Robust hypothesis test using Wasserstein uncertainty sets

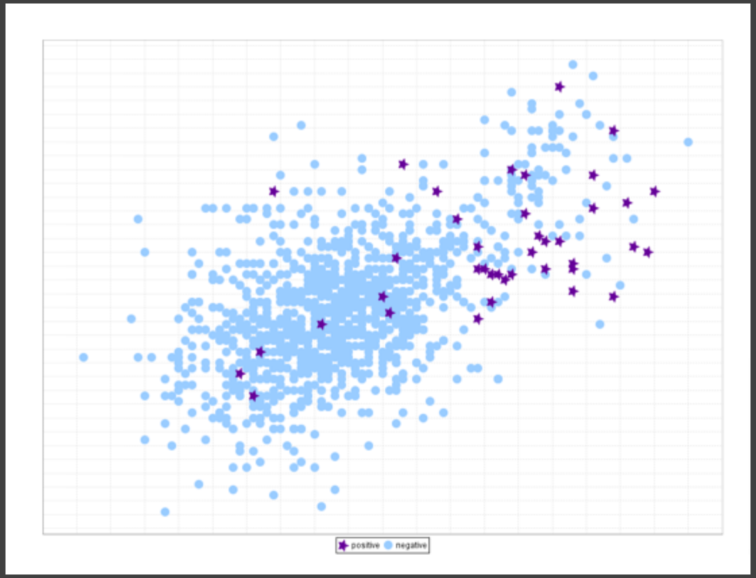
Yao Xie

Georgia Institute of Technology

Joint work with Rui Gao, Liyan Xie, Huan Xu

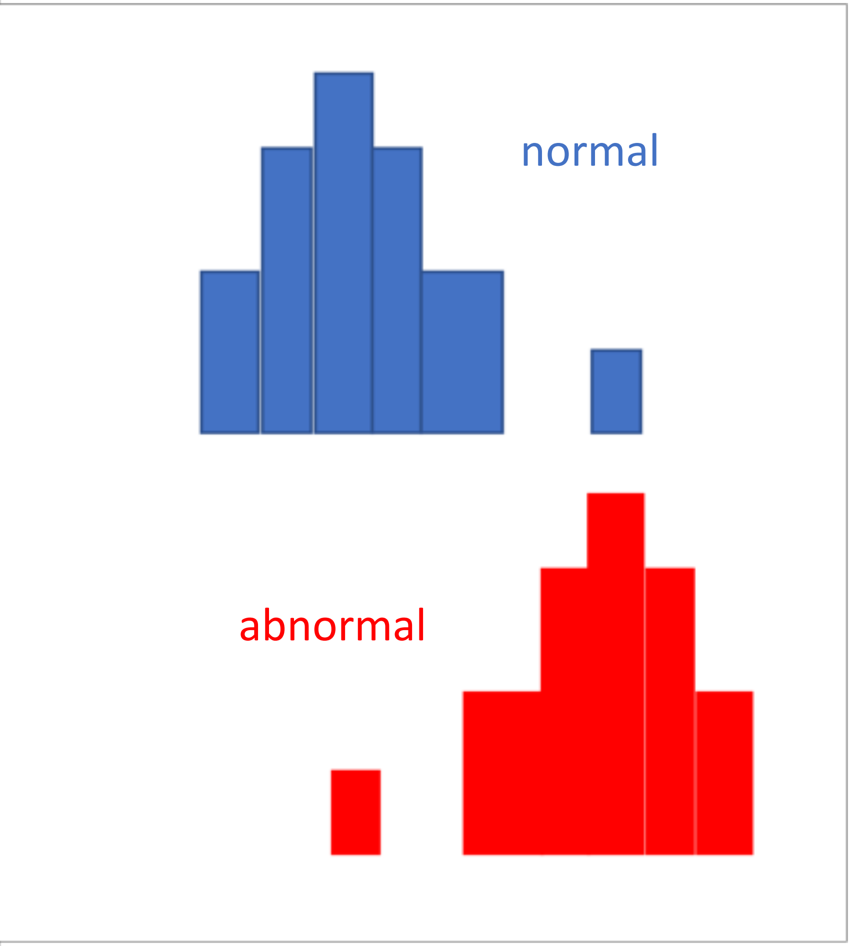
Classification with unbalanced data

fewer data for several classes



- **Anomaly detection:** self-driving car, network intrusion detection, credit fraud detection, online detection with fewer samples

