

Minimax center to extract a common subspace from multiple datasets

Emilie Renard¹, P.-A. Absil¹ and Kyle A. Gallivan² *

1- Université catholique de Louvain - ICTEAM Institute
Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve - Belgium

2- Florida State University - Department of Mathematics
208 Love Building - 1017 Academic Way - Tallahassee FL 32306-4510, USA

Abstract. We address the problem of extracting common information from multiple datasets. More specifically, we look for a common subspace minimizing the maximal dissimilarity with all datasets and we propose an algorithm derived from the first order necessary conditions of optimality. On synthetic datasets the proposed method gives as good results as a Riemannian based approach, but also provides an evaluation on how far the iterate is from a critical point.

1 Introduction

Extracting a subset of components common to multiple datasets is a common problem, appearing in bioinformatics [1, 2], signal processing [3], ecological data analysis [4] or strategic management [5]. A typical example can be found in bioinformatics, where there often exist multiple studies measuring the treatment and progress of the same disease. Each corresponding dataset was measured on its own set of patients and with its own experimental conditions, that should be taken into account in any analysis. It is crucial to extract the information of interest common to all datasets (for example, linked to a disease) while bypassing possible stronger effects specific to each dataset.

Let $\{X_1, \dots, X_m\}$ be a collection of m datasets where $X_i \in \mathbb{R}^{p \times n_i}$ represents a dataset of p features and n_i samples. For example, dealing with gene expressions, each $X_i(r, s)$ can represent the expression level of gene r for patient s in dataset i . The simplest method to extract a common component is to concatenate all the datasets X_1, \dots, X_m into a larger dataset $X = [X_1 \dots X_m] \in \mathbb{R}^{p \times (n_1 + \dots + n_m)}$ and apply standard methods such as principal components analysis on X . However, this approach does not use the dataset membership information, which can help improve the extraction of components common to all the datasets. When dealing with only two datasets, the best known method is canonical correlation analysis (CCA) [6]. CCA looks for two linear combinations of each dataset features, such that the correlation between both linear combinations is maximized. Deflation

*This work was supported by (i) the Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160, (ii) “Communauté française de Belgique - Actions de Recherche Concertées” (contract ARC 14/19-060). Part of this work was performed while the third author was a visiting professor at UCLouvain, funded by the Science and Technology Sector, and with additional support by the Netherlands Organization for Scientific Research.

is then used to compute the next best linear combinations. Another well-known method, partial least square regression [7], maximizes covariance. Extensions of such methods to more than two datasets usually try to maximize the sum of correlations (or covariances) between all pairs of datasets, with orthogonality constraints on the new features or on the linear combination coefficients [2, 8, 9].

If the stronger variations present in the X_i 's correspond to the common components, then any CCA-like method should work. However the common components may be hidden by other variations in the datasets, stronger but not present in all the X_i 's. For instance, if we are dealing with a set of datasets all of which are similar except one that was measured using a different technology then maximizing the covariance can lead to components representing very well technical features strongly present in all the similar datasets but absent from the last one. Considering the column space $\mathcal{X}_i = \{X_i\alpha : \alpha \in \mathbb{R}^{n_i}\}$ instead of the matrix X_i , for $i = 1, \dots, m$, facilitates giving all of the variations the same importance.

A second difficulty may arise if the common component in the X_i 's is present in a noisy version. We should be careful to not recover another signal which is less noisy, but not present in all X_i 's. To avoid this, we propose to compute an approximate common subspace, represented by $U \in \mathbb{R}^{p \times K}$, by minimizing the maximal dissimilarity d between U and all datasets $X_i \in \mathbb{R}^{p \times n_i}$:

$$U^* = \arg \min_U \max_i d(U, X_i). \quad (1)$$

This approach is similar to the minimum enclosing ball problem, that finds the center of the smallest-radius sphere enclosing all X_i (also called 1-center problem or minimax optimization problem). This kind of optimization problem is well-studied, especially in Euclidean space [10, 11, 12]. However, here each X_i represents a subspace and not a point in \mathbb{R}^p , which requires the use of an adapted dissimilarity. We want $d(U, X_i) = d(\mathcal{U}, \mathcal{X}_i)$ with \mathcal{U} and \mathcal{X}_i denoting the subspaces generated by the columns of U and X_i . An algorithm to solve the minimum enclosing ball problem when all points belong to the same Riemannian manifold was proposed in [13]. This method could be used when the column space of each X_i has the same dimension K , because in this case the m subspaces belong to the same Grassmann manifold $\mathcal{G}(p, K)$, i.e., the set of all K -dimensional subspaces in \mathbb{R}^p . Here, however, the points X_i may have different dimensions, and the desired center of the ball, U , is of lower dimension than the X_i 's. An adaptation of [13] to the case with various dimensions was proposed in [14]. We propose here another approach that has the advantage of also giving an estimate of how far the iterate is from a critical point.

