Distinguishing Distributions with Interpretable Features

Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, Arthur Gretton

Gatsby Unit, University College London

Summary

- Two semimetrics, ME and SCF, on distributions are based on the differences of analytic functions evaluated at spatial or frequency locations (i.e., **features**).
- **Proposal**: choose the features so as to maximize the distinguishability of the distributions, by optimizing a lower bound on test power for a statistical test using these features.
- **Result**: powerful, linear-time, nonparametric, interpretable two-sample test. **Performance comparable to the quadratic-time MMD test**.

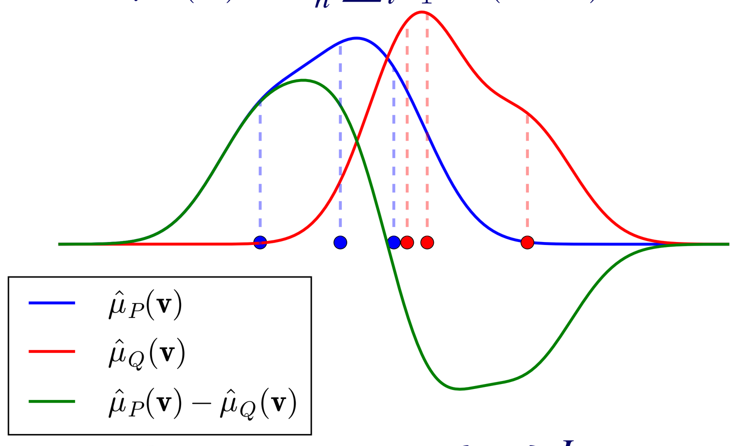
ME and SCF Tests

- Observe $\mathsf{X} := \{\mathbf{x}_i\}_{i=1}^n \sim P$ and $\mathsf{Y} := \{\mathbf{y}_i\}_{i=1}^n \sim Q$ in $\mathbb{R}^d$.
- Test $H_0 : P = Q$ v.s. $H_1 : P \neq Q$. Calculate a statistic $\hat{\lambda}_n$, and reject H_0 if $\hat{\lambda}_n > T_\alpha = (1-\alpha)$-quantile of the null distribution.

Mean Embedding (ME) Test:

$$\hat{\mu}_P(\mathbf{v}) := \frac{1}{n}\sum_{i=1}^n k_\sigma(\mathbf{x}_i, \mathbf{v})$$

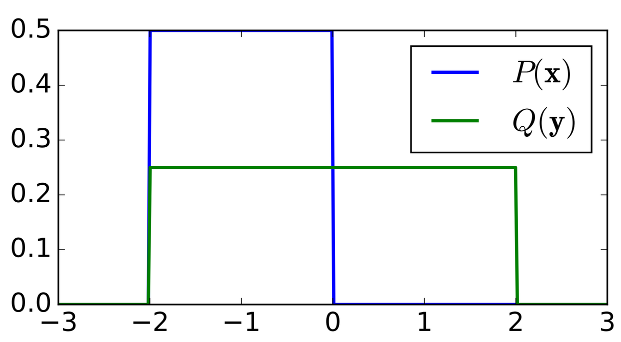
Test statistic: $\hat{\lambda}_n := n\bar{\mathbf{z}}_n^\top(\mathbf{S}_n + \gamma_n I)^{-1}\bar{\mathbf{z}}_n$,

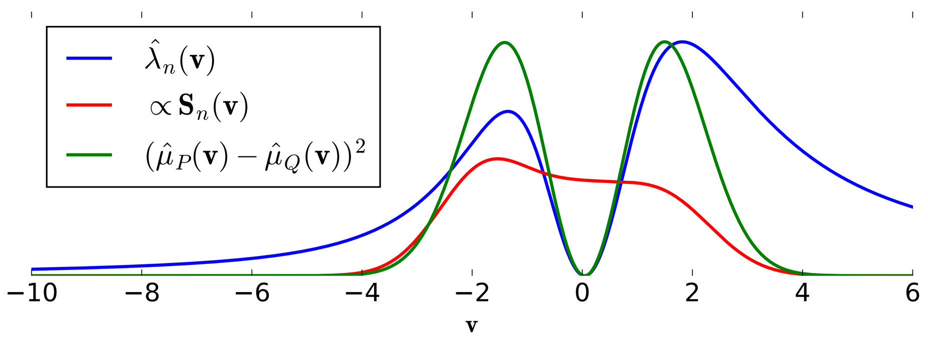


- $\bar{\mathbf{z}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$,
- $\mathbf{S}_n := \frac{1}{n-1}\sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}}_n)(\mathbf{z}_i - \bar{\mathbf{z}}_n)^\top$,
- $\mathbf{z}_i := (k_\sigma(\mathbf{x}_i, \mathbf{v}_j) - k_\sigma(\mathbf{y}_i, \mathbf{v}_j))_{j=1}^J \in \mathbb{R}^J$,
- γ_n is a regularizer.
- Need a positive definite kernel k_σ, and spatial features $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J$.