- **Health care:** many negative samples, not many positive samples



Non-parametric hypothesis test with **unbalanced and limited data**

- empirical distribution may *not* have common support

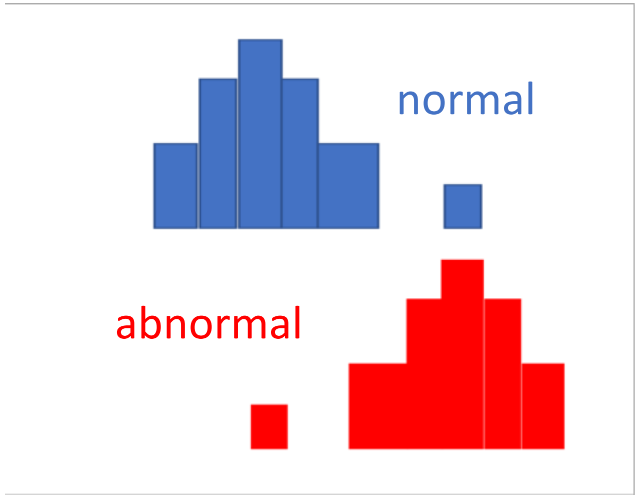
- no possible to use *likelihood ratio*: optimal by well-known Neyman-Pearson.

Hypothesis test using Wasserstein uncertainty sets

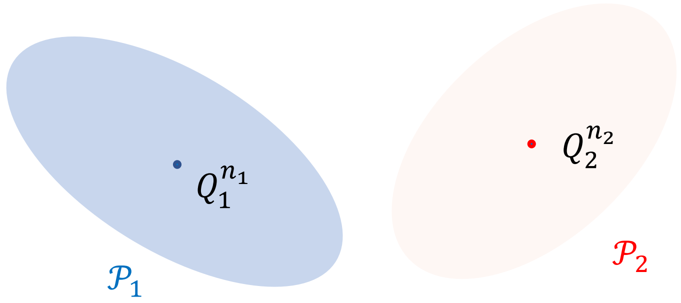
- Test two hypothesis

$$\begin{aligned} H_1 &: \quad \omega \sim P_1, \quad P_1 \in \mathcal{P}_1 \\ H_2 &: \quad \omega \sim P_2, \quad P_2 \in \mathcal{P}_2 \end{aligned}$$

- **Wasserstein uncertainty sets** for distributional robustness



Wasserstein metrics can deal with distributions with different support, better than K-L divergence



- Goal: find optimal detector, minimizes worst-case type-I + type-II errors

$$\inf_{\phi:\Omega\to\mathbb{R}} \sup_{P_1\in\mathcal{P}_1, P_2\in\mathcal{P}_2} \mathbb{E}_{P_1}[\ell \circ (-\phi)(\omega)] + \mathbb{E}_{P_2}[\ell \circ \phi(\omega)]$$

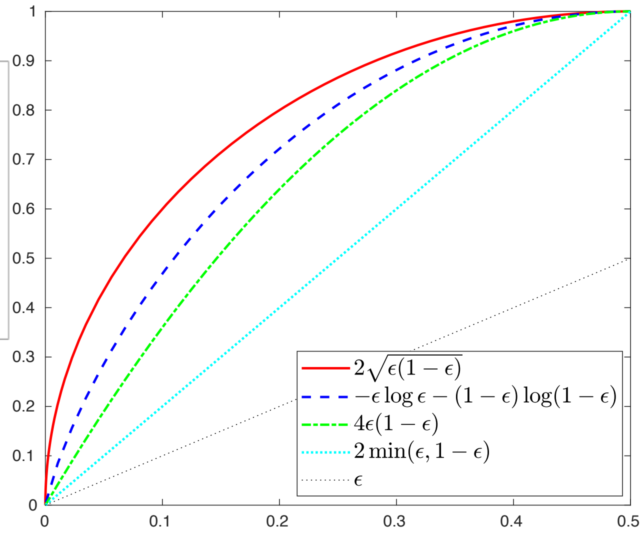
Main results

Distributionally robust nearly-optimal detector

- **Theorem**: General distributionally robust detector has nearly-optimal detector has risk bounded by small constant

