# Linear-Time Divergence Measures with Applications in Hypothesis Testing

Zoltán Szabó (CMAP, École Polytechnique)





Joint work with Wittawat Jitkrittum, Kacper Chwialkowski, Wenkai Xu, Arthur Gretton, Kenji Fukumizu

Tao Seminar
Feb. 13, 2018

# Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

$$KL(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log \left[ \frac{p(x)}{q(x)} \right] \mathrm{d}x.$$

# Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

$$KL(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log \left[ \frac{p(x)}{q(x)} \right] \mathrm{d}x.$$

- Mutual information:

$$I(\mathbb{P}) = KL(\mathbb{P}, \mathbb{P}_1 \otimes \mathbb{P}_2).$$

# Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

$$KL\left(\mathbb{P},\mathbb{Q}\right) = \int_{\mathbb{R}^d} p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

- Mutual information:

$$I\left(\mathbb{P}\right) = KL\left(\mathbb{P},\mathbb{P}_1\otimes\mathbb{P}_2\right).$$

Properties:

1. $I(\mathbb{P}) \geqslant 0$. $I(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{P}_1\otimes\mathbb{P}_2$.

# Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

$$KL(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log\left[\frac{p(x)}{q(x)}\right] \mathrm{d}x.$$

- Mutual information:

$$I(\mathbb{P}) = KL(\mathbb{P},\mathbb{P}_1 \otimes \mathbb{P}_2).$$

Properties:

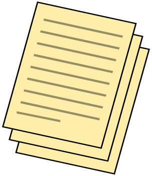
1. $I(\mathbb{P}) \geqslant 0$. $I(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$.

It can be hard to estimate them. Alternatives? Applications?

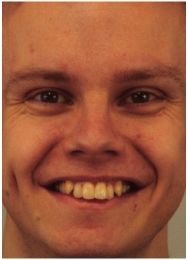
# Motivating Examples

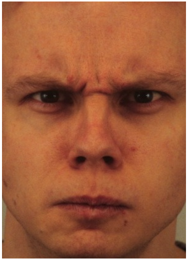
# NLP

- Given: two categories of documents (Bayesian inference, neuroscience).
- Task:
  - test their distinguishability,
  - most discriminative words $\rightarrow$ interpretability.



# Computer Vision





- Given: two sets of faces (happy, angry).
- Task:
  - check if they are different,
  - determine the most discriminative features/regions.

# Phrased as a Two-Sample Testing Task

- Given:
  - $X = \{\mathbf{x}_i\}_{i=1}^n \stackrel{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \stackrel{i.i.d.}{\sim} \mathbb{Q}$.
  - Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.

# Phrased as a Two-Sample Testing Task

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

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

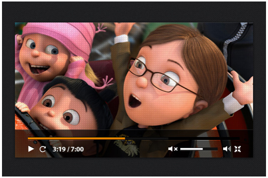
# Dependency Testing of Media Annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

# Dependency Testing of Media Annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

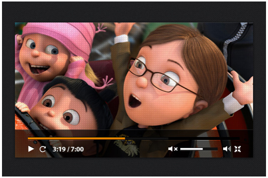
  - (video, caption) pairs



# Dependency Testing of Media Annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

  - (video, caption) pairs



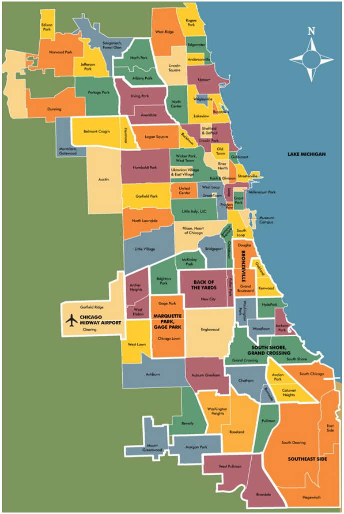

- $\{(x_i, y_i)\}_{i=1}^n \xrightarrow{?} H_0 : \mathbb{P}_{xy} = \mathbb{P}_x\mathbb{P}_y,\ H_1 : \mathbb{P}_{xy} \neq \mathbb{P}_x\mathbb{P}_y.$

Given:

- Density/model: $p$.

