Representing evidence for attribute privacy : bayesian updating, compositional evidence and calibration

Paul-Gauthier Noé

AVIGNON UNIVERSITÉ

THÈSE DE DOCTORAT D'AVIGNON UNIVERSITÉ

École Doctorale n°536
Agrosciences et Sciences

Mention de doctorat :
Informatique

Laboratoire Informatique d'Avignon

Présentée par
Paul-Gauthier Noé

Représentation de la preuve pour le respect de la vie privée

Inférence bayésienne, preuve compositionnelle et calibration

Soutenue publiquement le 26 avril 2023 devant le jury composé de :

Frédéric Bimbot	Directeur de recherche	CNRS/IRISA	Rapporteur
Daniel Ramos	Professeur associé	Université Autonome de Madrid	Rapporteur
Pierre-Michel Bousquet	Professeur agrégé, docteur	Avignon Université	Examinateur
Corinne Fredouille	Professeure	Avignon Université	Présidente du jury
David Lovell	Professeur	Université de Technologie du Queensland	Examinateur
Isabel Trancoso	Professeure	Université de Lisbonne	Examinatrice
Junichi Yamagishi	Professeur	Institut National d'Informatique, Tokyo	Examinateur
Jean-François Bonastre	Professeur	Avignon Université	Directeur de thèse
Driss Matrouf	Maître de conférence, HDR	Avignon Université	Co-encadrant

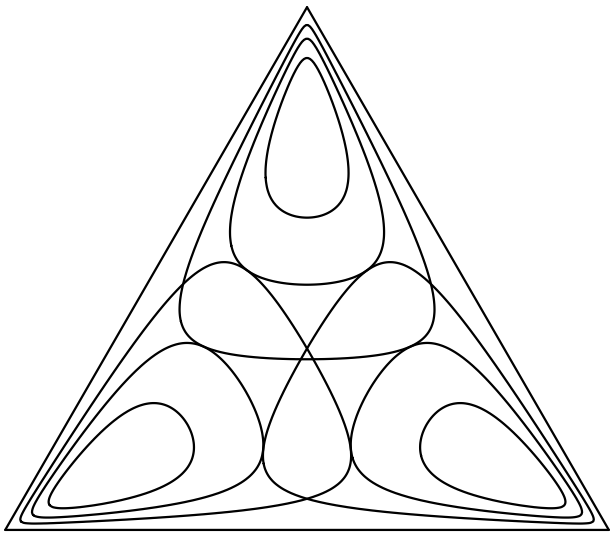
LABORATOIRE INFORMATIQUE D'AVIGNON

REPRESENTING EVIDENCE FOR ATTRIBUTE PRIVACY: BAYESIAN UPDATING, COMPOSITIONAL EVIDENCE, AND CALIBRATION

REPRÉSENTATION DE LA PREUVE POUR LE RESPECT DE LA VIE PRIVÉE : INFÉRENCE BAYÉSIENNE, PREUVE COMPOSITIONNELLE ET CALIBRATION

PAUL-GAUTHIER NOÉ

PhD thesis



Supervised by Jean-François Bonastre and Driss Matrouf

Laboratoire Informatique d'Avignon
Avignon Université
September 2023

Paul-Gauthier Noé: *Representing evidence for attribute privacy: Bayesian updating, compositional evidence, and calibration,* Représentation de la preuve pour le respect de la vie privée : Inférence Bayésienne, preuve compositionnelle et calibration, © September 2023

To my Grandmother Anne-Marie.

ABSTRACT

Attribute privacy in multimedia technology aims to hide only one or a few personal characteristics, or attributes, of an individual rather than the full identity. To give a few examples, these attributes can be the sex, nationality, or health state of the individual. When the attribute to hide is discrete with a finite number of possible values, the attacker's belief about the attribute is represented by a discrete probability distribution over the set of possible values. The Bayes' rule is known as an information acquisition paradigm and tells how the likelihood function is changing the prior belief into a posterior belief.

In the binary case—i. e. when there are only two possible values for the attribute—the likelihood function can be written in the form of a Log-Likelihood-Ratio (LLR). This has been known as the *weight-of-evidence* and is considered a good candidate to inform which hypothesis the data is supporting and how strong. The Bayes' rule can be written as a sum between the LLR and the log-ratio of prior probabilities decoupling therefore the evidence provided by the data and the initial personal belief.

