Mixture Proportion Estimation

Clay Scott

EECS and Statistics
University of Michigan

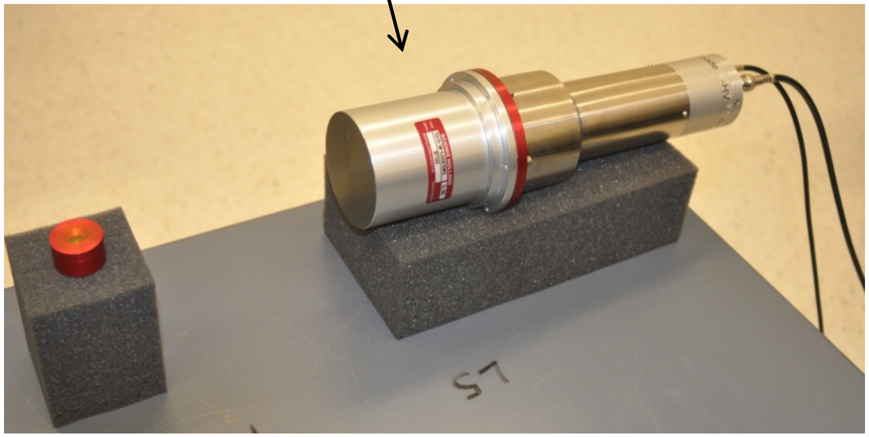
Nuclear Nonproliferation

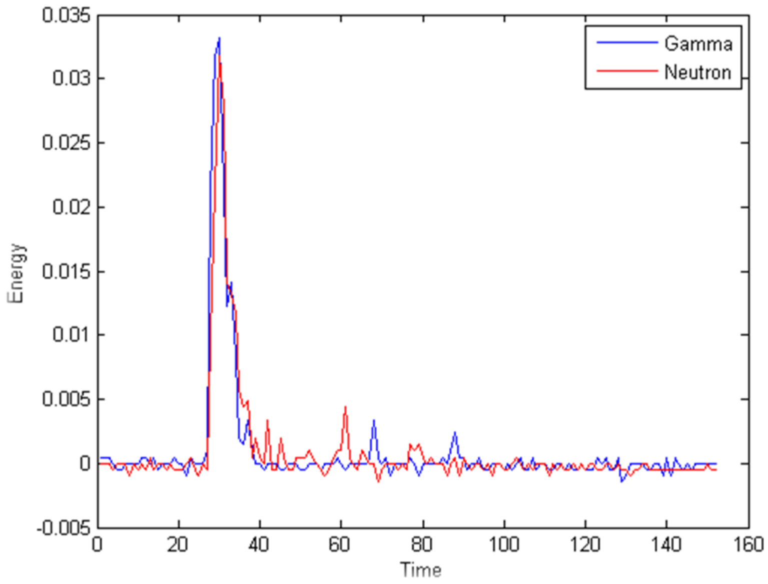
- Radioactive sources are characterized by distribution of neutron energies
- Organic scintillation detectors: prominent technology for neutron detection