Given:

- Density/model: $p$.
- Samples: $X = \{x_i\}_{i=1}^n \sim q$ (unknown).

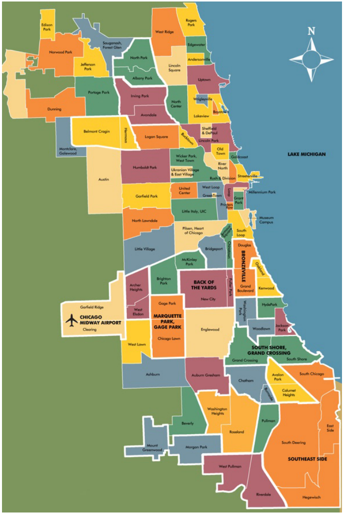


Given:

- Density/model: $p$.
- Samples: $X = \{x_i\}_{i=1}^n \sim q$ (unknown).

Problem: using $p$, $X$ test

$$H_0 : p = q, \quad \text{vs}$$
$$H_1 : p \neq q.$$



# Dissemination, Code

- ITE toolbox:

  `https://bitbucket.org/szzoli/ite-in-python/`

# Dissemination, Code

- ITE toolbox:

  `https://bitbucket.org/szzoli/ite-in-python/`
- Linear-time testing
  - two-sample (NIPS-2016, oral):

    `https://github.com/wittawatj/interpretable-test`

# Dissemination, Code

- ITE toolbox:

  `https://bitbucket.org/szzoli/ite-in-python/`
- Linear-time testing
  - two-sample (NIPS-2016, oral):

    `https://github.com/wittawatj/interpretable-test`
  - independence (ICML-2017):

    `.../fsic-test`

# Dissemination, Code

- ITE toolbox:

  `https://bitbucket.org/szzoli/ite-in-python/`
- Linear-time testing
  - two-sample (NIPS-2016, oral):

    `https://github.com/wittawatj/interpretable-test`
  - independence (ICML-2017):

    `.../fsic-test`
  - goodness-of-fit (NIPS-2017, best paper award):

    `.../kernel-gof`

# Divergence & Independence Measures

# Distribution Representation: Examples

- Mean:

$$\mathbb{P} \mapsto \mathbb{E}_{x \sim \mathbb{P}}[x].$$

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z)$$

# Distribution Representation: Examples

- Mean:

$$\mathbb{P} \mapsto \mathbb{E}_{x\sim\mathbb{P}}[x].$$

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}} I_{(-\infty,z)}(x).$$

# Distribution Representation: Examples

- Mean:

$$\mathbb{P} \mapsto \mathbb{E}_{x\sim\mathbb{P}}[x].$$

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}} I_{(-\infty,z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(z) = \int e^{i\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

# Distribution Representation: Examples

- Mean:

$$\mathbb{P} \mapsto \mathbb{E}_{x \sim \mathbb{P}}[x].$$

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x \sim \mathbb{P}} I_{(-\infty, z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(z) = \int e^{i\langle z, x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z, x\rangle} \mathrm{d}\mathbb{P}\left(x\right).$$

# Distribution Representation: Examples

- Mean:

$$\mathbb{P} \mapsto \mathbb{E}_{x\sim\mathbb{P}}[x].$$

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}} I_{(-\infty,z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(z) = \int e^{i\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

**Pattern**

$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$

# Distribution Representation

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

# Distribution Representation

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

**Question**

- How to choose $\varphi$?

# Distribution Representation

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

**Question**

- How to choose $\varphi$?

- We use kernels. $\rightarrow$ Computational tractability: ✓
- $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$.

# Distribution Representation

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

**Question**

- How to choose $\varphi$?

- We use kernels. $\rightarrow$ Computational tractability: ✓
- $k(x,y) = \langle \varphi(x), \varphi(y) \rangle$. Examples ($\gamma > 0$, $p \in \mathbb{Z}^+$):

$$k_p(x,y) = (\langle x,y \rangle + \gamma)^p, \qquad k_G(x,y) = e^{-\gamma \|x-y\|_2^2},$$

$$k_e(x,y) = e^{-\gamma \|x-y\|_2}, \qquad k_C(x,y) = 1 + \frac{1}{\gamma \|x-y\|_2^2}.$$

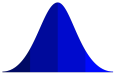
# KL Divergence and Mutual Information Alternatives