$$\psi(\epsilon) - \epsilon$$

$\ell(t)$	$\psi(\epsilon)$
$\exp(t)$	$2\sqrt{\epsilon(1-\epsilon)}$
$\log(1+\exp(t))/\log 2$	$H(\epsilon)/\log 2$
$(t+1)_+^2$	$4\epsilon(1-\epsilon)$
$(t+1)_+$	2ϵ



Computationally efficient

- Tractable convex reformulation
- Complexity independent of dimensionality, scalable to large dataset

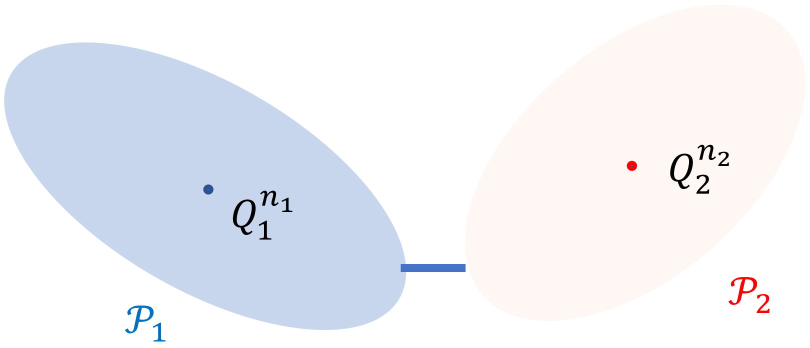
$$O(\ln(n_1) + \ln(n_2))$$

$$\begin{aligned}
\max_{\substack{p_1,p_2\in\mathbb{R}_+^{n_1+n_2}\\ \gamma_1,\gamma_2\in\mathbb{R}_+^{(n_1+n_2)}\times\mathbb{R}_+^{(n_1+n_2)}}} \quad & \sum_{l=1}^{n_1+n_2}(p_1^l+p_2^l)\psi\left(\frac{p_1^l}{p_1^l+p_2^l}\right)\\
\text{subject to} \quad & \sum_{l=1}^{n_1+n_2}\sum_{m=1}^{n_1+n_2}\gamma_k^{lm}\left\|\omega^l-\omega^m\right\| \le \theta_k,\ k=1,2,\\
& \sum_{m=1}^{n_1+n_2}\gamma_k^{lm} = Q_k^{n_k}(\omega^l),\ 1\le l\le n_1+n_2,\ k=1,2,\\
& \sum_{l=1}^{n_1+n_2}\gamma_k^{lm} = p_k^m,\ 1\le m\le n_1+n_2,\ k=1,2
\end{aligned}$$

Statistical interpretations

- Minimizes divergence between two distributions within two Wasserstein balls, centered around empirical distributions, and have common support on $n_1 + n_2$ data points

$$\inf_{\phi:\Omega\to\mathbb{R}} \sup_{P_1\in\mathcal{P}_1, P_2\in\mathcal{P}_2} \mathbb{E}_{P_1}[\ell\circ(-\phi)(\omega)] + \mathbb{E}_{P_2}[\ell\circ\phi(\omega)$$

