Notes for Week of 1/28/19 - 2/1/18

Alden Green

January 31, 2019

(WARNING: NOTATION IS NOT WELL-DEFINED. NO GUARANTEE OF CORRECTNESS. THESE ARE INTENDED ONLY AS RECORDS. ALL RELEVANT INFORMATION SHOULD BE TRANSFERRED TO A BETTER-WRITTEN DOCUMENT.)

We work with distributions $\mathbb{P}$ and $\mathbb{Q}$ with support over $\mathcal{D} \subset \mathbb{R}$. Recall that our *Laplacian smooth* test statistic can be written as

$$T_2(\boldsymbol{\ell}; G_{n,r}) = \sup_{\mathbf{f} \in \mathbb{R}^n : \|\mathbf{B}\mathbf{f}\|_2^2 \leq C_{n,r}} \frac{1}{n} \sum_{k=1}^{n} \ell_k f_k$$

for $G_{n,r}$ the neighborhood graph of radius r over data,, and $\mathbf{B} \in \mathbb{R}^{m \times n}$ corresponding incidence matrix.

Consider the set of functions

$$\mathcal{H}^{1,2}(\mathcal{D}, \nu) := \bigl\{ f : \mathcal{D} \to \mathbb{R} |\ f(0) = 0, \text{ and } f \text{ is absolutely continuous with } f' \in L^2(\mathcal{D}) \bigr\}$$

where f' is the derivative of f, and ν is the Lebesgue measure over $\mathbb{R}$. Equip $\mathcal{H}^{1,2}(\mathcal{D}, \nu)$ with the inner product

$$\langle f, g \rangle_{\mathcal{H},\nu}^{1,2} := \int_{\mathcal{D}} f'(z) g'(z) dz,$$

and denote the corresponding norm $\|\cdot\|_{\mathcal{H},\nu}^{1,2}$

Let us consider the test statistic

$$T_2(\boldsymbol{\ell}; \mathcal{W}^{1,2}) = \sup_{f \in \mathcal{W}^{1,2}(\mathcal{D},\nu)} \frac{1}{n} \sum_{k=1}^{n} \ell_k f(z_k)$$

where

$$\mathcal{W}^{1,2}(\mathcal{D}, \nu) = \bigl\{ f \in \mathcal{H}^{1,2}(\mathcal{D}, \nu) : \|f\|_{\mathcal{H}} \leq 1 \bigr\}$$

Hilbert-Sobolev Space. The Hilbert space $\mathcal{H}^{1,2}(\mathcal{D}, \nu)$ equipped with inner product $\langle f, g \rangle^{1,2}_{\mathcal{H},\nu}$ can be shown to be an RKHS with associated kernel

$$K_{\mathcal{H}}(x, z) = \min\{x, z\}$$

As a result, results for two-sample testing in the RKHS setup can be brought to bear. The following is a (slight restatement of) Theorem 8 from [1]. The restatement involved is because we the test statistic we consider is biased, whereas their results are stated only with respect to an bias-corrected test statistic.

Theorem 1 (Theorem 8 of [1]). *Let $m = n/2$. Under $H_0 : \mathbb{P} = \mathbb{Q}$,*

$$m \left| T_2(\boldsymbol{\ell}; \mathcal{W}^{1,2}) \right|^2 \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l [Z_l^2 - 2] + \mathbb{E}_{\mathbb{P}}[Z]$$

where $Z_l \sim N(0, 2)$ i.i.d, λ_i are the solutions to the eigenvalue equation

$$\int_{\mathcal{D}} \tilde{k}(x, x') \chi_i(x) d\mu(x) = \lambda_i \chi_i(x')$$

and $\tilde{k}(x, x') := k(x, x) - \mathbb{E}_{X \sim \mathbb{P}}[k(x, X)] - \mathbb{E}_{X \sim \mathbb{P}}[k(X, x')] + \mathbb{E}_{X, X' \sim \mathbb{P}}[k(X, X')]$.

Note that this is not distribution free, even asymptotically, under the null; the eigenvalues λ_l are solutions to linear equations which depend on μ.

REFERENCES

[1] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.