This thesis proposes to represent the sensitive information disclosed by the data by a likelihood function. In the binary case, the LLR is a good candidate for expressing the likelihood function. However, this appealing form of the Bayes' Rule can not be generalized straightforwardly to cases where more than two hypotheses are possible. In order to get around this issue, this thesis proposes to treat discrete probability distributions and likelihood functions as compositional data. The sample space of compositional data is a simplex on which a Euclidean vector space structure—known as the Aitchison geometry—can be defined. With the coordinate representation given by the Isometric-Log-Ratio (ILR) approach, the Bayes' rule is the translation of the prior distribution by the likelihood function. Within this space, the likelihood function—in the form of a ILR transformation of the likelihood vector (ILRL)—is considered in this thesis as the multiple hypotheses and multidimensional extension of the LLR. The norm of the Isometric-Log-Ratio-Likelihood (ILRL) is the *strength-of-evidence* and measures the distance between the prior distribution and the posterior distribution. This can be seen as a measure of the information disclosed by the data. This measure of information is referred as *evidence information*.

Perfect privacy—coming from Claude Shannon's *perfect secrecy*—is reached when the attacker's belief does not change when observing the data: its posterior probabilities remain equal to its prior ones. In other words, we want the data to provide no evidence about the value the attribute takes. This idea—also known as *zero-evidence*—is theoretically reached when the LLR is zero in a binary setting, and by extension when the ILRL is the zero vector in a non-binary case corresponding to no *strength-of-evidence*.

The information—contained in an observation—about an attribute, is represented by a ILRL. However, in order to properly represent the information, the ILRLs have to be *calibrated*. The concept of calibration has been mostly discussed for probabilities but can be applied to likelihood functions. The *idempotence* of calibrated LLRs and its constraint on the

distributions of normally distributed LLRs are well-known properties. In this thesis, these properties are generalized to the ILRL for multiple hypotheses applications.

Based on these properties and on the compositional nature of the likelihood function, a new discriminant analysis approach is proposed. First, for binary applications, the proposed discriminant analysis maps the input feature vectors into a space where the discriminant component forms a calibrated LLR. The mapping is learned through Normalizing Flow (NF) a cascade of invertible neural networks.

This discriminant analysis can be used for standard pattern recognition but also for privacy purposes. Since the mapping is invertible, the LLR can be set to zero—which is consistent with the zero-evidence formulation of privacy—and the data can then be mapped back to the feature space. This protection strategy is tested on the concealment of the speaker's sex in neural network-based speaker embeddings. The resulting protected embeddings are tested for Automatic Speaker Verification (ASV) and for voice conversion applications.

Since the properties of the LLR naturally extend to the ILRL thanks to the Aitchison geometry of the simplex, the proposed discriminant analysis is easily generalized to cases where more than two classes, or hypotheses, are involved. We call this new approach Compositional Discriminant Analysis (CDA). It maps the data into a space where the discriminant components form calibrated likelihood functions expressed by the ILRLs.

The family of invertible transformations given by the NF can be used to learn a calibration mapping for LLR. This is briefly discussed at the end of this thesis.

Although this work is presented first in the context of privacy preservation, we believe this opens several research directions in pattern recognition, calibration of probabilities and likelihoods for multiclass applications, and the learning of interpretable representation of information.

RÉSUMÉ

Le respect de la vie privée dans les technologies multimédia consiste généralement à dissimuler l'identité d'un individu. Cette thèse s'intéresse cependant au respect de la vie privée dit *orienté attributs*. Le but est de dissimuler l'information relative à un seul attribut de l'individu comme son sexe, sa nationalité ou son état de santé, tout en préservant les autres attributs ou caractéristiques de l'individu. Quand l'attribut à dissimuler ne peut prendre qu'une seule valeur parmi un ensemble fini de valeurs possibles, la connaissance d'un attaquant sur l'attribut est représentée par une distribution de probabilité discrète sur l'ensemble des valeurs possibles. L'inférence bayésienne décrit comment une connaissance a priori, c'est-à-dire avant d'avoir observé des données, est transformée en une connaissance a posteriori par une fonction de vraisemblance.

Dans le cas binaire, c'est-à-dire lorsque l'ensemble de valeurs possibles pour l'attribut ne contient que deux éléments, la fonction de vraisemblance peut être écrite comme le log-ratio des deux vraisemblances (LRV). Le LRV est connu en inférence bayésienne comme le *poids de la preuve* et informe quelle hypothèse (ou valeur de l'attribut) une observation appuie et à quel point. La formule de Bayes peut être écrite comme la somme du LRV et du log-ratio des probabilités a priori. De cette manière, la contribution de l'observation et la contribution de la connaissance a priori sont séparées dans le calcul des probabilités a posteriori.

