Classification in mutual contamination models

G. Blanchard

Universität Potsdam, Institut für Mathematik

Label Noise Workshop, Nancy 30/11/2017

Joint work with C. Scott, G. Handy, J. Katz-Samuels (U. Michigan)

PLAN

1. Contamination models
2. Binary classification case
   - One contaminated class
   - Mutual contamination
3. Multiclass case
   - One contaminated class
   - Mutual contamination

OUTLINE

STANDARD (GENERATIVE) SETTING FOR CLASSIFICATION

- $P_i \equiv P(X|Y=i)$: generating probability distributions for objects of class $1 \leq i \leq L$ on space $\mathcal{X}$.
- Observed: samples

$$S^i = (X_1^i, \ldots, X_{n_i}^i) \stackrel{i.i.d}{\sim} P_i$$

- **Goal:** estimate decision function $f : \mathcal{X} \to \{1, \ldots, L\}$
- Various performance error criteria: average classification error, min-max error, Neyman-Pearson error, ...

Standard Classification: General Principles

- Approximate P_i by corresponding empirical distribution $\widehat{P}_i$
- For all error criteria, key quantities to estimate for classifiers f are

$$R_i(f) := \mathbb{P}_i\left[f(X) \neq i\right] \to \widehat{R}_i(f) := \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{1}\{f(X_j^i) \neq i\}$$

- "Learning"/distribution-free philosophy:
  - don't want a specific (parametric) model for P_i.
  - (first) theoretical goal is universal consistency
- Basic strategy: uniform probabilistic control of $\left|R_i(f) - \widehat{R}_i(f)\right|$ over function/set classes $\mathcal{C}_k$
- Use structural risk minimization to choose adapted class $\mathcal{C}_k$

CONTAMINATION MODEL

- Assume the sample S_i is drawn according to a contaminated distribution:

$$S_i = (X_1^i, \ldots, X_{n_i}^i) \stackrel{i.i.d.}{\sim} \widetilde{P}_i = \sum_{i=1}^{L} \pi_{ij} P_j$$

or in short form

$$\widehat{\boldsymbol{P}} = \Pi \boldsymbol{P}$$

(Π: mixing matrix)

- Goal: find a classification function f that performs well for the true source distributions.
- Goal: estimate mixing weights Π and true sources (demixing)
- Can only access/ estimate

$$\widetilde{R}_i(f) := \widetilde{P}_i\,(f(X) \neq i)$$

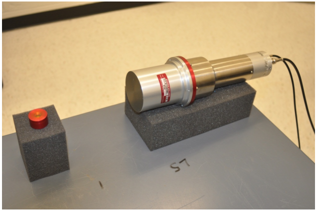
Equivalent Model: (Asymmetric) Random Label Noise Model

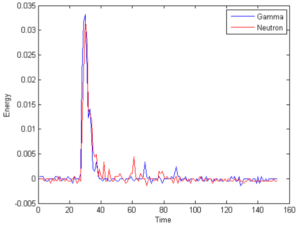
Assume

$$(X_i, Y_i) \overset{i.i.d.}{\sim} P;$$

- True labels Y_i unobserved, instead $\widetilde{Y}_i$
- Corrupted labels $\mathbb{P}\left[\widetilde{Y} = i | Y = j, X\right] = \zeta_{ij}$
- Label corruption assumed not to depend on X
- Label corruption not symmetric

MOTIVATING APPLICATION
ORGANIC SCINCILLATION DETECTOR





- Detect neutrons and gamma rays; need to classify between them
- Training using gamma ray source (e.g. Na-22) and neutron source (e.g. Cf-252)
- But: no pure neutron source – always mixed neutron/gamma ray
- Additionally, background radiation (both particles)

FURTHER SETTINGS AND GOALS

- Recover source distributions *up to permutation*: demixing problem.
  - Application: Topic models (each observed document is a mixture of topics; goal is to recover "pure" topic distribution themselves)

- Recover source distributions with the additional knowledge of the support of Π (positions of positive entries).
  - Application: Partial labels models (each object comes with a subset of labels)

UNDERSTANDING LABEL NOISE

- Assume P_0, P_1 have densities p_0, p_1
- Then $\widetilde{P}_0, \widetilde{P}_1$ have densities

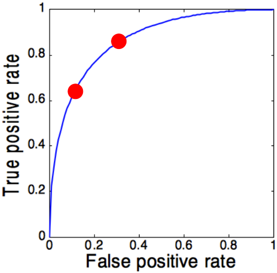
$$\begin{cases} \widetilde{p}_0 = (1-\kappa_0)p_0 + \kappa_0 p_1 \\ \widetilde{p}_1 = (1-\kappa_1)p_1 + \kappa_1 p_0 \end{cases}$$

Simple algebra:

$$\frac{p_1(x)}{p_0(x)} \lesseqgtr \lambda \Longleftrightarrow \frac{\widetilde{p}_1(x)}{\widetilde{p}_0(x)} \lesseqgtr \gamma,$$

where

$$\lambda(\gamma) = \frac{\kappa_1 + \gamma(1-\kappa_1)}{1-\kappa_0-\gamma\kappa_0}$$



Training a regular classifier on contaminated data leads to asymptotic bias and inconsistency except in very particular circumstances.