Difference to MMD's Witness Function

- $(\hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v}))^2 = \bar{\mathbf{z}}_n(\mathbf{v})^2$. Variance $\mathbf{S}_n(\mathbf{v})$ is high in overlapping regions.





Smooth Characteristic Function (SCF) Test:

$$\mathbf{z}_i := [\hat{l}_\sigma(\mathbf{x}_i)\exp(i\mathbf{x}_i^\top\mathbf{v}_j) - \hat{l}_\sigma(\mathbf{y}_i)\exp(i\mathbf{y}_i^\top\mathbf{v}_j)]_{j=1}^J \in \mathbb{R}^{2J}$$

- Check the difference of smoothed (by l_σ) characteristic functions.
- Need an analytic smoothing kernel l_σ, and frequency features $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J$.
- Both tests are consistent. Under H_0, $\hat{\lambda}_n$ asymptotically follow $\chi^2(\dim(\bar{\mathbf{z}}_n))$.

Test Power Lower Bound

Proposition. *The power $\mathbb{P}_{H_1}(\hat{\lambda}_n \geq T_\alpha)$ of the ME test is at least*

$$L(\lambda_n) = 1 - 2e^{-\frac{(\lambda_n - T_\alpha)^2}{3^2 \cdot 8B^2\bar{c}_2^2 Jn}} - 2e^{-\frac{(\gamma_n(\lambda_n - T_\alpha)(n-1) - 24B^2\bar{c}_1 Jn)^2}{3^2 \cdot 32B^4\bar{c}_1^2 J^2 n(2n-1)^2}} - 2e^{-\frac{((\lambda_n - T_\alpha)/3 - \bar{c}_3 n\gamma_n)^2\gamma_n^2}{32B^4J^2\bar{c}_1^2 n}}.$$

For large n, $L(\lambda_n)$ is increasing in λ_n.

- $\bar{c}_1, \bar{c}_2$ and $\bar{c}_3$ are constants. B bounds the kernel k pointwise.
- $\lambda_n := n\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is the population counterpart of $\hat{\lambda}_n$.
- $\boldsymbol{\mu} = \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1]$ and $\boldsymbol{\Sigma} = \mathbb{E}_{\mathbf{xy}}[(\mathbf{z}_1 - \boldsymbol{\mu})(\mathbf{z}_1 - \boldsymbol{\mu})^\top]$.

Proposal: Optimize $\mathcal{V}, \sigma = \arg\max_{\mathcal{V},\sigma} L(\lambda_n) = \arg\max_{\mathcal{V},\sigma} \lambda_n$.

- λ_n unknown. Use $\hat{\lambda}_{n/2}^{tr}$ instead (computed on a separate training set).

Theorem (convergence rate): If $\gamma_n = \mathcal{O}(n^{-1/4})$, then

$$\left|\sup_{\mathcal{V},k} \bar{\mathbf{z}}_n^\top(\mathbf{S}_n + \gamma_n I)^{-1}\bar{\mathbf{z}}_n - \sup_{\mathcal{V},k} \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right| = \mathcal{O}_p(n^{-1/4}),$$

implying that the objective converges as $n \to \infty$.

We thank the Gatsby Charitable Foundation for the financial support.

