Estimation of proportions under Open set Label Shift using Mahalanobis Projection

Bastien Dussap [1] & Gilles Blanchard [2] & Badr-Eddine Chérief-Abdellatif [3]

[1] *Laboratoire de Mathématiques d'Orsay, Inria, CNRS, Université Paris-Saclay, France, bastien.dussap@universite-paris-saclay.fr*

[2] *Laboratoire de Mathématiques d'Orsay, Inria, CNRS, Université Paris-Saclay, France, gilles.blanchard@universite-paris-saclay.fr*

[3] *CNRS, France, badr.eddine.cherief.abdellatif@gmail.com*

Résumé. Cette présentation aborde deux aspects clés de l'adaptation de domaine non supervisée : le Label Shift et son extension, l'open set label shift. Le Label Shift postule que la divergence entre les ensembles d'entraînement et de test réside uniquement dans la distribution des labels, tandis que l'open set label shift permet l'émergence de nouvelles classes dans la cible, tout en maintenant les distributions conditionnelles des classes invariantes. L'accent est mis sur l'estimation des proportions des labels dans l'échantillon test, un problème appelé quantification dans la littérature. S'appuyant sur des travaux antérieurs utilisant un embedding des classes via un classifier ou par des méthodes à noyaux, nous proposons d'utiliser la distance de Mahalanobis en estimant les matrices de covariances des différentes classes afin d'exploiter toute l'information disponible et non plus simplement les moyennes des embeddings. Dans cette présentation, nous mettons en avant deux théorèmes de consistance de cette méthode dans les deux contextes étudiés et appuyons nos résultats par des expériences numériques.

Mots-clés. Machine Learning, Label Shift, Quantification Learning, Adaptation de Domaine

Abstract. This presentation addresses two key aspects of unsupervised domain adaptation: Label Shift and its extension, the open set label shift. Label Shift posits that the discrepancy between training and testing sets lies solely in the distribution of labels, while open set label shift allows for the emergence of new classes in the target, while maintaining invariant conditional class distributions. The focus is on estimating label proportions in the test sample, a problem referred to as quantification in the literature. Building upon previous work utilizing class embedding through a classifier or kernel methods, we propose using the Mahalanobis distance by estimating the covariance matrices of different classes to leverage all available information, rather than just the means of embeddings. In this presentation, we highlight two consistency theorems of this method in the two studied contexts and support our findings by numerical experiments.

Keywords. Machine Learning, Label Shift, Quantification Learning, Adaptation Domain

1 Introduction

In this talk, we will focus on two particular instances of unsupervised domain adaptation: the estimation of the target label proportions under *Label Shift* [1, 8] and an extension proposed independently by Garg et al. [6] and Dussap et al. [2] named *Open set Label Shift*.
The **Label Shift** hypothesis states that the distributions of the training and test sets differ only in the marginal distribution of the label $p(y)$, while the conditional distributions of the covariates given the label $p(y|x)$, are assumed to be the same. For example, label shift occurs in infectious disease modelling where the covariates are the observed symptoms while the label is the underlying disease state. During an epidemic, the expected proportion of sick individuals is larger than usual, although the distribution of symptoms given the disease state remains unchanged.
Open set label shift, suppose that the label distribution changes arbitrarily and a new class emerges, but the class conditional distributions $p(y|x)$ remain domain invariant.

Several different objectives have been addressed in the literature under these assumptions, here we focused on on the estimation of the target label proportions. For label shift, this problem has many names such as *class ratio estimation* [7] or *posterior probability shift* [5] but the most commonly used is **quantification** [4]. To the best of our knowledge, only two papers have addressed the quantification problem under open set label shift [2, 6]. In this talk we continue the unification work we did and presented last year [2, 3].

Our main contribution is a variant of the method we proposed in our previous work where we embed the distribution using, for instance, a classifier or Kernel Mean Embedding [9]. We now rely on the Mahalanobis distance equipped with an estimate of the covariance matrices of the embeddings to exploit the information obtained from each point and not just the mean as we did. We obtain new theoretical results and show that this new method is still robust in the open set label shift setting.

Numerical experiments support our theory.