In this paper, we first propose a formulation of the problem and derive the associated first order necessary conditions of optimality in Section 2. Based on those conditions, we propose an algorithm in Section 3, and show in Section 4 that our approach gives results as good as those in [14] on synthetic datasets. Conclusion are presented in Section 5.

2 Formulation

Let $\{X_1, \dots, X_m\}$ be a collection of m datasets where $X_i \in \mathbb{R}^{p \times n_i}$ represents a dataset of p features and n_i samples. We want to solve problem (1), i.e., to find a subspace \mathcal{U} of given dimension K (via a representative $U \in \mathbb{R}_*^{p \times K}$) minimizing the maximal dissimilarity to all the subspaces \mathcal{X}_i associated with the X_i 's. To evaluate the dissimilarity between two subspaces \mathcal{X} and \mathcal{U} , we use

$$d(\mathcal{U}, \mathcal{X}) = d(U, X) = \sqrt{K - \text{Tr}(\check{U}^T \check{X} \check{X}^T \check{U})} = \sqrt{\sum_l \sin^2 \phi_l(U, X)}$$

with $\phi_l(U, X)$ the principal angles between \mathcal{U} and \mathcal{X} . So the $\cos \phi_l(U, X)$ are the singular values of $\check{U}^T \check{X}$, with \check{Y} denoting a matrix with columns that are an orthonormal basis of \mathcal{Y} . For $K = 1$, $\phi_1(U, X)$ is the angle between U and its orthogonal projection onto \mathcal{X} . Since $\min_U \max_i d(U, X_i)$ is equivalent to $\min_{U, \tau} \tau$ subject to $\tau \geq d^2(U, X_i)$ for all i , the problem is equivalent to:

$$\max_{U, \tau} \tau \text{ s.t.} \quad \tau - \sum_{l=1}^K u_l^T \check{X}_i \check{X}_i^T u_l \leq 0 \quad \forall i \quad (2a)$$

$$U^T U = 0. \quad (2b)$$

Associating Lagrange multipliers γ_i 's with constraints (2a), and M with constraint (2b), the KKT conditions can be written as:

$$\sum_i \gamma_i = 1 \quad (3a)$$

$$\left(\sum_i \gamma_i \check{X}_i \check{X}_i^T \right) U = U M \quad (3b)$$

$$U^T U = I \quad (3c)$$

$$\tau - \text{Tr}(U^T \check{X}_i \check{X}_i^T U) \leq 0 \quad \forall i = 1, \dots, m \quad (3d)$$

$$\gamma_i \geq 0 \quad \forall i = 1, \dots, m \quad (3e)$$

$$\gamma_i (\tau - \text{Tr}(U^T \check{X}_i \check{X}_i^T U)) = 0 \quad \forall i = 1, \dots, m. \quad (3f)$$

Due to the symmetry in the constraint (2b), M is symmetric. So $M = QDQ^T$ with D a diagonal matrix and Q an orthogonal matrix, and $(\sum_i \gamma_i \check{X}_i \check{X}_i^T) U Q = U Q D$ which means that $U Q$ is a matrix of eigenvectors of $\sum_i \gamma_i \check{X}_i \check{X}_i^T$.

Conditions (3b) and (3c) are equivalent to finding eigenvectors of the matrix $\sum_i \gamma_i \check{X}_i \check{X}_i^T$. Condition (3a) combined with (3b) and (3e) can be interpreted as finding optimal weights γ_i 's for the different subspaces. The optimal U is then given by eigenvectors of $\sum_i \gamma_i \check{X}_i \check{X}_i^T$, a convex combination of the $\check{X}_i \check{X}_i^T$'s. Condition (3f) implies that for each \check{X}_i , either $\gamma_i = 0$ which means that the corresponding dataset is not used, or $\tau = K - \text{Tr} U^T \check{X}_i \check{X}_i^T U$ which tightens the corresponding constraint (3d). Assume that we have γ_i 's satisfying condition (3f), and let $U_Y D_Y V_Y^T$ be the singular value decomposition of $Y = [\sqrt{\gamma_1} \check{X}_1, \sqrt{\gamma_2} \check{X}_2, \dots, \sqrt{\gamma_m} \check{X}_m] \in \mathbb{R}^{p \times N}$. A candidate solution of problem (2)

would then be $U^* = U_Y$. The difficulty is then to find γ_i such that condition (3f) holds.

