Interpretable Distribution Features with Maximum Testing Power

Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London

Summary

- **Have**: Two collections drawn from two unknown distributions.
- **Goal**: Learn distinguishing features indicating how they differ.
- **How**: Maximize a lower bound on test power for a two-sample test using these features.
- **Our methods are both**:

1. Understandable spatial and frequency feature extractors.
2. Linear-time, nonparametric, consistent, two-sample tests. (**Power matches the quadratic-time MMD test**).

- **Applications**: 1. Differentiate positive/negative emotions. 2. Distinguish articles from different categories.

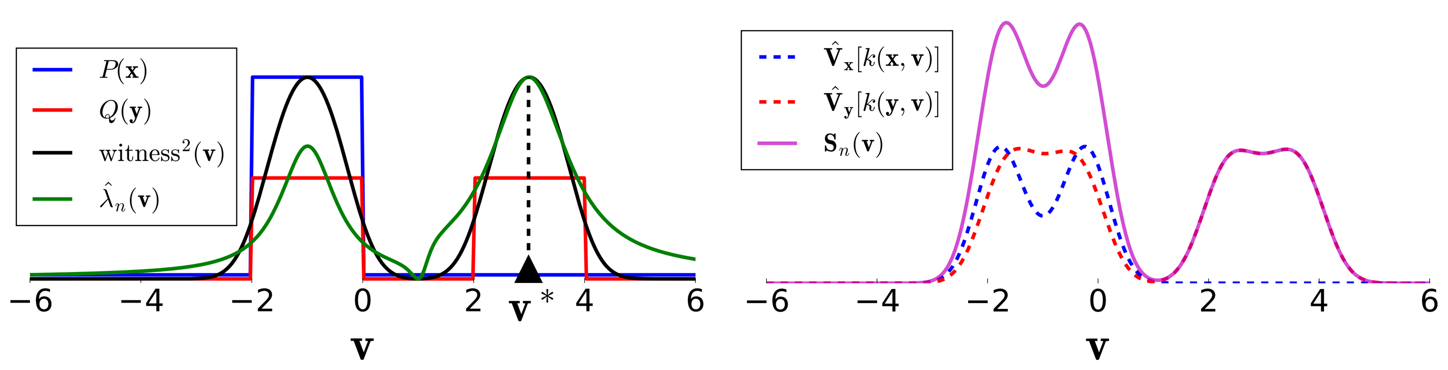
ME and SCF Tests

- Observe $X := \{\mathbf{x}_i\}_{i=1}^n \sim P$ and $Y := \{\mathbf{y}_i\}_{i=1}^n \sim Q$ 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.

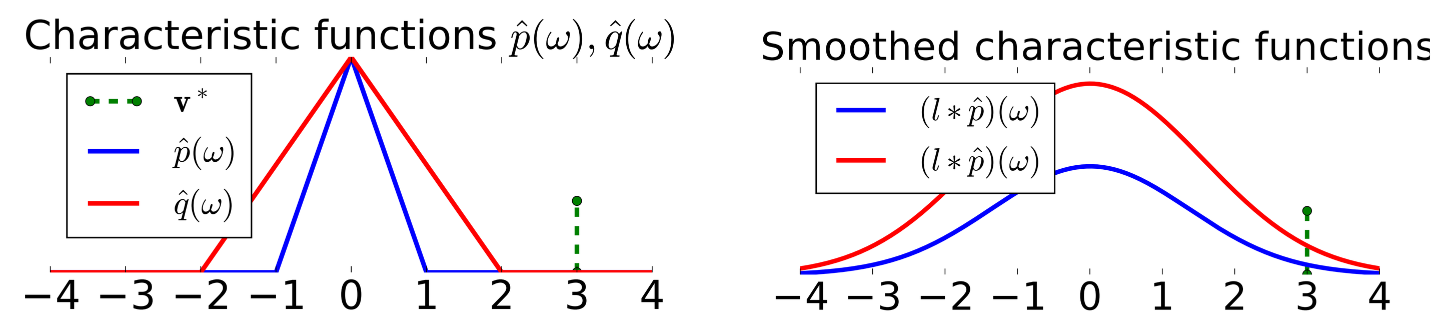
Mean Embedding (ME) Test:

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

- J spatial features (test locations): $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Regularizer γ_n. Gaussian kernel k_σ.
- Witness function: $\mathrm{witness}(\mathbf{v}) := \hat{\mathbb{E}}_\mathbf{x}[k_\sigma(\mathbf{x},\mathbf{v})] - \hat{\mathbb{E}}_\mathbf{y}[k_\sigma(\mathbf{y},\mathbf{v})]$.
- $\mathbf{w}_n := (\mathrm{witness}(\mathbf{v}_1), \ldots, \mathrm{witness}(\mathbf{v}_J))^\top \in \mathbb{R}^J$.
- $(\mathbf{S}_n)_{ij} = \widehat{\mathrm{cov}}_\mathbf{x}[k(\mathbf{x},\mathbf{v}_i), k(\mathbf{x},\mathbf{v}_j)] + \widehat{\mathrm{cov}}_\mathbf{y}[k(\mathbf{y},\mathbf{v}_i), k(\mathbf{y},\mathbf{v}_j)]$.
- Under H_0, $\hat{\lambda}_n$ asymptotically follows $\chi^2(J)$.



Smooth Characteristic Function (SCF) Test:



- Difference of smoothed (by l) characteristic functions.

Test Power Lower Bound

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

$$L(\lambda_n) = 1 - 2e^{-\xi_1(\lambda_n - T_\alpha)^2/n} - 2e^{-\frac{[\gamma_n(\lambda_n - T_\alpha)(n-1) - \xi_2 n]^2}{\xi_3 n(2n-1)^2}} - 2e^{-\frac{[(\lambda_n - T_\alpha)/3 - \bar{c}_3 n\gamma_n]^2\gamma_n^2}{\xi_4}}$$

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

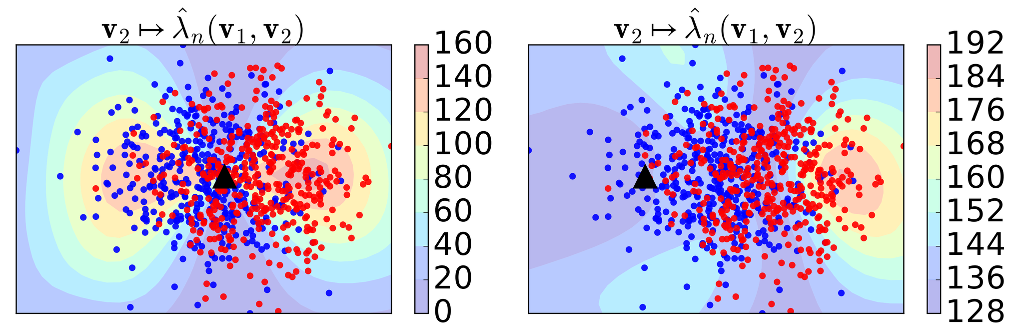
- λ_n is the population counterpart of $\hat{\lambda}_n$. Constants: $\bar{c}_3, \xi_1, \ldots, \xi_4 > 0$.

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

- **Key**: Parameters chosen to maximize the test power lower bound.
- Use a separate training set to estimate λ_n.

Informative Features

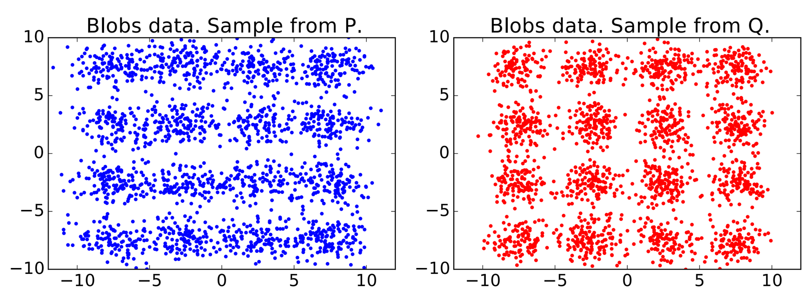
- Contour plot of $\hat{\lambda}_n$ as a function of $\mathbf{v}_2$ when $J = 2$. $\mathbf{v}_1$ fixed at ▲.



