Optimal Kernel Choice for Large-Scale Two-Sample Tests

Arthur Gretton,[1,3] Bharath Sriperumbudur,[1] Dino Sejdinovic,[1] Heiko Strathmann[2], Sivaraman Balakrishnan[4], Massimiliano Pontil[2], Kenji Fukumizu[5]

[1]Gatsby Unit and [2]CSD, CSML, UCL, UK; [3]MPI for Intelligent Systems, Germany; [4] Carnegie Mellon University; [5] Institute of Statistical Mathematics

Introduction

- Given:
  - m samples $\mathbf{X} := \{x_1, \dots, x_m\}$ drawn i.i.d. from $\mathbf{P}$
  - samples $\mathbf{Y}$ drawn from $\mathbf{Q}$
- Determine: are $\mathbf{P}$ and $\mathbf{Q}$ different?
- Kernel test based on MMD:
  - High dimensionality
  - Low sample size
  - Structured data (strings and graphs)
- How to choose the kernel?
  - Maximize the MMD with constraint on kernel class?
  - Cross validation based on classification view?
  - Explicitly maximize test power

Linear time vs quadratic time statistic:

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence...
- ...a much less powerful test for a given m

Advantages of the linear time MMD vs quadratic time MMD

- Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of χ^2)
- Both test statistic and threshold computable in $O(m)$, with storage $O(1)$.
- With enough data, a given Type II error can be attained with less computation

Definition of the MMD

Functions revealing difference in distribution:

- Idea: avoid density estimation when testing $\mathbf{P} \neq \mathbf{Q}$ [Fortet and Mourier, 1953]

$$D(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} f(\mathbf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathbf{y}) \right].$$

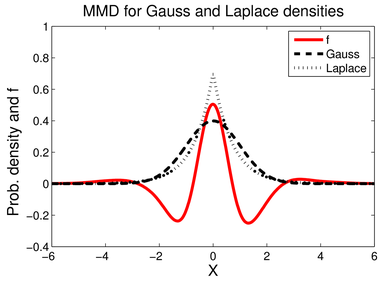
- $D(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq 1\}$ the unit ball in a characteristic RKHS $\mathcal{F}$ [Fukumizu et al., 2008, Sriperumbudur et al., 2008, Fukumizu et al., 2009]
  - These include Gaussian, Laplace on $\mathbb{R}^d$, universal kernels on compact domains [Steinwart, 2001] .

The (kernel) MMD:

$$\begin{aligned}
&\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) \quad := \quad (D(\mathbf{P}, \mathbf{Q}; F))^2 \\
&\underset{(1)}{=} \left(\sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2 \\
&\underset{(2)}{=} \|\mu_x - \mu_y\|^2_{\mathcal{F}} \\
&= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}} \\
&= \mathbf{E}_{\mathbf{P},\mathbf{P}'} k(\mathbf{x}, \mathbf{x}') + \mathbf{E}_{\mathbf{Q},\mathbf{Q}'} k(\mathbf{y}, \mathbf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathbf{x}, \mathbf{y}),
\end{aligned}$$

- $\mathbf{x}'$ is a R.V. independent of $\mathbf{x}$ with distribution $\mathbf{P}$
- $\mathbf{y}'$ is a R.V. independent of $\mathbf{y}$ with distribution $\mathbf{Q}$.

Illustration: Gauss vs Laplace



Results used:

1. Mean elements corresponding to $\phi(\mathbf{x})$ and $\phi(\mathbf{y})$:

$$\begin{aligned}
\mathbf{E}_{\mathbf{P}}(f(\mathbf{x})) &= \mathbf{E}_{\mathbf{P}} \left[\langle \phi(\mathbf{x}), f \rangle_{\mathcal{F}} \right] =: \langle \mu_x, f \rangle_{\mathcal{F}}, \\
\mathbf{E}_{\mathbf{Q}}(f(\mathbf{y})) &= \mathbf{E}_{\mathbf{Q}} \left[\langle \phi(\mathbf{y}), f \rangle_{\mathcal{F}} \right] =: \langle \mu_y, f \rangle_{\mathcal{F}}.
\end{aligned}$$

2. The norm is also written as

$$\|\mu\|_{\mathcal{F}} = \sup_{f \in F} \langle f, \mu \rangle_{\mathcal{F}}$$

Linear time unbiased estimate:

$$\check{\eta}_k = \frac{2}{m} \sum_{i=1}^{m/2} h_k(v_i),$$

- $v_i := [x_{2i-1},\ x_{2i},\ y_{2i-1},\ y_{2i}]$
- $h_k(v_i) := k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(x_{2i}, y_{2i-1})$