Dans cette thèse, il est proposé que l'information relative à l'attribut, révélée par une donnée, soit représentée par une fonction de vraisemblance. Dans le cas binaire, le LRV exprime de manière intuitive la fonction de vraisemblance. Cependant, cette manière d'écrire la formule de Bayes n'est pas directement généralisable aux cas avec plus de deux hypothèses, ou valeurs de l'attribut possibles. Cette thèse propose donc de traiter les distributions de probabilité et les fonctions de vraisemblance comme des données compositionnelles. La formule de Bayes peut ainsi être réécrite comme une somme entre la contribution des données et la connaissance a priori. Les données compositionnelles vivent sur le simplexe sur lequel un espace vectoriel euclidien, connu sous le nom de géométrie d'Aitchison, peut être défini. Avec le système de coordonnées défini par l'approche isometrique-log-ratio, l'inférence bayésienne est la translation de la distribution a priori par la fonction de vraisemblance. Dans cet espace, la fonction de vraisemblance, appelée Isométrique-Log-Ratio-Vraisemblance (ILRV), est considérée comme la généralisation multidimensionnelle et multi-hypothèses du LRV. La norme du ILRV est la *force de la preuve* et mesure la distance entre la distribution a priori et la distribution a posteriori ce qui peut être vu comme une mesure de l'information révélée par les données.

La notion de *secret parfait* introduite par Claude Shannon, peut être appliquée au respect de la vie privée. Le secret parfait correspond à la situation où la distribution a posteriori de l'attaquant est égale à sa distribution a priori. De cette manière, les données n'ont fourni aucune information à l'attaquant. Le secret parfait est atteint lorsque le LRV est zéro pour

les cas binaires et, par extension, lorsque le ILRV est égal au vecteur nul pour les cas non-binaires.

Pour que les ILRVs représentent correctement l'information révélée par les données, ils doivent être *calibrés*. Le concept de calibration est habituellement appliqué aux probabilités mais peut être appliqué aux vraisemblances. L'*idempotence* des LRVs calibrés et sa contrainte sur la distribution des LRVs normalement distribués sont des propriétés bien connues. Dans cette thèse, ces propriétés sont généralisées aux ILRVs pour des applications multi-hypothèses.

À partir de ces propriétés et de la nature compositionnelle des fonctions de vraisemblance, une nouvelle analyse discriminante est proposée. D'abord présentée pour des applications binaires, l'analyse discriminante plonge les vecteurs de caractéristiques en entrée dans un espace où la composante discriminante est un LRV calibré. La transformation est apprise avec un flot normalisant (normalizing flow) qui est une cascade de réseaux de neurones artificiels inversibles.

Dans cette thèse, nous proposons d'utiliser cette analyse discriminante pour le respect de la vie privée orienté attributs. La transformation étant inversible, le LRV peut être mis à zéro, avant de replonger les données dans l'espace des caractéristiques, respectant ainsi l'idée de secret parfait. Cette approche est testée sur la dissimulation du sexe du·de la locuteur·trice sur des représentations locuteur·trice issues de réseaux de neurones artificiels profonds. Une fois protégées, ces représentations sont testées sur une tâche de vérification automatique du·de la locuteur·trice et sur une tâche de conversion de la voix.

Les propriétés du LRV étant généralisables au ILRV grâce à la géométrie d'Aitchison, l'analyse discriminante proposée dans le cas binaire se généralise facilement aux cas non-binaires. De manière similaire au cas binaire, cette approche, que nous proposons et appelons Analyse Discriminante Compositionnelle, plonge les données dans un espace où les dimensions discriminantes forment une fonction de vraisemblance calibrée exprimée par l'ILRV.

L'idée d'utiliser un flot normalisant peut être expérimentée pour apprendre une transformation de calibration de LRV. Ce point est brièvement abordé à la fin de cette thèse.

Même si les travaux de cette thèse sont principalement présentés dans un contexte de sécurité des données personnelles, les notions abordées ouvrent des directions de recherche dans les domaines de la calibration des probabilités et des vraisemblances et dans l'apprentissage automatique, en particulier pour l'apprentissage de représentations interprétables de l'information.

PUBLICATIONS

First author publications:

- Paul-Gauthier Noé et al. "Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2023
- Paul-Gauthier Noé et al. "A Bridge between Features and Evidence for Binary Attribute-Driven Perfect Privacy." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2022, pp. 3094–3098
- Paul-Gauthier Noé et al. "Towards a unified assessment framework of speech pseudonymisation." In: *Computer Speech & Language* 72 (2022), p. 101299
- Paul-Gauthier Noé et al. "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation." In: *Proc. ISCA Interspeech*. 2021, pp. 1902–1906
- Paul-Gauthier Noé et al. "Speech Pseudonymisation Assessment Using Voice Similarity Matrices." In: *Proc. Interspeech*. 2020, pp. 1718–1722
- Paul-Gauthier Noé, Titouan Parcollet, and Mohamed Morchid. "CGCNN: Complex Gabor Convolutional Neural Network on Raw Speech." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2020, pp. 7724–7728

Non first author publications:

- Natalia Tomashenko et al. "Introducing the VoicePrivacy Initiative." In: *Proc. ISCA Interspeech*. 2020, pp. 1693–1697
- Natalia Tomashenko et al. "The VoicePrivacy 2020 Challenge: Results and findings." In: *Computer Speech & Language* 74 (2022), p. 101362
- Andreas Nautsch et al. "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment." In: *Proc. Interspeech*. 2020, pp. 1698–1702
- Jennifer Williams et al. "Revisiting Speech Content Privacy." In: *Proc. ISCA Symposium on Security and Privacy in Speech Communication*. 2021, pp. 42–46
- Jean-François Bonastre et al. "Benchmarking and challenges in security and privacy for voice biometrics." In: *Proc. ISCA Symposium on Security and Privacy in Speech Communication*. 2021, pp. 52–56
- Mohammad Mohammadamini, Driss Matrouf, and Paul-Gauthier Noé. "Denoising x-vectors for Robust Speaker Recognition." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2020, pp. 75–80

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to Frédéric Bimbot and Daniel Ramos for carefully reviewing my work, and to Pierre-Michel Bousquet, Corinne Fredouille, David Lovell, Isabel Trancoso and Junichi Yamagishi for being members of the jury. I have really appreciated your presence and your guidance. I will keep a wonderful memory of the day of the viva.

Words cannot express my gratitude to my supervisors Jean-François Bonastre and Driss Matrouf. More than a good professional relationship, I have spent friendly times with you. You believed in me throughout these years. Driss, your office was always open for discussion, not only to talk about our work but also to talk about everyday life, jokes and Moroccan food. The wholesome relationship I have had with you two is representative of the global atmosphere present at the Laboratoire Informatique d'Avignon (LIA). The LIA is a great place and I would like to extend my sincere thanks to all my colleagues and friends from there. I will keep wonderful memories of Avignon thanks to all of you. A very special thanks to La team Vezzo[1], La team RU, the pétanque lovers, and my Emacs Master[2].

I am also thankful to Junichi Yamagishi and his whole team for welcoming me so warmly at the National Institute of Informatics in Tokyo. It was a great pleasure to work with you. My experience in Japan was wonderful both professionally and humanly thanks to you. Thanks should also go to my roommates and friends in Nakano and Kōenji as well as all the go players I had the opportunity to play with.

I owe my deep gratitude to Nicolas Obin for introducing me to Jean-François—when I was looking for a PhD research project—and for being interested in my work throughout these years. I am grateful to all the people who were interested in my work and with whom I had great discussions: Nicholas Evans, Pierre-Michel, Themos Stafylakis and Itshak Lapidot. In particular, I would like to thank Andreas Nautsch, without whom this thesis would not be what it is. The friendly hours of discussion we had and the beers we shared were so inspiring.

I would like to thank all my friends in Grenoble, Marseille, Dunkerque, Lille, and Avignon for the great moments we spent together. I would also like to thank in particular my friends and roommates Amélie, Éléonore and Léo.

I heartily thank my parents, my brother and my sisters for their love and support during my studies. I wish to thank my dear Aphélie for her love, patience and attention throughout these years. Your continuous encouragement has made this thesis possible. I would also like to thank your parents and brother for bearing me throughout these years. And finally, big up to our cute little rats Peanut and Maroilles!

This work was supported by the VoicePersonae project ANR-18-JSTS-0001.

1 Vezzo is dead by the way.

2 No Gods, No Masters... except my Emacs Master.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ASV Automatic Speaker Verification

ASR Automatic Speech Recognition

CDA Compositional Discriminant Analysis

CDF Cumulative Distribution Function

CNN Convolutional Neural Network

DNN Deep Neural Network

ECE Empirical Cross-Entropy

EER Equal Error Rate

FA Factor Analysis

GAN Generative Adversarial Network

GMM Gaussian Mixture Model

ILR Isometric-Log-Ratio

ILRL Isometric-Log-Ratio-Likelihood

LDA Linear Discriminant Analysis

LLR Log-Likelihood-Ratio

LPC Linear Predictive Coding

LR Likelihood-Ratio

MAP Maximum A Posteriori

MFCC Mel-Frequency Cepstral Coefficients

NF Normalizing Flow

PAVA Pool Adjacent Violators Algorithm

PCA Principal Component Analysis

PLDA Probabilistic Linear Discriminant Analysis

PSR Proper Scoring Rule

QDA Quadratic Discriminant Analysis

TDNN Time Delay Neural Network

TDPSOLA Time Domain Pitch Synchronous Overlap Add

UBM Universal Background Model

UMAP Uniform Manifold Approximation and Projection

VAD Voice Activity Detection

VC Voice Conversion

WER Word Error Rate

ZEBRA Zero Evidence Biometric Recognition Assessment

zLLR zero-Log-Likelihood-Ratio

1

INTRODUCTION

With the advent of the digital revolution, data is being shared more freely than ever before. Consequently, privacy concerns are increasing, as evidenced by the attempt to create legislation[1]. People are using "smart" multimedia devices for multiple tasks like listening to music, online shopping, or even for medical consultations. It has also become natural for many people to control their digital devices through voice commands to such an extent that certain people are not aware that they are disclosing personal information. However, the manner a user is speaking and his or her voice embed information that could be considered private such as the sex, age, health and emotional state, or some information about the native language of the user [1, 64, 65, 72, 103, 114]. Private information can be misused by someone who intercepted the data, unbeknown to the user. Or, this can be misused by someone to whom the data is intentionally given in exchange for a service[2]. Therefore, privacy systems are required for the protection of the sensitive information in the data.

