Gradient of Multiview MDS stress function

1 Some matrix calculus

Let X be an $n \times p$ matrix variable and let $y = f(X)$, where $f : \mathbb{R}^{n \times p} \to \mathbb{R}$ is a function of X. The derivative of y with respect to X is the $p \times n$ matrix given by

$$\frac{dy}{dX} := \left[\frac{\partial f}{\partial X_{ji}}\right]_{ij},$$

where X_{ij} is the (i, j) entry of X. The gradient of y with respect to X is the $n \times p$ matrix given by

$$\nabla_X y := \left[\frac{\partial f}{\partial X_{ij}}\right]_{ij} = \left(\frac{dy}{dX}\right)^T.$$

The first notation is useful when deriving differentiation rules, the second will be used for optimization.

Note that

$$\begin{array}{rcl} y(X_0 + \Delta X) & = & y(X_0) + \sum_{i,j} \left(\nabla_X y|_{X_0}\right)_{ij} (\Delta X)_{ij} + \text{higher order terms} \\ & = & y(X_0) + \operatorname{tr}\left(\frac{dy}{dX}(X_0)\Delta X\right) + \text{higher order terms} \end{array},$$

which we can write in terms of differentials as

$$dy = \operatorname{tr}\left(\frac{dy}{dX}(X_0)dX\right).$$

We can then use properties of differentials and the trace function to derive many differentiability rules for matrix calculus.

If $g : \mathbb{R} \to \mathbb{R}$ and $z = g(f(X))$, the chain rule says that

$$\frac{dz}{dX}(X) = g'(f(X))\frac{df}{dX}(X).$$

If P is a fixed $p \times p$ matrix and $y(X) = f(XP^T)$, we have

$$\begin{array}{rcl} y(X + \Delta X) & = & f\left((X + \Delta X)P^T\right) \\ & = & f\left(XP^T + \Delta XP^T\right) \\ & = & f(XP^T) + \operatorname{tr}\left(\frac{df}{dX}(XP^T)\Delta XP^T\right) + \mathcal{O}(\|\Delta XP^T\|^2) \\ & = & f(XP^T) + \operatorname{tr}\left(P^T\frac{df}{dX}(X_0P^T)\Delta X\right) + \mathcal{O}(\|\Delta X\|^2) \end{array},$$

so that

$$\frac{d}{dX}\left(f(XP^T)\right) = P^T \frac{df}{dX}(XP^T)$$

and

$$\nabla_X \left(f(XP^T)\right) = \nabla f(XP^T)P.$$

If we differentiate with respect to P instead, set $y(P) = f(XP^T)$ and note that

$$\begin{aligned} y(P+\Delta P) &= f\left(X(P+\Delta P)^T\right) \\ &= f(XP^T) + \operatorname{tr}\left(\tfrac{df}{dX}(XP^T)X\Delta P^T\right) + \mathcal{O}(\|X\Delta P^T\|^2) \\ &= f(XP^T) + \operatorname{tr}\left(X^T\left(\tfrac{df}{dX}(XP^T)\right)^T \Delta P\right) + \mathcal{O}(\|\Delta P^T\|^2) \end{aligned},$$

and therefore

$$\frac{d}{dP}\left(f(XP^T)\right) = X^T\left(\frac{df}{dX}(XP^T)\right)^T$$

and

$$\nabla_P \left(f(XP^T)\right) = \left(\nabla f(XP^T)\right)^T X.$$

If we write $P = QQ^T$ and differentiate with respect to Q, we have $y(Q) = f(XQQ^T)$ and

$$\begin{aligned} y(Q+\Delta Q) &= f\left(X(Q+\Delta Q)(Q+\Delta Q)^T\right) \\ &= f\left(XQQ^T + X\left(Q\Delta Q^T + \Delta QQ^T + \Delta Q\Delta Q^T\right)\right) \\ &= f(XQQ^T) + \operatorname{tr}\left(\tfrac{df}{dX}(XQQ^T)X\left(Q\Delta Q^T + \Delta QQ^T + \Delta Q\Delta Q^T\right)\right) + \mathcal{O}(\|\Delta Q\|^2) \\ &= f(XQQ^T) + \operatorname{tr}\left(\tfrac{df}{dX}(XQQ^T)XQ\Delta Q^T\right) + \operatorname{tr}\left(\tfrac{df}{dX}(XQQ^T)X\Delta QQ^T\right) + \mathcal{O}(\|\Delta Q\|^2) \\ &= f(XQQ^T) + \operatorname{tr}\left(Q^TX^T\left(\tfrac{df}{dX}(XQQ^T)\right)^T \Delta Q\right) + \operatorname{tr}\left(Q^T\tfrac{df}{dX}(XQQ^T)X\Delta Q\right) + \mathcal{O}(\|\Delta Q\|^2) \\ &= f(XQQ^T) + \operatorname{tr}\left(Q^T\left(X^T\left(\tfrac{df}{dX}(XQQ^T)\right)^T + \tfrac{df}{dX}(XQQ^T)X\right)\Delta Q\right) + \mathcal{O}(\|\Delta Q\|^2) \end{aligned},$$

so it follows that

$$\frac{d}{dQ}\left(f(XQQ^T)\right) = Q^T\left(X^T\left(\frac{df}{dX}(XQQ^T)\right)^T + \frac{df}{dX}(XQQ^T)X\right)$$

and

$$\nabla_Q \left(f(XQQ^T)\right) = \left(\left(\nabla f(XQQ^T)\right)^T X + X^T\nabla f(XQQ^T)\right)Q$$