1.1 Notations

Formally, consider a covariate space $\mathcal{X}$, typically a subset of $\mathbb{R}^d$, and two label spaces $\mathcal{Y} = \{1, \dots, c\}$ and $\tilde{\mathcal{Y}} = \{0, \dots, c\}$. We define the *source* noted $\mathbb{P}$ and the *target* noted $\mathbb{Q}$, as different probability distributions over the covariate label space pair $\mathcal{X} \times \mathcal{Y}$ for the source and $\mathcal{X} \times \tilde{\mathcal{Y}}$ for the target.

The target label distribution is denoted $\alpha^* = (\alpha_i^*)_{i=0}^c$ while each class-i conditional target distribution is denoted $\mathbb{Q}_i$. Similarly, the source label distribution is denoted $\beta^* = (\beta_i^*)_{i=1}^c$ while each class-i conditional source distribution is denoted $\mathbb{P}_i$. We assume that the Open Set Label Shift hypothesis holds:

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i + \alpha_0^* \mathbb{Q}_0 \qquad (\mathcal{OSLS})$$
$$\forall i = 1, \dots, c\,, \quad \mathbb{P}_i = \mathbb{Q}_i.$$

We recover the label shift setting when $\alpha_0^* = 0$:

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i \qquad (\mathcal{LS})$$
$$\forall i = 1, \dots, c\,, \quad \mathbb{P}_i = \mathbb{Q}_i,$$

and in that case $\tilde{\mathcal{Y}} = \mathcal{Y}$.

A source dataset $\{(x_j, y_j)\}_{j\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n$ and a target dataset $\{x_{n+j}\}_{j\in[m]} \in \mathcal{X}^m$ are given. All data points from the source (respectively the target) dataset are sampled independently from the source (resp. the target) domain. We have access to the source labels y_j but not to the unobserved target labels. We denote by $\hat{\mathbb{P}}_i \coloneqq \sum_{j\in[n]:y_j=i} \delta_{x_j}(\cdot)/n_i$ the empirical conditional distribution of class i from the source, where n_i denotes the number of instances labelled i in the source dataset. Note that $n_1 + \ldots + n_c = n$. We denote by $\tilde{\beta}$ the empirical proportions of each class in the source dataset, i.e. $\tilde{\beta}_i \coloneqq n_i/n$. Likewise, we denotes by $\hat{\mathbb{Q}} \coloneqq \sum_{j\in[m]} \delta_{x_{n+j}}(\cdot)/m$ the empirical distribution of the target.

Finally, for any function $\Phi : \mathcal{X} \to \mathcal{F}$, let's note $V = \left[\Phi(\hat{\mathbb{P}}_1), \cdots, \Phi(\hat{\mathbb{P}}_c)\right]$, $\boldsymbol{G} = V^T V$ the gram matrix of the source empirical embeddings and for any matrix (or operator) M, let's note $\boldsymbol{G}^M = V^T M^T M V$ the gram matrix associated with the Mahalanobis distance.

2 Distribution Feature Matching

In this section we present our previous work [3].

Let $\Phi : \mathcal{X} \to \mathcal{F}$ be a fixed feature mapping from $\mathcal{X}$ into a Hilbert space $\mathcal{F}$ (possibly $\mathcal{F} = \mathbb{R}^D$). We extend the mapping Φ to probability distributions on $\mathcal{X}$ by taking the expectation, i.e. $\Phi\colon \mathbb{P} \mapsto \mathbb{E}_{X\sim\mathbb{P}}[\Phi(X)] \in \mathcal{F}$. Thus it holds $\Phi(\hat{\mathbb{P}}_i) = n_i^{-1} \sum_{j\in[n]:y_j=i} \Phi(x_j)$, and similarly for $\Phi(\hat{\mathbb{Q}})$.
We call *Distribution Feature Matching* (DFM) any estimation procedure that can be formulated as the minimiser of the following problem:

$$\hat{\alpha} = \arg\min_{\alpha\in\Delta^c} \left\| \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2 \qquad (\mathcal{P})$$

where Δ^c is the (c-1)-dimensional simplex.

This general formulation allows us to encompass other algorithms in the literature. Two in particular: *Kernel Mean Matching* [7] that rely on Kernel Mean Embedding and *BBSE* [8] that use the output of a classifier to embed the data.