Originally, privacy approaches have been focused on the sanitization of datasets before their publication. Approaches like k-*anonymity* [153] aim in making each row of a dataset indistinguishable from at least $k-1$ other rows. This reduces the granularity of the dataset and can be seen as a group-based approach: within a group, some samples are made indistinguishable. This approach has been introduced in the context of dataset publishing, especially for tabular data where the columns represent independent factors. This thesis is rather concerned with multimedia data, like speech, where a sample[3] is commonly represented by a large dimensional vector in which the different factors of variation are entangled.

Many privacy approaches are based on randomized mechanisms incorporating noise in the data. Differential privacy[4] [52] is a theoretical guarantee of the level of privacy provided by these mechanisms. Starting with the idea that for the data release to be useful, some leak of sensitive information is inevitable, the level of the noise is adjusted to find a trade-off between the privacy and the utility of the data.

Our approach to privacy is different. While standard approaches are mostly concerned with indistinguishability between individuals, this thesis is concerned with what is called *attribute privacy*[5]. Attribute privacy aims for the concealment of a personal *attribute* of a

1 As the General Data Protection Regulation in European Union.
2 For instance, this can be a banking or insurance company that uses authentication by voice, or a company producing smart speakers.
3 Here, the term "sample" does not refer to one sample of a waveform. It rather refers to an observation like a speech utterance, an image, etc.
4 Originally developed in the context of statistical database release too, differential privacy generalizations have been proposed for other applications [32].
5 Not to be confused with *attribute privacy* as defined in [165].

person which can be for instance the sex, nationality or health state[6]. Attribute privacy is of interest in a situation where the user is concerned with the protection of only one of their personal attribute[7]. In such a situation, since the private attribute explains only a part of the variability in the data, the complete concealment of the attribute can be considered while keeping the rest of the variability for the data to remain useful. This is substantially different from approaches that consider that the level of privacy protection is at the expense of the utility of the data and therefore seek a trade-off between privacy and utility.

The privacy approach discussed in this thesis aims to provide *no* information about the sensitive attribute when a sample is released. Therefore, the first natural question which drives this thesis is:

How to represent the information–contained in a sample–related to a personal attribute?

Thinking about how to represent the sensitive information contained in the data is considered here as a crucial step before considering its concealment. The Bayes' rules tells how an individual's[8] belief about an attribute, represented by a discrete probability distribution over the set of values the attribute can take, changes when observing new data [14, 42, 60, 93]. In this way, the Bayes's rule can be used as an information acquisition paradigm. When the attribute is binary, i. e. when it has only two possible values, the concept of *weight-of-evidence* [69, 102], is a good candidate for representing how the data is supporting one value against the other. However, extending this concept to cases where the attribute has more than two possible values is not straightforward and requires a new paradigm. By treating the vectors of probabilities and likelihoods as *compositional data* [128], this thesis presents an extension of the concept of weight-of-evidence to non-binary attributes.

Even if a formal way of representing the evidence is adopted, one has to make sure that this can not be misinterpreted. This is where the concept of *calibration* [22, 40, 43, 137] steps in. In a nutshell, calibration insures that the information is properly represented avoiding potential misinterpretation. This concept will be essential all through this thesis.

Once it has been stated how to represent the sensitive information in the data, i. e. the evidence, the manipulation of this information for privacy purpose can be considered. This thesis is secondly concerned by the following question:

How to accurately model and manipulate the information related to an attribute?

These last years, *machine learning* methods appeared to be really effective for data modeling and pattern recognition [16, 109]. However, their behavior often suffers from a lack of interpretability, especially since the advent of what is called *deep learning* [90]. For critical applications like privacy, it is crucial to manipulate the data in an accurate and interpretable manner. *Discriminant analysis* approaches aim to model the data and transform

6 To be more precise, this can be any attribute representable by a *discrete* variable with a *finite* number of possible values.

7 This situation has also been called *user-configurable* privacy [7].

8 In the context of privacy preservation, the individual refers to an attacker, i. e. the person who seeks to infer the value of the private attribute.

it into a space that maximizes the information related to the classes (i. e. the possible values of the attribute) by maximizing the separability between the classes. They are usually based on assumptions about the distribution of the data or are not designed to represent the information in a calibrated and interpretable manner. This thesis will propose a new discriminant analysis for modeling and manipulating the evidence in the data for privacy purposes. The data is mapped into a space where the first dimensions accurately represent the evidence while the others represent the remaining variability.