Related Work, Previous Assumptions

- Previous work on related topics include:
  - Learning on positive and unlabeled data (LPUE)
  - Co-training
  - Label noise models and PAC learning
- Generally the following is assumed:
  - P_0, P_1 have non-overlapping support ($\leftrightarrow$ deterministic target concept)
  - symmetric label noise
  - criterion is probablity of error
- We do not assume the above here
- Main asumption: label noise independent of X – no adversarial noise

- We can surely estimate $\widetilde{R}_i(f)$ from its empirical counterpart

$$\widehat{\widetilde{R}}_i(f) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}\{f(X_j^i) \neq i\},$$

uniformly in f in a limited complexity classifier class $\mathcal{C}_K$

- Observe

$$\begin{aligned}
\widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1 \Longrightarrow \quad & \widetilde{R}_0(f) = (1-\kappa_0)R_0(f) + \kappa_0 R_1(f) \\
\widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \Longrightarrow \quad & \widetilde{R}_1(f) = (1-\kappa_1)R_1(f) + \kappa_1 R_0(f)
\end{aligned}$$

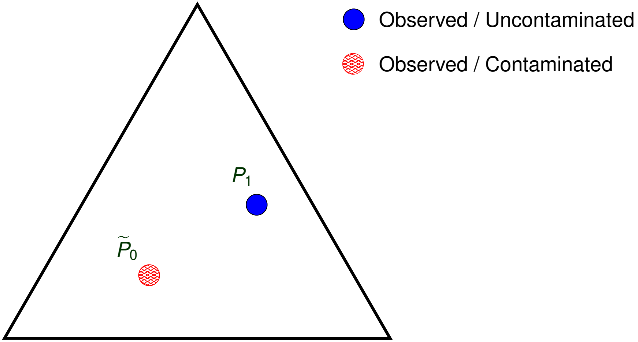
implying

$$\begin{aligned}
R_0(f) &= \frac{(1-\kappa_1)\widetilde{R}_0(f) - \kappa_0 \widetilde{R}_1(f)}{1-(\kappa_0+\kappa_1)}, \\
R_1(f) &= \frac{(1-\kappa_0)\widetilde{R}_1(f) - \kappa_1 \widetilde{R}_0(f)}{1-(\kappa_0+\kappa_1)}
\end{aligned}$$

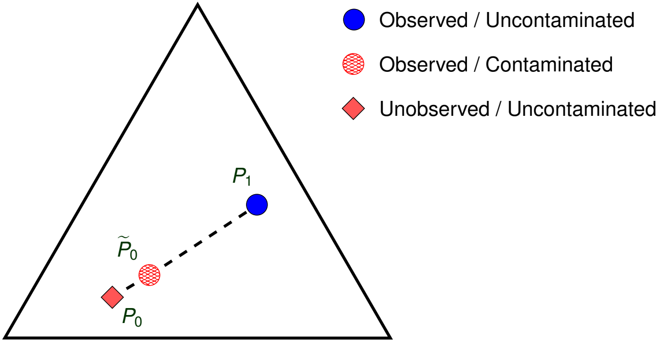
- Key point: estimation of contamination proportions κ_0, κ_1.

Outline

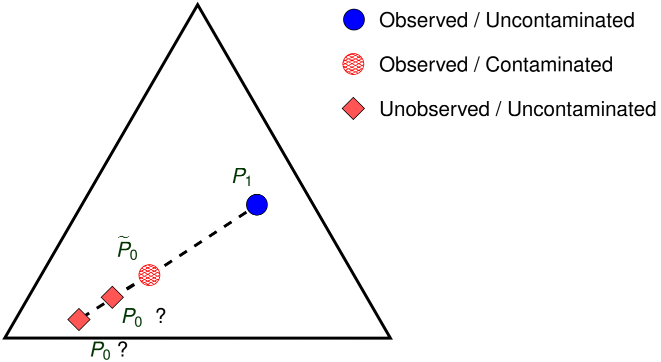
The Binary Case
Only One Contaminated Distribution



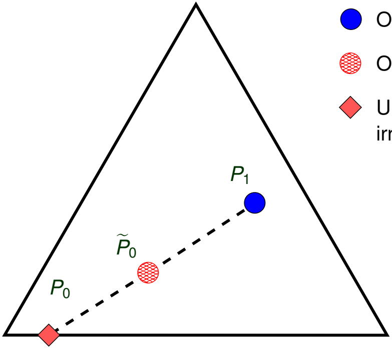
The Binary Case
Only One Contaminated Distribution



The Binary Case
Only One Contaminated Distribution



The Binary Case
Only One Contaminated Distribution



- Observed / Uncontaminated
- Observed / Contaminated
- Unobserved / Uncontaminated, irreducible wrt. P_1

Only $\widetilde{P}_0$ Contaminated: Identifiablity

$$\begin{cases} (X_1^0, \ldots, X_{n_0}^0) \overset{i.i.d.}{\sim} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1 \\ (X_1^1, \ldots, X_{n_1}^1) \overset{i.i.d.}{\sim} P_1 \end{cases}$$

- Define the "maximum proportion of source H in F"

$$\kappa^*(F|H) = \max\left\{\kappa \in [0,1] \middle| \exists \text{ a distribution } G \text{ s.t.} F = (1-\kappa)G + \kappa H\right\};$$

- The following holds:

$$\kappa_0 = \kappa^*(\widetilde{P}_0|P_1) \Leftrightarrow \kappa^*(P_0|P_1) = 0 \qquad (P_0 \text{ is irreducible wrt.} P_1)$$

ONLY $\widetilde{P}_0$ CONTAMINATED: ESTIMATION

- F, H distributions; Lebesgue decomposition:

$$F = F_H + F_H^{\perp},$$

with $F_H \ll H$ and $(F_H^{\perp}, H)$ mutually singular;