Collaborators: Sara Pozzi, Marek Flaska @ UM Nuclear Engineering

Organic Scintillation Detector

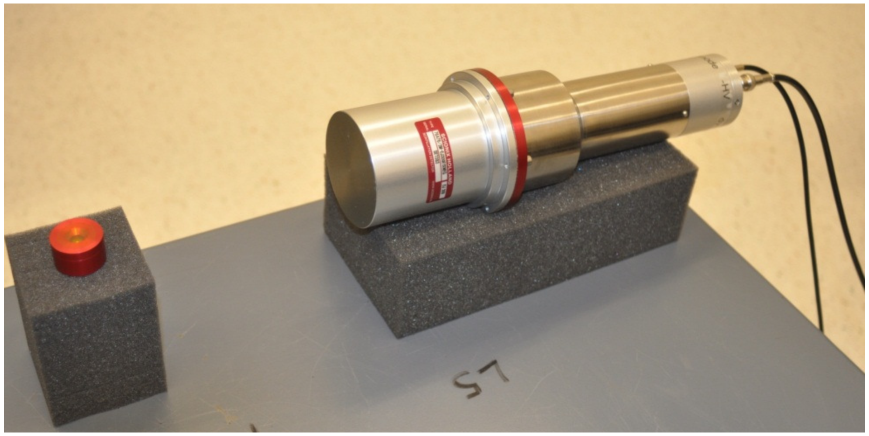
Source material





- Detects both neutrons and gamma rays
- Need to classify neutrons and gamma rays

Nuclear Particle Classification



Source material

- $X \in \mathbb{R}^d$, $d =$ signal length
- Training data:

$$\begin{aligned} X_1, \ldots, X_m &\overset{iid}{\sim} P_0 \qquad \text{(from gamma ray source, e.g. Na-22)} \\ X_{m+1}, \ldots, X_{m+n} &\overset{iid}{\sim} P_1 \qquad \text{(from neutron source, e.g. Cf-252)} \end{aligned}$$

- $P_0, P_1 =$ class-conditional distributions; don't want to model

Reality: No Pure Neutron Sources

- Contamination model for training data:

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1-\pi)P_1 + \pi P_0$$

- π unknown
- P_0, P_1 may have overlapping supports (nonseparable problem)
- Nonparametric approach desired
- Can be viewed as an **anomaly detection** problem
  - P_0 = nominal distribution
  - P_1 = anomalous distribution
  - $\tilde{P}_1$ = distribution of test data, to be classified

Measuring Performance

- Classifier:
$$f : \mathbb{R}^d \to \{0, 1\}$$

- False positive/negative rates:
$$R_0(f) := P_0(f(X) = 1)$$
$$R_1(f) := P_1(f(X) = 0)$$
$$\tilde{R}_1(f) := \tilde{P}_1(f(X) = 0)$$

- Estimating false negative rate:
$$\tilde{P}_1 = (1 - \pi)P_1 + \pi P_0$$
$$\Downarrow$$
$$\tilde{R}_1(f) = (1 - \pi)R_1(f) + \pi(1 - R_0(f))$$
$$\Downarrow$$
$$R_1(f) = \frac{\tilde{R}_1(f) - \pi(1 - R_0(f))}{1 - \pi}$$