Nevertheless, even if the main motivation of this work is for privacy application, the concepts encountered throughout this thesis are more general and can be applied to decision theory, pattern recognition, and any field interested in the extraction and manipulation of useful and interpretable representation of the information contained in some data.

1.1 ORGANISATION OF THIS THESIS

This thesis is organized as follows:

- **Chapter 2** presents some basics of Bayesian decision theory and the subjective interpretation of probabilities. It introduces the Bayes' rule, the Log-Likelihood-Ratio (LLR), also known as *weight-of-evidence*, and the concept of calibration of probabilities and LLRs. In addition, Proper Scoring Rule (PSR), its connection with information theory, and how this can be used to measure the goodness of probabilities and LLRs are briefly presented. This chapter also introduces, in Section 2.6, properties of calibrated LLRs: the *idempotence* property and its constraint on the distribution's parameters of normally distributed LLRs.

- **Chapter 3** introduces compositional data analysis. Discrete probability distribution lives on the *probability simplex* and can therefore be treated as compositional data. It presents how vectors of likelihoods can also be treated as such and that within the *Aitchison geometry of the simplex*, the likelihood vector naturally extends the concept of LLR to non-binary cases. This generalization of the LLR is called the Isometric-Log-Ratio-Likelihood (ILRL), or *evidence function*.

- **Chapter 4** extends Claude Shannon's definition of *perfect secrecy* to privacy. *Perfect privacy* is reached when no evidence is given to the attacker, i. e. when all likelihoods are equal corresponding to a zero vector on the Aitchison simplex: this is *zero-evidence*. The *expected privacy disclosure* from the Zero Evidence Biometric Recognition Assessment (ZEBRA) framework is then presented. It is a measure of the amount of sensitive information disclosed by a two classes recognizer. In addition, this chapter discusses two different views of the information. One refers to the information *received* by an observer and necessarily depends on his or her prior belief; and the other, called *evidence information*, refers to the information *disclosed* by an observation and does not depend on the observer.

- **Chapter 5** presents how the idempotence property's constraint on the distributions of the LLR can be used to design a two classes discriminant analysis which maps the

data into a space where the first dimension represents the evidence (i. e. the sensitive information in a privacy context) expressed by the LLR. This chapter also shows how the proposed discriminant analysis can be used for attribute privacy, applying it to the concealment of the speaker's sex in speaker embeddings and speech utterances.

- **Chapter 6** shows that the idempotence property applies also to calibrated ILRLs. This property leads to a constraint on the parameters of the distributions of normally distributed ILRLs. This generalizes the properties of the LLR presented in Section 2.6 to the ILRL.
- The constraint on the distributions of the LLRs has been used to design the two classes discriminant analysis presented in Chapter 5. Since this constraint generalizes also to the ILRL, i. e. the multiclass extension of the LLR, **Chapter 7** naturally extends the discriminant analysis to more than two classes.
- **Chapter 8** discusses a potential new generative calibration approach for the transformation of binary classifier scores into calibrated LLRs.
- **Chapter 9** is the conclusion. It summarizes the main findings of this thesis, gives additional thoughts, and discusses a few potential directions for future works.

1.2 CONTRIBUTIONS

Novel technical contributions of this thesis are listed above:

- The treatment of probability vectors as compositional data in the context of Bayesian updating has been discussed in [55, 58]. However, Chapter 3 presents the ILRL as the multidimensional extension of the LLR. Moreover, we provide a simple proof of the general formula of Equation 45 that recursively gives the component of the ILR transformation of a composition when the Aitchison basis is obtained through the Gram-Schmidt procedure[9].
- Section 4.3.2 provides original thoughts about uncertainty and evidence information.
- Chapter 2.6 presents properties of the LLR known for decades but lays the foundations of what will constitute, in Chapter 6, their generalization to ILRL. Chapter 6 indeed provides new results: it shows the *idempotence* property of the ILRL and its constraint on the distribution of normally distributed ILRL written in Proposition 2.
- Chapter 5 and 7 present a new approach to discriminant analysis. The main originality of this approach is that—taking advantage of the properties of calibrated likelihood vectors within the Aitchison geometry of the simplex—it is explicitly designed such that the discriminant components form a calibrated likelihood function.
- Finally, Chapter 8 is an attempt to design a new approach to generative calibration.

9 Even if this formula can be found in [56], at the beginning of our work, we were not aware of this result, thus we proofed it.