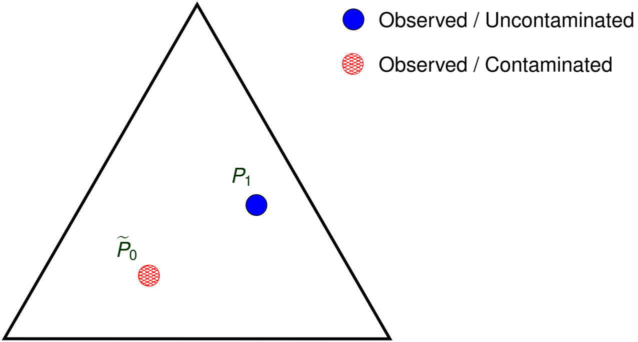
$$\kappa^*(F|H) = \mathrm{Ess.Inf.} \frac{dF_H}{dH} = \inf_{C:H(C)>0} \frac{F(C)}{H(C)}$$

- Suggests the estimator

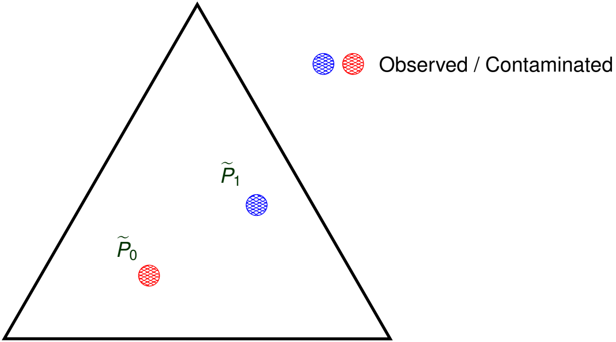
$$\widehat{\kappa}(\widehat{\widetilde{P}}_0|\widehat{P}_1) = \inf_{C \in \mathcal{C}_k} \frac{\widehat{\widetilde{P}}_0(C) + \varepsilon_k}{\left(\widehat{P}_1(C) - \varepsilon_k\right)_+}$$

- $\widehat{\kappa}(\widehat{\widetilde{P}}_0|\widehat{P}_1) \geq \kappa^*(\widetilde{P}_0|P_1)$ with high probability
- Appropriate choice of ε_k + take inf. over sequence of nested classes $\mathcal{C}_1 \subset \mathcal{C}_2 \subset \ldots$ with universal approximation property yields universally consistent estimator

MUTUAL CONTAMINATION



Mutual Contamination



Mutual Contamination

$$\begin{cases} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ \widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases}$$

Proposition (Decoupled Representation)

Assume $P_0 \neq P_1$ *and*

$$(\mathbf{A}) \qquad \kappa_1 + \kappa_2 < 1\,;$$

then $\widetilde{P}_0 \neq \widetilde{P}_1$*, and there exist unique* $0 \leq \widetilde{\kappa}_0, \widetilde{\kappa}_1 < 1$ *such that*

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0 \widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1 \widetilde{P}_0. \end{cases}$$

with

$$\widetilde{\kappa}_0 = \frac{\kappa_0}{1-\kappa_1} < 1; \qquad \widetilde{\kappa}_1 = \frac{\kappa_1}{1-\kappa_0} < 1.$$

IDENTIFIABILITY

Decoupled model:

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0\widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1\widetilde{P}_0. \end{cases}$$

From the results on mixture proportion estimation: we can estimate $\widetilde{\kappa}_0$ consistently if $\kappa(P_0, \widetilde{P}_1) = 0$

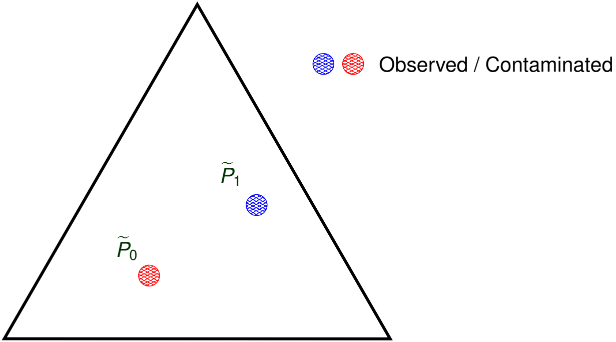
Lemma

Under assumption **(A)***:* $\kappa_0 + \kappa_1 < 1$*, it holds*

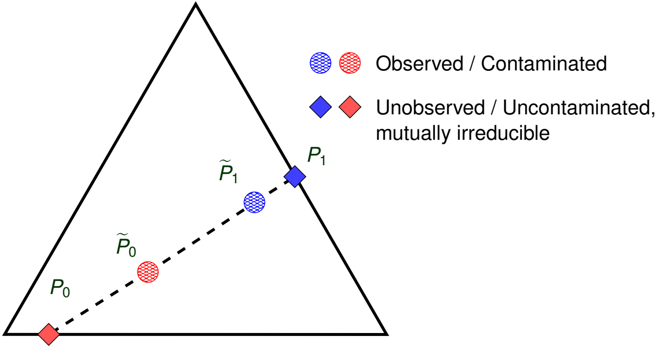
$$(\mathbf{B}) \begin{Bmatrix} \kappa(P_0|\widetilde{P}_1) = 0 \\ \kappa(P_1|\widetilde{P}_0) = 0 \end{Bmatrix} \Longleftrightarrow \begin{Bmatrix} \kappa(P_0|P_1) = 0 \\ \kappa(P_1|P_0) = 0 \end{Bmatrix} (\mathbf{C})$$