- Suffices to estimate π

Mixture Proportion Estimation

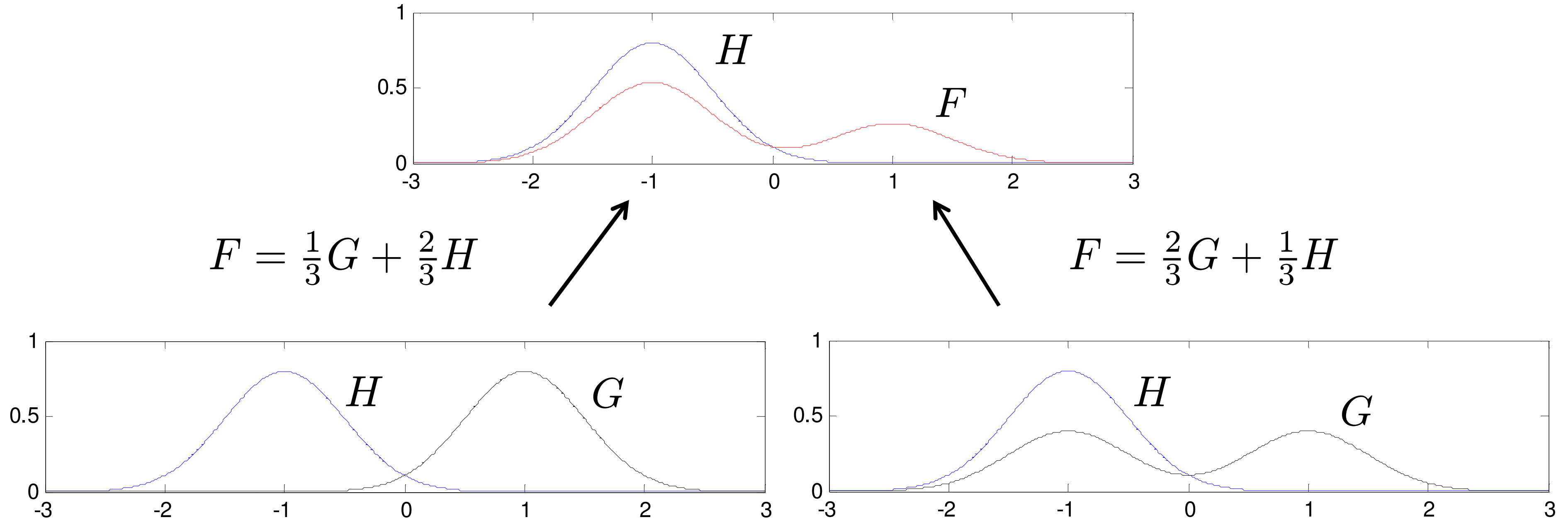
- Consider

$$Z_1, \ldots, Z_m \overset{iid}{\sim} H$$

$$Z_{m+1}, \ldots, Z_{m+n} \overset{iid}{\sim} F = (1-\nu)G + \nu H$$

- Need consistent estimate of ν

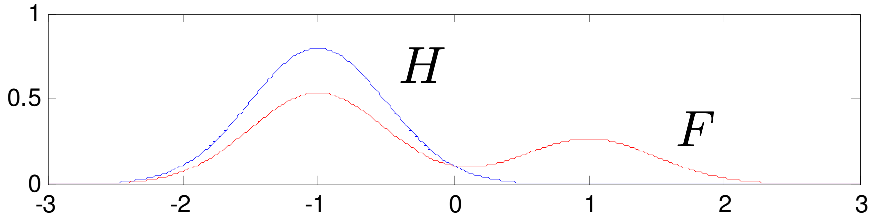
- Note: ν not identifiable in general



$F = \frac{1}{3}G + \frac{2}{3}H$

$F = \frac{2}{3}G + \frac{1}{3}H$

Mixture Proportion Estimation



- Given two distributions F, H, define

$$\nu^*(F,H) = \max\{\alpha \in [0,1] : \exists G' \text{ s.t. } F = (1-\alpha)G' + \alpha H\}$$

- Blanchard, Lee, S. (2010) give universally consistent estimator

$$\widehat{\nu}(\{Z_i\}_{i=1}^m, \{Z_i\}_{i=m+1}^{m+n}) \xrightarrow{a.s.} \nu^*(F,H)$$

- When is $\nu = \nu^*(F,H)$?

Identifiability Condition

- If

$$F = (1-\nu)G + \nu H$$

then

$$\nu = \nu^*(F,H) \Longleftrightarrow \boxed{\nu^*(G,H) = 0}$$

- Apply to anomaly detection problem

$$X_1,\dots,X_m \overset{iid}{\sim} P_0$$

$$X_{m+1},\dots,X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1-\pi)P_1 + \pi P_0$$

- Need

$$\nu^*(P_1,P_0) = 0$$

In words: Can't write P_1 as a (nontrivial) mixture of P_0 and some other distribution