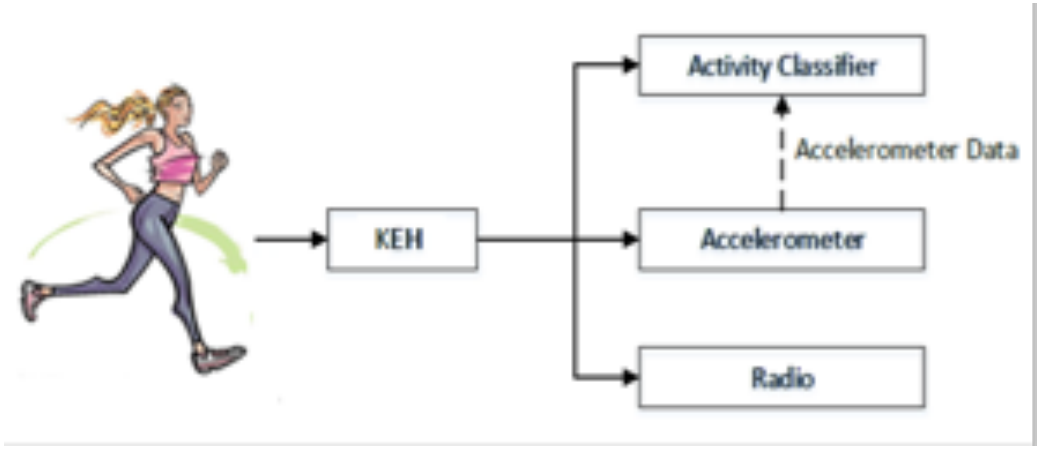


Generating function	Auxiliary function	Optimal detector	Detector risk
$\ell(t)$	$\psi(p)$	ϕ^*	$1 - 1/2\inf_\phi \Phi(\phi; P_1, P_2)$
$\exp(t)$	$2\sqrt{p(1-p)}$	$\ln\sqrt{p_1/p_2}$	$H^2(P_1, P_2)$
$\log(1+\exp(t))/\log 2$	$-H(p)/\log 2$	$\log(p_1/p_2)$	$JS(P_1, P_2)/\log 2$
$(t+1)_+^2$	$4p(1-p)$	$1 - 2\frac{p_1}{p_1+p_2}$	$\chi^2(P_1, P_2)$
$(t+1)_+$	$2\min(p, 1-p)$	$\mathrm{sgn}(p_1 - p_2)$	$TV(P_1, P_2)$

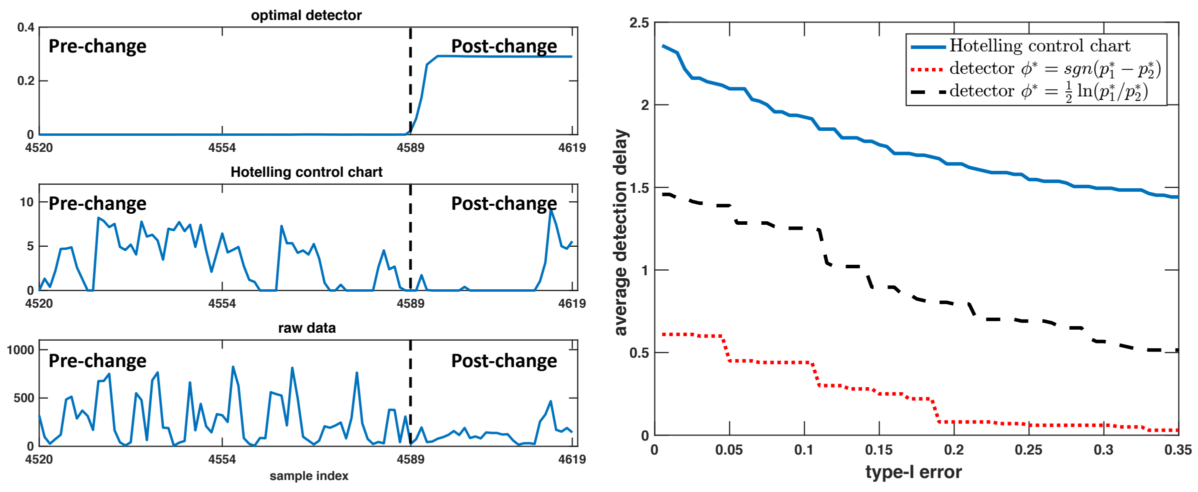
(Juditsky, Nemirovski, 2015)

Human activity detection

arXiv



Credit: CSIRO Research



(a) (b)

Figure: Jogging vs. Walking, the average is taken over 100 sequences of data.