- Mean embedding:

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \, \mathrm{d}\mathbb{P}(x)$$

- Mean embedding:

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k = \overline{span}\left(k(\cdot,x) : x \in \mathcal{X}\right).$$



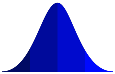
- Maximum mean discrepancy:

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

# KL Divergence and Mutual Information Alternatives

- Mean embedding:

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k = \overline{span}\left(k(\cdot,x) : x \in \mathcal{X}\right).$$



- Maximum mean discrepancy:

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

- Hilbert-Schmidt independence criterion, $k = k_1 \otimes k_2$:

$$\mathrm{HSIC}_k(\mathbb{P}) = \mathrm{MMD}_k\left(\mathbb{P}, \mathbb{P}_1 \otimes \mathbb{P}_2\right),$$
$$(k_1 \otimes k_2)\left((x,y),(x',y')\right) = k_1(x,x')k_2(y,y').$$

# Estimation of MMD and HSIC

$$\widehat{MMD}^2 = \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)}_{\text{within-block similarity}} - \underbrace{\frac{2}{n^2}\sum_{i,j=1}^{n} k(x_i, y_j)}_{\text{between-block similarity}} .$$

$$\widehat{HSIC}^2 = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F$$

# Estimation of MMD and HSIC

$$\widehat{MMD}^2 = \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)}_{\text{within-block similarity}} - \underbrace{\frac{2}{n^2}\sum_{i,j=1}^{n} k(x_i, y_j)}_{\text{between-block similarity}}.$$

$$\widehat{HSIC}^2 = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F, \mathbf{G}_x = [k_1(x_i, x_j)]_{i,j=1}^n$$

# Estimation of MMD and HSIC

$$\widehat{MMD}^2 = \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)}_{\text{within-block similarity}} - \underbrace{\frac{2}{n^2}\sum_{i,j=1}^{n} k(x_i, y_j)}_{\text{between-block similarity}} .$$

$$\widehat{HSIC}^2 = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F, \mathbf{G}_x = [k_1(x_i, x_j)]_{i,j=1}^{n}, \ \tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}, \ \mathbf{H} = \mathbf{I}_n - \frac{\mathbf{E}}{n}.$$

**Bottleneck**

Computational time: $\mathcal{O}\left(n^2\right)$.

# Linear-Time 'MMD'

## Idea [Chwialkowski et al., 2015a]

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

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

**Idea [Chwialkowski et al., 2015a]**

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

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

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

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

# Linear-Time 'MMD'

**Idea [Chwialkowski et al., 2015a]**

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

Plug-in estimate: $\mathcal{O}(n)$-time. Whitened test statistic: $\chi^2_J$ null.

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

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

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

# Linear-Time 'MMD'

**Idea [Chwialkowski et al., 2015a], [Jitkrittum et al., 2016]**

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

Plug-in estimate: $\mathcal{O}(n)$-time. Whitened test statistic: $\chi^2_J$ null. Power opt.

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

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

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

$$(\sigma^*,\mathcal{V}^*) = \arg\max_{\sigma,\mathcal{V}}\lambda, \qquad \lambda = n\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}.$$

Use different norm of the witness function ($u$):

$$HSIC(x,y) = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

# Linear-Time 'HSIC' [Jitkrittum et al., 2017]

Use different norm of the witness function ($u$):

$$HSIC(x,y) = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$FSIC(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

# Linear-Time 'HSIC' [Jitkrittum et al., 2017]

Use different norm of the witness function ($u$):

$$HSIC(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$FSIC(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

$$= \|u\|_{L^2(\mathcal{V})}.$$

# Linear-Time 'HSIC' [Jitkrittum et al., 2017]

Use different norm of the witness function ($u$):

$$HSIC(x,y) = \left\| \mu_{xy} - \mu_x \otimes \mu_y \right\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$FSIC(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

$$= \|u\|_{L^2(\mathcal{V})}.$$

- Whitening $\Rightarrow \chi^2_J$ null. Computation: $\mathcal{O}(n)$. Power optimization.