Idea:

$$\begin{aligned}
\widehat{\mathbf{E}}_{\mathbf{P}} k(x, x') &= \frac{m}{2} \left[k(x_1, x_2) + k(x_3, x_4) + \dots \right] \\
&= \frac{m}{2} \sum_{i=1}^{m/2} k(x_{2i-1}, x_{2i})
\end{aligned}$$

Hypothesis test, optimal kernel choice

Asymptotic distribution when $\mathbf{P} = \mathbf{Q}$:

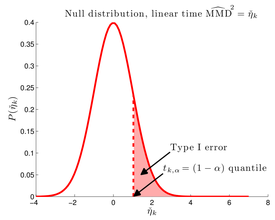
By central limit theorem, whether $\mathbf{P} \neq \mathbf{Q}$ or $\mathbf{P} = \mathbf{Q}$,

$$m^{1/2} \left(\check{\eta}_k - \eta_k(p, q) \right) \xrightarrow{D} \mathcal{N}(0, 2\sigma_k^2)$$

- assuming $0 < \mathbf{E}(h_k^2) < \infty$ (true for bounded k)
- $\sigma_k^2 = \mathbf{E}_v h_k^2(v) - \left[\mathbf{E}_v (h_k(v)) \right]^2$.

Hypothesis test of asymptotic level α:

$$t_{k,\alpha} = m^{-1/2} \sigma_k \sqrt{2} \Phi^{-1}(1 - \alpha)$$



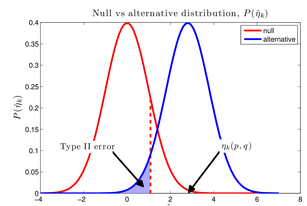
Optimal kernel choice:

Type II error: $\check{\eta}_k$ falls below $t_{k,\alpha}$ and $\eta_k(p, q) > 0$.
Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi \left(\Phi^{-1}(1 - \alpha) - \frac{\eta_k(p, q)\sqrt{m}}{\sigma_k \sqrt{2}} \right)$$

Since Φ monotonic, best kernel choice to minimize Type II error prob. is:

$$k_* = \arg\max_{k \in \mathcal{K}} \eta_k(p, q) \sigma_k^{-1},$$



Kernel class: linear combinations

Define the family of kernels as follows:

$$\mathcal{K} := \left\{ k : k = \sum_{u=1}^{d} \beta_u k_u,\ \|\beta\|_1 = D,\ \beta_u \geq 0,\ \forall u \in \{1, \dots, d\} \right\}.$$

The squared MMD becomes

$$\eta_k(p, q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p, q),$$

where $\eta_u(p, q) := \mathbf{E}_v h_u(v)$.
Denote:

- $\beta = (\beta_1, \beta_2, \dots, \beta_d)^\top \in \mathbb{R}^d$,
- $h = (h_1, h_2, \dots, h_d)^\top \in \mathbb{R}^d$,
- $\eta = \mathbf{E}_v(h) = (\eta_1, \eta_2, \dots, \eta_d)^\top \in \mathbb{R}^d$.

Quantities for test:

$$\eta_k(p, q) = \mathbf{E}(\beta^\top h) = \beta^\top \eta \qquad \sigma_k^2 := \beta^\top \mathrm{cov}(h) \beta.$$

Kernel optimization

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I \right) \beta},$$

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \hat{\sigma}_k$ on data to be tested

Objective:

$$\begin{aligned}
\hat{\beta}^* &= \arg\max_{\beta \succeq 0} \hat{\eta}_k(p, q) \hat{\sigma}_{k,\lambda}^{-1} \\
&=: \alpha(\beta; \hat{\eta}, \hat{Q})
\end{aligned}$$

Assume: $\hat{\eta}$ has at least one positive entry

Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier quadratic program for β^*:

$$\min\{\beta^\top \left(\hat{Q} + \lambda_m I \right) \beta : \beta^\top \hat{\eta} = 1,\ \beta \succeq 0\}$$

Consistency of kernel selection: Assume bounded kernel, σ_k, bounded away from 0. If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P \left(m^{-1/3} \right).$$

Proof idea:

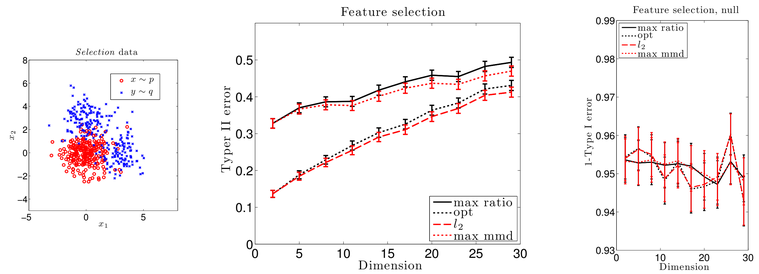
$$\begin{aligned}
&\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| \leq \sup_{k \in \mathcal{K}} \left| \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \eta_k \sigma_{k,\lambda}^{-1} \right| + \sup_{k \in \mathcal{K}} \left| \eta_k \sigma_{k,\lambda}^{-1} - \eta_k \sigma_k^{-1} \right| \\
&\leq \frac{\sqrt{d}}{D \sqrt{\lambda_m}} \left(C_1 \sup_{k \in \mathcal{K}} |\hat{\eta}_k - \eta_k| + C_2 \sup_{k \in \mathcal{K}} \left| \hat{\sigma}_{k,\lambda} - \sigma_{k,\lambda} \right| \right) + C_3 D^2 \lambda_m.
\end{aligned}$$

Experiments

Key to plots:

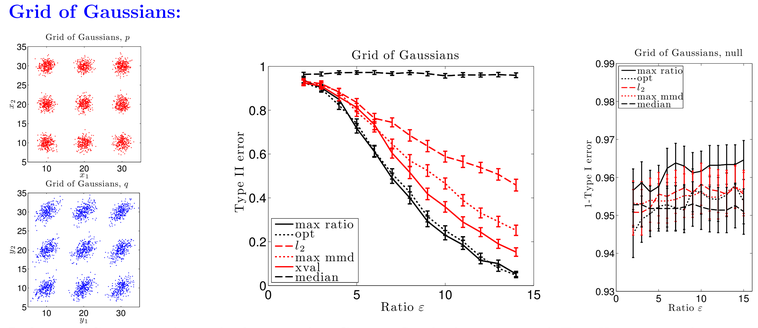
opt - kernel from $\mathcal{K}$ that maximizes $\hat{\eta}_k / \hat{\sigma}_{k,\lambda}$; ***max-ratio*** - single base kernel k_u with largest $\hat{\eta}_u / \hat{\sigma}_{u,\lambda}$; ***max-mmd*** - single base kernel k_u with largest $\hat{\eta}_u$; ℓ_2 - maximizes $\hat{\eta}_k$ subject to $\|\beta\|_2 \leq 1$; ***xval***, kernel from $\{k_u\}_{u=1}^d$ chosen via five-fold cross-validation [Sugiyama et al., 2011].
Asymptotic test level was $\alpha = 0.05$.

Feature selection:



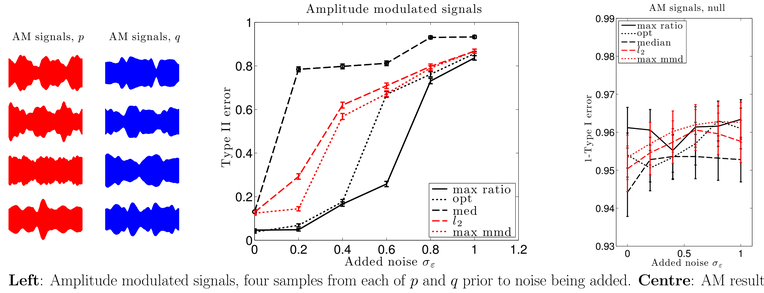
Left: Feature selection data for 2 dimensions (for higher d, remaining dimensions filled with i.i.d. zero mean Gaussian noise). **Centre:** Feature selection results, Type II error vs number of dimensions, average over 5000 trials, $m = n = 10^4$. **Right:** Type I error.

Grid of Gaussians:



Left: 3×3 Gaussian grid, samples from p and q. **Centre:** Results for a 5×5 grid, Type II error vs ε, the eigenvalue ratio for the covariance of the Gaussians in q; average over 1500 trials, $m = n = 10^4$. **Right**: Type I error.

Amplitude modulated signals:



Left: Amplitude modulated signals, four samples from each of p and q prior to noise being added. **Centre**: AM results, Type II error vs added noise, average over 5000 trials, $m = n = 10^4$. **Right**: Type I error.

References

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS 20*, pages 489–496, 2008.

K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In *NIPS 21*, pages 473–480, 2009.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. In *COLT 21*, pages 111–122, 2008.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2001.

M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.