A unified framework for label shift quantification

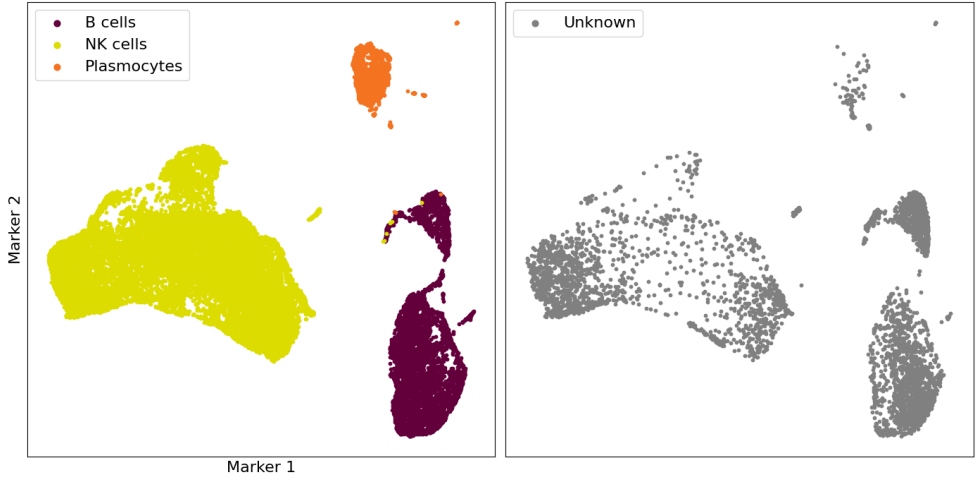
Cadre unifié pour la quantification label shift

Dussap Bastien

Laboratoire de mathématiques d'Orsay
Université Paris-Saclay, Inria

Tuesday 1st October, 2024

Introductory example



Domain adaptation

Model

- $\mathcal{X}$: The data space, in our case $\mathbb{R}^d$.
- $\mathcal{Y}$: The label space, $\{1, \cdots, c\}$.
- $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$: Two distributions on $\mathcal{X} \times \mathcal{Y}$.
- $\mathbb{P}$ is the Source.
- $\mathbb{Q}$ is the Target.

Training set

$$((x_1, y_1), \cdots, (x_n, y_n)) \sim \mathbb{P}$$

Testing set

$$(x_{n+1}, \cdots, x_{n+m}) \sim \mathbb{Q}_X$$

Label shift

Definition 2.1

We say that $\mathbb{P}$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, follow the label shift assumption if $\forall i \in [c]$ and $\forall x \in \mathcal{X}$

$$\mathbb{P}(x|Y = i) = \mathbb{Q}(x|Y = i), \tag{1}$$

the distribution $\mathbb{Q}_X$ can therefore be decomposed as follows

$$\mathbb{Q}_X(x) = \sum_{i=1}^{c} \underbrace{\mathbb{Q}_Y(i)}_{\alpha_i^*} \mathbb{P}(x|Y = i) \tag{2}$$

Label Shift Quantification

Goal: Estimate proportions

Estimate the proportions $\alpha^* \coloneqq \mathbb{Q}_Y(\cdot)$.


Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José Del Coz *"A review on quantification learning"*. In *ACM Computing Surveys*, 2017.


Andrea Esuli, Alessandro Fabris, Alejandro Moreo and Fabrizio Sebastiani *"Learning to Quantify"*. In *Springer Nature*, 2023

Notations

Notations

- $\mathbb{P}_i := \mathbb{P}(x|Y = i)$: List of c distributions.
- $\mathbb{Q}_i := \mathbb{Q}(x|Y = i)$: List of c distributions.
- $\alpha^* \in \Delta^c$: Class proportions, where $\Delta^c := \left\{x \in \mathbb{R}_+^c : \sum_{i=1}^c x_i = 1\right\}$.

Notations

Training set

$$(x_1^i, \cdots x_{n_i}^i) \sim \mathbb{P}_i \in \mathcal{X}$$

$$\hat{\mathbb{P}}_i := \frac{1}{n_i} \sum_{j \in [n]: y_j = i} \delta_{x_j}(\cdot)$$

$$n = \sum n_i.$$

Testing set

$$(x_{n+1}, \cdots, x_{n+m}) \sim \mathbb{Q}$$

$$\hat{\mathbb{Q}} := \frac{1}{m} \sum_{j=1}^{m} \delta_{x_{n+j}}(\cdot)$$

Label Shift Quantification

Goal: Estimate proportions

Use $(\hat{\mathbb{P}}_i)_{i=1}^{c}$ and $\hat{\mathbb{Q}}$ to estimate the proportions $\alpha^* := \mathbb{Q}_Y(\cdot) \in \Delta^c$.


Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José Del Coz *"A review on quantification learning"*. In *ACM Computing Surveys*, 2017.


Andrea Esuli, Alessandro Fabris, Alejandro Moreo and Fabrizio Sebastiani *"Learning to Quantify"*. In *Springer Nature*, 2023

Classify and Count (CC)

Classify and Count

Let $\hat{f}$ be a classifier trained on the training set.

$$\hat{\alpha}_{\mathrm{cc}} = \left(\frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\left\{ \hat{f}(x_{n+j}) = i \right\} \right) \xrightarrow[m \to \infty]{} \alpha_{cc} := \left(\mathbb{Q}_X(\hat{f}(x) = i) \right)_i$$

Problem : Label Shift

$$\alpha_{cc} = C_{\hat{y}|y} \times \alpha^*, \tag{3}$$

with $\left(C_{\hat{y}|y}\right)_{i,j} = \mathbb{P}(\hat{f}(x) = i | y = j)$.

Adjusted Classify and Count (ACC)

Black-Box Shift Estimation (BBSE)

$$\hat{\alpha}_{\mathrm{BBSE}} = \hat{C}_{\hat{y}|y}^{-1} \hat{\alpha}_{\mathrm{cc}} \tag{4}$$

George Forman. *"Counting positives accurately despite inaccurate classification"*. In *ECML*, 2005.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. *"Detecting and correcting for label shift with black box predictor"*. In *ICML*, 2018.