# Linear-Time 'HSIC' [Jitkrittum et al., 2017]

Use different norm of the witness function ($u$):

$$HSIC(x,y) = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$FSIC(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

$$= \|u\|_{L^2(\mathcal{V})}.$$

- Whitening $\Rightarrow \chi^2_J$ null. Computation: $\mathcal{O}(n)$. Power optimization.
- Alternative view: $u(\mathbf{v},\mathbf{w}) = cov_{\mathbf{xy}}(k_1(\mathbf{x},\mathbf{v}), k_2(\mathbf{y},\mathbf{w})) = (\mathbf{v},\mathbf{w})^{th}$ entry of

$$C_{xy} = \mathbb{E}_{xy}\left[\varphi_1(x) \otimes \varphi_2(y)\right] - \mu_x \otimes \mu_y.$$

# Until Now

We

- assumed analytic, characteristic, bounded kernels.
- replaced the RKHS norm with $L^2(\mathcal{V})$ norm.

In linear-time 'MMD' and 'HSIC', respectively:

$$\mathbb{P} = \mathbb{Q} \Leftrightarrow \qquad \mu_{\mathbb{P}-\mathbb{Q}} = 0,$$

$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \Leftrightarrow \qquad \mu_{\mathbb{P}-\mathbb{P}_1 \otimes \mathbb{P}_2} = 0.$$

# Goodness-of-Fit

Let $d = 1$. Stein operator of $p$

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

# Goodness-of-Fit

Let $d = 1$. Stein operator of $p$

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

# Goodness-of-Fit

Let $d = 1$. Stein operator of $p$

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Let us take the unit ball of $\mathcal{H}_k$:

$$\sup_{\|f\|_{\mathcal{H}_k}\leqslant 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \underbrace{\|g\|_{\mathcal{H}_k}}_{g \text{ is the argsup}}, \quad g(v) = \mathbb{E}_{x\sim q}\frac{\partial_x[p(x)k(x,v)]}{p(x)}.$$

# Goodness-of-Fit [Chwialkowski et al., 2016, Liu et al., 2016]

Let $d = 1$. Stein operator of $p$

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Let us take the unit ball of $\mathcal{H}_k$:

$$\sup_{\|f\|_{\mathcal{H}_k}\leqslant 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \underbrace{\|g\|_{\mathcal{H}_k}}_{g \text{ is the argsup}}, \quad g(v) = \mathbb{E}_{x\sim q}\frac{\partial_x[p(x)k(x,v)]}{p(x)}.$$

For universal $k$:

$$\boxed{p = q \Leftrightarrow g = 0 \text{ (witness)}}.$$

# Goodness-of-Fit [Jitkrittum et al., 2017], [Chwialkowski et al., 2016, Liu et al., 2016]

Let $d = 1$. Stein operator of $p$

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Let us take the unit ball of $\mathcal{H}_k$:

$$\sup_{\|f\|_{\mathcal{H}_k}\leqslant 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \underbrace{\|g\|_{\mathcal{H}_k}}_{g \text{ is the argsup}}, \quad g(v) = \mathbb{E}_{x\sim q} \frac{\partial_x [p(x)k(x,v)]}{p(x)}.$$

For universal $k$:

$$\boxed{p = q \Leftrightarrow g = 0 \text{ (witness)}}.$$

$L^2(\mathcal{V})$ trick goes through.

# Numerical Illustrations

# 2-Sample Testing: Parameter Settings

- Gaussian kernel ($\sigma$). $\alpha = 0.01$. $J = 1$. Repeat 500 trials.
- Report rejection rate of $H_0$
- Compare 4 methods
  - **ME-full**: Optimize $\mathcal{V}$ and $\sigma$.
  - **ME-grid**: Optimize $\sigma$. Random $\mathcal{V}$ [Chwialkowski et al., 2015b].
  - **MMD-quad**: Test with quadratic-time MMD [Gretton et al., 2012].
  - **MMD-lin**: Test with linear-time MMD [Gretton et al., 2012].
- Optimize kernels to power in MMD-lin, MMD-quad.

# NLP: Discrimination of Document Categories

- 5903 NIPS papers (1988-2015).
- Keyword-based category assignment into 4 groups:
  - Bayesian inference, Deep learning, Learning theory, Neuroscience
- $d = 2000$ nouns. TF-IDF representation.

