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

Marianna Bolla[a*]

[a]Institute of Mathematics, Budapest University of Technology and Economics, Hungary

Keywords: Conditional expectation, singular value decomposition, low-dimensional representation, correspondence analysis, compromise factors

Introduction

A general (usually not linear) statistical model of the latent variable approach of PLS is stated in terms of joint distributions, and solved by the spectral or singular value decomposition (SD or SVD) of the Hilbert–Schmidt operator taking conditional expectation between the margins. In the possession of an i.i.d. sample from a multivariate distribution, the components of which are divided into two parts (in a symmetric or asymmetric way), we make inference via the empirical covariance and cross-covariance matrices (in the multivariate Gaussian case) or the normalized contingency table (in the case of categorical data). Compromise factors of several data sets are also looked for, based on a novel algorithm for maximizing the sum of heterogeneous quadratic forms.

1 Representation of joint distributions

Let (ξ,η) be a pair of random variables defined over the product space $\mathcal{X}\times\mathcal{Y}$ and having joint distribution $\mathbb{W}$ with margins $\mathbb{P}$ and $\mathbb{Q}$, respectively. Assume that the dependence between ξ and η is regular, i.e., their joint distribution $\mathbb{W}$ is absolutely continuous with respect to the product measure $\mathbb{P}\times\mathbb{Q}$, w denotes the Radon–Nikodym derivative (notation of A. Rényi[1], see [1]). In the spirit of [2], $H=L^2(\xi)$ and $H'=L^2(\eta)$ denote the sets of random variables which are functions of ξ and η, and have zero expectation and finite variance with respect to $\mathbb{P}$ and $\mathbb{Q}$. Both H and H' are Hilbert spaces with the covariance as inner product; further, they are embedded as subspaces into the L^2 space defined likewise by the (ξ,η) pair over the product space.

Let $P_{\mathcal{X}}:H'\to H$ and $P_{\mathcal{Y}}:H\to H'$ be the integral operators taking *conditional expectation* between the two margins, $P_{\mathcal{X}}^*=P_{\mathcal{Y}}$. Provided $\int_{\mathcal{X}}\int_{\mathcal{Y}}w^2(x,y)\mathbb{Q}(dy)\mathbb{P}(dx)<\infty$, $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are Hilbert–Schmidt operators with SVD

$$P_{\mathcal{X}}=\sum_{i=1}^{\infty}s_i\langle .,\phi_i\rangle_{H'}\psi_i,\quad P_{\mathcal{Y}}=\sum_{i=1}^{\infty}s_i\langle .,\psi_i\rangle_{H}\phi_i,\qquad P_{\mathcal{X}}\phi_i=s_i\psi_i,\quad P_{\mathcal{Y}}\psi_i=s_i\phi_i\quad(i=1,2,\dots),\tag{1}$$

where for the singular values $1>s_1\geq s_2\geq\cdots\geq 0$ holds, since the operators $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are in fact orthogonal projections from one margin onto the other, and $\psi_i\in H$, $\phi_i\in H'$ are the corresponding function pairs.

*E-mail: marib@math.bme.hu. Address: 1 Egry József, Bldg H5/2, 1111 Budapest, Hungary. Research supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project.

[1]Alfréd Rényi, founder of the Mathematical Institute of the Hungarian Academy of Sciences, in 1959 published two seminal papers on the maximal correlation in the Acta Math. Acad. Sci. Hung. with titles "On measures of dependence" and "New version of the probabilistic generalization of the large sieve".

1.1 When the role of the two spaces is symmetric

The pair $(\mathbf{X},\mathbf{Y})$ of k-dimensional random vectors with components in H and H', respectively, is called *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 $\mathbb{E}_{\mathbb{Q}}\mathbf{Y}\mathbf{Y}^T=\mathbf{I}_k$.
Theorem 1. With the notation of (1), the minimum of the cost $\mathbb{E}_{\mathbb{W}}\|\mathbf{X}-\mathbf{Y}\|^2$ of this representation is $2\sum_{i=1}^k(1-s_i)$, and it is attained with the k-dimensional representation $\mathbf{X}^*=(\psi_1,\dots,\psi_k)$ and $\mathbf{Y}^*=(\phi_1,\dots,\phi_k)$.