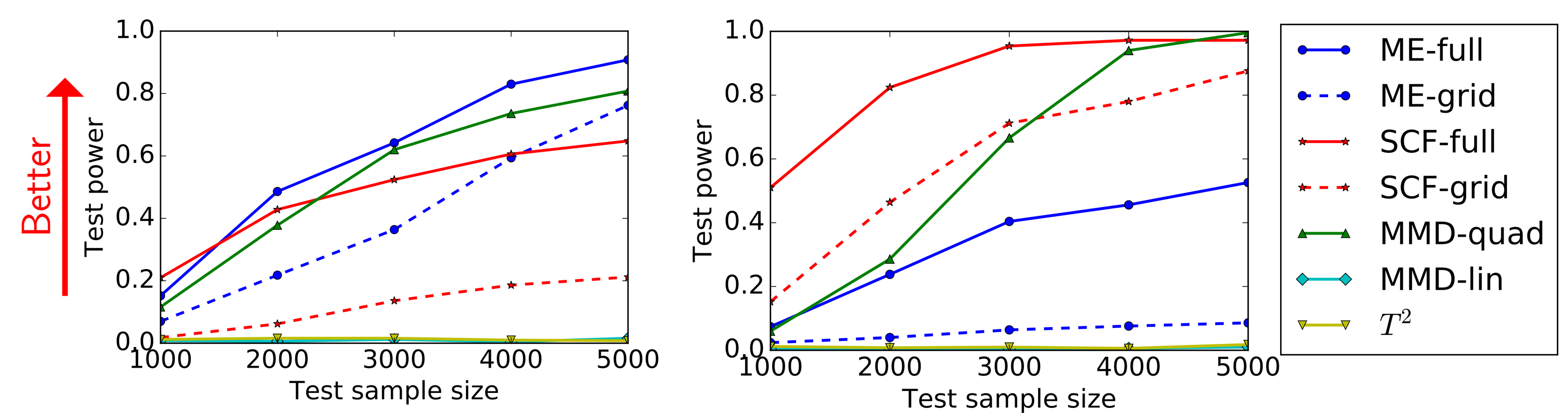
- $P: \mathcal{N}([0,0], \mathbf{I})$ vs. $Q: \mathcal{N}([1,0], \mathbf{I})$.
- $\hat{\lambda}_n$ is high in the regions that reveal the difference.
- Nonconvexity indicates many informative ways to detect the differences.

Test Power vs. Sample Size

Problem	P	Q
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 →	



- ME-full, SCF-full = Proposed methods. Full optimization. $J = 5$.
- ME-grid, SCF-grid = Random $\mathcal{V}$. Grid search for σ.
- MMD-quad, MMD-lin = Quadratic and linear-time MMD tests.