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

- Performance of ME-full $[\mathcal{O}(n)]$ is comparable to MMD-quad $[\mathcal{O}(n^2)]$.

# NLP: Most/Least Discriminative Words

- Aggregating over trials; example: 'Bayes-Neuro'.
- Most discriminative words:

  spike, markov, cortex, dropout, recurr, iii, gibb.

  - learned test locations: highly interpretable,
  - 'markov', 'gibb' ($\Leftarrow$ Gibbs): Bayesian inference,
  - 'spike', 'cortex': key terms in neuroscience.

# NLP: Most/Least Discriminative Words

- Aggregating over trials; example: 'Bayes-Neuro'.

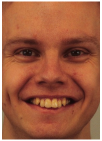
- Least discriminative ones:

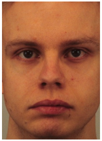
  circumfer, bra, dominiqu, rhino, mitra, kid, impostor.

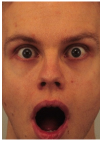
# Distinguish Positive/Negative Emotions

- Karolinska Directed Emotional Faces (KDEF) [Lundqvist et al., 1998].
- 70 actors = 35 females and 35 males.
- $d = 48 \times 34 = 1632$. Grayscale. Pixel features.

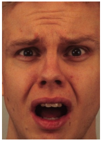
$+$ :

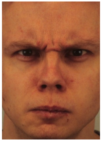

happy

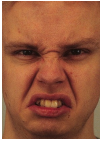

neutral


surprised

$-$ :


afraid


angry


disgusted

| Problem | $n^{te}$ | **ME-full** | ME-grid | MMD-quad | MMD-lin |
|---|---|---|---|---|---|
| $\pm$ vs. $\pm$ | 201 | .010 | .012 | .018 | .008 |
| $+$ vs. $-$ | 201 | .998 | .656 | 1.00 | .578 |

- Learned test location (averaged) = 

# Independence Testing: Parameters

- $k_1$, $k_2$: Gaussian. $J = 10$.
- Report: rejection rate of $H_0$.
- Compare 6 methods:

| Method | Description | Tuning | Test size | Complexity |
|---|---|---|---|---|
| **NFSIC-opt** | Studied | Gradient descent | $n/2$ | $\mathcal{O}(n)$ |
| NFSIC-med | No tuning | Random locations | $n$ | $\mathcal{O}(n)$ |
| QHSIC | Full HSIC | Median heuristic | $n$ | $\mathcal{O}(n^2)$ |
| NyHSIC | Nyström + HSIC | Median heuristic | $n$ | $\mathcal{O}(n)$ |
| FHSIC | RFF + HSIC | Median heuristic | $n$ | $\mathcal{O}(n)$ |
| RDC | RFF + CCA | Median heuristic | $n$ | $\mathcal{O}(n \log n)$ |

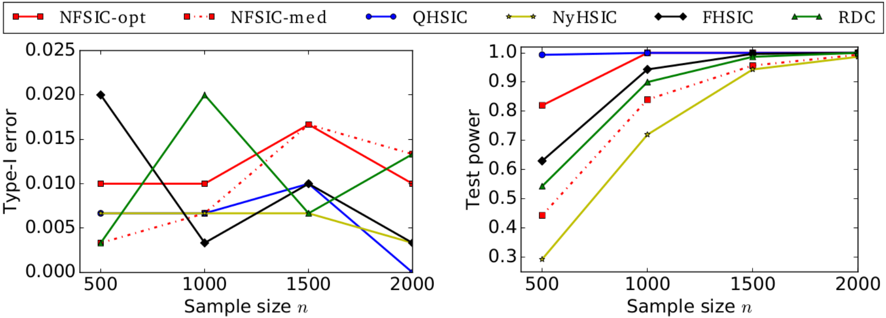
# Demo-1: Million Song Data

Song $(x)$ vs. year of release $(y)$.

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $x \in \mathbb{R}^{90}$: audio features.
- Left: break $(x, y)$ pairs, i.e. $H_0$; right: $H_1$ is true.

Song ($x$) vs. year of release ($y$).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $x \in \mathbb{R}^{90}$: audio features.
- Left: break $(x, y)$ pairs, i.e. $H_0$; right: $H_1$ is true.