(C): P_0 and P_1 are mutually irreducible

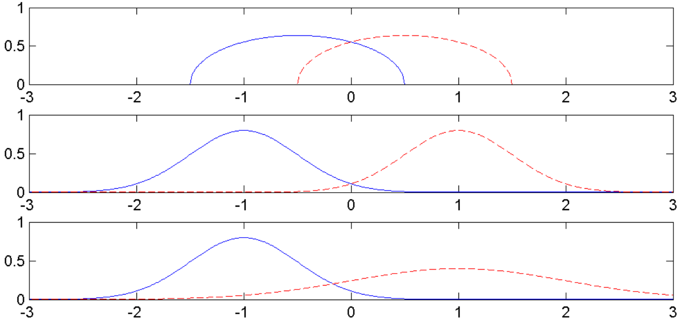
IDENTIFIABILITY



IDENTIFIABILITY



Mutual Irreducibility



- Top: mutually irreducible
- Middle: mutually irreducible
- Bottom: P_1 irreducible wrt P_0, but P_0 not irreducible wrt P_0.

MUTUAL IRREDUCIBILITY

Under joint distribution model

$$(X, Y) \sim \mathbb{P}_{XY}, \qquad \eta(x) = \mathbb{P}_{XY}\left[Y = 1 | X = x\right]$$

Then:

$$\left.\begin{array}{r}\kappa(P_0|P_1) = 0\\ \kappa(P_1|P_0) = 0\end{array}\right\} \Leftrightarrow \left\{\begin{array}{r}\mathrm{Ess.Sup.}_x\eta(x) = 1,\\ \mathrm{Ess.Inf.}_x\eta(x) = 0,\end{array}\right.$$

Interpretation of the Irreducible Solution

For given observed contaminated $\widetilde{P}_0 \neq \widetilde{P}_1$, let Δ be the convex set of quadruples $(\kappa_0, \kappa_1, P_0, P_1)$ satisfying **(A)** and solution of:

$$\begin{cases} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ \widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases} \tag{1}$$

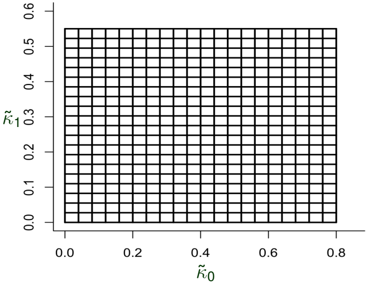
Proposition

The solution $(\kappa_0^*, \kappa_1^*, P_0^*, P_1^*)$ is characterized as either of:

- *the unique quadruple for which (P_0, P_1) are mutually irreducible;*
- *the unique maximizer over Λ of $\|P_0 - P_1\|_{TV}$.*
- *the unique minimizer over Λ of the Bayes error for classifying P_0 vs. P_1 with equal a priori proportions*

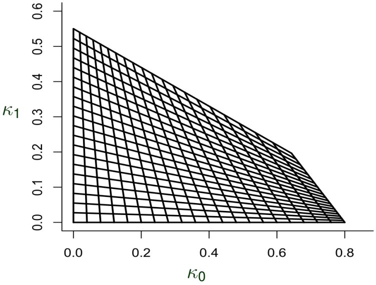
Interpretation: maximal denoising

THE TWO REPRESENTATIONS



Decoupled representation

$\leftrightarrows$

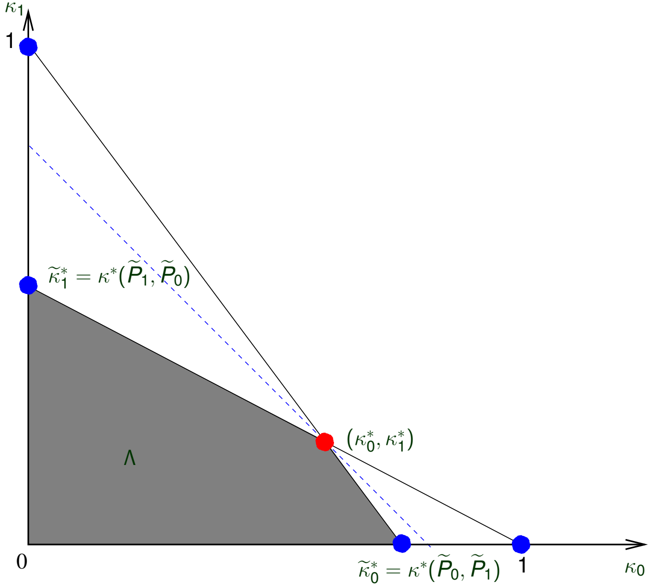


Original representation

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0 \widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1 \widetilde{P}_0. \end{cases}$$

$$\begin{cases} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ \widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases}$$

GEOMETRY OF SOLUTIONS



Consistent Estimation of Contamination Proportions

Decoupled representation:

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0 \widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1 \widetilde{P}_0. \end{cases}$$

- (P_0, P_1) mutually irreducible $\Rightarrow$ P_0 irreducible wrt $\widetilde{P}_1$, and P_1 irreducible wrt. $\widetilde{P}_1$
- leverage case of only one contaminated distribution (twice):

$$\widehat{\widetilde{\kappa}}_0 = \widehat{\kappa}(\widehat{\widetilde{P}}_0 | \widehat{\widetilde{P}}_1); \qquad \widehat{\widetilde{\kappa}}_1 = \widehat{\kappa}(\widehat{\widetilde{P}}_1 | \widehat{\widetilde{P}}_0)$$

- Then

$$\widehat{\kappa}_0 = \frac{\widehat{\widetilde{\kappa}}_0(1-\widetilde{\kappa}_1)}{1-\widetilde{\kappa}_0\widetilde{\kappa}_1}; \qquad \widehat{\kappa}_1 = \frac{\widehat{\widetilde{\kappa}}_1(1-\widetilde{\kappa}_0)}{1-\widetilde{\kappa}_0\widetilde{\kappa}_1}$$

are universally consistent estimators of κ_0, κ_1 under **(A)**, **(C)**.