Theoretical guarantees

We make the following hypothesis on the mapping Φ:

$$\sum_{i=1}^{c} \beta_i \Phi(\mathbb{P}_i) = 0 \iff \beta_i = 0 \ \forall i \tag{$\mathcal{A}_1$}$$

$$\|\Phi(x)\| \leq C \text{ for all } x. \tag{$\mathcal{A}_2$}$$

We can state the main theorem of [2]:

Theorem 2.1. *If the Label Shift hypothesis $\mathcal{LS}$ holds, and if the mapping Φ verifies the assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$*
Then with probability greater than $1-\delta$:

$$\|\hat{\alpha} - \alpha^*\|_2 \leq \frac{2CR_{c/\delta}}{\sqrt{\lambda_{min}}} \left(\sqrt{\frac{\|w\|_1}{n}} + \frac{1}{\sqrt{m}} \right) \tag{1}$$

$$\leq \frac{2CR_{c/\delta}}{\sqrt{\lambda_{min}}} \left(\frac{1}{\sqrt{\min_i n_i}} + \frac{1}{\sqrt{m}} \right), \tag{2}$$

where $R_x = 1 + \sqrt{2\log(2/x)}$, $w_i = \frac{\alpha_i^*}{\hat{\beta}_i}$ and $\lambda_{min} := \lambda_{min}(\boldsymbol{G}^M)$ is the smallest eigenvalue of $\boldsymbol{G}^M$.

In other words, Theorem 2.1 states that under label shift and mild conditions on the embedding Φ, the DFM procedure converges to the true proportions α^* with speed $\mathcal{O}\left((\min_i n_i)^{-1/2}\right)$ in the worst case and $\mathcal{O}\left((c/n)^{1/2}\right)$ in the best case scenario where the proportions of the source and the target haven't changed.

2.1 Open Set Label Shift

In our previous work, we proposed a new procedure called *soft*-DFM to deal with the Open Set Label Shift setting $\mathcal{OSLS}$. Since the proportions α^* we want to estimate no longer sum to one, the "hard" condition $\sum_i \alpha_i = 1$ in $\mathcal{P}$ is no longer needed:

$$\underset{\alpha \in \text{int}(\Delta^c)}{\arg\min} \left\| \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2, \tag{$\mathcal{P}_2$}$$

where $\text{int}(\Delta^c) := \{x \in \mathbb{R}^c \colon x \geq 0, \ \sum x_i \leq 1\}$.

We had a theorem for *soft*-DFM procedure:

Theorem 2.2. *Denote by $\hat{\alpha}_{\text{soft}}$ the minimiser of the soft-DFM problem $\mathcal{P}_2$. Assume the Open Set Label Shift hypothesis ($\mathcal{OSLS}$) holds. If the mapping Φ verifies Assumptions ($\mathcal{A}_1$) and ($\mathcal{A}_2$). Then, with probability greater than $1-\delta$:*

$$\|\hat{\alpha}_{\text{soft}} - \alpha^*\|_2 \leq \frac{1}{\sqrt{\lambda_{min}}}\Big(3\epsilon_n + \varepsilon_m + \sqrt{2\alpha_0\ \epsilon_n\ \|\Phi(\mathbb{Q}_0)\|_{\mathcal{F}}} + \|\Pi_V(\Phi(\mathbb{Q}_0))\|_{\mathcal{F}}\Big),$$

with:

$$\epsilon_n = \frac{R_{\delta/\log c}}{\sqrt{\min_i n_i}}; \qquad \varepsilon_m = \frac{R_\delta}{\sqrt{m}};$$

and Π_V the orthogonal projection on V.