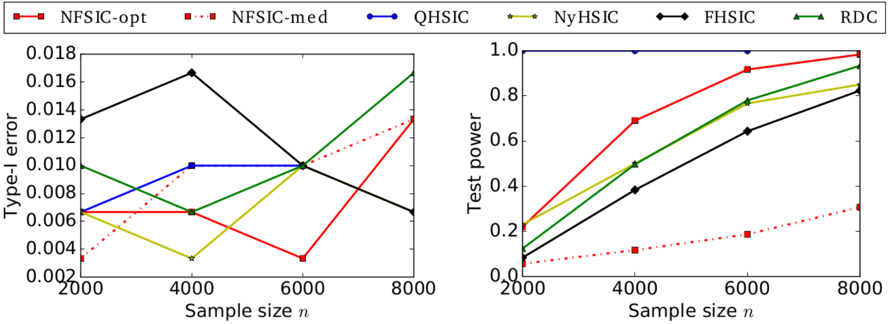
# Demo-2: Videos and Captions

Youtube video $(x)$ vs. caption $(y)$.

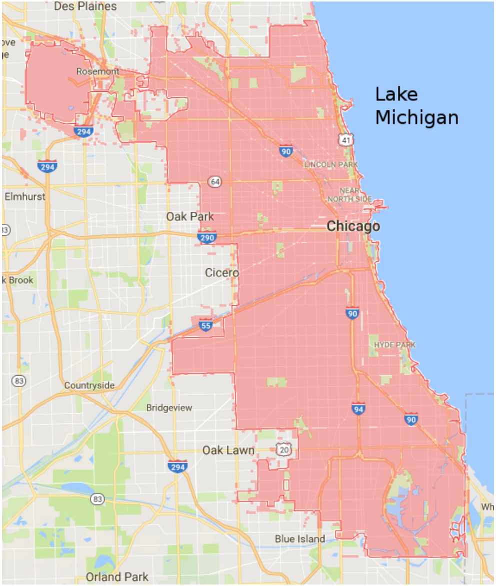
- VideoStory46K [Habibian et al., 2014]
- $x \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $y \in \mathbb{R}^{1878}$: bag of words. TF.
- Left: break $(x, y)$ pairs, i.e. $H_0$; right: $H_1$ is true.

Youtube video $(x)$ vs. caption $(y)$.

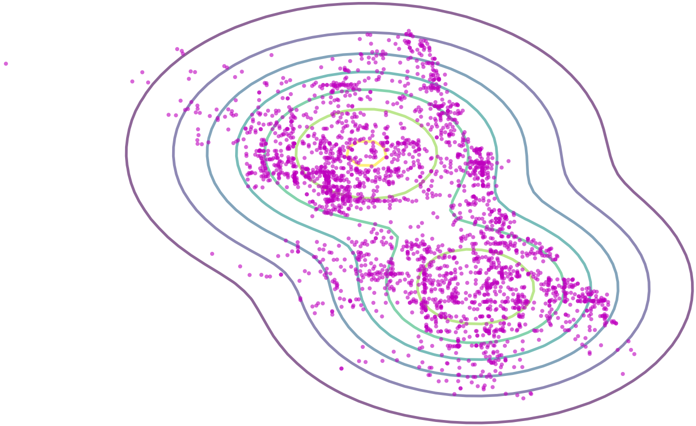
- VideoStory46K [Habibian et al., 2014]
- $x \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $y \in \mathbb{R}^{1878}$: bag of words. TF.
- Left: break $(x, y)$ pairs, i.e. $H_0$; right: $H_1$ is true.