CONSISTENT ESTIMATION OF RISK

- Construction of estimator for type II error:

$$\widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0\widetilde{P}_1 \Rightarrow R_0(f) = \frac{\widetilde{R}_0(f) - \widetilde{\kappa}_0(1-\widetilde{R}_1(f))}{1-\widetilde{\kappa}_0}$$

$$\rightarrow \widehat{R}_0(f) = \frac{\widehat{\widetilde{R}}_0(f) - \widehat{\widetilde{\kappa}}_0(1-\widehat{\widetilde{R}}_1(f))}{1-\widehat{\widetilde{\kappa}}_0}$$

- Uniform convergence over e.g. VC-Classes of classifiers f
- Can apply SRM principle to choose appropriate model
- Can construct universally consistent estimators for various error measures

OUTLINE

ONLY $\widetilde{P}_0$ CONTAMINATED

$$\begin{cases} (X_1^0, \ldots, X_{n_0}^0) \overset{i.i.d.}{\sim} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 \left(\sum_{i=1}^{M} \mu_i P_i \right) \qquad \text{with } \sum_i \mu_i = 1 \\ (X_1^i, \ldots, X_{n_1}^i) \overset{i.i.d.}{\sim} P_i\,; \qquad i = 1, \ldots, M \end{cases}$$

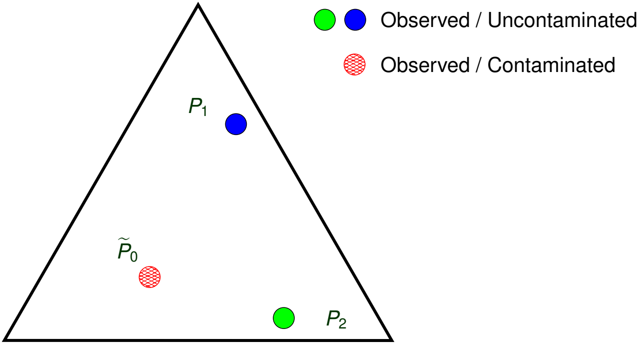
- Maximum collective proportion of $H_1, \ldots, H_M$ in F ?

$$\kappa^*(F|H_1, \ldots, H_M) = \max_{\boldsymbol{\mu} \in S_M} \kappa^*(F|H_{\boldsymbol{\mu}})$$

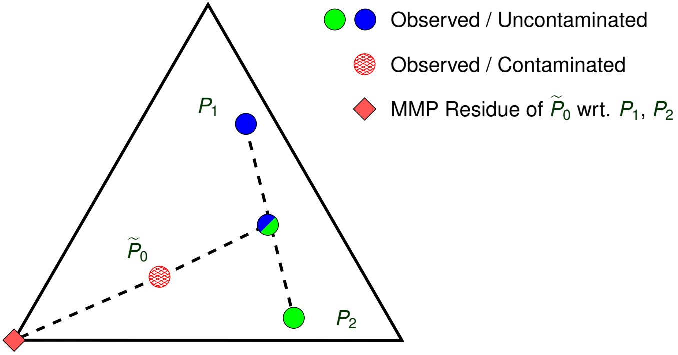
where

- S_M: $(M-1)$-dimensional simplex
- For $\boldsymbol{\mu} \in S_M$: $H_{\boldsymbol{\mu}} = \sum_{i=1}^{M} \mu_i H_i$
- Interpretation: attained for "projection" of F onto convex hull of $\{H_1, \ldots, H_M\}$ for the separation distance $1 - \kappa^*(F|\bullet)$

MAXIMAL MIXTURE PROPORTION



Maximal Mixture Proportion



Maximal Mixture Proportion Estimation (MMPE)

$$\kappa^*(F|H_1,\ldots,H_M) = \max_{\boldsymbol{\mu}\in S_M} \kappa^*(F|H_{\boldsymbol{\mu}})$$

- Estimator:

$$\widehat{\kappa}(\widehat{P}_0|\widehat{P}_1,\ldots,\widehat{P}_M) = \max_{\boldsymbol{\mu}\in S_M} \inf_{C\in\mathcal{C}_k} \frac{\widehat{P}_0(C)+\varepsilon_{0,k}}{\left(\widehat{P}_{\boldsymbol{\mu}}(C)-\sum_i \mu_i\varepsilon_{i,k}\right)},$$

  for $(\mathcal{C}_k)$ sequence of VC-classes
- $\widehat{\kappa} \geq \kappa^*$ with high probability
- Universally consistent if the VC sequence is universally approximating
- $\widehat{\boldsymbol{\mu}}$ attaining the max converges to the population maximum $\boldsymbol{\mu}$, whenever the latter is unique

IDENTIFIABILITY

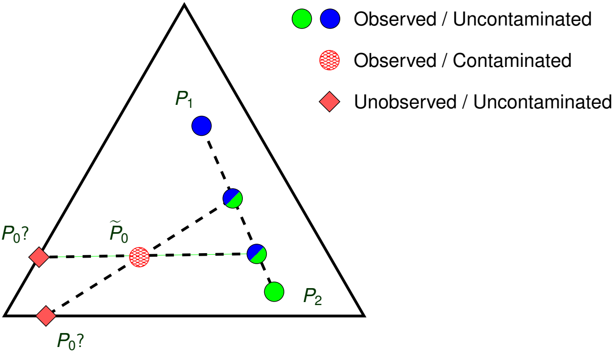
$$\widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 \left(\sum_{i=1}^{M} \mu_i P_i \right) \qquad \text{with } \sum_i \mu_i = 1$$

- When is it the case that $\kappa_0 = \kappa^*(\widetilde{P}_0 | P_1, \ldots, P_M)$?