Variant of BBSE

Black-Box Shift Estimation (BBSE+)

$$\hat{\alpha}_{\text{BBSE+}} = \underset{\alpha \in \Delta^c}{\arg\min} \left\| \hat{\alpha}_{\text{cc}} - \hat{C}_{\hat{y}|y} \alpha \right\|_2 \tag{5}$$

Daniel J. Hopkins and Gary King. *"A method of automated nonparametric content analysis for social science."*. In *American Journal of Political Science*, 2010.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang and Geoffrey J. Gordon. *"Domain adaptation with conditional distribution matching and generalized label shift"*. In *Advances in Neural Information Processing Systems*, 2020.

Kernel methods

Positive-definite kernel

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if it is symmetric, i.e. $k(x, y) = k(y, x)$, and if the Gram matrix is positive:

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \tag{6}$$

for all $n \in \mathbb{N}$, any choice of $x_1, \ldots, x_n \in \mathcal{X}$ and any $c_1, \ldots, c_n \in \mathbb{R}$.

Kernel methods

Kernel Methods

For a positive-definite kernel k, there exists a functional Hilbert space $\mathcal{H}_k$ and an embedding $\phi_k : \mathcal{X} \mapsto \mathcal{H}_k$ such that:

$$k(x, y) = \langle \phi_k(x), \phi_k(y) \rangle_{\mathcal{H}_k}$$

Kernel

Classical kernels

- $k(x, y) = x^T y$, linear.
- $k(x, y) = (\gamma x^T y + c_0)^d$, polynomial.
- $k(x, y) = \tanh(\gamma x^T y + c_0)$, sigmoid.
- $k(x, y) = \exp(-\gamma \|x - y\|_2^2)$, Gaussian.
- $k(x, y) = \exp(-\gamma \|x - y\|_1)$, Laplacian.
- $k(x, y) = \|x\| + \|y\| - \|x - y\|$, energy.
- $k(x, y) = \left(1 + \frac{\|x-y\|^2}{\sigma^2}\right)^{-1}$, Cauchy.

Kernel Mean Embedding

Kernel Mean Embedding

$$\Phi_k \colon \mathcal{P}(\mathcal{X}) \to \mathcal{H}_k$$
$$\mathbb{P} \mapsto \mathbb{E}_{\mathbb{P}}[\phi_k(x)].$$

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Schölkopf *"Kernel mean embedding of distributions: A review and beyond"*. In *Foundations and Trends in Machine Learning*, 2017.

Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD)

The function :

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

is a pseudo-distance on $\mathcal{P}(\mathcal{X})$, called the Maximum Mean Discrepancy.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf and Alex Smola. *"A kernel method for the two-sample problem"*. In *Advances in neural information processing systems*, 2006.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet and Bernhard Schölkopf. *"Injective Hilbert space embeddings of probability measures"*. In *21st annual conference on learning theory (COLT)*, 2008.

Kernel Mean Matching

Kernel Mean Matching (KMM)

$$\hat{\alpha}_{\mathrm{KMM}} = \underset{\alpha \in \Delta^c}{\arg\min}\ \mathrm{MMD}\left(\sum_{i=1}^{c} \alpha_i \hat{\mathbb{P}}_i, \hat{\mathbb{Q}}\right), \tag{7}$$

where $\Delta^c := \{x \in \mathbb{R}_+^c : \sum x_i = 1\}$.

Arun Iyer, Saketh Nath, and Sunita Sarawagi. *"Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection"*. In *ICML*, 2014.

Hideko Kawakubo, Marthinus Christoffel du Plessis and Masashi Sugiyama. *"Computationally efficient class-prior estimation under class balance change using energy distance"*. In *IEICE Transactions on Information and Systems*, 2016.

Unification

- Aykut Firat. *"Unified framework for quantification"*. In *arXiv preprint arXiv:1606.00868*, 2016.
- Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. *"A unified view of label shift estimation"*. In *Advances in Neural Information Processing Systems*, 2020.
- Mirko Bunse. *"Unification of algorithms for quantification and unfolding"*. In *INFORMATIK*, 2022.
- Mirko Bunse. *"Qunfold: Composable quantification and unfolding methods in python"*. In *Proceedings of the 3rd International Workshop on Learning to Quantify*, 2023.

Distribution Matching

Definitions

- $\Phi\colon \mathcal{P}(\mathcal{X}) \mapsto \mathcal{Z}$.
- $D\colon \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}^+$, a distance (or pseudo-distance) on $\mathcal{Z}$.

Distribution Matching

$$\hat{\alpha}_{\mathrm{DM}} = \underset{\alpha \in \Delta^c}{\arg\min}\ D\left(\Phi\left(\sum_{i=1}^{c} \alpha_i \hat{\mathbb{P}}_i\right), \Phi(\hat{\mathbb{Q}})\right) \tag{8}$$

Mean Vectorisation or Mean Embedding

$\rightarrow$ Let $\phi\colon \mathcal{X} \to \mathcal{F}$ be a fixed feature mapping from $\mathcal{X}$ into a Hilbert space $\mathcal{F}$ (possibly $\mathcal{F} = \mathbb{R}^D$).

Mean Embedding

$$\Phi(\mathbb{P}) := \mathbb{E}_{X\sim\mathbb{P}}[\phi(X)] \in \mathcal{F}$$

Distribution Feature Matching (DFM)

DFM

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

where $\Delta^c := \{x \in \mathbb{R}_+^c : \sum x_i = 1\}$.

Examples

Examples

- $\phi(x) = \phi_k(x) \in \mathcal{H}_k$, Kernel Mean Matching.
- $\phi(x) = (1\{\hat{f}(x) = i\})_{i=1,\dots,c} \in \mathbb{R}^c$. For $\hat{f}$ a classifier, BBSE+.

BBSE+ as Distribution feature matching

Proposition 3.2.