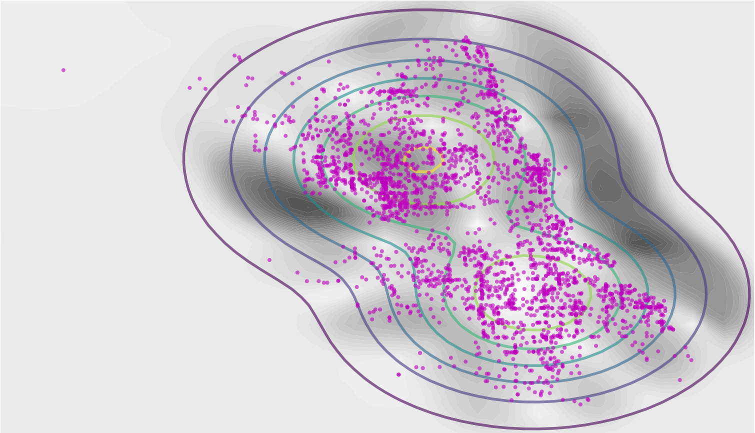
# Goodness-of-Fit Demo


Des Plaines
Rosemont
Lake Michigan
LINCOLN PARK
NEAR NORTH SIDE
Elmhurst
Oak Park
Chicago
Cicero
HYDE PARK
Countryside
Bridgeview
Oak Lawn
Blue Island
Orland Park

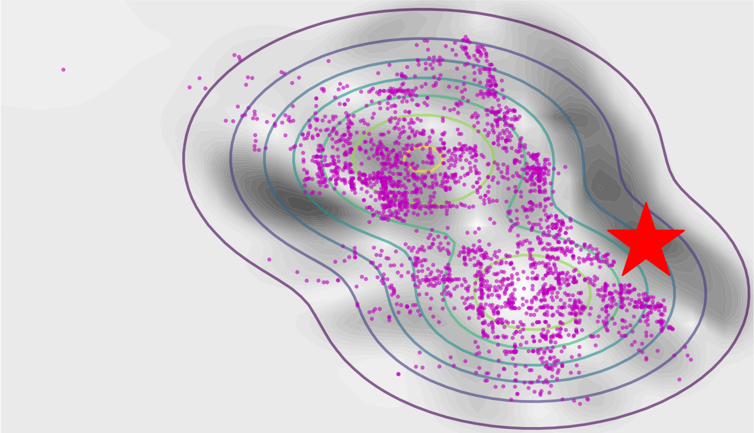
Robbery events (lat/long coordinates) $\sim q$.



Model $p$: 2-component Gaussian mixture.

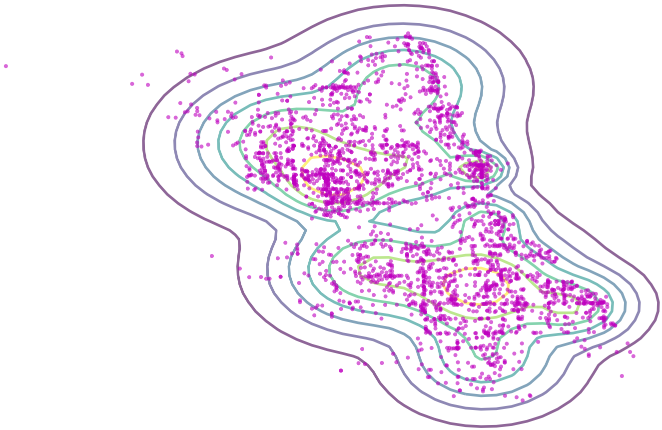


Score surface

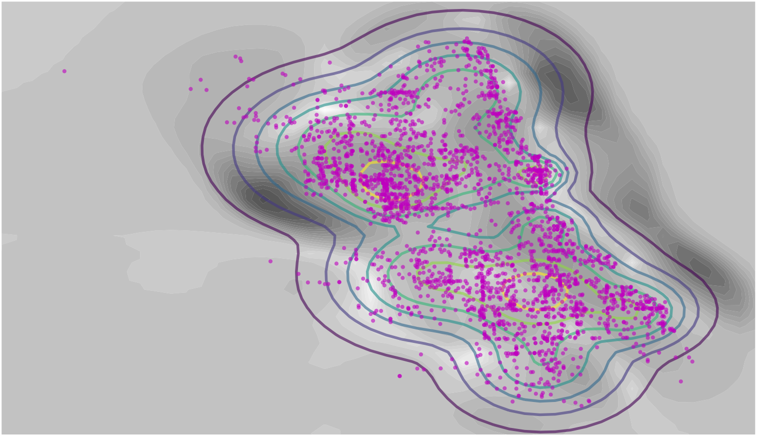


★ = optimized **v**.