Observe that the bound of theorem 2.2 shows robustness of DFM procedures against perturbations $\Phi(\mathbb{Q}_0)$ that are orthogonal to the source embeddings.
In particular if we use a classifier to embed the source distributions, the feature space has the same dimension as the number of sources, so under the condition ($\mathcal{A}_1$), V will coincide with E_1, the affine space of vectors summing to one. Since any distribution will also be mapped to E_1, the orthogonal component will always be 0. Thus, we do not expect any particular robustness property for BBSE methods. On the other hand, if we use Kernel Mean Embedding with a translation invariant kernel i.e. $k(x,y) = \varphi(x-y)$, then for any distributions $\mathbb{P}, \mathbb{P}'$ it holds $\langle \Phi(\mathbb{P}), \Phi(\mathbb{P}')\rangle = \mathbb{E}_{(X,Y)\sim\mathbb{P}\otimes\mathbb{P}'}[\varphi(X-Y)]$. Thus, if φ decays rapidly (e.g. Gaussian kernel), the feature mappings $\Phi(\mathbb{P})$ and $\Phi(\mathbb{P}')$ will be nearly orthogonal (have a scalar product close to 0) whenever the distributions $\mathbb{P}$ and $\mathbb{P}'$ are well separated. From this analysis, we expect that embedding the data using the Gaussian kernel will lead to a robust *soft*-DFM procedure for contamination distributions $\mathbb{Q}_0$ whose main mass is far from the source distributions.

3 Mahalanobis Distance

We now propose a new method where instead of minimising the L_2 distance between the embeddings as we did ($\mathcal{P}$), we minimise the Mahalanobis distance associated with an operator M:

$$\hat{\alpha} = \underset{\alpha\in\Delta^c}{\operatorname{argmin}} \left\| M\left(\sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}})\right)\right\|_{\mathcal{F}}^2 \qquad (\mathcal{MP})$$

To study the approximation error of our proposed estimator, we rely on a Bernstein inequality for Hilbert space-valued independent random variables due to Wolfer et al. [11], based on seminal work by Pinelis [10] and Yurinsky [12].

Theorem 3.1 (Bernstein)**.** *Let $\mathcal{H}$ be a Hilbert space, and let (z_i) be n independent random variables (not necessarily identically distributed) with values in $\mathcal{H}$. Suppose that for each i, $\|z_i\| \leq C$ almost surely, where $C < \infty$. Denote $\overline{\Sigma} \coloneqq \frac{1}{n}\sum \Sigma_{z_i}$, where Σ_{z_i} denotes the covariance matrix of z_i. Then for any $0 < \delta < 1$, with confidence $1-\delta$ it holds that,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} M(z_i - \mathbb{E}z_i) \right\| \leq \frac{2}{3}\sigma_1(M)\frac{CL_{1/\delta}}{n} + \sqrt{\frac{2L_{1/\delta}}{n} Tr(M\overline{\Sigma}M^\top)}, \tag{3}$$

where $L_x = \log(2x)$.

We can now state our main theorem: For any matrix M (or any operator M if $\mathcal{H}$ is an infinite dimensional space), under label shift we have

Theorem 3.2. *If the Label Shift hypothesis holds $\mathcal{LS}$ and if the mapping Φ verifies Assumptions $(\mathcal{A}_1)$, $(\mathcal{A}_2)$, then for any $\delta \in (0,1)$, with probability greater than $1-\delta$, the solution $\hat{\alpha}$ of $(\mathcal{MP})$ satisfies:*

$$\begin{aligned}\|\hat{\alpha} - \tilde{\alpha}\| \leq R_1(\delta, c)\frac{\|M\|_{\text{op}} C}{\sqrt{\lambda_{min}(\boldsymbol{G}^M)}}\left(\frac{\|w\|_1}{n} + \frac{1}{m}\right) \\ + R_2(\delta, c)\sqrt{\frac{Tr(M\Sigma_{\tilde{\alpha}}M^\top)}{\lambda_{min}(\boldsymbol{G}^M)}}\left(\sqrt{\frac{\|w\|_1}{n}} + \frac{1}{\sqrt{m}}\right),\end{aligned}$$

with $w = \frac{\tilde{\alpha}_i}{\tilde{\beta}_i}$, $R_1(\delta, c) = 4/3\log(4c/\delta)$, $R_2 = 2\sqrt{2\log(4c/\delta)}$, $\Sigma_{\tilde{\alpha}} = \sum_{i=1}^{c} \tilde{\alpha}\Sigma_i$ *and* Σ_i *the covariance matrix of* $\Phi(\mathbb{P}_i)$.

We have the same order of convergence as in theorem 2.1, but the constant before the term in $\mathcal{O}\left(n^{-1/2}\right)$ no longer depends on C but on the trace of the covariance matrix.

The matrix M which minimises the upper bound minimises :

$$\frac{Tr(M\Sigma_{\tilde{\alpha}}M^\top)}{\lambda_{\min}(\boldsymbol{G}^M)}. \tag{4}$$

The following theorem gives the close-form expression of this optimal M.