When $\phi(x) = (1\{\hat{f}(x) = i\})_{i=1,\dots,c} \in \mathbb{R}^c$:

$$\Phi(\mathbb{P}_j)_i := (\mathbb{E}_{X\sim\mathbb{P}_j}[\phi(x)])_i = \mathbb{P}(\hat{f}(x) = i|Y = j),$$

and

$$\Phi(\mathbb{Q}_X)_i := (\mathbb{E}_{X\sim\mathbb{Q}_X}[\phi(x)])_i = \mathbb{Q}_X(\hat{f}(x) = i) = (\alpha_{cc})_i,$$

so that :

$$\left\|\sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}})\right\|_2 = \left\|\alpha \hat{C}_{\hat{y}|y} - \hat{\alpha}_{\mathrm{cc}}\right\|_2$$

Main Theorem under Label Shift

Assumptions

$$\exists C > 0 : \ \|\phi(x)\|_{\mathcal{F}} \leq C \text{ for all } x \qquad (\mathcal{A}_1)$$

and

$$\sum_{i=1}^{c} \beta_i \Phi(\mathbb{P}_i) = 0 \iff \beta = 0. \qquad (\mathcal{A}_2)$$

Main Theorem under Label Shift

Theorem 3.1.

For any $\delta \in (0,1)$, with probability greater than $1-\delta$:

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

$$\leq \frac{2CR_{c/\delta}}{\sqrt{\Delta_{\min}(\hat{G})}} \left(\frac{1}{\sqrt{\min_i n_i}} + \frac{1}{\sqrt{m}} \right), \tag{$\mathcal{E}_2$}$$

where $R_x = 1 + \sqrt{2\log(2/x)}$, $w_i = \alpha_i^*/(n_i/n)$ and $\hat{G}_{i,j} = \left\langle \Phi(\hat{\mathbb{P}}_i), \Phi(\hat{\mathbb{P}}_j) \right\rangle$.

Properties of $\Delta_{\min}$

Proposition 3.3

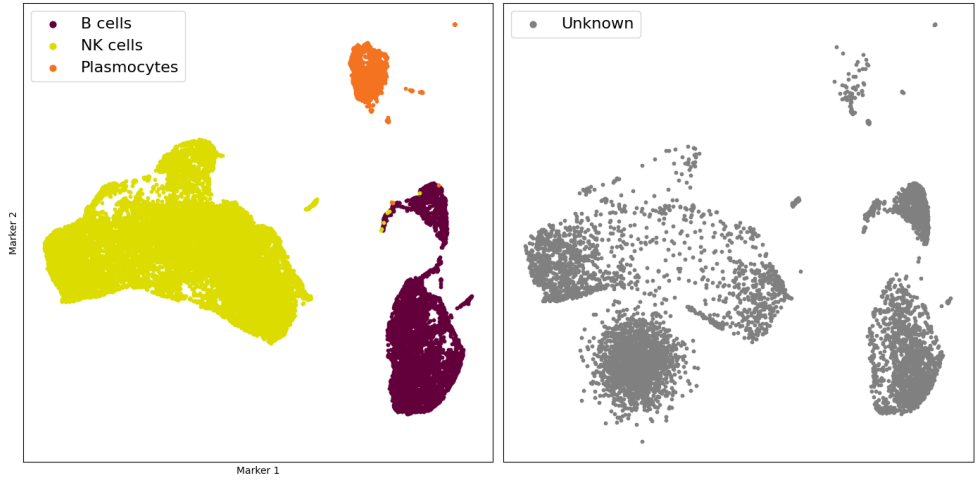
For any number of classes c, $\Delta_{\min}(\hat{\boldsymbol{G}})$ is equal to $\min_{\substack{\|u\|=1 \\ \mathbf{1}^T u=0}} u^T \hat{\boldsymbol{G}} u$.

$\rightarrow$ $\Delta_{\min}(\hat{\boldsymbol{G}})$ is always greater than the smallest eigenvalue of the Gram matrix.

Corollary 3.1

In particular for two classes, $\sqrt{\Delta_{\min}(\hat{\boldsymbol{G}})} = \frac{1}{\sqrt{2}} \left\| \Phi(\hat{\mathbb{P}}_1) - \Phi(\hat{\mathbb{P}}_2) \right\|$.

Introductory example



Open Set Label Shift

Definition

We say that $\mathbb{P}$ and $\mathbb{Q}$, follow the open set label shift assumption if $\forall i \in \{1, \cdots, c\}$ and $\forall x \in \mathcal{X}$

$$\mathbb{P}_i(x) = \mathbb{Q}_i(x),$$

the distribution $\mathbb{Q}_X$ can therefore be decomposed as follows

$$\mathbb{Q}_X(x) = \sum_{i=1}^{c} \underbrace{\mathbb{Q}_Y(i)}_{\alpha_i^*} \mathbb{P}_i(x) + \mathbb{Q}_Y(0)\mathbb{Q}_0(X) \tag{9}$$

Open Set Label Shift Quantification

Goal: Estimate proportions

Estimate the proportions $\alpha^* := (\mathbb{Q}_Y(1), \cdots, \mathbb{Q}_Y(c)) \in \bar{\Delta}^c$.
Where $\bar{\Delta}^c := \{x \in \mathbb{R}_+^c : \sum x_i \leq 1\}$


Saurabh Garg, Sivaraman Balakrishnan and Zachary Lipton. *"Domain adaptation under open set label shift"*. In *Advances in Neural Information Processing Systems*, 2022.


Bastien Dussap, Gilles Blanchard and Badr-Eddine Chérief-Abdellatif. *"Label Shift Quantification with Robustness Guarantees via Distribution Feature Matching"*. In *ECML/PKDD*, 2023

soft-DFM

soft-DFM

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

Main theorem under Open Set Label Shift

Corollary 3.2.