IDENTIFIABILITY

$$\widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 \left(\sum_{i=1}^{M} \mu_i P_i \right) \qquad \text{with } \sum_i \mu_i = 1$$

- When is it the case that $\kappa_0 = \kappa^*(\widetilde{P}_0 | P_1, \ldots, P_M)$?



P_0 irreducible w.r.t. all $P_{\boldsymbol{\mu}}, \boldsymbol{\mu} \in S_M$ is **not** sufficient for identifiability

- $\rightarrow$ joint irreducibility of (P_i)...

Joint Irreducibility

Call a family of distributions $Q_1, \ldots, Q_L$ jointly irreducible under either of the equivalent conditions:

- For any $I \subset \{1, \ldots, L\}$; $1 \le |I| \le (L-1)$:
  any distribution in $\mathrm{ConvHull}\{Q_i, i \in I\}$ is irreducible with respect to any distribution in $\mathrm{ConvHull}\{Q_i, i \in I^c\}$
- If $\sum_{i=1}^{L} \gamma_i Q_i$ is a distribution, then $\gamma_i \ge 0$ for all i.
- If $\mathcal{M}_1(\mathcal{X})$ is the set of all probability distributions on $\mathcal{X}$,

$$\mathcal{M}_1(\mathcal{X}) \cap \mathrm{Span}\{Q_i, 1 \le i \le n\} = \mathrm{ConvHull}\{Q_i, 1 \le i \le n\}$$

JOINT IRREDUCIBILITY – INTERPRETATION

Assume:

- $(P_1, \ldots, P_L)$ are jointly irreducible;
- $\widetilde{P}_i = \boldsymbol{\pi}_i^T \boldsymbol{P}$, with $\boldsymbol{\pi}_i$ (rows of the mixing matrix Π) linearly independent

Then:

$$\kappa^*(\widetilde{P}_k | (\widetilde{P}_i)_{i \in I}) = \kappa^*(\boldsymbol{\pi}_k | (\boldsymbol{\pi}_i)_{i \in I}),$$

and there is a one-to-one correspondance between the set of residues.

RECOVERABILITY

Recall the general contamination model:

$$\widetilde{P}_i = \sum_{i=1}^{L} \pi_{ij} P_j \qquad \Longleftrightarrow \qquad \widetilde{\boldsymbol{P}} = \Pi \boldsymbol{P}$$

Call mixing weight matrix Π *recoverable* under either of the equivalent conditions:

- Π^{-1} has strictly positive diagonal entries and nonpositive off-diagonal entries
- For all ℓ, $\kappa^*(\boldsymbol{\pi}_\ell | \{\boldsymbol{\pi}_j, j \neq \ell\}) = \kappa_\ell$ is uniquely attained for decomposition

  $$\boldsymbol{\pi}_\ell = (1 - \kappa_\ell)\boldsymbol{e}_\ell + \kappa_\ell \boldsymbol{\pi}'_\ell, \qquad (*)$$

  where $\boldsymbol{\pi}_\ell$ is ℓ-th row of Π and $\boldsymbol{e} = \ell$-th canonical basis vector, $\boldsymbol{e}_\ell = (0, \ldots, 0, 1, 0, \ldots, 0)$

Decontamination under the recoverability assumption

- Recoverability implies $\boldsymbol{\pi}_\ell = (1-\kappa_\ell)\boldsymbol{e}_\ell + \kappa_\ell \boldsymbol{\pi}'_\ell$, unique maximal decomposition
- Irreducibility implies one-to-one correspondance, therefore

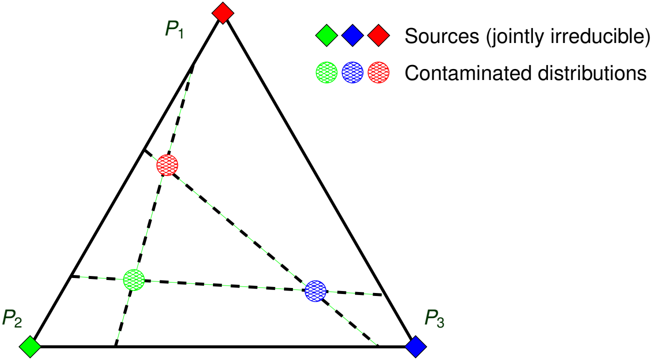
$$\widetilde{P}_\ell = (1-\kappa_\ell)P_\ell + \sum_{j \neq \ell} \nu_{\ell j} \widetilde{P}_j \,;$$

unique maximal decomposition.