3 Algorithm

Based on the interpretation of KKT conditions, we can build an algorithm to compute a candidate U^* . The main difficulty is to find good γ_i 's. Since $\sum_i \gamma_i = 1$, we have that

$$\sum_i \gamma_i d^2(U, X_i) \leq \sum_i \gamma_i \max_j d^2(U, X_j) = \max_j d^2(U, X_j) \quad (4)$$

with equality at optimality since condition (3f) must hold. If we stop when $\delta_f = \max_i d(U, X_i) - \sqrt{\sum_i \gamma_i d^2(U, X_i)} \leq \epsilon$ with ϵ small and positive, condition (3f) should nearly hold.

Ideally we want both to increase $\sum_i \gamma_i d^2(U, X_i)$ and decrease $\max_i d^2(U, X_i)$ at each iteration t . With $U^{(t)} = U(\gamma^{(t)})$ fixed for iteration t , maximizing the first term is setting to 0 all γ_i 's except for those associated with the X_i 's at maximal dissimilarity to U . However, taking directly $\gamma_i^{(t+1)} = \alpha$ if $\max_j d(U^{(t)}, X_j) - d(U^{(t)}, X_i) \leq \epsilon$, and 0 otherwise, would eventually ignore the other X_i 's and probably not decrease $\max_i d^2(U(\gamma^{(t+1)}), X_i)$. To smooth the γ_i 's evolution, we propose to take:

$$\gamma_i^{(t+1)} = \left(\gamma_i^{(t)} + \frac{1}{t+1} r_i^{(t)} \right) \beta \quad (5)$$

with $r_i^{(t)} = 1$ if $\max_j d(U^{(t)}, X_j) - d(U^{(t)}, X_i) \leq \epsilon$, $r_i^{(t)} = 0$ otherwise, and β a normalizing constant ensuring that $\sum_i \gamma_i = 1$.

The proposed algorithm (Algorithm 1) enforces conditions (3a) to (3e) at each iteration and tries to achieve (3f) in the limit. The approach is reminiscent of the one proposed in [11] for the minimax problem in \mathbb{R}^n .

Algorithm 1 KKT approach for minimax problem (2).

- 1: $t \leftarrow 1$
 - 2: $\gamma_i^{(t)} = \frac{1}{m}$ for all i
 - 3: **while** Stopping criterion not met **do**
 - 4: $Y = [\sqrt{\gamma_1} \check{X}_1, \sqrt{\gamma_2} \check{X}_2, \dots, \sqrt{\gamma_m} \check{X}_m]$
 - 5: $U^{(t)} \leftarrow$ singular vectors for the K largest singular values of Y
 - 6: $r_i^{(t)} = 1$ if $\max_j d(U^{(t)}, X_j) - d(U^{(t)}, X_i) \leq \epsilon$, $r_i^{(t)} = 0$ otherwise,
 - 7: $\gamma_i^{(t+1)} \leftarrow \left(\gamma_i^{(t)} + \frac{1}{t+1} r_i^{(t)} \right) / \sum_j \left(\gamma_j^{(t)} + \frac{1}{t+1} r_j^{(t)} \right)$
 - 8: $t \leftarrow t + 1$
 - 9: **end while**
-

4 Results

To study the behavior of the algorithm, we generated datasets in the following way. We first generated three orthonormal matrices U_1, U_2, U_3 of dimension

$p \times K$ with orthogonal column spaces. Data matrices $X_i \in \mathbb{R}^{p \times n_i}$ were generated by combining information from the matrices U_1 , U_2 , and U_3 and adding noise and additional directions. Each of the K generating vectors of U_1 is present in each X_i with moderate noise, each of the K generating vectors of U_2 is present in 80% of the X_i 's with small noise, and each of the K generating vectors of U_3 is present in 50% of the X_i 's with small noise. We also added s_i orthonormal vectors to each X_i , with $s_i \sim U_{[5 \ 11]}$ uniformly distributed between 5 and 11. The noise is represented as the angle ϕ_{ijl} between the component l of U_j and X_i . We took $\phi_{ijl} \sim U_{\frac{\pi}{2}[0.25 \ 0.45]}$ for $j = 1$ and $\phi_{ijl} \sim U_{\frac{\pi}{2}[0 \ 0.1]}$ for $j = 2, 3$. Finally, we multiplied by 2 the U_3 related vectors to give them more weight.