Theorem 3.3. *If the mapping Φ verifies Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_1)$. The matrix M that minimise 4 verify:*

$$MM^\top = \Sigma_{\tilde{\alpha}}^{-1/2}\left(\Sigma_{\tilde{\alpha}}^{-1/2} V V^\top \Sigma_{\tilde{\alpha}}^{-1/2}\right)^{+} \Sigma_{\tilde{\alpha}}^{-1/2}, \tag{$\mathcal{M}$}$$

Where the notation $()^+$ designate the Moore–Penrose inverse or Pseudo-inverse of a matrix.

4 is then equals to

$$\text{Tr}\left(\left(\Sigma_{\tilde{\alpha}}^{-1/2} V V^\top \Sigma_{\tilde{\alpha}}^{-1/2}\right)^{+}\right) \tag{5}$$

Figure 1 shows the two bounds from theorem 2.1 and theorem 3.2 with respect to the number of points in the source and target. As we can see, this new bound is better than the previous one, especially for small numbers of points.

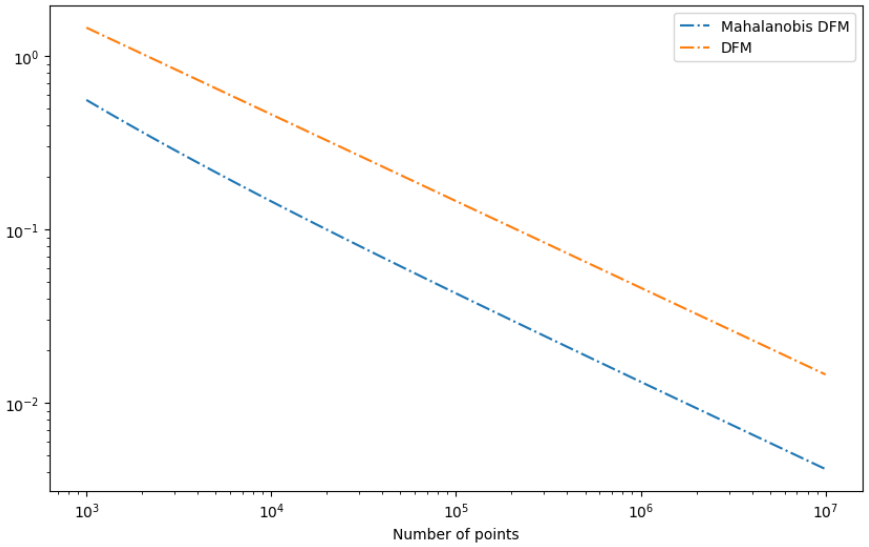

Figure 1: In blue the bound of theorem 3.2 and in yellow the bound of theorem 2.1. On the x-axis the number of points (in logarithmic scale) in the source and target, on the y-axis the value of both bounds (in logarithmic scale).

The default choice for M would be to take $\tilde{M} = \Sigma_{\tilde{\alpha}}$. Without the constraint $\alpha \in \Delta^c$, the solution of $\mathcal{MP}$ for the optimal M and $\tilde{M}$ are the same.

With the constraint however, the two solutions are different. For the optimal M the solution is $\mathrm{Proj}_{\Delta^c}\Big(V^+ \mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big)$ while for $M = \Sigma^{-1/2}$ the solution is $V^+ \mathrm{Proj}_{\mathcal{C},\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)$, where $\mathcal{C}$ denote the convex hull of the empirical source embeddings, i.e. $\mathcal{C} \coloneqq \Big\{\sum \beta_i \Phi(\hat{\mathbb{P}}_i) \colon \beta \in \Delta^c\Big\}$. One can show that :

$$\mathrm{Proj}_{\Delta^c}\Big(V^+ \mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big) = V^+ \mathrm{Proj}_{\mathcal{C}}\Big(\mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big),$$
$$V^+ \mathrm{Proj}_{\mathcal{C},\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big) = V^+ \mathrm{Proj}_{\mathcal{C},\Sigma^{-1/2}}\Big(\mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big).$$

Put simply, the difference between the optimal choice of M and the standard choice M^* concerns only the renormalisation of the proportions given by $\mathrm{Proj}_{V,\Sigma^{-1/2}}\Big(\Phi(\hat{\mathbb{Q}})\Big)$. In the first case, the coordinates are projected orthogonally onto the simplex, while in the second case the coordinates are projected along the metric induced by $\Sigma^{-1/2}$. In particular, if $\mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big) \in \mathcal{C}$ then the two solutions coincide.
This explains why in practice we don't see any difference between the two M.