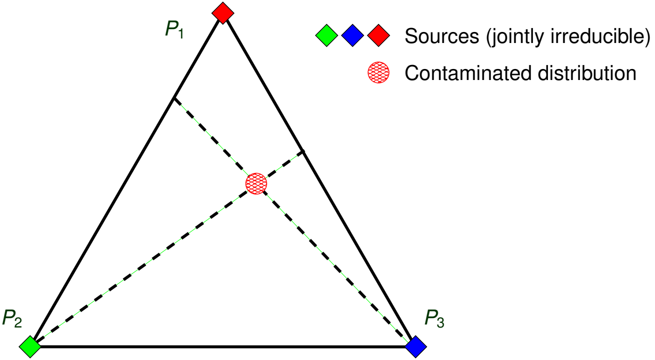
Decontamination under the recoverability assumption

- Recoverability implies $\boldsymbol{\pi}_\ell = (1-\kappa_\ell)\boldsymbol{e}_\ell + \kappa_\ell \boldsymbol{\pi}'_\ell$, unique maximal decomposition
- Irreducibility implies one-to-one correspondance, therefore
$$\widetilde{P}_\ell = (1-\kappa_\ell)P_\ell + \sum_{j\neq\ell} \nu_{\ell j}\widetilde{P}_j\,;$$
unique maximal decomposition.
- Conclusion: κ_ℓ can be estimated consistently by MMPE estimators $\widehat{\kappa}(\widehat{\widetilde{P}}_\ell | \{\widehat{\widetilde{P}}_j, j\neq\ell\})$
- We estimate also consistently the sources P_ℓ (residues), and further $\nu_{\ell j}$, Π^{-1} and finally Π

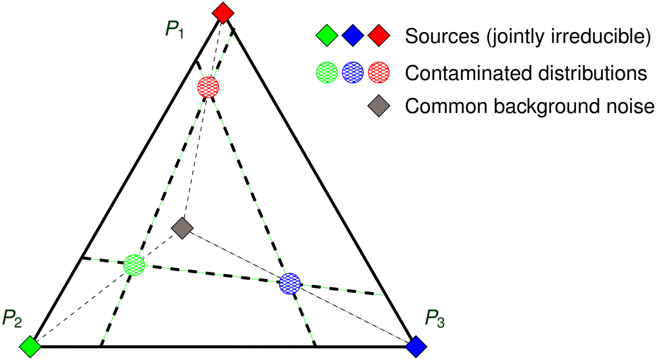
WHEN DOES RECOVERABILITY HOLD?



When Does Recoverability Hold?



When Does Recoverability Hold?



Consistent Estimation of Risk

From

$$\widetilde{P}_\ell = (1-\kappa_\ell)P_\ell + \sum_{j\neq\ell} \nu_{\ell j}\widetilde{P}_j$$

we get, denoting $\widetilde{R}_{ij}(f) = \widetilde{\mathbb{P}}_i(f(X)\neq j)$

$$R_\ell(f) = \frac{\widetilde{R}_{\ell\ell}(f) - \sum_{j\neq\ell}\nu_{\ell j}\widetilde{R}_{\ell j}}{1-\kappa_\ell} \qquad \longrightarrow \qquad \widehat{R}_\ell(f) = \frac{\widehat{\widetilde{R}}_{\ell\ell}(f) - \sum_{j\neq\ell}\widehat{\nu}_{\ell j}\widehat{\widetilde{R}}_{\ell j}}{1-\widehat{\kappa}_\ell}$$

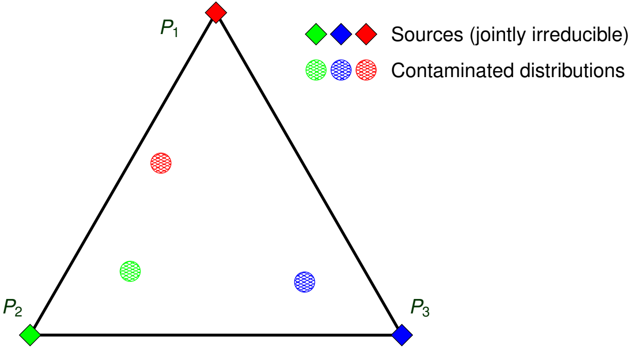
Then it holds:

$$\sup_{f\in\mathcal{F}_{k(\boldsymbol{n})}} \left|\widehat{R}_\ell(f) - R_\ell(f)\right| \to 0 \text{ in probability,}$$

as $\boldsymbol{n} = \min(n_1,\ldots,n_L)\to\infty$., for VC-classes $\mathcal{F}_k$ of dimension V_k, provided $\frac{V_{k(\boldsymbol{n})}\log \boldsymbol{n}}{\boldsymbol{n}} \to 0$

Demixing without recoverability

- Goal: estimating sources up to permutation (demixing problem)
- Try to "reduce dimension":

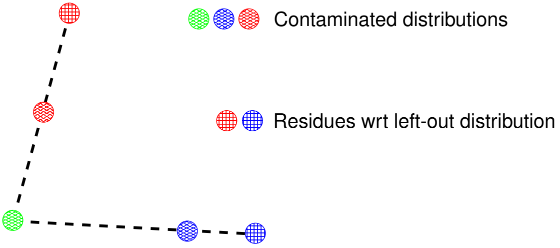


DEMIXING WITHOUT RECOVERABILITY

- Goal: estimating sources up to permutation (demixing problem)
- Try to “reduce dimension”:

DEMIXING WITHOUT RECOVERABILITY

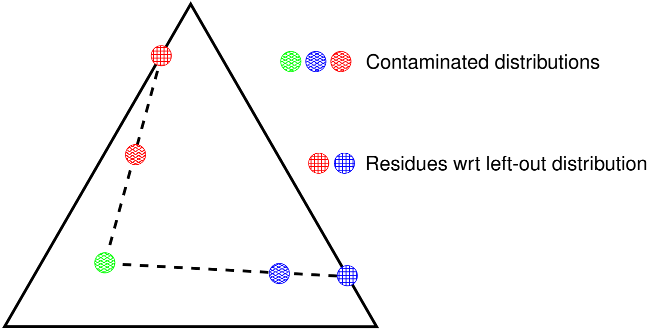
- Goal: estimating sources up to permutation (demixing problem)
- Try to “reduce dimension”:



- Residues always belong to the boundary

DEMIXING WITHOUT RECOVERABILITY

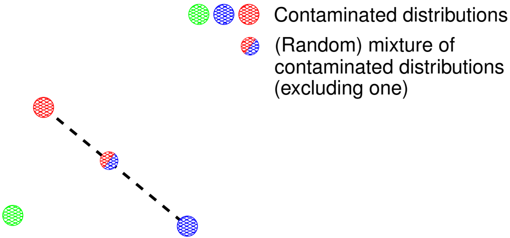
- Goal: estimating sources up to permutation (demixing problem)
- Try to "reduce dimension":



- Residues always belong to the boundary
- Need a test of whether the residues belong to the same "face"

Demixing without recoverability

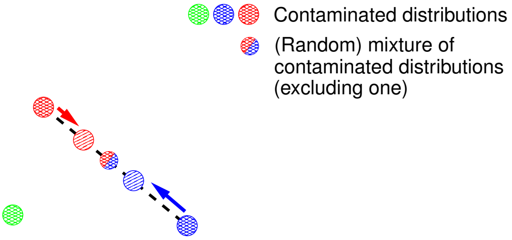
- Goal: estimating sources up to permutation (demixing problem)
- Try to "reduce dimension":



- Residues always belong to the boundary
- Need a test of whether the residues belong to the same "face"

DEMIXING WITHOUT RECOVERABILITY

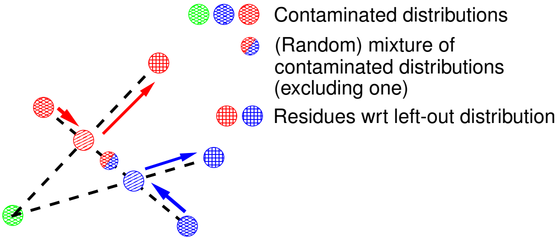
- Goal: estimating sources up to permutation (demixing problem)
- Try to “reduce dimension”:



- Residues always belong to the boundary
- Need a test of whether the residues belong to the same “face”

Demixing without recoverability

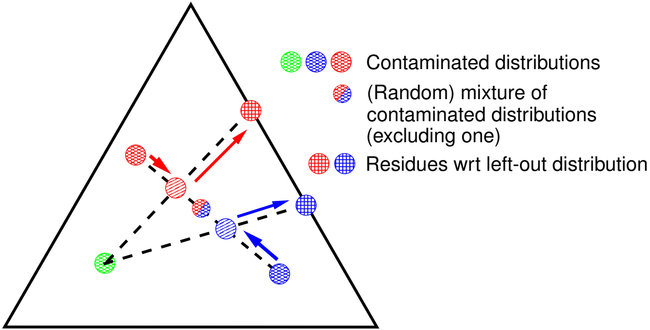
- Goal: estimating sources up to permutation (demixing problem)
- Try to "reduce dimension":



- Residues always belong to the boundary
- Need a test of whether the residues belong to the same "face"

DEMIXING WITHOUT RECOVERABILITY

- Goal: estimating sources up to permutation (demixing problem)
- Try to "reduce dimension":



- Residues always belong to the boundary
- Need a test of whether the residues belong to the same "face"
- If test does not reject, apply algorithm recursively

Demixing without recoverability

- Advantage: only need estimator $\widehat{\kappa}$ for two distributions (much simpler to implement)
- Advantage: only need full column rank (weaker than recoverability) to establish population consistency
- Disadvantage: need more iterations/retries, theoretical consistency of estimation only established under the stronger assumption of

$$\forall i \qquad \mathrm{Supp}(P_i) \not\subseteq \bigcup_{j \neq i} \mathrm{Supp}(P_j)$$

- Extension: If support $\boldsymbol{S}$ of Π is known, and all columns of $\boldsymbol{S}$ are unique, can recover the specific sources by support matching.

CONCLUSIONS

Contributions:

- Nonparametric/distribution-free point of view
- 2-class case: characterization of irreducible solution and consistent estimation
- Multiclass case:
  - Consistent maximal mixture proportions estimation
  - Consistent de-contamination under irreducibility + recoverability
  - Consistent de-mixing (up to permutation) under support irreducibility + full column rank
  - Consistent de-contamination under the same conditions as the previous point, if support of mixing weights known

THANK YOU FOR YOUR ATTENTION!

References

G. Blanchard, G. Lee, C. Scott.
Semi-Supervised Novelty Detection.
Journal of Machine Learning Research 11: 2973-3009, 2010.

C. Scott, G. Blanchard, G. Handy.
Classification with Asymmetric Label Noise: Consistency and Maximal Denoising.
COLT, 2013

G. Blanchard, C. Scott.
Decontamination of mutually contaminated models.
AISTATS, 2014

C. Scott.
A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels.
AISTATS, 2015.

J. Katz-Samuels, G. Blanchard, C. Scott.
Decontamination of Mutual Contamination Models
ArXiV: 1710.01167