Assume the mapping Φ satisfies Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$. Then, with probability greater than $1-\delta$:

$$\|\hat{\alpha}_{\text{soft}} - \alpha^*\|_2 \lesssim \lambda_{\min}(\hat{\boldsymbol{G}})^{-1/2} \left(\sqrt{\left\|\Phi(\mathbb{Q}_0)\right\|} \min_i n_i^{-1/4} + \left\|\Pi_V(\Phi(\mathbb{Q}_0))\right\| \right), \quad (\mathcal{E}_3)$$

where n_i is the number of points in the i-th class of the training set and Π_V is the orthogonal projector on $V = \text{Span}\{\Phi(\mathbb{P}_i)\}$.

Orthogonality

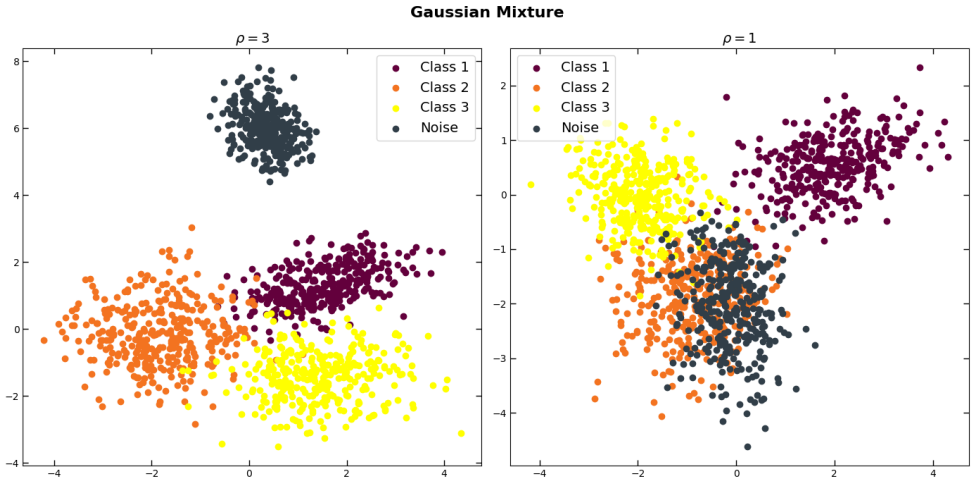
Translation-invariant kernel

For translation-invariant kernel, i.e. $k(x, y) = \gamma(x - y)$, it holds

$$\left\langle \Phi(\mathbb{P}), \Phi(\mathbb{P}') \right\rangle_{\mathcal{H}_k} = \mathbb{E}_{(X,Y)\sim\mathbb{P}\otimes\mathbb{P}'}[\gamma(X - Y)].$$

For rapidly decaying functions γ, two embeddings $\Phi(\mathbb{P})$ and $\Phi(\mathbb{P}')$ will be nearly orthogonal if $\mathbb{P}$ and $\mathbb{P}'$ are sufficiently separated.
We expect DFM with the Gaussian kernel to be robust against contamination distributions $\mathbb{Q}_0$ concentrated far from the source distributions $(\mathbb{P}_i)_i$.

Experiments : Gaussian Mixture



Gaussian mixture : $\rho = 10$

Percentage of contamination $\mathbb{Q}_Y(0)$	**Quantifier**	Number of classes = 5		
		dim = 2	dim = 5	dim = 10
0 %	CC	1.50	0.78	0.58
0 %	BBSE+	1.45	0.45	**0.37**
0 %	KMM (Gaussian)	**0.65**	**0.30**	**0.32**
0 %	KMM (Energy)	1.21	0.39	0.38
20 %	CC	5.23	4.59	4.41
20 %	BBSE+	9.90	5.93	5.36
20 %	KMM (Gaussian)	**1.17**	**0.69**	**0.70**
20 %	KMM (Energy)	13.83	12.43	9.43
50 %	CC	10.87	10.34	10.25
50 %	BBSE+	17.42	14.77	13.53
50 %	KMM (Gaussian)	**2.19**	**1.66**	**1.81**
50 %	KMM (Energy)	19.99	18.83	17.14
70 %	CC	14.47	14.22	14.16
70 %	BBSE+	19.78	18.41	17.93
70 %	KMM (Gaussian)	**2.54**	**2.13**	**2.62**
70 %	KMM (Energy)	20.65	19.93	19.32

Gaussian mixture : $\rho = 1$

Percentage of contamination $\mathbb{Q}_Y(0)$	**Quantifier**	Number of classes = 5		
		dim = 2	dim = 5	dim = 10
0 %	CC	1.50	0.78	0.58
0 %	BBSE+	1.45	0.45	**0.37**
0 %	KMM (Gaussian)	**0.65**	**0.30**	**0.32**
0 %	KMM (Energy)	1.21	0.39	0.38
20 %	CC	**4.94**	**4.30**	**4.06**
20 %	BBSE+	8.54	4.59	4.22
20 %	KMM (Gaussian)	**5.09**	**4.34**	4.49
20 %	KMM (Energy)	9.72	7.40	5.61
50 %	CC	**10.68**	**10.06**	**10.01**
50 %	BBSE+	15.36	11.36	10.56
50 %	KMM (Gaussian)	**9.76**	**10.68**	11.34
50 %	KMM (Energy)	15.68	14.94	13.05
70 %	CC	**14.33**	**14.03**	**14.00**
70 %	BBSE+	17.85	15.49	14.39
70 %	KMM (Gaussian)	**12.82**	**14.39**	**14.43**
70 %	KMM (Energy)	18.29	16.80	15.65

Mahalanobis Distribution Feature Matching

Definition 4.2

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

Bernstein-based theorem for M-DFM

Theorem 4.2. (Bernstein-based)

For any $\delta \in (0,1)$, with probability greater than $1-\delta$:

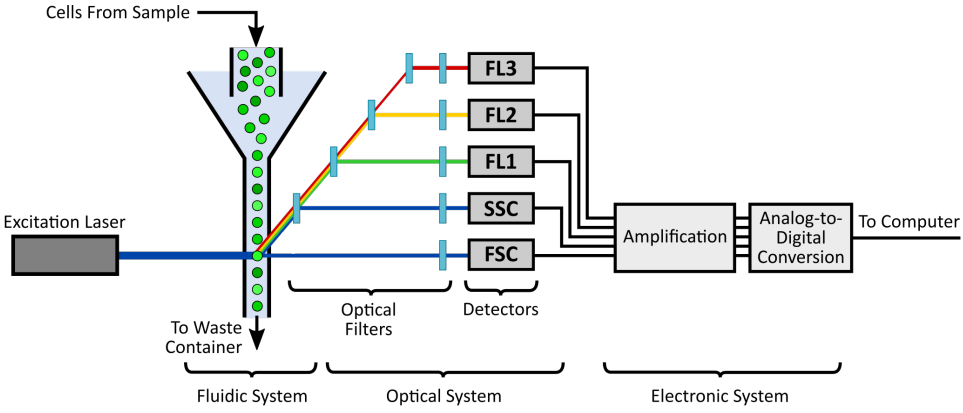
$$\|\hat{\alpha}_M - \tilde{\alpha}\|_2 \leq R_1(\delta, c) \frac{\|M\|_{\text{op}} C}{\sqrt{\Delta_{\min}(\hat{\boldsymbol{G}}^M)}} \left(\frac{\|w\|_1}{n} + \frac{1}{m} \right) \tag{$\mathcal{E}_3$}$$

$$+ R_2(\delta, c) \sqrt{\frac{Tr(M \Sigma_{\tilde{\alpha}} M^\top)}{\Delta_{\min}(\hat{\boldsymbol{G}}^M)}} \left(\sqrt{\frac{\|w\|_1}{n}} + \frac{1}{\sqrt{m}} \right), \tag{$\mathcal{E}_4$}$$

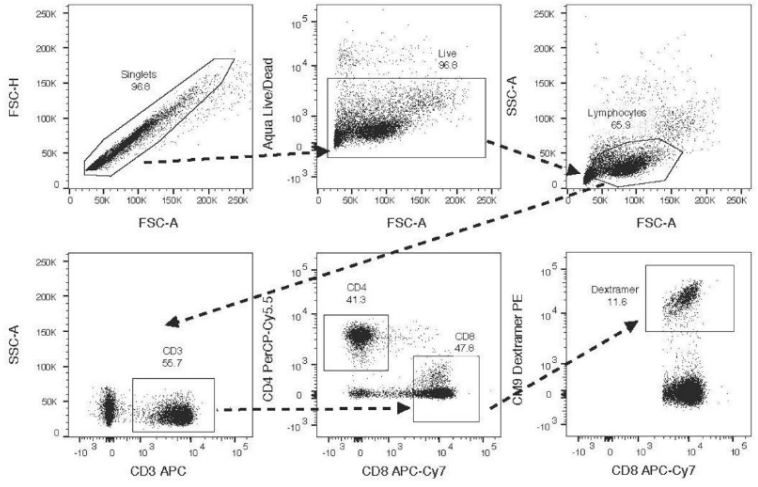
with $w = \tilde{\alpha}/\tilde{\beta}$, $R_1(\delta, c) = \frac{4}{3}\log(4c/\delta)$, $R_2(\delta, c) = 2\sqrt{2\log(4c/\delta)}$ and $\Sigma_{\tilde{\alpha}} = \sum_{i=1}^{c} \tilde{\alpha} \Sigma_i$.

Flow Cytometry

Cytometer


Cells From Sample
FL3
FL2
FL1
SSC
FSC
Excitation Laser
Amplification
Analog-to-Digital Conversion
To Computer
To Waste Container
Optical Filters
Detectors
Fluidic System
Optical System
Electronic System

Gating


FSC-H
FSC-A
Singlets
96.8
Aqua LiveDead
Live
96.8
SSC-A
Lymphocytes
65.9
SSC-A
CD3
55.7
CD3 APC
CD4 PerCP-Cy5.5
CD4
41.3
CD8
47.8
CD8 APC-Cy7
CM9 Dextramer PE
Dextramer
11.6
CD8 APC-Cy7

Flow Cytometry

Flow Cytometry

- $\mathbb{P}_i$ is the distribution of a given population of cells.
- $\mathbb{Q}_0$ is the set of unlabelled cell populations.
- The embedding of a node is the average of the embeddings of its children.

Problem : Kernel Mean Embedding

For a given distribution $\mathbb{P}$, $\Phi_k(\mathbb{P}) \in \mathcal{H}_k$.

Random Fourier Features

Random Fourier Features

For every translation-invariant kernel k, i.e. $k(x, y) = \gamma(x - y)$, there exists a distribution Λ_k such that for every sample $(\omega_i)_{i=1}^{D/2}$ i.i.d. from Λ_k

$$z_\omega(x) = \sqrt{\frac{2}{D}} \left[\cos(\omega_i^T x),\ \sin(\omega_i^T x)\right]_{i=1}^{D/2} \in \mathbb{R}^D$$

such that :

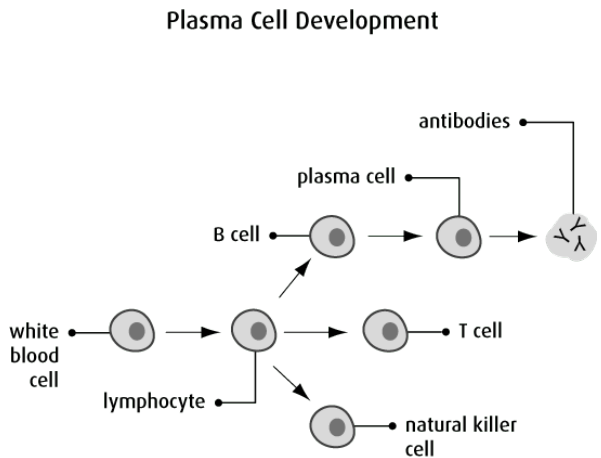
$$k(x, y) = \mathbb{E}_\omega[z_\omega(x)^T z_\omega(y)].$$

