Metric learning for multivariate linear models

Charles Zheng and Yuval Benjamini

October 25, 2015

1 Introduction

Let $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} \subset \mathbb{R}^q$ be random vectors with a joint distribution, and let $d_F(\cdot, \cdot)$ be a distance on probability measures.

Let F_x denote the conditional distribution of Y given $X = x$ (and assume that such conditional distributions can be constructed.) Define the *induced metric* on $\mathcal{X}$ by

$$d_{\mathcal{X}}(x_1, x_2) = d_F(F_{x_1}, F_{x_2})$$

We are interested in the problem of estimating the induced metric $d_{\mathcal{X}}$ based on iid observations $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from the joint distribution of (X, Y). We define the loss function for estimation as follows. Let $\hat{d}$ (suppressing the subscript) denote the estimate of $d_{\mathcal{X}}$, and let G denote the marginal distribution of X. Then the loss is defined as

$$\mathcal{L}(d_{\mathcal{X}}, \hat{d}) = 1 - \text{Cor}_{X,X' \sim G}[d_{\mathcal{X}}(X, X'), \hat{d}(X, X')]$$

where the correlation is taken over independent random pairs (X, X') drawn from $G \times G$.

Now we make the following additional assumptions. Let us assume that $X \sim N(0, \Sigma_X)$ and that the conditional distribution of $Y|X = x$ is given by

$$F_x = N(B^T x + \eta, \Sigma_\epsilon)$$

for some $p \times q$ coefficient matrix B, $p \times p$ covariance matrix Σ_X, $q \times q$ covariance matrix Σ_ϵ, and $q \times 1$ vector η.

In the special case that both arguments of the KL divergence are are multivariate gaussian distributions with the same covariance matrix, the KL divergence reduces to a multiple of the Mahalanobis distance. Hence given our assumptions it is natural to adopt a multiple of the KL divergence as the error metric d_F:

$$d_F(\mu_1, \mu_2) = (\mu_1 - \mu_2)\Sigma_\epsilon^{-1}(\mu_1 - \mu_2)$$

Therefore, we obtain the following induced metric:

$$d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^T B \Sigma_\epsilon^{-1} B^T (x_1 - x_2)$$

This is a function only of $\delta = x_1 - x_2$. So defining the positive-semidefinite matrix norm

$$||x||_A = \sqrt{x^T A x}$$

we have

$$d_{\mathcal{X}}(x_1, x_2) = ||x_1 - x_2||^2_{B\Sigma_\epsilon^{-1}B^T}$$

2 Estimation

Since $d_{\mathcal{X}}$ is completely specified by B and Σ_ϵ, the problem of *metric learning for a multivariate linear models* (MLMLM) reduces to the problem of jointly estimating B and Σ_ϵ, under a loss function $\tilde{L}$ defined by

$$\tilde{\mathcal{L}}(B, \Sigma_\epsilon; \hat{B}, \hat{\Sigma}_\epsilon) = 1 - \text{Cor}_{\delta \sim N(0, \Sigma_X)}(||\delta||^2_{B\Sigma_\epsilon^{-1}B^T}, ||\delta||^2_{\hat{B}\hat{\Sigma}_\epsilon^{-1}\hat{B}^T})$$

One can verify that

$$\tilde{\mathcal{L}}(B, \Sigma_\epsilon; \hat{B}, \hat{\Sigma}_\epsilon) = \mathcal{L}(d_{\mathcal{X}}, \hat{d})$$

where

$$\hat{d}(x_1, x_2) = (x_1 - x_2)^T \hat{B} \hat{\Sigma}_\epsilon^{-1} \hat{B}^T (x_1 - x_2).$$

The loss function $\tilde{L}$ looks complicated at first, but perhaps we can find a simplified approximation.

2.1 Approximating $\tilde{L}$

Let δ be multivariate normal $N(0, \Sigma_X)$. Then $X = \Sigma^{-1/2}\delta$ has distribution $N(0, I_p)$ and

$$||\delta||^2_{B\Sigma_\epsilon^{-1}B^T} = \delta^T B \Sigma_\epsilon^{-1} B^T \delta = X^T \Sigma_X^{1/2} B \Sigma_\epsilon^{-1} B^T \Sigma_X^{1/2} X = ||X||^2_{\Sigma_X^{1/2} B \Sigma_\epsilon^{-1} B^T \Sigma_X^{1/2}}$$

Defining the $p \times p$ matrices Γ and $\hat{\Gamma}$ by

$$\Gamma = \Sigma_X^{1/2} B \Sigma_\epsilon^{-1} B^T \Sigma_X^{1/2}$$

$$\hat{\Gamma} = \Sigma_X^{1/2} \hat{B} \hat{\Sigma}_\epsilon^{-1} \hat{B}^T \Sigma_X^{1/2},$$

we have

$$\tilde{L} = 1 - \mathrm{Cor}_{z \sim N(0,I)}(||Z||_\Gamma^2, ||Z||_{\hat{\Gamma}}^2)$$