2

BAYESIAN UPDATING OF BELIEF

If someone asks you whether it will rain tomorrow or not in Avignon, do you feel able to say "Yes, it will rain" or "No, it will not rain"? You may prefer to say something like "I'm not sure, but I believe that..." because you feel uncertain about it. Even the weather forecasters are not certain about the coming weather. Instead of affirming whether it will rain or not, they provide a probability that it will rain. You may have noticed that we wrote *"a"* rather than *"the"* probability. Indeed, probabilities represent here a personal belief that can differ from one another because we all have different knowledge and background. Weather forecasters may have for instance different atmospheric models, sensor positions, etc. This interpretation of probability is referred to as *subjectivist*. Probability allows dealing with uncertainty by representing an individual's belief about something.

Decision theory is the study of rational decision-making in presence of uncertainty[1]. An individual has to decide what action to do based on their belief and some potential consequences represented by *costs*. Let's consider a winegrower who has to decide whether to harvest or not based on his or her probability that it will be freezing tonight. If it freezes while he or she had decided to not harvest, all the grapes will be lost and the cost of this loss will be bigger than if he or she had harvested even if it did not freeze. Decision theory provides a framework for minimizing the expected cost also called the *Bayes risk*.

In this chapter, we recall some of the basics of subjective probability and Bayesian decision theory[2]. The first section presents the application of Bayes' rule as the updating of an individual's belief when receiving new information. It then discusses the binary hypothesis testing and the likelihood-ratio for representing the *evidence*'s effect i. e. how the data is changing the belief of the observer. The application of the Bayesian decision framework to Automatic Speaker Verification (ASV) is then briefly discussed. Then, the concept of calibration of probabilities is introduced before presenting Proper Scoring Rule (PSR), how it relates to information theory [34], and its use to evaluate the goodness of probabilities. Then, the concept of calibration is extended to the likelihood-ratios rather than probabilities. Finally, we briefly mention how the Bayesian paradigm can be used to define *perfect privacy*, a concept that will be discussed in more detail in Chapter 4.

1 To be more precise, we refer here to *epistemic* uncertainty i. e. an uncertainty that is due to a lack of knowledge rather than some random phenomenon [125].

2 For further readings on subjective probability and decision theory, the reader can refer to Bruno de Finetti's *Theory of Probability* [60], *Understanding Uncertainty* by Dennis V. Lindley [93], *Bayesian Theory* by José M. Bernardo and Adrian F. M. Smith [14] and *Optimal Statistical Decisions* by Morris H. DeGroot [42].

2.1 UPDATING OF BELIEF & THE LIKELIHOOD-RATIO

The state of knowledge of an individual about a set of hypotheses is represented by a probability distribution over these hypotheses. This set can be seen as the set of possible events or outcomes of a coming phenomenon as for instance {"rain", "no rain"}, it can be a set of possible values for the unknown height of the Mont Blanc[3] or it can be for instance a set of models that pretend to describe a physical phenomenon where some experiments and observations may support some models against some others.

Let's consider one protagonist in Japan who is aware that Avignon is located in the southeast of France, a region that he or she knows has low rainfall. He or she may first go in favor of the "no rain" hypothesis. However, after having a call with a colleague at Avignon University who said that the weather is currently bad and rainy in the south of France, the belief of our protagonist may change in favor of the other hypothesis.

Personal belief is changing when new information is obtained.

2.1.1 *The Bayes' rule*

Bayes' formula provides a natural way to revise an individual's belief when observing new information. It results in a posterior probability distribution that represents the individual's new belief in the light of the observed data also called *evidence*. Let $\mathcal{H} = \{H_1, H_2, \dots H_N\}$ be a set of possible events, the posterior probability for an event $H \in \mathcal{H}$ after observing an evidence E is given by:

$$P(H \mid E, K) = \frac{P(E \mid H, K)P(H \mid K)}{P(E \mid K)}. \tag{1}$$

K represents the prior knowledge which is all information available to the individual other than the evidence E. $P(H \mid K)$ refers to the prior belief i. e. when the evidence is not considered and $P(H \mid E, K)$ refers to the posterior belief i. e. where all the information is taken into account including the evidence. From now, K will be taken for granted and dropped from the equation. We rewrite the Bayes' rule as:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \tag{2}$$

$P(E \mid H)$ is a quantity that informs on the *likelihood* of observing E when H appears to be true. This should not be seen as a probability but rather as a quantity that is multiplied by the prior probability to obtain the posterior probability up to a scaling factor. The latter is $P(E)$ which is not zero and can be written using the law of total probability as:

$$P(E) = \sum_{H \in \mathcal{H}} P(E \mid H)P(H). \tag{3}$$

3 The highest mountain in Europe.

The Bayes' rule can therefore be read as:

"*posterior probability* $\propto$ *likelihood* $\times$ *prior probability*",

where $\propto$ means "proportional to".