If $k=1$, Theorem 1 gives the solution of the *maximal correlation* problem of A. Rényi: $\max_{\psi\in H,\,\phi\in H'}\mathrm{Corr}_{\mathbb{W}}(\psi,\phi)=\max_{\|\psi\|=\|\phi\|=1}\mathrm{Cov}_{\mathbb{W}}(\psi,\phi)=s_1$; or equivalently, $\min_{\|\psi\|=\|\phi\|=1}\|\psi-\phi\|^2=2(1-s_1)$, and both are attained on the ψ_1,ϕ_1 pair. When $\mathcal{X}$ and $\mathcal{Y}$ are finite sets, the solution corresponds to the SVD of the normalized contingency table $\mathbf{P}^{-1/2}\mathbf{W}\mathbf{Q}^{-1/2}$, and the representatives are obtained by the correspondence vector pairs. Since numerical algorithms are capable to find singular vector pairs orthogonal in Euclidean norm, when performing *correspondence analysis*, we have to normalize the underlying contingency table $\mathbf{W}$ with the diagonal matrices $\mathbf{P}$ and $\mathbf{Q}$ containing the probabilities, corresponding to $\mathbb{P}$ and $\mathbb{Q}$, in their main diagonals. If ξ and η are multivariate Gaussian, then their maximum correlation is the largest *canonical correlation*, realized by appropriate linear combinations of them. We can find canonical correlations successively; the procedure relies on the SVD of the matrix $\mathbf{C}_{11}^{-1/2}\mathbf{C}_{12}\mathbf{C}_{22}^{-1/2}$, where the covariance- and cross-covariance matrices $\mathbf{C}_{11}$, $\mathbf{C}_{22}$, and $\mathbf{C}_{12}$ are estimated from a sample.

1.2 When the role of the two spaces is asymmetric

Let ψ be the response, and ϕ be the predictor, and only $\|\psi\|=1$ is assumed when $\|\psi-\phi\|^2$ is minimized. Now the minimum is $1-s_1^2$, attained with the $\psi_1,s_1\phi_1$ pair. This is the nonlinear regression problem, and based on a sample, the iteration of the ACE (Alternating Conditional Expectation) algorithm of [2] converges to the solution.

1.3 The case of a symmetric joint distribution

Here $\mathbb{W}$ is a symmetric measure with margin $\mathbb{P}$, H and H' are isomorphic, and $P_{\mathcal{X}}$ is selfadjoint with SD $P_{\mathcal{X}}=\sum_{i=1}^{\infty}\lambda_i\langle .,\psi_i'\rangle_{H'}\psi_i$, where $|\lambda_i|\le 1$, and $P_{\mathcal{X}}\psi_i'=\lambda_i\psi_i$, where ψ_i and ψ_i' are identically distributed $(i=1,2,\dots)$.
Theorem 2. Assume that $1>\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_k>0$. Then the minimum of $\mathbb{E}_{\mathbb{W}}\|\mathbf{X}-\mathbf{X}'\|^2$ subject to $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T=\mathbf{I}_k$, where $\mathbf{X}$ and $\mathbf{X}'$ are identically distributed, is $2\sum_{i=1}^k(1-\lambda_i)$, and it is attained with $\mathbf{X}^*=(\psi_1,\dots,\psi_k)$.

A finite $\mathcal{X}$ belongs to the vertex-set of a weighted graph, with edge-weights $w_{ii}=0$, $w_{ij}=w_{ji}\ge 0$ $(i\ne j)$ summing to 1. The operator $P_{\mathcal{X}}$ corresponds to the *normalized modularity matrix* of the graph, and based on the low-dimensional representatives of the vertices, spectral clustering techniques are to be used, see [1].

2 Compromise factors of independent samples and conclusions

Having k independent samples with underlying n-dimensional random vectors $\mathbf{X}_1,\dots,\mathbf{X}_k$ $(k\le n)$, we are looking for *compromise factors*, i.e., linear combinations $\mathbf{a}_i^T\mathbf{X}_i$ that maximize $\mathrm{Var}(\sum_{i=1}^k\mathbf{a}_i^T\mathbf{X}_i)=\sum_{i=1}^k\mathbf{a}_i^T\mathbf{C}_i\mathbf{a}_i$ subject to $\mathbf{a}_i^T\mathbf{a}_j=\delta_{ij}$ $(i,j=1,\dots,k)$, where $\mathbf{C}_i$ is the covariance matrix of the i-th sample. In [3], we introduced a novel algorithm to find the *compromise vectors* $\mathbf{a}_i$'s, the coordinates of which are loadings characterizing the samples in relation to the others; an application for finding compromise factors of three nephrotic stages will be presented.

To sum up, via the above representations we are modeling the relations between two or more sets of observed variables. Our theory extends to the non-Gaussian case, and the number of latent variables depends on the spectral properties of the underlying operators. Sometimes we use preliminary regularization to treat non-linearities in the data. This theory has applications in non-parametric regression (see [2]), correspondence analysis, and spectral clustering of social networks (see [1]). For the SD or SVD, fast numerical algorithms are at our disposal (e.g., the Lánczos method), which usually make use of the conjugate gradient method, a well-known PLS technique.

References

[1] M. Bolla, *Spectral Clustering and Biclustering*, Wiley, Chichester, 2013.

[2] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Assoc.* **80**, pp. 580–619, 1985.

[3] M. Bolla, G. Michaletzky, G. Tusnády, and M. Ziermann, "Extrema of sums of heterogeneous quadratic forms," *Linear Algebra Appl.* **269**, pp. 331–365, 1998.