Open Set Label Shift and Mahalanobis

Similarly to $\mathcal{P}_2$, we propose *soft*-MDFM:

$$\underset{\alpha \in \text{int}(\Delta^c)}{\arg\min} \left\| M \left(\sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right) \right\|_{\mathcal{F}}^2, \qquad (\mathcal{MP}_2)$$

Theorem 2.2 is still true when applied to *soft*-MDFM, except that the orthogonality condition that ensures robustness to the new class is now with respect to the Mahalanobis metric.

4 Experiments

In our experiment, the source consists of a list of c Gaussian distributions: $\mathbb{P}_1, \cdots, \mathbb{P}_c$ in $\mathbb{R}^D$. The important parameter is ρ and it controls how far the source is from the target. We tested with $c = 5$ and $c = 2$ classes, $D = 2, 5$ and 10, and we tested $\rho = 1$ we called it the "close setting" because the source and the target are close, and $\rho = 10$ the "far setting" where the source and the target are far from each other.
We refer to BBSE as the method that uses a classifier to embed the data, RFFM as the method that uses the Gaussian kernel, and MahalanobisRFFM as the method that uses the Gaussian kernel and a Mahalanobis distance with $M = \left(\hat{\Sigma} + \lambda I\right)$.

RFFM

Percentage of noise ϵ	**Quantifier**	Number of classes = 5		
		dim = 2	dim = 5	dim = 10
0.0	BBSE	4.11 ; 3.0	1.23 ; 3.0	0.97 ; 3.0
0.0	RFFM	1.75 ; 2.0	0.82 ; 2.0	0.84 ; 2.0
0.0	MahalanobisRFFM	**1.55 ; 1.0**	**0.71 ; 1.0**	**0.71 ; 1.0**
0.2	BBSE	26.90 ; 3.0	15.43 ; 3.0	12.42 ; 2.5
0.2	RFFM	**16.20 ; 2.0**	12.76 ; 2.0	12.22 ; 2.5
0.2	MahalanobisRFFM	**17.38 ; 2.0**	**11.69 ; 1.0**	**10.94 ; 1.0**
0.5	BBSE	52.43 ; 3.0	39.42 ; 3.0	31.95 ; 3.0
0.5	RFFM	**30.70 ; 1.0**	31.65 ; 2.0	30.51 ; 2.0
0.5	MahalanobisRFFM	33.98 ; 2.0	**29.88 ; 1.0**	**27.88 ; 1.0**
0.7	BBSE	67.25 ; 3.0	52.79 ; 3.0	44.55 ; 3.0
0.7	RFFM	**45.76 ; 1.0**	44.09 ; 2.0	41.21 ; 2.0
0.7	MahalanobisRFFM	52.71 ; 2.0	**42.76 ; 1.0**	**39.50 ; 1.0**

Table 1: **Gaussian Mixture: Comparison of BBSE, RFFM and MahalanobisRFFM when the new class is close**. The value before the semicolon is the geometric mean of the L2 error (multiplied by 100 for clarity) over 50 repetitions. Value after the semicolon is the median rank of a method relative to the other method in the same setting (same dimension, number of classes and percentage of noise). A bold value in a group is not significantly different from the best-performing method in the group (also in bold), as measured by a paired Wilcoxon test at $p < 0.01$.

