September 30, 2019

1 GD on Manifold

1.1 Understanding the algorithm

Denote by O_n the collection of n-dimensional orthogonal matrices. Let $G_{n,p}$ denotes the Grassmann manifold $G_{n,p} = O_n / (O_p \times O_{n-p})$ — the collection of p-dimensional linear subspaces of the n-dimensional space. More specifically, a point in $G_{n,p}$ is an equivalent class

$$[Q] = \left\{ Q \begin{pmatrix} Q_p & 0 \\ 0 & Q_{n-p} \end{pmatrix} : Q_p \in O_p, Q_{n-p} \in O_{n-p} \right\}.$$

According to [1], for $Y \in O_n$, the tangent directions at Y has the form

$$\Delta = YA + (I - YY^\top)C,$$

where A is $p \times p$ skew-symmetric, C is $p \times p$ arbitrary.
To obtain the tangent space of the quotient space, we need to decompose the tangent space of O_n into two complementary subspaces: vertical space and horizontal space. The vertical space consists of directions along which the matrix does not move in the quotient space; the horizontal space is defined to be orthogonal to the vertical space. Specifically, for Grassmann manifold, the vertical space at a point $[Q]$ is the set of matrices of the form

$$\Phi = Q \begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix},$$

where A is $p \times p$ skew-symmetric and C is $(n-p) \times (n-p)$ skew-symmetric. The horizontal space at a point $[Q]$ is the set of matrices of the form

$$\Delta = Q \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix}.$$

The above horizontal space gives the tangents to the Grassmann manifold.

Having the tangent space, we could now consider the geodesics. The orthogonal group geodesic

$$Q(t) = Q(0) \exp t \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix} \qquad [DEFINITION?]$$

has horizontal tangent

$$\dot{Q}(t) = Q(t) \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix}$$

at every point along the curve $Q(t)$; therefore, they are geodesics on the Grassmann manifold as well.

Theorem 1 (Theorem 2.3 in [1]) *If $Y(t) = Y(0) \exp \left\{ t \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix} \right\} I_{n,p}$, with $Y(0) = Y, \dot{Y}(0) = H$, then*

$$Y(t) = \begin{pmatrix} YV & U \end{pmatrix} \begin{pmatrix} \cos \Sigma t \\ \sin \Sigma t \end{pmatrix} V^T, \qquad [PROOF?]$$

where $H = U \Sigma V^\top$ is the compact singular value decomposition of H.

This theorem provides a useful method for computing the geodesic given the initial gradient. Before deriving the gradient, we need to introduce the canonical metric on the Grassmann manifold. Suppose

$$\Delta_i = Q \begin{pmatrix} 0 & -B_i^\top \\ B_i & 0 \end{pmatrix}, \qquad i = 1, 2.$$

The canonical metric on the Grassmann manifold is defined [FOR SOME REASON] as

$$g_c\left(\Delta_1, \Delta_2\right) = \operatorname{tr} B_1^T B_2.$$

It is shown that this metric is essentially equivalent to the Euclidean metric (up to multiplication by 1/2).

Now we derive the gradient on the Grassmann manifold. The gradient of F at a point $[Y]$ is defined to be the tangent vector ∇F such that

$$\operatorname{tr}(F_Y^T \Delta) \equiv g_e(F_Y, \Delta) = g_c(\nabla F, \Delta) \equiv \operatorname{tr}(\nabla F^T \Delta)$$

for any tangent vectors at Y, where $(F_Y)_{ij} = \frac{\partial F}{\partial Y_{ij}}$. The solution of the above equation turns out to be

$$\nabla F = F_Y - YY^T F_Y.$$

References

[1] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

1.2 Algorithm (from last report)

Unlike previous methods, here we try to use SVD to estimate the complete matrix. That is, minimize the loss function as follows

$$\operatorname{minimize} F(X, Y) \qquad \text{s.t. } X \in G_{n_1,p}, Y \in G_{n_2,p},$$

where

$$F(X, Y) := \frac{1}{2} \min_{S \in \mathbb{R}^{r \times r}} \left\| \mathcal{P}_\Omega\left(M^* - XSY^T\right) \right\|_{\mathrm{F}}^2.$$

Taking gradients over the Grassmann manifold (Keshavan, R. H., Montanari, A., & Oh, S. (2009). Matrix Completion from a Few Entries. Retrieved from http://arxiv.org/abs/0901.3150) yields

$$\nabla F_X(X, Y) = \left(I - XX^T\right) P_\Omega\left(XSY^T - M\right) YS^T,$$
$$\nabla F_Y(X, Y) = \left(I - YY^T\right) P_\Omega\left(XSY^T - M\right)^T XS.$$

Let $-\nabla F_X(X, Y) = U_t D_t V_t^T$ be its compact SVD, then the geodesic on the manifold along the gradient direction is given by

$$X_t(\eta_t) = \left[X_t V_t \cos\left(D_t \eta_t\right) + U_t \sin\left(D_t \eta_t\right)\right] V_t^T.$$

A similar expression holds for $Y(t)$.

2 ADMM may not be suitable

This algorithm is always applied to solve the problem

$$\min_x f(x) + g(x)$$

by solving its equivalent problem

$$\min_{x,y} f(x) + g(y), \qquad \text{subject to } x = y.$$

That is, ADMM integrates the augmented Lagrangian method and the partial update method.
If the objective function is not separable in x and y, such as in the matrix completion problem, ADMM is almost equivalent to AltMin.
One possible way using ADMM is to solve

$$\min_{X,Y,Z} \frac{1}{2}\|XY - Z\|_F^2 + \Lambda \bullet \mathcal{P}_\Omega(Z - M) + \frac{\alpha}{2}\|\mathcal{P}_\Omega(Z - M)\|_F^2,$$

but the projector is still a problem when updating Z.

3 Presenting the results

To figure out why GD on manifold does not perform well:

- Show the results of the optimization over S
- Show the results when fixing the learning rate
- Show the results when using other possible definition of $\cos(A)$

To present the matrix and compare M^* and $\hat{M}$:

- use image() for low-dimensional matrix
- generate the ground truth whose image shows a meaningful pattern (e.g. a zebra) [HOW TO MAKE IT RANDOM WHILE MEANINGFUL?]