Mixture Proportion Estimation

- Assume F, H have densities f and h

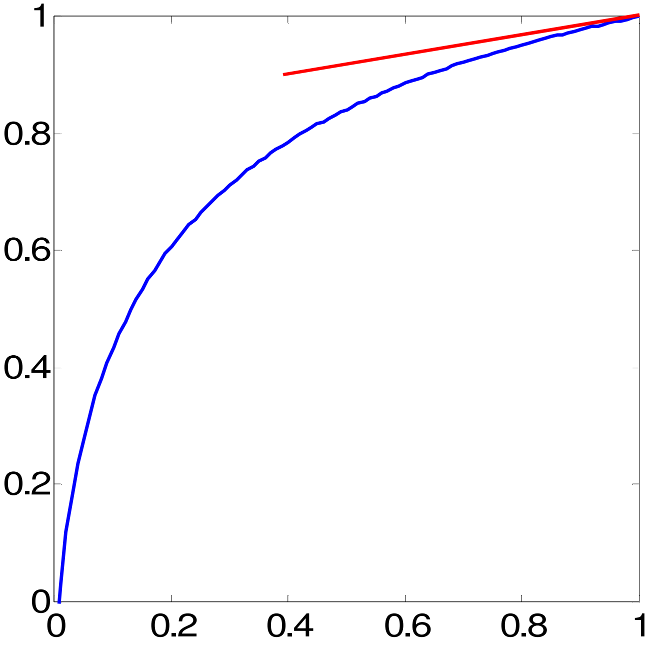
- Easy to show:

$$\nu^*(F,H) = \inf_{x:h(x)>0} \frac{f(x)}{h(x)}$$

- Consider ROC of LRT

$$\frac{f(x)}{h(x)} \gtrless \gamma$$

  Slope of ROC corresponding to threshold γ is γ



- Combine previous two facts:

$$\nu^*(F,H) = \text{slope of ROC of } f/h \text{ at right end-point}$$

- Remark: $1 - \nu^*(F,H) =$ "separation distance" between F and H

Classification with Label Noise

- Contaminated training data:

$$X_1, \ldots, X_m \overset{iid}{\sim} \tilde{P}_0 = (1 - \pi_0)P_0 + \pi_0 P_1$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \pi_1)P_1 + \pi_1 P_0$$

- P_0, P_1 **unknown**
- P_0, P_1, may have **overlapping supports**
- π_0, π_1 **unknown**
- **Asymmetric** label noise: $\pi_0 \neq \pi_1$

- **Random** label noise, as opposed to adversarial, or feature-dependent

Understanding Label Noise

- Assume P_0, P_1 have densities $p_0(x), p_1(x)$
- Then $\tilde{P}_0, \tilde{P}_1$ have densities

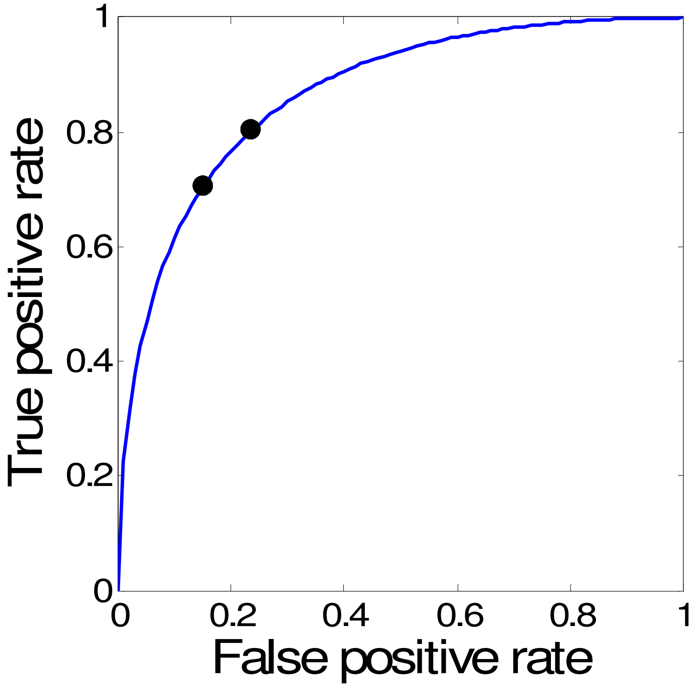
$$\begin{aligned}\tilde{p}_0(x) &= (1-\pi_0)p_0(x) + \pi_0 p_1(x) \\ \tilde{p}_1(x) &= (1-\pi_1)p_1(x) + \pi_1 p_0(x)\end{aligned}$$