Percentage of noise ϵ	**Quantifier**	Number of classes = 5		
		dim = 2	dim = 5	dim = 10
0.0	BBSE	4.11 ; 3.0	1.23 ; 3.0	0.97 ; 3.0
0.0	RFFM	1.75 ; 2.0	0.82 ; 2.0	0.84 ; 2.0
0.0	MahalanobisRFFM	**1.55 ; 1.0**	**0.71 ; 1.0**	**0.71 ; 1.0**
0.2	BBSE	32.88 ; 3.0	23.96 ; 3.0	21.67 ; 3.0
0.2	RFFM	**3.17 ; 1.5**	**1.85 ; 1.0**	**1.87 ; 1.0**
0.2	MahalanobisRFFM	**3.27 ; 1.5**	**1.92 ; 2.0**	**2.03 ; 2.0**
0.5	BBSE	66.21 ; 3.0	60.31 ; 3.0	55.56 ; 3.0
0.5	RFFM	**6.00 ; 1.0**	**4.42 ; 1.0**	**4.78 ; 1.0**
0.5	MahalanobisRFFM	**6.56 ; 2.0**	**4.50 ; 2.0**	**4.89 ; 2.0**
0.7	BBSE	82.88 ; 3.0	77.70 ; 3.0	75.12 ; 3.0
0.7	RFFM	**6.74 ; 1.0**	**5.64 ; 1.0**	**6.88 ; 1.0**
0.7	MahalanobisRFFM	**7.03 ; 2.0**	5.94 ; 2.0	**6.94 ; 2.0**

Table 2: **Gaussian Mixture: Comparison of BBSE, RFFM and MahalanobisRFFM when the new class is far**. The value before the semicolon is the geometric mean of the L2 error (multiplied by 100 for clarity) over 50 repetitions. Value after the semicolon is the median rank of a method relative to the other method in the same setting (same dimension, number of classes and percentage of noise). A bold value in a group is not significantly different from the best-performing method in the group (also in bold), as measured by a paired Wilcoxon test at $p < 0.01$.

RFFM and its Mahlanobis version are both robust to noise that is far from the source, but not when it is close, whereas BBSE is never robust. This confirms our theoretical analysis. What we can also see is that even in the close setting, RFFM and MahalanobisRFFM both outperform BBSE, and MahalanobisRFFM is often slightly better than RFFM.

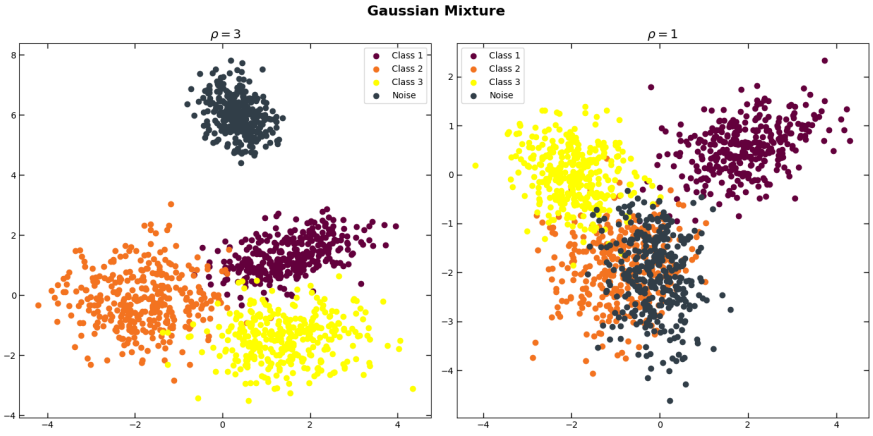

Figure 2: **Plot of the two settings in dimension 2 and with 3 classes.** On the left the new class in black is far from the sources in plum, orange and yellow. On the right the new class in black is close. Given of our theoretical analysis, we expect robustness for RFFM for the left setting and no robustness for the right setting.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[2] Bastien Dussap, Gilles Blanchard, and Badr-Eddine Chérief-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 69–85. Springer Nature Switzerland, 2023.

[3] Bastien Dussap, Gilles Blanchard, and Badr-Eddine Chérief-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching. `https://bastiendussap.github.io/assets/files/slides/JdS2023.pdf`, 2023.

[4] Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. *Learning to Quantify*, volume 47. Springer Nature, 2023.

[5] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the saerens-latinne-decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–34, 2020.

[6] Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, 35:22531–22546, 2022.

[7] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538. PMLR, 2014.

[8] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[9] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[10] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[11] Geoffrey Wolfer and Pierre Alquier. Variance-aware estimation of kernel mean embedding. *arXiv preprint arXiv:2210.06672*, 2022.

[12] Vadim Yurinsky. *Sums and Gaussian vectors*. Springer, 2006.