Metric estimation for multivariate linear models

Charles Zheng and Yuval Benjamini

October 25, 2015

1 Introduction

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ 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

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

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