Common subspace extraction using a Grassmannian minimum enclosing ball approach

Emilie Renard [1] **Kyle A. Gallivan** [2] **P.-A. Absil** [1]

We study the problem of finding a subspace representative of multiple datasets by minimizing the maximal dissimilarity between this subspace and all the subspaces generated by those datasets (Renard et al., 2018). Extracting common information from multiple datasets is crucial, a typical example can be found in bioinformatics when dealing with various datasets measuring the same disease on different sets of patients, but corresponding to different studies and different experimental conditions that should be taken into account in further analysis.

1. Problem formulation

Beside the basic possibility to simply concatenate all the datasets X_1,...,X_m into a larger dataset $X = [X_1 \ \dots \ X_m]$ and apply methods such as PCA on X, more specific approaches exist to extract common components present in the datasets (Ponnapalli et al., 2011; Hotelling, 1936; Wold, 1985; Meng et al., 2014; Tenenhaus & Tenenhaus, 2011). A central question when using more than two datasets is the importance to give to the different datasets. A common approach is to give all datasets the same importance. To avoid obtaining components representing very well a set of similar datasets but not being representative at all of others, we minimize the maximal dissimilarity d between the common component $U \in \mathbb{R}^{p\times K}$ and all datasets $X_i \in \mathbb{R}^{p\times n_i}$:

$$U^* = \arg \min_{U\in\mathbb{R}^{p\times K}} \max_i d(U, X_i).$$

This can be viewed as looking for the center of a minimum enclosing ball. As U represents a subspace we want to use a dissimilarity d that is invariant under basis selection. Let $\mathcal{U}$ and $\mathcal{X}_i$ represent the subspaces generated by the columns of U and X_i. Let $\bar{U}$ and $\bar{X}_i$ be orthonormal basis of $\mathcal{U}$ and $\mathcal{X}_i$, and $\sigma_k = \cos\phi_k(U, X_i)$ be the kth singular value of $\bar{U}^\top \bar{X}_i$. To preserve $d(U, X_i) = 0$ when $\mathcal{U} \subset \mathcal{X}_i$, we consider the following dissimilarity:

$$d(X_i, U) = \sqrt{\min(n_i, n_u) - \sum_{k}^{\min(n_i,n_u)} \cos^2(\phi_k)}$$

with n_u and n_i dimensions of subspaces $\mathcal{U}$ and $\mathcal{X}_i$.

2. Proposed approach

In (Badoiu & Clarkson, 2003), a procedure is proposed to compute the minimum enclosing ball center of data points in a Euclidean space. The procedure is extended to arbitrary Riemannian manifolds in (Arnaudon & Nielsen, 2013). A candidate solution $U^{(t)}$ is initialized with a data point, and is iteratively updated as $U^{(t+1)} = \text{Geodesic}\left(U^{(t)}, X_f^{(t)}, \frac{1}{t+1}\right)$ where $X_f^{(t)}$ is the farthest data point to $U^{(t)}$. $\text{Geodesic}(p, q, t)$ represents the intermediate point m on the geodesic passing through p and q such that $\text{dist}(p, m) = \text{dist}(p, q)$. This approach can be used to solve our problem, but requires some adaptations. Since we are interested in finding the best subspace of dimension K in $\mathbb{R}^p$, our solution U belongs to the Grassmann manifold $\mathcal{G}(K, p)$. Moreover, we are dealing with points representing subspaces of different dimensions n_i and therefore belonging to different manifolds $\mathcal{G}(n_i, p)$: the usual Grassmaniann distance cannot be used to determine $X_f^{(t)}$. To preserve $d(U, X_i) = 0$ when $\mathcal{U} \subset \mathcal{X}_i$, we use a dissimilarity which becomes a metric if the two subspaces belong to the same Grassmannian. Another adaptation is that $X_f^{(t)}$ must be projected on $\mathcal{G}(K, p)$ to allow the use of a geodesic. Given $\mathcal{X}_f \in \mathcal{G}(n_f, p)$ and $\mathcal{U} \in \mathcal{G}(K, p)$ with $n_f \geq K$, we compute $\mathcal{Y}_f \in \mathcal{G}(K, p)$ included in $\mathcal{X}_f$ that minimizes the distance to $\mathcal{U}$. We can then update U using the corresponding geodesic.

Tested on generated synthetic data, the proposed method is promising. We also compared it to a K-truncated SVD on X and on $\bar{X} = [\bar{X}_1 ... \bar{X}_n]$. As expected, the SVD on $\bar{X}$ tends to recover the mean while the Grassmaniann approach tends to recover the center. On the criterion minimized, the Grassmaniann approach gives the best results (see (Renard et al., 2018) for more details).

---

[1] ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium, Supported by EOS Project no 30468160 [2] Department of Mathematics, Florida State University, Tallahassee, USA. Correspondence to: Emilie Renard <emilie.renard@uclouvain.be>.

Working Notes of a ICML/FAIM 2018 Workshop, Stockholm, Sweden, 2018.

References

Arnaudon, M. and Nielsen, F. On approximating the Riemannian 1-center. *Computational Geometry*, 46(1):93–104, 2013.

Badoiu, M. and Clarkson, K. L. Smaller core-sets for balls. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 801–802. Society for Industrial and Applied Mathematics, 2003.

Hotelling, H. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 00063444. doi: 10.2307/2333955.

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, 15(1):162, 2014.

Ponnapalli, S. P., Saunders, M. A., Van Loan, C. F., and Alter, O. A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms. *PloS one*, 6(12):e28072, 2011.

Renard, E., Gallivan, K. A., and Absil, P.-A. A Grassmannian minimum enclosing ball approach for common subspace extraction. In *14th international conference on Latent Variable Analysis and Signal Separation*, 2018. URL https://sites.uclouvain.be/absil/2018.02.

Tenenhaus, A. and Tenenhaus, M. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.

Wold, H. Partial least squares. *Encyclopedia of statistical sciences*, 6:581–591, 1985.