Locally-Adaptive Kernel Tests*

Zoltán Szabó, Éric Moulines

CMAP, École Polytechnique

Quick summary

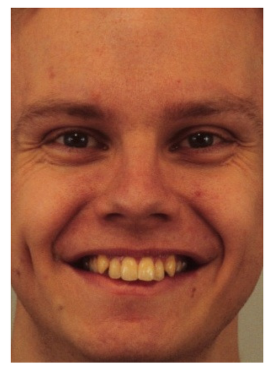
- Focus:
  - hypothesis testing $\xrightarrow{\text{example}}$ 2-sample testing.
  - Check if $\mathbb{P} = \mathbb{Q}$ based on samples.
- Fast analytic kernel-based tests [1]:
  - Adaptivity [3]: parameters optimized for 'power'. $\mathbb{P}$, $\mathbb{Q}$: fixed.
- Challenge:
  - $\mathbb{P}$: fixed. Sequence of alternatives: $\mathbb{Q}_n \xrightarrow{n\to\infty} \mathbb{P}$.
  - Adaptivity: realizable? Objective function?

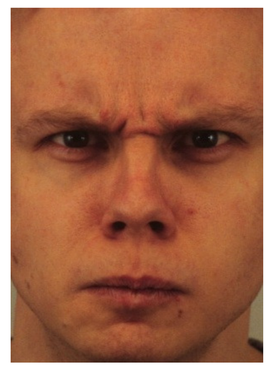
Two-sample testing

- Given: $X = \{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Task: using X, Y test

$$H_0 : \mathbb{P} = \mathbb{Q}, \text{ vs } H_1 : \mathbb{P} \neq \mathbb{Q}.$$

- Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.





- Challenge (intuition): difference in emotions $\to 0$.

Test power

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Decision: H_0 is rejected if $\hat{\lambda}_n$ is 'large'.
- $P(\text{we say } H_1 \,|\, H_0) \leq 1 - \alpha$. Typically $\alpha = 0.01$.
- Power = 1 - $\underbrace{P(H_0 \text{ is accepted} \,|\, H_1)}_{\text{Type-II error}} \to \max$.

Representation of distributions

- Mean embedding:

$$\mu_{\mathbb{P}} := \underbrace{\int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x)}_{\text{Bochner integral}} \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$MMD(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

In two-sample testing [2]:

- $\hat{\lambda}_n = \widehat{MMD}(\mathbb{P}_n, \mathbb{Q}_n)$. Computation: $\mathcal{O}(n^2)$.

Metric with analytic kernels

- Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$

$$\rho(\mathbb{P}, \mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^J [\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

$\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J$: random locations. It is metric a.s. [1].

- Plug-in estimate, $\mathcal{O}(n)$-time

$$\hat{\rho}(\mathbb{P}, \mathbb{Q}) = \frac{\bar{\mathbf{z}}_n^T \bar{\mathbf{z}}_n}{J}, \quad \bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^n \underbrace{[k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j)]_{j=1}^J}_{=:\mathbf{z}(\mathbf{x}_i, \mathbf{y}_i)}.$$

- Modified test statistic, χ_J^2 null:

$$\hat{\lambda}_n = n \bar{\mathbf{z}}_n^T \mathbf{\Sigma}_n^{-1} \bar{\mathbf{z}}_n, \qquad \mathbf{\Sigma}_n = \widehat{cov}\left(\{\mathbf{z}(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n\right).$$

Optimize for test power

- Power proxy [3]:

$$\lambda_n = n\mathbf{m}^T \mathbf{\Sigma}^{-1} \mathbf{m}, \ \mathbf{m} = \mathbb{E}_{\mathbf{xy}}[\mathbf{z}(\mathbf{x}, \mathbf{y})], \ \mathbf{\Sigma} = cov_{\mathbf{xy}}[\mathbf{z}(\mathbf{x}, \mathbf{y})].$$

- Objective function:

$$(k^*, \mathcal{V}^*) := \arg\max_{k, \mathcal{V}} \mathbf{m}^T \mathbf{\Sigma}^{-1} \mathbf{m}.$$

Locally-adaptive test

- Instead of fixed $\mathbb{P}$, $\mathbb{Q}$:

$$\mathbb{P}, \qquad \mathbb{Q}_n = (1 - \alpha_n)\mathbb{Q} + \alpha_n \mathbb{P} \xrightarrow{n\to\infty} \mathbb{P}, \quad \alpha_n \to 1.$$

- By linearity: with $\alpha_n = 1 - \frac{1}{\sqrt{n}}$

$$\mathbf{r}_{\mathbb{Q}_n} - \mathbf{r}_{\mathbb{P}} = \frac{\boldsymbol{\delta}}{\sqrt{n}}, \qquad \boldsymbol{\delta} = \mathbf{r}_{\mathbb{Q}-\mathbb{P}}, \qquad \mathbf{r}_{\mathbb{M}} := [\mu_{\mathbb{M}}(\mathbf{v}_j)]_{j=1}^J.$$

- Assume: $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, $\{\mathbf{y}_{i,n}\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}_n$. Let

$$\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}, \mathbf{y}\sim\mathbb{P}}\left[\mathbf{z}(\mathbf{x}, \mathbf{y})\mathbf{z}^T(\mathbf{x}, \mathbf{y})\right],$$

$$\hat{\mathbf{\Sigma}}_n = \widehat{cov}(\mathbf{z}(\mathbf{x}_i, \mathbf{y}_{i,n})), \ \bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{z}(\mathbf{x}_i, \mathbf{y}_{i,n}).$$

- After some algebra: $\theta := (k, \mathcal{V})$

$$n\bar{\mathbf{z}}_n^T \hat{\mathbf{\Sigma}}_n^{-1} \bar{\mathbf{z}}_n \xrightarrow{w} \chi_J^2\Big(\underbrace{\boldsymbol{\delta}^T \mathbf{\Sigma}^{-1} \boldsymbol{\delta}}_{=:\lambda(\boldsymbol{\delta}, \theta)}\Big).$$

- $\Rightarrow$ Power against $(\mathbb{Q}_n)_{n=1}^\infty$ goes to

$$\beta(\boldsymbol{\delta}, \theta) = P\left(\chi_J^2(\lambda) \geq q\right) = M_{\frac{J}{2}}\left(\sqrt{\lambda}, q\right).$$

Objective

Maximize the worst-case local power,

$$(k^*, \mathcal{V}^*) := \arg\max_{\theta} \min_{\boldsymbol{\delta}} \lambda(\boldsymbol{\delta}, \theta).$$

References

[1] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pages 1972–1980, 2015.

[2] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

[3] W. Jitkrittum, Z. Szabó, K. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *NIPS*, pages 181–189, 2016.

*Data Learning and Inference (DALI), Tenerife, Canary Islands, Spain, 17-20 April 2017. Acknowledgements: This work was supported by the Data Science Initiative.