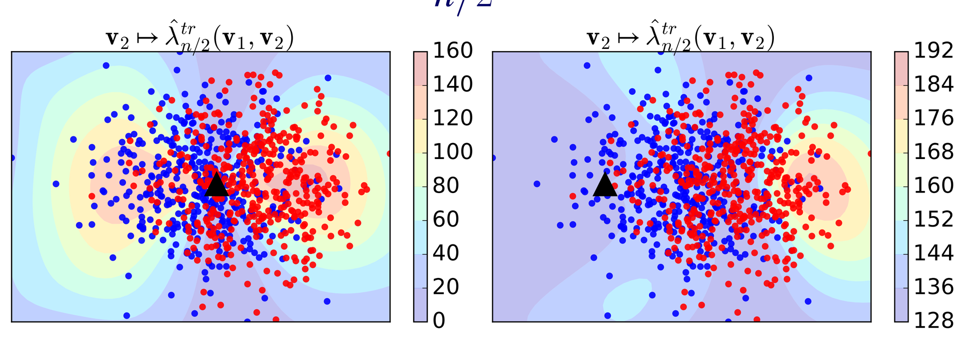
Contact: wittawat@gatsby.ucl.ac.uk

Code: github.com/wittawatj/interpretable-test

Paper: http://arxiv.org/abs/1605.06796

Informative Features

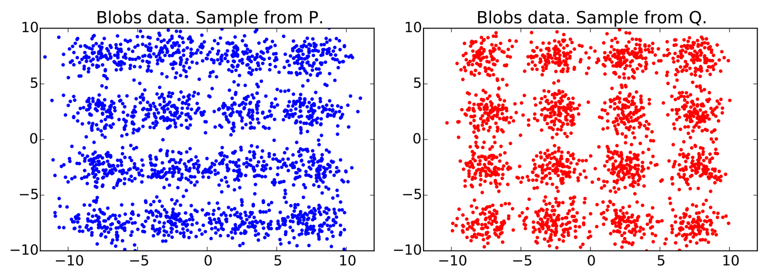
- Contour plot of $\hat{\lambda}_{n/2}^{tr}$ as a function of $\mathbf{v}_2$ when $J = 2$. $\mathbf{v}_1$ fixed at ▲.



- $\hat{\lambda}_{n/2}^{tr}$ is high in the regions that reveal the difference.
- Nonconvexity indicates many ways to detect the differences.

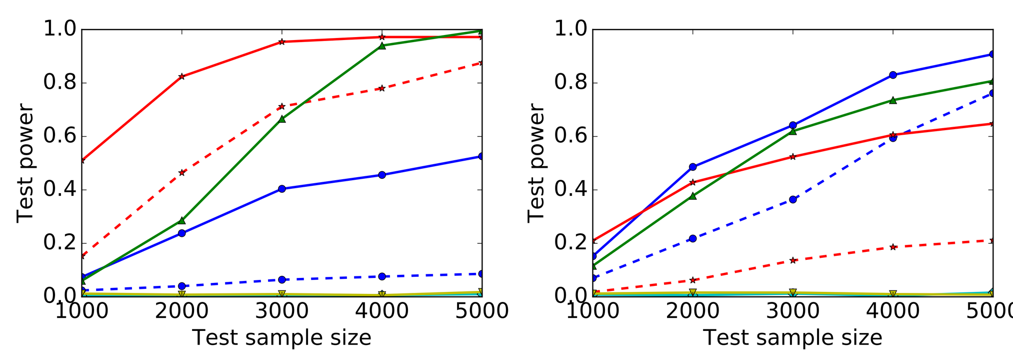
Test Power vs. n and d

Problem	P	Q
SG	$\mathcal{N}(\mathbf{0}_d, I_d)$	$\mathcal{N}(\mathbf{0}_d, I_d)$
GVD	$\mathcal{N}(\mathbf{0}_d, I_d)$	$\mathcal{N}(\mathbf{0}_d, \mathrm{diag}(2, 1, \ldots, 1))$
Blobs	Mixture of 16 Gaussians in $\mathbb{R}^2$. See →	



- Use Gaussian kernel $k_\sigma(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2)$.
- ME-full, SCF-full = Proposed methods with full optimization. $J = 5$.
- ME-grid, SCF-grid = Fixed $\mathcal{V}$. Optimize kernel parameter σ.
- MMD-quad, MMD-lin = Quadratic and linear-time MMD tests.

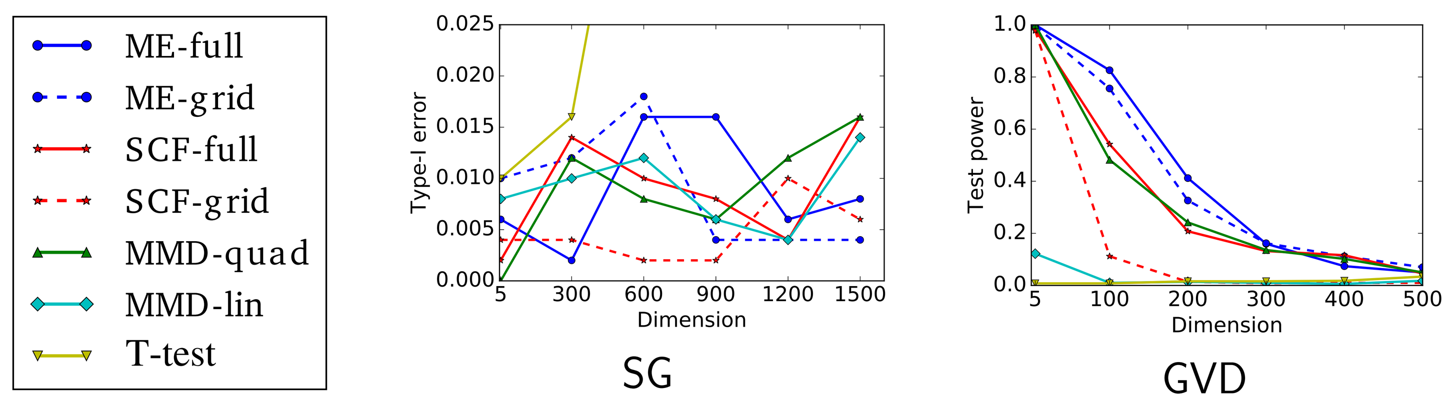
$\mathbb{P}(\text{reject } H_0)$ vs. test sample size. 500 trials. $\alpha = 0.01$.



