

Testing with Graphs

Alden Green

January 30, 2019

Two-sample non-parametric hypothesis-testing problem. For fixed integers $n_1 + n_2 = n$, let $\mathbf{X} = \{x_1, \dots, x_{n_1}\} \subset \mathbb{R}^d$ and $\mathbf{Y} = \{y_1, \dots, y_{n_2}\}$ be sampled i.i.d from distributions \mathbb{P} and \mathbb{Q} with density functions p and q , respectively, both with support on $D \subset \mathbb{R}^d$. Our statistical problem is testing the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ vs. the alternative $H_1 : \mathbb{P} \neq \mathbb{Q}$, where our knowledge of \mathbb{P} and \mathbb{Q} come from the samples \mathbf{X} and \mathbf{Y} .

Integral Probability Metric. Given \mathcal{F} a class of real-valued bounded measurable functions on D , the *integral probability metric* between \mathbb{P} and \mathbb{Q} with respect to \mathcal{F} is

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_D f d\mathbb{P} - \int_D f d\mathbb{Q} \right|$$

One such IPM that has not received close attention until surprisingly recently is the (*weighted*) *Sobolev IPM*. For $f \in L^2(\mathcal{D})$ and ρ a density function over \mathcal{D} , the (ρ^2 -weighted) Sobolev 1,2 norm of f is given by

$$\|f\|_{1,2,\rho^2} := \begin{cases} \int_{\mathcal{D}} \|\nabla_x f(x)\|^2 \rho^2 dx, & f \in H^1(\mathcal{D}) \\ \infty, & f \in L^2(\mathcal{D}) \setminus H^1(\mathcal{D}) \end{cases}$$

where $H^1(\mathcal{D})$ is the Sobolev space of $L^2(\mathcal{D})$ functions with weak derivative $\nabla_x f(x) \in L^2(\mathcal{D})$.

Consider the unit ball of $\|\cdot\|_{1,2,\rho^2}$,

$$\mathcal{W}^{1,2}(\mathcal{D}, \rho^2) := \left\{ f : \|f\|_{1,2,\rho^2} \leq 1 \right\}.$$

The weighted Sobolev IPM is simply $\gamma_{\mathcal{W}^{1,2}(\mathcal{D}, \rho^2)}(\mathbb{P}, \mathbb{Q})$.

Neighborhood graph. Let $G_{n,r_n} = (V, E, w)$ denote a weighted, undirected graph constructed from the samples $\mathbf{Z} = \{z_1, \dots, z_n\} = (\mathbf{X}, \mathbf{Y})$ where $V = \{1, \dots, n\}$, and $w_{uv} = K(z_u, z_v) := k\left(\frac{\|z_u - z_v\|}{\epsilon_n}\right) \geq 0$ for $u, v \in V$, and a particular kernel function k . Here $(u, v) \in E$ if and only if $w_{uv} > 0$.

Motivated by the integral probability metric, our test statistics will be of the form

$$\gamma_{\mathcal{F}_n}(\mathbb{P}_n, \mathbb{Q}_n) = \sup_{f_n \in \mathcal{F}_n} \left| \int_D f_n d\mathbb{P}_n - \int_D f_n d\mathbb{Q}_n \right|$$

where

$$\mathbb{P}_n := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{x_i}, \quad \mathbb{Q}_n := \frac{1}{n_2} \sum_{i=1}^{n_2} \delta_{y_i}$$

are the empirical distributions of \mathbf{X} and \mathbf{Y} , respectively, and \mathcal{F}_n is a **class of functions** $f_n : \mathbf{Z} \rightarrow \mathbb{R}$ **exhibiting some regularity with respect to the neighborhood graph** $G_{n,r}$.