Random Fourier Features

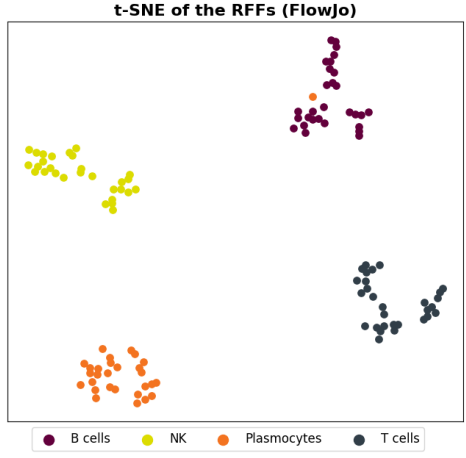
Examples

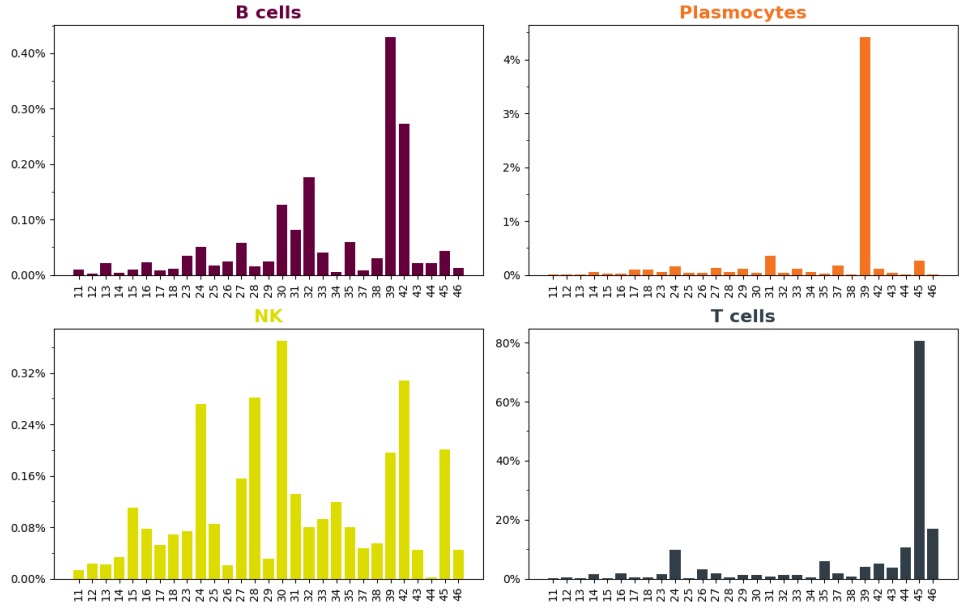
- $\phi(x) = \phi_k(x) \in \mathcal{H}_k$, Kernel Mean Matching.
- $\phi(x) = (1\{\hat{f}(x) = i\})_{i=1,\dots,c} \in \mathbb{R}^c$. For $\hat{f}$ a classifier, BBSE+.
- $\phi(x) = z_\omega(x)$, Random Fourier Features Matching.

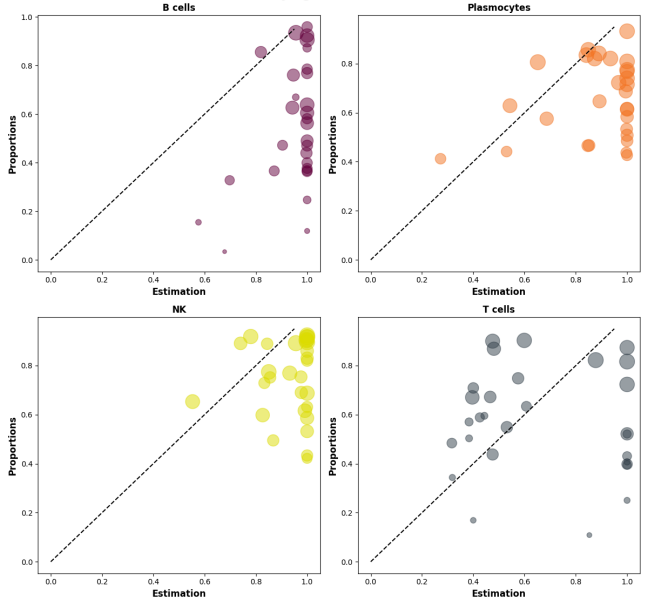
Myeloma


Plasma Cell Development
antibodies
plasma cell
B cell
white blood cell
T cell
lymphocyte
natural killer cell

t-SNE




Proportion of cells in the samples for each patients
B cells
0.40%
0.30%
0.20%
0.10%
0.00%
Plasmocytes
4%
3%
2%
1%
0%
NK
0.32%
0.24%
0.16%
0.08%
0.00%
T cells
80%
60%
40%
20%
0%
11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46


Comparison proportions best clusters
When FlowJo gates are unknown
B cells
Plasmocytes
NK
T cells
Proportions
Estimation
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

- A literature review on quantification
- A general framework for quantification under Label Shift and Open Set Label Shift.
- An extension using Mahalanobis-type distance.
- Application of DFM to flow cytometry.
- Application of RFF embeddings to Metaflow.

Perspectives

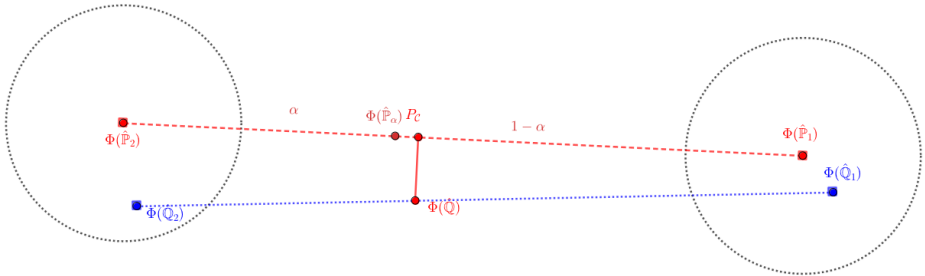
- Extension of DFM for hypothesis testing.
- Data-dependent theorem for M-DFM.
- Type of shift in Flow Cytometry.