Blobs. GVD. $d = 50$.

- Blobs: Best performance by SCF-full.
- GVD: Best performance by ME-full.

$\mathbb{P}(\text{reject } H_0)$ vs. dimension d. $n = 10000$.



SG GVD

Distinguishing NIPS Articles

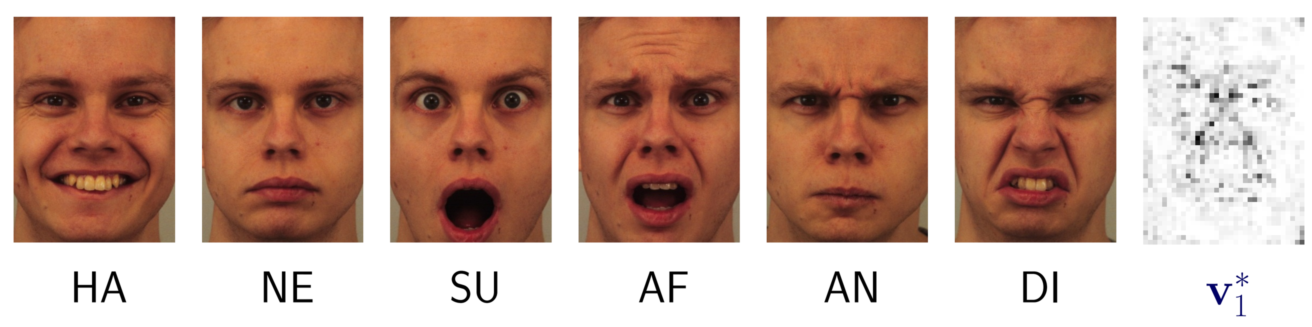
- **Task:** distinguishing 2 categories of NIPS papers (1988–2015).
- Stemmed $d = 2000$ nouns. TF-IDF representation. $J = 1$. $\alpha = 0.01$.

Problem	n^{te}	**ME-full**	ME-grid	**SCF-full**	SCF-grid	MMD-quad	MMD-lin
1. Bayes-Bayes	215	.012	.018	.012	.004	.022	.008
2. Bayes-Deep	216	**.954**	.034	.688	.180	.906	.262
3. Bayes-Learn	138	.990	.774	.836	.534	**1.00**	.238
4. Bayes-Neuro	394	**1.00**	.300	.828	.500	.952	.972
5. Learn-Deep	149	**.956**	.052	.656	.138	.876	.500
6. Learn-Neuro	146	.960	.572	.590	.360	**1.00**	.538

- In (4), words with highest weights as ranked by the learned $\mathbf{v}_1$:
  spike, markov, cortex, dropout, recurr, iii, gibb, basin, circuit.
- ME-full, SCF-full: high powers, correct type-I errors, and interpretable.

Distinguishing Pos. & Neg. Emotions

- **Task:** distinguishing images of positive and negative facial expressions.
- (+): { happy (HA), neutral (NE), surprised (SU) }
  vs. (−): { afraid (AF), angry (AN), disgusted (DI). }
- $d = 48 \times 34 = 1632$ pixels. Grayscale. $J = 1$.



HA NE SU AF AN DI $\mathbf{v}_1^*$

Problem	n^{te}	**ME-full**	ME-grid	**SCF-full**	SCF-grid	MMD-quad	MMD-lin
± vs. ±	201	.010	.012	.014	.002	.018	.008
+ vs. −	201	.998	.656	**1.00**	.750	**1.00**	.578

- ME-full achieves a high test power and gives an interpretable feature $\mathbf{v}_1^*$.
- $\mathbf{v}_1^*$ = average across trials of the learned test locations.