Fast Two Sample Testing

Matthew Zhao
Hyppo - Fast Two Sample Testing Notes
Pseudocode

1

Use difference in analytic functions (mean-embeddings) as test statistic:

$$\sqrt{n}\sum_{j=1}^{J}(\hat{\mu}_P(t_j) - \hat{\mu}_Q(t_j))^2$$

Where $\hat{\mu}_P$ and $\hat{\mu}_Q$ are empirical mean-embeddings given by:

$$\frac{1}{n}\sum_{i=1}^{n} k(X_i, \cdot)$$

The former expression converges to a sum of chi-squared random variables. However, it is very difficult to compute. Thus, a parallel is drawn with Hotelling's T^2 statistic, which is distributed as a chi-squared variable with J degrees of freedom:

$$T^2 = W\Sigma^{-1}W$$

Here W is a Gaussian vector with J entries and Σ is the covariance matrix. If we replace W with the difference of normalized mean-embeddings we can define Z_i:

$$Z_i = (k(X_i, T_1) - k(Y_i, T_1), \ldots, k(X_i, T_J) - k(Y_i, T_j))$$

Where T_j are test points and $k(\cdot, \cdot)$ is the Gaussian kernel. If we define the following:

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

$$\Sigma_n = \frac{1}{n} Z Z^T$$

Then, using the Hotelling method, we can construct a new test statistic:

$$S_n = nW_n\Sigma_n^{-1}W_n$$