Thank you for your attention.

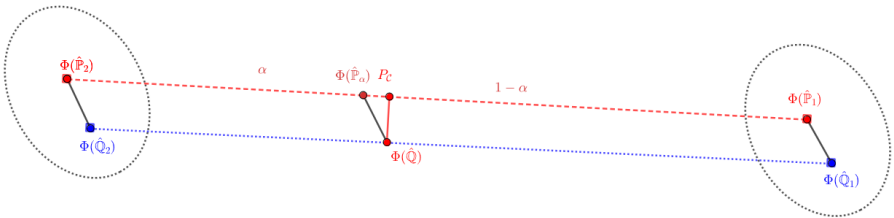
Overview of the Distribution Matching framework

Method	Embedding function Φ	Distance function $\mathcal{D}$	DFM
BBSE	$\mathbf{1}\{f(x) = i\}$	L_2	Yes
Threshold policy	$\mathbf{1}\{f(x) = i\}$	L_2	Yes
PACC	$f(x)$	L_2	Yes
HDy	Histogram of $f(x)$	Hellinger distance	No
HDx	Histogram of x	Hellinger distance	No
Dys	Histogram of $f(x)$	Any distance	No
Dys	Histogram of $f(x)$	L_2	Yes
FMM	cumsum of the cdf of f	L_1	No
SORD	$f(x)$	W_2	No
KDEy	KDE of $f(x)$	Multiple Choice	No
KDEy	KDE of $f(x)$	L_2	Yes
ReadMe	Special embedding	L_2	Yes
Kernel methods	KME	MMD	Yes
Wasserstein	Identity function	Regularised W_2	No
Sinkhorn	Identity function	Sinkhorn	No

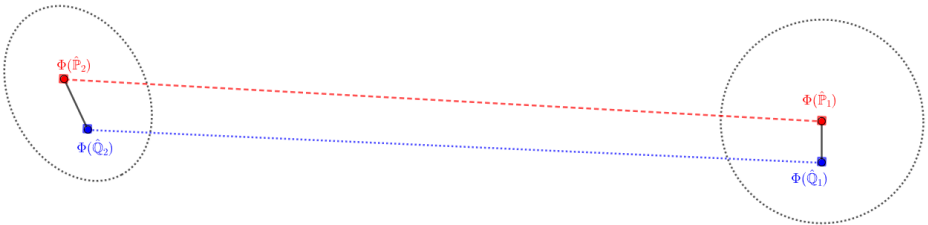
Distribution Feature Matching (DFM)



M-DFM


$\Phi(\hat{\mathbb{P}}_2)$
$\Phi(\hat{\mathbb{Q}}_2)$
α
$\Phi(\hat{\mathbb{P}}_\alpha)$
P_C
$\Phi(\hat{\mathbb{Q}})$
$1-\alpha$
$\Phi(\hat{\mathbb{P}}_1)$
$\Phi(\hat{\mathbb{Q}}_1)$

M-DFM


$\Phi(\hat{\mathbb{P}}_2)$
$\Phi(\hat{\mathbb{Q}}_2)$
$\Phi(\hat{\mathbb{P}}_1)$
$\Phi(\hat{\mathbb{Q}}_1)$

Definition of Gamma

Definition 3.3

$$\Gamma(b_1, \cdots, b_c) \coloneqq \min_{(I_1, I_2) \in \mathcal{P}_2(c)} d^2(C_{I_1}, C_{I_2}),$$

with the following:

$$\mathcal{P}_2(c) = \left\{ I_1, I_2 \subset \{1, \cdots, c\} |\ |I_1 \cap I_2| = 0, |I_1 \cup I_2| = c, |I_1| \text{ and } |I_2| > 0 \right\}$$

$$C_I = \left\{ \sum_{j \in I} \lambda_j b_j \mid \lambda \in \Delta^{|I|} \right\}$$

$$d^2(A, B) = \inf_{\substack{x \in A \\ y \in B}} \|x - y\|_2^2$$

Properties of Δ_{min}

Theorem 3.2

For any number of classes c and any vectors $\{b_1, \cdots, b_c\}$:

$$K_{\max(c)}\ \Gamma(b_1, \cdots, b_c) \geq \Delta_{\min}(b_1, \cdots, b_c) \geq \frac{1}{2}\ \Gamma(b_1, \cdots, b_c), \qquad (10)$$

where $K_{\max(c)} = \frac{c}{4}$ if c is even and $\frac{(c+1)(c-1)}{4c}$ if c is odd.

Remark

If $c = 2$, $K_{\max(2)} = 1/2$ and $\Delta_{\min}(b_1, b_2) = 1/2\Gamma(b_1, b_2)$.

Optimisation problem

DFM

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

where $\Delta^c := \{x \in \mathbb{R}_+^c : \sum x_i = 1\}$.

Solving $(\mathcal{P})$ amounts to solving a Quadratic Programming (QP) in dimension c. Indeed, we can rewrite the problem as:

$$\begin{aligned} &\text{minimise } \frac{1}{2}\alpha^T \hat{\boldsymbol{G}} \alpha + q^T \alpha \qquad (\text{QP}) \\ &\text{subject to } \alpha \succeq 0_c \text{ and } \mathbf{1}_c^T \alpha = 1, \end{aligned}$$

with $q = \left(\langle \phi(\hat{\mathbb{P}}_i), \phi(\hat{\mathbb{Q}}) \rangle \right)_{i=1}^{c}$. This is a c-dimensional QP problem, which can be solved efficiently.

