed stage.

Compromise factor loadings for three nephrotic stages of 100 patients

	no symptoms	intermediate	nephrotic
AT	-0.104339	-0.151711	-0.068392
PC	-0.151864	+0.060398	+0.062981
KO2	-0.355027	**-0.662945**	**-0.423931**
TG	-0.134190	-0.372486	**+0.781611**
HK	-0.241672	+0.194526	**+0.421601**
LK	**+0.496214**	**-0.543357**	+0.149016
PROT	**+0.522984**	+0.194241	-0.027665
URIN	**-0.493607**	+0.155758	+0.001543
NAK	-0.014336	+0.005123	+0.001286

One may conclude that mainly measurements with high loadings in absolute value have to be considered seriously in the diagnosis.

Conclusions

- In contrast with the algorithmic approaches to PLS methods, a rather theory-oriented description of latent variable models is presented, where the number of latent variables depends on the spectral properties of the underlying operators.
- Via these so-called representation techniques, we establish a common outline structure for the contents of each algorithm based on traditional or novel matrix decompositions, together with appropriate normalizations.
- For treating non-linearities, methods of kernelization also fit into this setup, as they merely work with the covariances.
- For the SD or SVD, fast numerical algorithms are at our disposal, which usually make use of the conjugate gradient method, a well-known PLS technique. Randomized methods are also useful when we want to find only some leading eigenvalues or singular values, and we suggest to confine the numerical efforts to this part.