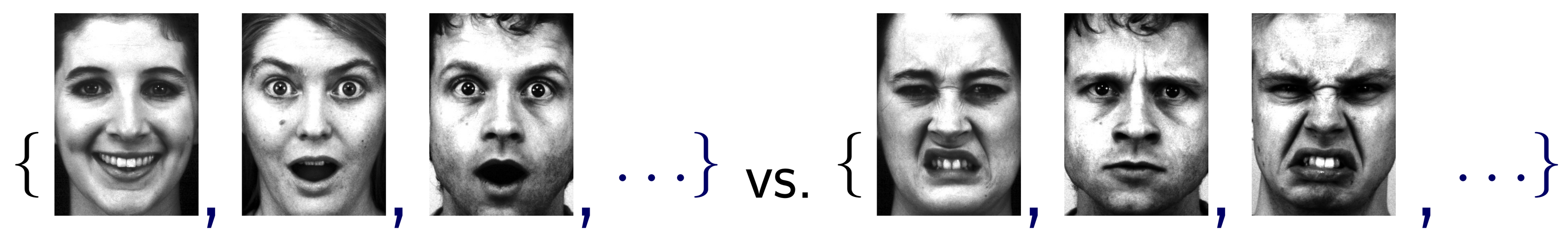


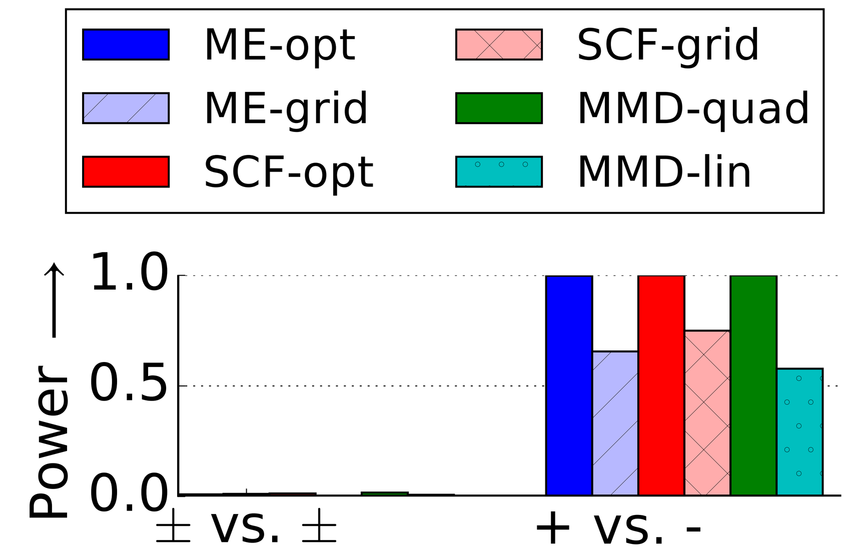
GVD. 50 dimensions. Blobs.

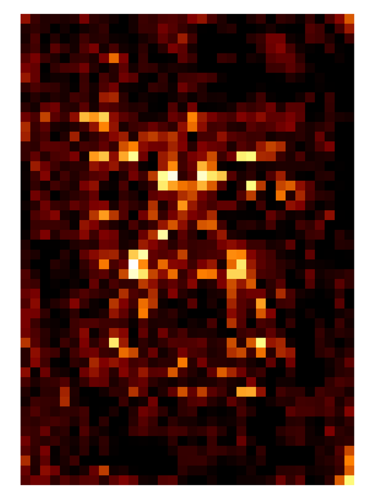
- GVD: Best performance by ME-full. Spatial differences.
- Blobs: Best performance by SCF-full. Frequency differences.

Distinguishing Pos. & Neg. Emotions

- **Task:** distinguish positive and negative facial expressions.
- $d = 48 \times 34 = 1632$ pixels. Use raw pixels. One feature ($J = 1$).





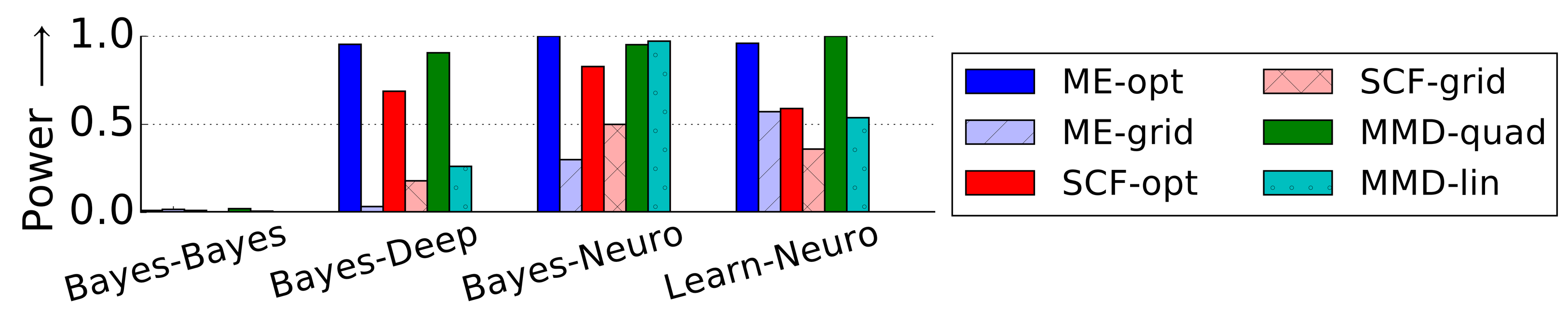


Learned feature

- ME-full, SCF-full achieves high test power.
- ME-full learned an informative feature.

Distinguishing NIPS Articles

- **Task:** distinguish two categories of NIPS papers (1988–2015).
- Stemmed $d = 2000$ nouns. TF-IDF representation. $J = 1$.



- ME-full: high powers comparable to MMD-quad; but faster.

Learned documents by ME-full show distinguishing keywords.

- **Bayes-Deep**: infer, Bayes, Monte Carlo, adaptor, motif, haplotype, ECG
- **Bayes-Neuro**: spike, Markov, cortex, dropout, recurrent, iii, Gibbs, basin
- **Learn-Neuro**: policy, interconnect, hardware, decay, histolog, EDG, period

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