2 Multiview MDS using gradient descent

2.1 Gradients of distance function

Given an $n \times p$ matrix X, containing the coordinates of n points in $\mathbb{R}^p$, the distance between points i and j is

$$d_{ij}(X) = \|X^Te_i - X^Te_j\|_2 = \|X^T(e_i - e_j)\|_2,$$

where $e_i, e_j \in \mathbb{R}^n$ are the ith and jth (column) basis vectors.

The square distance can be written as

$$\begin{array}{rcl} d_{ij}^2(X) & = & \|X^T(e_i - e_j)\|_2^2 \\ & = & \mathrm{tr}\left(X^T(e_i - e_j)(e_i - e_j)^T X\right) \\ & = & \mathrm{tr}\left(X^T A_{ij} X\right) \end{array},$$

where

$$A_{ij} := (e_i - e_j)(e_i - e_j)^T.$$

Note that A_{ij} is symmetric. The square distance can be writen more compactly as

$$d_{ij}^2(X) = (e_i - e_j)^T X X^T (e_i - e_j),$$

but the first form is easier to work with. Note that

$$\begin{array}{rcl} d\mathrm{tr}\left(X^T A_{ij} X\right) & = & \mathrm{tr}\left(d\left(X^T A_{ij} X\right)\right) \\ & = & \mathrm{tr}\left(dX^T A_{ij} X + X^T A_{ij} dX\right) \\ & = & \mathrm{tr}\left(X^T A_{ij}^T dX + X^T A_{ij} dX\right) \\ & = & \mathrm{tr}\left(\left(2X^T A_{ij}\right) dX\right) \end{array},$$

and so

$$\frac{dd_{ij}^2}{dX}(X) = 2X^T A_{ij}.$$

It then follows that

$$\begin{array}{rcl} \frac{dd_{ij}}{dX}(X) & = & \frac{d}{dX}\sqrt{d_{ij}^2(X)} \\ & = & \frac{1}{2\sqrt{d_{ij}^2(X)}} \frac{dd_{ij}^2}{dX}(X) \\ & = & \frac{1}{d_{ij}(X)} X^T A_{ij} \end{array}$$

and that

$$\nabla d_{ij}(X) = \frac{1}{d_{ij}(X)} A_{ij} X$$

If P is a $p \times p$ matrix, we have

$$\begin{array}{rcl} \nabla_X\left(d_{ij}(XP^T)\right) & = & \nabla_X d_{ij}(XP^T) P \\ & = & \left(\frac{1}{d_{ij}(XP^T)} A_{ij}\left(XP^T\right)\right) P \\ & = & \frac{1}{d_{ij}(XP^T)} A_{ij} X P^T P \end{array}.$$

If we differentiate with respect to P instead, we obtain

$$\begin{array}{rcl} \nabla_P\left(d_{ij}(XP^T)\right) & = & \left(\nabla d_{ij}(XP^T)\right)^T X \\ & = & \left(\frac{1}{d_{ij}(XP^T)} A_{ij}\left(XP^T\right)\right)^T X \\ & = & \frac{1}{d_{ij}(XP^T)} P X^T A_{ij} X \end{array}.$$

Finally, if we set $P = QQ^T$ and differentiate with respect to Q,

$$\begin{array}{rcl} \nabla_Q\left(d_{ij}(XQQ^T)\right) & = & \left(\left(\nabla d_{ij}(XQQ^T)\right)^T X + X^T \nabla d_{ij}(XQQ^T)\right) Q \\ & = & \left(\left(\frac{1}{d_{ij}(XQQ^T)} A_{ij} XQQ^T\right)^T X + X^T \frac{1}{d_{ij}(XQQ^T)} A_{ij} XQQ^T\right) Q \\ & = & \frac{1}{d_{ij}(XQQ^T)}\left(QQ^T X^T A_{ij} X + X^T A_{ij} XQQ^T\right) Q \end{array}.$$