Laplacian smoothing and Total Variation denoising. For convenience, we number the edges $E = (e_1, \dots, e_m)$. We denote by $\mathbf{B} \in \mathbb{R}^{m \times n}$ the edge incidence matrix of $G_{n,r}$, which for k th edge $e_k = (u, v)$ has k th row $\mathbf{B}_k = (0, \dots, -w_{uv}, \dots, w_{uv}, \dots, 0)$ with a $-w_{uv}$ in the u th location, and a w_{uv} in the v th location. The random (unnormalized) Laplacian matrix is then $\mathbf{L} = \mathbf{B}^T \mathbf{B}$. We also introduce a *label vector*, given by $\boldsymbol{\ell} = (\ell_1, \dots, \ell_n)$ with

$$\ell_k = \begin{cases} \frac{n}{n_1}, & z_k \in \mathbf{X} \\ -\frac{n}{n_2}, & z_k \in \mathbf{Y} \end{cases} \quad (1)$$

Our test statistics $T_1(\boldsymbol{\ell}; G_{n,r})$ and $T_2(\boldsymbol{\ell}; G_{n,r})$ are defined as follows:

$$T_1(\boldsymbol{\ell}; G_{n,r}) := \sup_{\mathbf{f} \in \mathbb{R}^n : \|\mathbf{B}\mathbf{f}\|_1 \leq C_{n,r}} \frac{1}{n} \sum_{k=1}^n \ell_k f_k$$

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

where $C_{n,r} =$ and we write $\mathbf{f} = (f_1, \dots, f_n)$. We note that, as promised, these satisfy the form

$$T_1(\boldsymbol{\ell}; G_{n,r}) = \sup_{f_n \in TV_n} \left| \int_D f_n d\mathbb{P}_n - \int_D f_n d\mathbb{Q}_n \right|$$

$$T_2(\boldsymbol{\ell}; G_{n,r}) = \sup_{f_n \in \mathcal{W}_n} \left| \int_D f_n d\mathbb{P}_n - \int_D f_n d\mathbb{Q}_n \right|$$

where $TV_n = \{\mathbf{f} : \|\mathbf{B}\mathbf{f}\|_1 \leq C_{n,r}\}$ and $\mathcal{W}_n = \{\mathbf{f} : \|\mathbf{B}\mathbf{f}\|_2 \leq C_{n,r}\}$.

1 Consistency under fixed alternative

Binomialized data model. For technical reasons, we would like z_1, \dots, z_n to be independent and identically distributed. We consider the following generative model, which we term the *binomialized data model*:

Fix $n \in \mathbb{N} > 0$, and $n_1 \sim \text{Bin}(n, 1/2)$, $n_2 = n - n_1$. Then, let $x_1, \dots, x_{n_1} \in \mathbb{R}^d$ be a sequence of i.i.d random points chosen according to \mathbb{P} , and $y_1, \dots, y_{n_2} \in \mathbb{R}^d$ a separate sequence of i.i.d random points chosen according to \mathbb{Q} , with $x_j \perp y_k$ for all j, k . Fix $\tilde{\mathbf{Z}} = (\tilde{z}_1, \dots, \tilde{z}_n) := (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2})$. Finally, for a permutation $\pi : [n] \rightarrow [n]$ chosen uniformly at random among all such permutations, let $\mathbf{Z} = (z_1, \dots, z_n) = (\tilde{z}_{\pi(1)}, \dots, \tilde{z}_{\pi(n)})$.

The label vector ℓ remains defined as in (1) with respect to \mathbf{Z} . Note that now $z_i \stackrel{i.i.d}{\sim} \frac{\mathbb{P}}{2} + \frac{\mathbb{Q}}{2}$, as we desired, with density function $\mu(x) := \frac{p(x)+q(x)}{2}$.

Theorem 1. *Let $d \geq 2$ and let $\mathcal{D} \subset \mathbb{R}^d$ be an open, bounded, connected set with Lipschitz boundary. Let μ satisfy*

$$m \leq \mu(x) \leq M \quad (\forall x \in D)$$

for some $0 < m \leq M$. Let (r_n) be a sequence of positive numbers converging to 0 and satisfying

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{(\log n)^{3/4}}{n^{1/2}} \frac{1}{r_n} &= 0 \quad \text{if } d = 2 \\ \lim_{n \rightarrow \infty} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{r_n} &= 0 \quad \text{if } d \geq 3 \end{aligned}$$