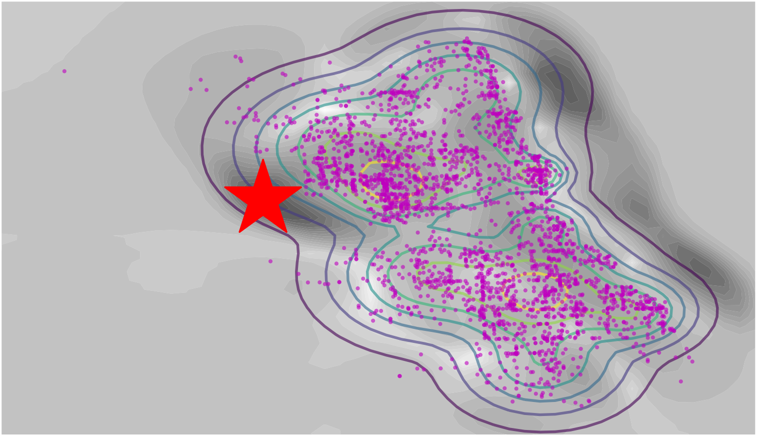
No robbery in Lake Michigan.



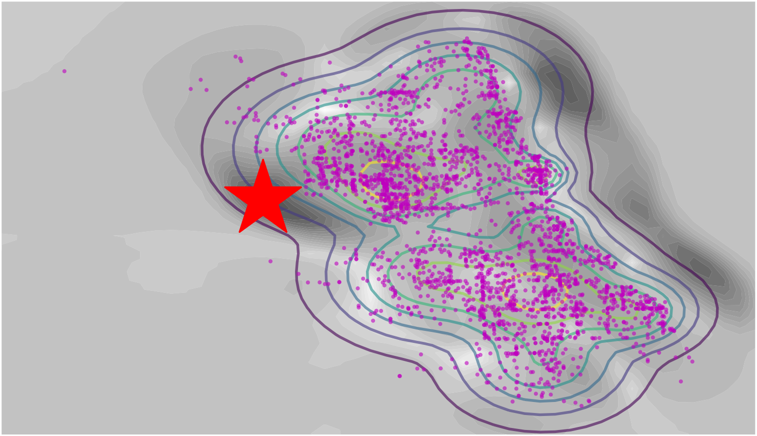
Model $p$: 10-component Gaussian mixture.



Capture the right tail better.



Still, does not capture the left tail.



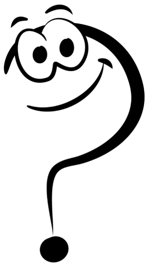
Still, does not capture the left tail.

**Sharp boundary (geography of Chicago) $\neq$ Gaussian tails. $\rightarrow$ interpretable features**

# Summary

- Hypothesis testing:
  - two-sample, independence, goodness-of-fit.
- MMD, HSIC: expensive $\Rightarrow$ proposed methods
  - linear-time.
  - adaptive: power/Bahadur-efficiency $\rightarrow$ max.
- Applications:
  - NLP, computer vision,
  - song-year, video-caption,
  - criminal data analysis.

Thank you for the attention!



Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011).
The million song dataset.
In *International Conference on Music Information Retrieval (ISMIR)*.

Chwialkowski, K., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015a).
Fast two-sample testing with analytic representations of probability measures.
In *Advances in Neural Informaton Processing Systems (NIPS)*, pages 1972–1980.

Chwialkowski, K., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015b).
Fast Two-Sample Testing with Analytic Representations of Probability Measures.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1981–1989.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *International Conference on Machine Learning (ICML)*, pages 2606–2615.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
*Journal of Machine Learning Research*, 13:723–773.

Habibian, A., Mensink, T., and Snoek, C. G. (2014).
Videostory: A new multimedia embedding for few-example recognition and translation of events.
In *ACM International Conference on Multimedia*, pages 17–26.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016).
Interpretable distribution features with maximum testing power.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 181–189.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017).
An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML)*, pages 1742–1751.

Liu, Q., Lee, J., and Jordan, M. (2016).
A Kernelized Stein Discrepancy for Goodness-of-fit Tests.
In *International Conference on Machine Learning (ICML)*, pages 276–284.

Lundqvist, D., Flykt, A., and Öhman, A. (1998).
The Karolinska directed emotional faces-KDEF.
Technical report, ISBN 91-630-7164-9.

Wang, H. and Schmid, C. (2013).
Action recognition with improved trajectories.

In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558.