- Simple algebra:

$$\frac{p_1(x)}{p_0(x)} > \gamma \iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} > \lambda,$$

where

$$\lambda = \frac{\pi_1 + \gamma(1-\pi_1)}{1-\pi_0+\gamma\pi_0}.$$



Modified Contamination Model

- Recall contaminaton model:

$$\begin{aligned} X_1, \ldots, X_m &\overset{iid}{\sim} \tilde{P}_0 = (1-\pi_0)P_0 + \pi_0 P_1 \\ X_{m+1}, \ldots, X_{m+n} &\overset{iid}{\sim} \tilde{P}_1 = (1-\pi_1)P_1 + \pi_1 P_0 \end{aligned}$$

- **Proposition:** If $\pi_0 + \pi_1 < 1$ holds and $P_0 \neq P_1$, then

$$\begin{aligned} \tilde{P}_0 &= (1-\tilde{\pi}_0)P_0 + \tilde{\pi}_0 \tilde{P}_1 \\ \tilde{P}_1 &= (1-\tilde{\pi}_1)P_1 + \tilde{\pi}_1 \tilde{P}_0 \end{aligned}$$

  where

$$\tilde{\pi}_0 = \frac{\pi_0}{1-\pi_1}, \quad \tilde{\pi}_1 = \frac{\pi_1}{1-\pi_0}$$

Error Estimation

- Focus on $R_0(f)$

$$\tilde{P}_0 = (1 - \tilde{\pi}_0) P_0 + \tilde{\pi}_0 \tilde{P}_1$$

$$\Downarrow$$

$$\tilde{R}_0(f) = (1 - \tilde{\pi}_0) R_0(f) + \tilde{\pi}_0 (1 - \tilde{R}_1(f))$$

$$\Downarrow$$

$$R_0(f) = \tfrac{\tilde{R}_0(f) - \tilde{\pi}_0 (1 - \tilde{R}_1(f))}{1 - \tilde{\pi}_0}$$

- Can estimate $\tilde{R}_0(f), \tilde{R}_1(f)$ accurately from data

- Suffices to estimate $\tilde{\pi}_0$

MPE for Label Noise

- Modified contamination model

$$X_1, \dots, X_m \overset{iid}{\sim} \tilde{P}_0 = (1 - \tilde{\pi}_0) P_0 + \tilde{\pi}_0 \tilde{P}_1$$

$$X_{m+1}, \dots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \tilde{\pi}_1) P_1 + \tilde{\pi}_1 \tilde{P}_0$$

- Need consistent estimates of $\tilde{\pi}_0$, $\tilde{\pi}_1$ → MPE
- Identifiability: Need

