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

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

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

2.1 Approximating $\tilde{\mathcal{L}}$

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

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

In the appendix we show that

$$\mathrm{Cor}(z^T A z, z^T B z) = \frac{\mathrm{tr}[AB]}{\sqrt{\mathrm{tr}[A^2]\mathrm{tr}[B^2]}}$$

for any positive semidefinite symmetric A, B. Hence

$$\tilde{\mathcal{L}} = 1 - \frac{\mathrm{tr}[\Gamma\hat{\Gamma}]}{\sqrt{\mathrm{tr}[\Gamma^2]\mathrm{tr}[\hat{\Gamma}^2]}}.$$

3 Connection to Frobenius norm estimation

Now we claim that under the condition that $||\Gamma||_F$ is large, the problem of estimation under the loss $\tilde{L}$ reduces to the problem of entrywise mean-squared estimation of Γ, with loss

$$||\Gamma - \hat{\Gamma}||_F^2.$$

Suppose that we have an estimator $\hat{\Gamma}$ with

$$\mathbf{E}||\Gamma - \hat{\Gamma}||_F^2 < r.$$

Then it follows (from Cauchy-Schwarz) that

$$r > \mathbf{E}||\Gamma - \hat{\Gamma}||_F^2 \geq \mathbf{E}(||\Gamma||_F - ||\Gamma||_F)^2,$$

therefore,

$$||\Gamma||_F - \sqrt{r} < ||\hat{\Gamma}||_F < ||\Gamma||_F + \sqrt{r}.$$

Now,

$$
\begin{aligned}
\mathbf{E}\tilde{\mathcal{L}} &= \mathbf{E}\left[1 - \frac{\mathrm{tr}[\Gamma\hat{\Gamma}]}{\sqrt{\mathrm{tr}[\Gamma^2]\mathrm{tr}[\hat{\Gamma}^2]}}\right] \\
&= \mathbf{E}\left[1 - \frac{\mathrm{tr}[\Gamma\hat{\Gamma}]}{||\Gamma||_F||\hat{\Gamma}||_F}\right] \\
&= \frac{1}{2}\mathbf{E}\left\|\frac{\Gamma}{||\Gamma||_F} - \frac{\hat{\Gamma}}{||\hat{\Gamma}||_F}\right\|_F^2 \\
&= \frac{1}{2}\mathbf{E}\left\|\frac{\Gamma - \hat{\Gamma}}{||\Gamma||_F} + (||\Gamma|| - ||\hat{\Gamma}||)\frac{\hat{\Gamma}}{||\hat{\Gamma}||_F}\right\|_F^2 \\
&\leq \frac{1}{2}\left[\mathbf{E}\left[\frac{\|\Gamma - \hat{\Gamma}\|_F^2}{||\Gamma||_F^2}\right] + \mathbf{E}\left[(||\Gamma||_F - ||\hat{\Gamma}||_F)^2\right] + 2\sqrt{\mathbf{E}\left[\left\|\frac{\Gamma - \hat{\Gamma}}{||\Gamma||_F}\right\|^2\right]\mathbf{E}\left[\left\|\frac{(||\Gamma|| - ||\hat{\Gamma}||)\hat{\Gamma}}{||\hat{\Gamma}||_F}\right\|^2\right]}\right] \\
&\leq \frac{1}{2}\left[\frac{r}{||\Gamma||_F^2} + r + \frac{2}{||\Gamma||_F}r\right] \\
&= \left(\frac{||\Gamma||_F^{-2} + 2||\Gamma||_F^{-1} + 1}{2}\right)r
\end{aligned}
$$

i.e. $\mathcal{L}$ is bounded by $\mathbf{E}||\Gamma - \hat{\Gamma}||_F^2$, times a decreasing function of $||\Gamma||_F$.

Meanwhile, root-mean-squared estimation of $\Gamma = \Sigma_X^{1/2}B\Sigma_\epsilon^{-1}B^T\Sigma_X^{1/2}$ can be reduced to estimating B and Σ_ϵ^{-1} separately, with respect to root-mean-squared error.

For simplicity let us first assume $\Sigma_X = 1$. Note the following lemma:

Lemma. *For constant matrices A, B and random matrices $\hat{A}, \hat{B}$,*

$$\mathbf{E}||AB - \hat{A}\hat{B}||_F \leq ||A||_F\mathbf{E}[||B - \hat{B}||_F] + ||B||_F\mathbf{E}[||A - \hat{A}||_F] + \mathbf{E}[||B - \hat{B}||_F||A - \hat{A}||_F]$$

4 Appendix

4.1 Covariance formula

Lemma. *Let $Z \sim N(0, I)$. Then for any PSD A, B we have*

$$Cov(Z^T AZ, Z^T BZ) = 2tr[AB]$$

and

$$Cor(Z^T AZ, Z^T BZ) = \frac{tr[AB]}{\sqrt{tr[A^2]tr[B^2]}}$$

Proof. From Fujikoshi (2010) theorems 2.2.5. and 2.2.6. we have that if $X \sim W_p(n, \Sigma)$

$$\mathbf{E}\mathrm{tr}[AW] = n\mathrm{tr}[A\Sigma]$$

and

$$\mathbf{E}\mathrm{tr}[AWBW] = n\mathrm{tr}[A\Sigma B^T\Sigma] + n\mathrm{tr}[A\Sigma]\mathrm{tr}[B\Sigma] + n^2\mathrm{tr}[A\Sigma B\Sigma]$$

for any $p \times p$ matrices A, B.

Now, since $ZZ^T \sim W_p(1, I_p)$, since A and B are symmetric, we have

$$\begin{aligned}\mathrm{Cov}(Z^T AZ, Z^T BZ) &= \mathbf{E}\mathrm{tr}[AZZ^T BZZ^T] - (\mathbf{E}\mathrm{tr}[AZZ^T])(\mathbf{E}\mathrm{tr}[BZZ^T]) \\ &= 2\mathrm{tr}[AB] + \mathrm{tr}[A]\mathrm{tr}[B] - \mathrm{tr}[A]\mathrm{tr}[B] = 2\mathrm{tr}[AB].\end{aligned}$$

Hence

$$\mathrm{Cor}(Z^T AZ, Z^T BZ) = \frac{\mathrm{Cov}(Z^T AZ, Z^T BZ)}{\sqrt{\mathrm{Cov}(Z^T AZ)\mathrm{Cov}(Z^T BZ)}} = \frac{2\mathrm{tr}[AB]}{\sqrt{2\mathrm{tr}[A^2]2\mathrm{tr}[B^2]}},$$

completing the proof.