An example of the evolution of the objective function (2) for the Grassmannian approach [14] and the proposed KKT approach is shown on Figure 1, with the lower bound (4) associated with the KKT approach. Evolution of the maximal dissimilarity is smoother for the Grassmannian approach, however the KKT approach gives us a clear indication on the proximity to a critical point. In Figure 2 we compare the results obtained for 50 randomly generated datasets when applying the SVD on the X_i 's (SVD) or on the \tilde{X}_i 's (SVD_o) concatenated, and the minimax KKT or Grassmannian approaches. We stopped the minimax algorithms after ten seconds, then took the best result in the 100 last iterations. As expected, SVD recovers U_3 , the subspace with high weight and SVD_o tends to recover U_2 , present in many X_i 's with small noise. The minimax approaches give the best results in terms of objective function values, with similar performance for both algorithms. They tend to recover U_1 , which by construction is close to the center U^* . The mean and standard deviation of the corresponding δ_f are 0.0027 ± 0.0044 for a mean objective function value of 0.7076.

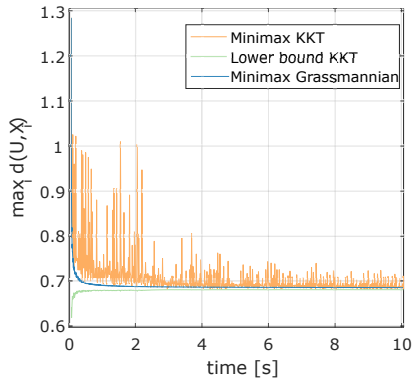


Fig. 1: Example of the evolution of the objective function for the Grassmannian and the KKT minimax approaches.

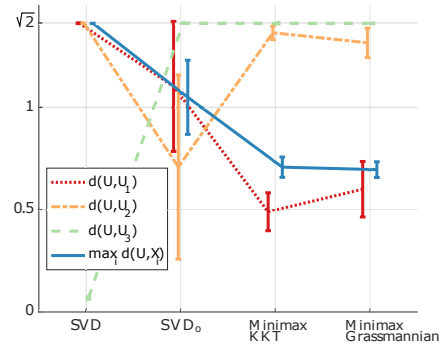


Fig. 2: Comparisons of the four methods for $m = 10$, $K = 2$, $p = 1000$ (the maximal possible dissimilarity of a subspace of dimension K is \sqrt{K}).

5 Conclusion

We proposed a formulation to compute a subspace that minimizes the maximal dissimilarity to each subspace in a set, and a simple algorithm based on the associated KKT conditions to solve it. Compared to simpler SVD-based approaches, we showed that this formulation better recovers such a central subspace in the presence of components with less noise, in most but not all subspaces. Compared to the Grassmannian approach, our KKT-based method gives similar performances but has the advantage of providing an estimate of how far the current iterate is from a critical point, which can be used as a stopping criterion. An important next step would be to improve the proposed γ_i 's update in order to ensure convergence.

References

- [1] F. Rohart, A. Eslami, N. Matigian, S. Bougeard, and K.-A. Le Cao. MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC bioinformatics*, 18(1):128, 2017.
- [2] S. P. Ponnappalli, M. A. Saunders, C. F. Van Loan, and O. Alter. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. *PLoS one*, 6(12):e28072, 2011.
- [3] A. Dogandzic and A. Nehorai. Generalized multivariate analysis of variance-A unified framework for signal processing in correlated noise. *IEEE Signal Processing Magazine*, 20(5):39–54, 2003.
- [4] S. Dray, D. Chessel, and J. Thioulouse. Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84(11):3078–3089, 2003.
- [5] J. F. Hair, M. Sarstedt, T. M. Pieper, and C. M. Ringle. The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. *Long Range Planning*, 45(5-6):320–340, 2012.
- [6] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN: 00063444.
- [7] H. Wold. Partial least squares. *Encyclopedia of Statistical Sciences*, 6:581–591, 1985.
- [8] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.
- [9] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.
- [10] D. W. Hearn and J. Vijay. Efficient algorithms for the (weighted) minimum circle problem. *Operations Research*, 30(4):777–795, 1982.
- [11] C. L. Lawson. The smallest covering cone or sphere. *SIAM Review*, 7(3):415–417, 1965.
- [12] M. Badoiu and K. L. Clarkson. Smaller core-sets for balls. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2003, 801–802.
- [13] M. Arnaudon and F. Nielsen. On approximating the Riemannian 1-center. *Computational Geometry*, 46(1):93–104, 2013.
- [14] E. Renard, K. A. Gallivan, and P.-A. Absil. A Grassmannian Minimum Enclosing Ball Approach for Common Subspace Extraction. In: *Latent Variable Analysis and Signal Separation*. Vol. 10891. Lecture Notes in Computer Science. Springer, 2018, 69–78.