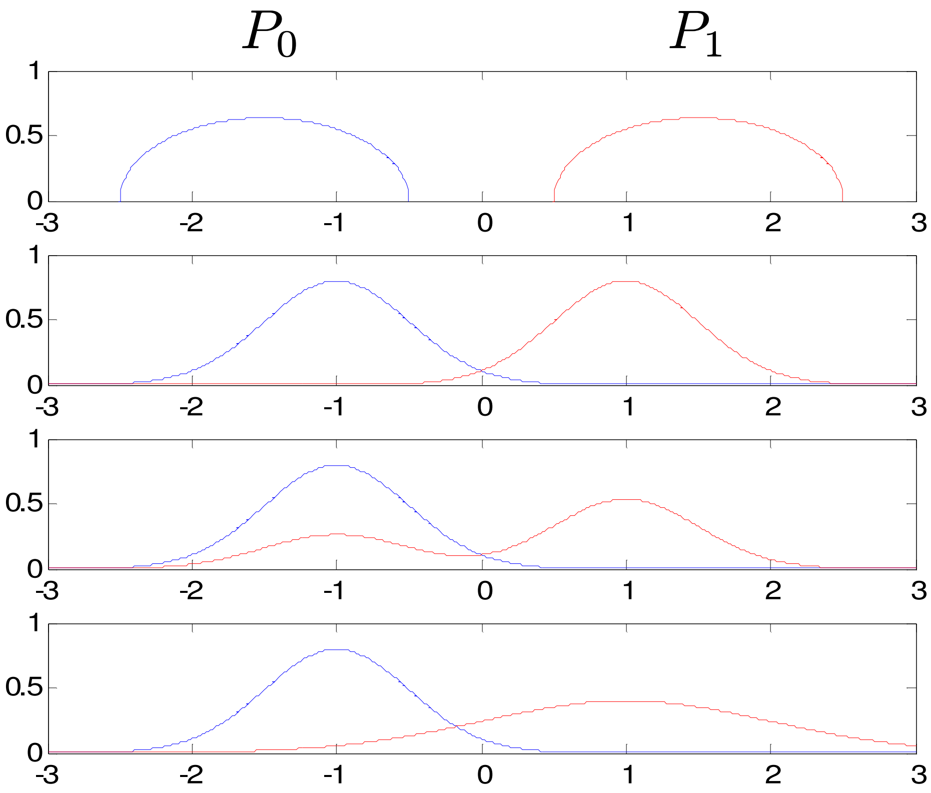
$$\nu^*(P_0, \tilde{P}_1) = 0 \text{ and } \nu^*(P_1, \tilde{P}_0) = 0$$

or equivalently (it can be shown)

$$\nu^*(P_0, P_1) = 0 \text{ and } \nu^*(P_1, P_0) = 0$$

Identifiability Condition

$$\nu^*(P_0, P_1) = 0 \text{ and } \nu^*(P_1, P_0) = 0$$



satisfied

satisfied

not satisfied

not satisfied

$$\left.\begin{array}{c}\\ \\ \\ \end{array}\right\} \begin{array}{c}\nu^*(P_0, P_1) = 0 \\ \text{but} \\ \nu^*(P_1, P_0) > 0\end{array}$$

Class Probability Estimation

- Assume joint distribution on (X, Y)

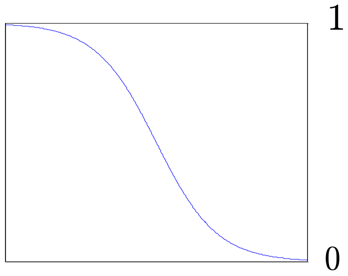
$$(X_i, Y_i) \overset{iid}{\sim} P_{XY}, \qquad Y_i \in \{0, 1\}$$

- Posterior probability

$$\eta(x) := P_{XY}(Y = 1 \,|\, X = x)$$

- Goal: Estimate η from training data
- Standard approach: logistic regression

$$\widehat{\eta}(x) = \frac{1}{1 + \exp(w^T x + b)}$$



- Let

$$\eta_{\max} := \sup_x \eta(x), \qquad \eta_{\min} := \inf_x \eta(x)$$

- Fact: **B** holds $\iff \eta_{\max} = 1$ and $\eta_{\min} = 0$

Classification with Unknown Class Skew

- Binary classification training data

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} P_1$$

- Test data:

$$Z_1, \ldots, Z_k \overset{iid}{\sim} P_{\text{test}} = \pi P_0 + (1 - \pi) P_1$$

- π **unknown**

- π needs to be known for several performance measures (probability of error, precision)

- MPE: If $\nu^*(P_1, P_0) = 0$ then $\pi = \nu^*(P_{\text{test}}, P_0)$

$$\rightarrow \quad \widehat{\pi} = \widehat{\nu}(\{X_i\}_{i=1}^m, \{Z_i\}_{i=1}^k)$$

Classification with Reject Option

- Binary classification training data

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} P_1$$

- Test data:

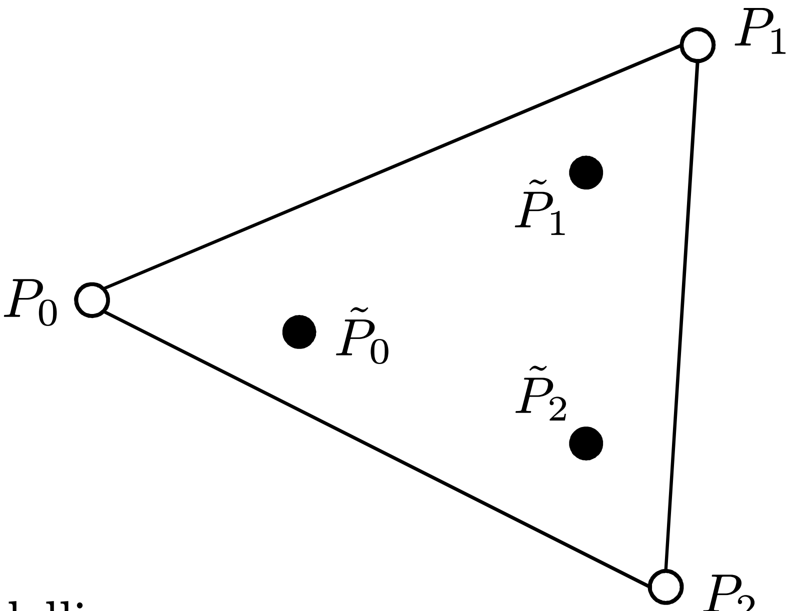
$$Z_1, \ldots, Z_k \overset{iid}{\sim} P_{\text{test}} = \pi_0 P_0 + \pi_1 P_1 + (1 - \pi_0 - \pi_1) P_2$$

- $P_2 =$ distribution of everything else (reject)
- π_0, π_1 **unknown**
- Use MPE (twice) to estimate π_0, π_1
  $\Longrightarrow$ estimate probability of class 2 error
  $\Longrightarrow$ design a classifier

Multiclass Label Noise

- Training distributions:

$$\tilde{P}_0 = (1 - \pi_{01} - \pi_{02})P_0 + \pi_{01}P_1 + \pi_{02}P_2$$
$$\tilde{P}_1 = \pi_{10}P_0 + (1 - \pi_{10} - \pi_{12})P_1 + \pi_{12}P_2$$
$$\tilde{P}_2 = \pi_{20}P_0 + \pi_{21}P_1 + (1 - \pi_{20} - \pi_{21})P_2$$



- Similar to topic modelling

Conclusion

- Mixture proportion estimation can be used to solve
  - Anomaly detection
  - Classification with label noise
  - Classification with unknown class skew
  - Classification with reject option
  - Multiclass label noise
  - Learning with partial labels
  - Two-sample problem
  - Multiple testing (interesting generalization of univariate p-value approach)
  - ???

Collaborators

- Gilles Blanchard
- Gregory Handy, Tyler Sanderson
- Marek Flaska, Sara Pozzi

Suppose Densities are Known

Problem of interest

$$\begin{aligned} H_0 : X &\sim p_0 \\ H_1 : X &\sim p_1 \end{aligned} \quad \Longrightarrow \quad \lambda \gtrless \frac{p_1(x)}{p_0(x)}$$

Surrogate problem

$$\begin{aligned} H_0 : X &\sim p_0 \\ \tilde{H}_1 : X &\sim \tilde{p}_1 \end{aligned} \quad \Longrightarrow \quad \lambda \gtrless \frac{\tilde{p}_1(x)}{p_0(x)} = \frac{(1-\pi)p_1(x) + \pi p_0(x)}{p_0(x)}$$

$$= (1-\pi)\frac{p_1(x)}{p_0(x)} + \pi$$

$$\Longrightarrow \quad \lambda' \gtrless \frac{p_1(x)}{p_0(x)}$$

Surrogate LR is monotone function of optimal test statistic → UMP test

- Data-based approach: Classification with prescribed false positive rate

- Challenges: Criteria other than Neyman-Pearson; estimating false negative rate