Assume the kernel k satisfies conditions:

$$k(0) > 0 \text{ and } k \text{ is continuous at } 0. \quad (\mathbf{K1})$$

$$k \text{ is non-increasing.} \quad (\mathbf{K2})$$

$$\text{The integral } \int_0^\infty k(r) r^{d+1} dr \text{ is finite.} \quad (\mathbf{K3})$$

Then with probability one the following statement holds: For (z_1, \dots, z_n) chosen under the binomialized data model

$$\lim_{n \rightarrow \infty} T_2(\ell; G_{n, r_n}) = \gamma_{\mathcal{W}^{1,2}(\mathcal{D}, \mu^2)}(\mathbb{P}, \mathbb{Q}).$$

2 Proofs

Fix $\mu_n = \frac{d\mathbb{P}_n + d\mathbb{Q}_n}{2}$. We will show a variational form of convergence of μ_n to μ .

2.1 Gamma convergence of constraint

Definition 2.1 (TL^2 convergence). Denote by $\mathfrak{B}(\mathcal{D})$ the Borel σ -algebra of \mathcal{D} and $\mathcal{P}(\mathcal{D})$ the set of all Borel probability measures on \mathcal{D} . Given a Borel map $T : \mathcal{D} \rightarrow \mathcal{D}$, the push-forward of μ by T is given by

$$T_*\mu(\mathcal{A}) := \mu(T^{-1}(\mathcal{A})), \quad \mathcal{A} \in \mathfrak{B}(\mathcal{D}).$$

Given $\tilde{\mu} \in \mathcal{P}(\mathcal{D})$, we say that T is a *transportation map* between μ and $\tilde{\mu}$ if $T_*\mu = \tilde{\mu}$. If for a sequence (T_n) of transportation maps

$$\int_{\mathcal{D}} |x - T_n(x)|^2 d\mu(x) \rightarrow 0, \text{ as } n \rightarrow \infty$$

we refer to the sequence as *stagnating*.

Take $f \in L^2(\mu)$ and a sequence (f_n) with $f_n \in L^2(\mu_n)$ for $n = 1, 2, \dots$. If there exists a stagnating sequence of transportation maps (T_n) such that

$$\int_{\mathcal{D}} |f - T_n(f_n(x))|^2 d\mu(x) \rightarrow 0, \text{ as } n \rightarrow \infty$$

we say that (f_n) converges TL^2 to f , and write $f_n \xrightarrow{TL^2} f$.

We restate Theorem 1.4 of [1], changing notation to match the rest of this paper.

Theorem 2 (Theorem 1.4 of [1]). *Under the setup and conditions of Theorem 1, with probability one the following statements hold:*

- Liminf inequality: *For all $f \in L^2(\mu)$ and all sequences (f_n) with $f_n \in L^2(\mu_n)$ and $f_n \xrightarrow{TL^2} f$,*

$$\liminf_{n \rightarrow \infty} \frac{1}{n^2 r_n^{d+2}} \|Bf\|_2^2 \geq \sigma_k \|f\|_{1,2,\rho^2}$$

- Limsup inequality: *For all $f \in L^2(\mu)$, there exists a sequence (f_n) with $f_n \in L^2(\mu_n)$ and $f_n \xrightarrow{TL^2} f$ such that*

$$\limsup_{n \rightarrow \infty} \frac{1}{n^2 r_n^{d+2}} \|Bf\|_2^2 \leq \sigma_k \|f\|_{1,2,\rho^2}$$

- Compactness property: *Every sequence (f_n) with $f_n \in L^2(\mu_n)$ satisfying*

$$\sup_{n \in \mathbb{N}} \frac{1}{n^2 r_n^{d+2}} \|Bf\|_2^2 < \infty$$

is precompact in TL^2 , that is, every subsequence of (f_n) has a further subsequence which converges in the TL^2 -sense to an element of $L^2(\mathcal{D})$.

2.2 Continuity of risk functional

2.3 Proof of Theorem 1

REFERENCES

- [1] Nicolas Garcia Trillos and Dejan Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018.