Concentrating norm in Hilbert spaces

Theorem A.1. (Hoeffding)

Let $Z_1, \ldots, Z_n$ be independent (not necessarily identically distributed) random variables taking values in an Hilbert space $\mathcal{H}$. Suppose that $\forall i \in [n] : \|Z_i\| \leq C < \infty$. Then with probability greater than $1 - \delta$:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \leq C \frac{(1 + \sqrt{2 \log(1/\delta)})}{\sqrt{n}}, \tag{11}$$

Proof Theorem 3.1

$$\left\| \sum_{i=1}^{c} \hat{\alpha}_i \Phi(\hat{\mathbb{P}}_i) - \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) \right\|^2 = \left\| \sum_{i=1}^{c} (\hat{\alpha}_i - \alpha_i) \Phi(\hat{\mathbb{P}}_i) \right\|^2$$

$$= (\hat{\alpha} - \alpha)\ \hat{\boldsymbol{G}}\ (\hat{\alpha} - \alpha)$$

$$\overset{(\dagger)}{\geq} \left(\min_{\substack{\|u\|=1 \\ \mathbf{1}^T u = 0}} u^T \hat{\boldsymbol{G}} u \right) \|\hat{\alpha} - \alpha\|^2$$

$$\overset{(\ddagger)}{=} \Delta_{\min}(\hat{\boldsymbol{G}})\ \|\hat{\alpha} - \alpha\|^2,$$

Random Fourier Features

Theorem (Bochner's Theorem)

A continuous function φ on $\mathbb{R}^d$ defines a positive definite kernel $k(x, y) = \varphi(x - y)$ if and only if φ is the Fourier transform of a non-negative measure.

Fourier transform

$$\begin{aligned} k(x, y) &= \varphi(x - y) \\ &= \mathbb{E}_{\omega \sim \Lambda_k}\left[e^{i\omega^T(x-y)}\right] \\ &= \mathbb{E}_{\omega \sim \Lambda_k}\left[\cos\left(\omega^T(x - y)\right)\right] \end{aligned}$$

Random Fourier Features

Using a sample $(\omega_i)_{i=1}^{D/2}$ i.i.d. from Λ_k:

$$z_\omega(x) = \sqrt{\frac{2}{D}}\left[\cos(\omega_i^T x),\ \sin(\omega_i^T x)\right]_{i=1}^{D/2}$$

is such that :

$$k(x, y) = \mathbb{E}_\omega[z_\omega(x)^T z_\omega(y)].$$

Main theorem under no Label Shift

Theorem 3.3

With probability greater than $1-\delta$:

$$\|\hat{\alpha}-\alpha^*\|_2 \leq \frac{1}{\sqrt{\Delta_{\min}(\hat{\boldsymbol{G}})}}\Big(3\epsilon_n+\varepsilon_m+\sqrt{2\epsilon_n}B^{\perp}+B^{\|}\Big), \tag{12}$$

with:

$$\epsilon_n = C\frac{R_{c/\delta}}{\sqrt{\min_i n_i}}; \qquad \varepsilon_m = C\frac{R_{1/\delta}}{\sqrt{m}}; \tag{13}$$

$$B^{\perp} = B^{\perp}(\mathbb{P},\mathbb{Q}) = \sqrt{\|\Phi(\mathbb{Q})-\Pi_{\mathcal{C}}(\Phi(\mathbb{Q}))\|_{\mathcal{F}}};$$
$$B^{\|} = B^{\|}(\mathbb{P},\mathbb{Q}) = \max_i \|\Phi(\mathbb{P}_i)-\Pi_{\bar{V}}(\Phi(\mathbb{Q}_i))\|_{\mathcal{F}}.$$

Main theorem under Open Set Label Shift

Corollary 3.2.

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 $\epsilon_n, \varepsilon_m$ defined as in (13).

Proof Theorem 3.3

$$\|\Pi_{\mathcal{C}_n}(\Phi(\mathbb{Q})) - \Pi_{\mathcal{C}}(\Phi(\mathbb{Q}))\| \leq ??$$

Definition

Let X and Y be two non-empty subsets of a metric space (M, d). We define their Hausdorff distance $H(X, Y)$ by:

$$H(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\}$$

Using Theorem 3.6 and Remark 3.7 of *Albert and Notik, 1993*:

$$\|\Pi_{\mathcal{C}_n}(\mathbb{Q}) - \Pi_{\mathcal{C}}(\Phi(\mathbb{Q}))\| \leq \sqrt{2H(\mathcal{C}, \mathcal{C}_n) \times \max\{\text{dist}(\Phi(\mathbb{Q}), \mathcal{C}), \text{dist}(\Phi(\mathbb{Q}), \mathcal{C}_n)\}},$$

Concentrating norm in Hilbert spaces

Theorem A.3. (Bernstein)

Let $Z_1, \dots, Z_n$ be independent (not necessarily identically distributed) random variables taking values in an Hilbert space $\mathcal{H}$. Suppose that $\forall i \in [n] : \|Z_i\| \leq C < \infty$. Denote $\overline{\Sigma} := \frac{1}{n} \sum \Sigma_{Z_i}$, where Σ_{Z_i} is the covariance matrix of Z_i. Then with probability greater than $1 - \delta$:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right\| \leq \frac{2}{3} \frac{C \log(2/\delta)}{n} + \sqrt{\frac{2 \log(2/\delta)}{n} Tr(\overline{\Sigma})}. \tag{14}$$

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

Iosif Pinelis *"Optimum bounds for the distributions of martingales in Banach spaces"*. In *The Annals of Probability*, 1994.

Optimal M

Theorem

Let us write $W = \Sigma_{\tilde{\alpha}}^{-1/2} \hat{V}$ *and* PDQ *the SVD decomposition of* W
For any given feature map Φ *that satisfies the conditions* $\mathcal{A}_1$ *and* $\mathcal{A}_2$*, the matrices* M *that minimise the criterion:*

$$\frac{\mathrm{Tr}(M\Sigma_{\tilde{\alpha}}M^\top)}{\lambda_{min}(\hat{\boldsymbol{G}}^M)}, \tag{15}$$

satisfy up to a multiplicative factor :

$$M^\top M = \Sigma_{\tilde{\alpha}}^{-1/2}\Big(WW^\top\Big)^+\Sigma_{\tilde{\alpha}}^{-1/2}. \tag{$\mathcal{M}$}$$