2.2 Gradient of MDS stress

For a fixed $n \times n$ distance matrix D, the MDS stress is defined by

$$\sigma^2(X;D) = \sum_{i<j} \left(d_{ij}(X) - D_{ij}\right)^2 .$$

Its gradient is

$$\begin{array}{rcc} \nabla\sigma^2(X;D) & = & \nabla_X \sum_{i<j} \left(d_{ij}(X) - D_{ij}\right)^2 \\ & = & \sum_{i<j} 2\left(d_{ij}(X) - D_{ij}\right) \nabla_X \left(d_{ij}(X) - D_{ij}\right) \\ & = & 2\sum_{i<j} \left(d_{ij}(X) - D_{ij}\right) \nabla_X d_{ij}(X) \\ & & 2\sum_{i<j} \left(d_{ij}(X) - D_{ij}\right) \frac{1}{d_{ij}(X)} A_{ij} X \\ & = & \left(2\sum_{i<j} \frac{(d_{ij}(X) - D_{ij})}{d_{ij}(X)} A_{ij}\right) X \\ & := & B(X;D)X \end{array} ,$$

where

$$B(X;D) := 2\sum_{i<j} \frac{(d_{ij}(X) - D_{ij})}{d_{ij}(X)} A_{ij}.$$

2.3 Gradient of MDS stress with fixed projections

If P is a $p \times p$ matrix (such as a projection matrix), then the action of P on the rows of X is the matrix XP^T. This is an $n \times p$ matrix giving the new coordinates (e.g. after projecting). The gradient of the MDS stress function w.r. to X is

$$\begin{array}{rcc} \nabla_X \left(\sigma^2(XP^T;D)\right) & = & \nabla\sigma^2(XP^T;D)P \\ & = & B(XP^T;D)XP^TP \end{array} .$$

If $\{(P_k, D_k)\}_{k=1}^K$ are k pairs of $p \times p$ transformations and $n \times n$ distance matrices, then the multiview MDS stress function is

$$\begin{array}{rcc} \sigma_m^2\left(X; \{(P_k, D_k)\}_{k=1}^K\right) & := & \sum_{k=1}^K \sigma^2(XP_k^T; D_k) \\ & = & \sum_k \sum_{i<j} \left(d_{ij}(XP_k^T) - (D_k)_{ij}\right)^2 \end{array} ,$$

and its gradient is

$$\begin{array}{rcc} \nabla_X \sigma_m^2\left(X; \{(P_k, D_k)\}_{k=1}^K\right) & = & \sum_k \nabla_X \sigma^2(XP_k^T; D_k) \\ & = & \sum_k B(XP_k^T; D_k) XP_k^T P_k \end{array} .$$

2.4 Gradient with respect to transformations

The gradient of the MDS stress function w.r. to P is

$$\begin{array}{rcc} \nabla_P \sigma^2(XP^T;D) & = & \left(\nabla\sigma^2(XP^T;D)\right)^T X \\ & = & \left(B(XP^T;D)XP^T\right)^T X \\ & = & PX^T B(XP^T;D)X \end{array} .$$

The gradient of the multiview MDS stress function with respect to one of the transformations is

$$\begin{aligned}\nabla_{P_k}\sigma_m^2\left(X;\{(P_k,D_k)\}_{k=1}^K\right) &= \nabla_{P_k}\sigma^2(XP_k^T;D_k)\\ &= P_kX^TB(XP_k^T;D_k)X\end{aligned}.$$

2.5 Gradient with respect to orthogonal projections

A rank-q, $p\times p$ orthogonal projection matrix is one of the form $P_A = QQ^T$, where Q is an $p\times q$ orthogonal matrix (that is, its q columns are orthonormal). We want to restrict optimization of multi-MDS stress to this type of transformations. The gradient of the MDS stress function with respect to Q is

$$\begin{aligned}\nabla_Q\sigma^2(XQQ^T;D) &= \left(\left(\nabla\sigma^2(XQQ^T;D)\right)^T X + X^T\nabla\sigma^2(XQQ^T;D)\right)Q\\ &= \left(\left(B(XQQ^T;D)XQQ^T\right)^T X + X^TB(XQQ^T;D)XQQ^T\right)Q\\ &= \left(QQ^TX^T\left(B(XQQ^T;D)\right)^T X + X^TB(XQQ^T;D)XQQ^T\right)Q\end{aligned}$$

The gradient for the multiview MDS stress function with respect to one of the Q matrices is

$$\nabla_{Q_k}\sigma_m^2\left(X;\left\{\left(Q_kQ_k^T,D_k\right)\right\}_{k=1}^K\right) = \left(Q_kQ_k^TX^T\left(B(XQ_kQ_k^T;D_k)\right)^T X + X^TB(XQ_kQ_k^T;D_k)XQ_kQ_k^T\right)Q_k\ .$$

The projection of a $p\times p$ matrix B into the subspace of rank-q orthogonal matrices is given by UI_qV^T, where $U\Sigma V^T$ is the singular-value decomposition of B and $I_q = [e_1\cdots e_q 0\cdots 0]$. That is, the matrix UI_qV^T minimizes $\|U\Sigma V^T - C\|_F^2$ over all q-rank orthogonal matrices C.

Since $\tilde{Q}_k^{(i+1)} = Q_k^{(i)} + \alpha\nabla_{Q_k}\sigma_m^2\left(X;\left\{\left(Q_k^{(i)}Q_k^{(i)T},D_k\right)\right\}_{k=1}^K\right)$ is not guaranteed to be a rank-q orthogonal matrix, We set $Q_k^{(i+1)} = \mathcal{P}_q\left(\tilde{Q}_k^{(i+1)}\right)$, where $\mathcal{P}_q$ is the projection described above.