2.1.2 *The binary case, likelihood ratio and evidence*

When there are only two competing hypotheses H_1 and H_2, like in the {"rain","no rain"} example, the Bayes' rule can be rewritten in an appealing way considering the ratios, named odds, of the probabilities for each event:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_2)} \frac{P(H_1)}{P(H_2)}, \tag{4}$$

which can be read as:

"*posterior odds* = *likelihood-ratio* $\times$ *prior odds*".

When one hypothesis appears to be true, the other is necessarily false such that $P(H_1) = 1 - P(H_2)$ and $P(H_1 \mid E) = 1 - P(H_2 \mid E)$. Applying the logarithm function leads to:

$$\begin{aligned} \log \frac{P(H_1 \mid E)}{1 - P(H_1 \mid E)} &= \log \frac{P(E \mid H_1)}{P(E \mid H_2)} + \log \frac{P(H_1)}{1 - P(H_1)}, \\ \operatorname{logit} P(H_1 \mid E) &= \log \frac{P(E \mid H_1)}{P(E \mid H_2)} + \operatorname{logit} P(H_1), \end{aligned} \tag{5}$$

where the logit is the inverse function of the sigmoid and is defined as $\operatorname{logit}(p) = \log \frac{p}{1-p}$ for $p \in]0, 1[$. The Bayes' rule is here expressed as a sum and can be read as:

"*posterior log-odds* = *log-likelihood-ratio* + *prior log-odds*".

This is a sum between a *subjective* quantity: a prior log-odds which depends only on the individual's prior knowledge, and an *objective* quantity in the sense that it does not depend on the prior probabilities: the Log-Likelihood-Ratio (LLR)[4].

The logit function is transforming the segment $]0, 1[$, which is actually the 1-simplex $\mathcal{S}^2$ that will be defined in the next chapter, into the real line $\mathbb{R} =]-\infty, +\infty[$. Figure 1 illustrates this transformation. When the logit function is applied to the probabilities p and $1 - p$, the probability distribution lives on the unbounded real line[5].

4 To be more precise, the LLR is not "fully" objective. Indeed, likelihoods are not necessarily available and sometimes need to be computed from a probabilistic model also known as *generative model* in pattern recognition. The parameters of this model have to be estimated from a *training* set of observations with their corresponding hypothesis (also called *class* in pattern recognition). This model can be seen as a part of the initial knowledge K in Equation 1 and because they may differ from one another, the computation of the likelihoods is to this extent subjective. However, we will consider here the LLR as an objective quantity.

5 We want here to share an informal intuition. On the left side of Figure 1, the space on which the probability distributions are living is bounded (by the vertices represented by the small black points). The translation

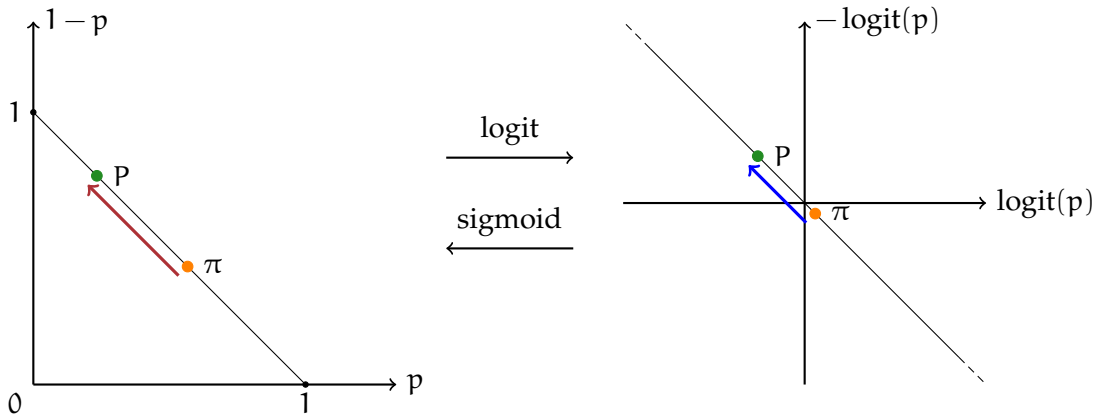


Figure 1: Logit transformation of the 1-simplex into the real line. The arrows represent the Bayesian updating of a prior probability distribution π into a posterior probability distribution P.

Example:

To better understand the application of the Bayes' rule, let's consider the following example highly inspired by Dennis V. Lindley's book [93]. Someone gave you an urn that contains a lot of black balls and red balls. The person only told you that the urn either contains two-thirds of black balls or two-thirds of red balls. Let's call these hypotheses B and R respectively. Because you do not have more information about the urn, let's say that your prior belief about the kind of the urn is represented by the prior probability $P(B) = 1 - P(R) = 0.5$ i. e. the prior odds $o(B) = P(B)/P(R) = 1$ corresponding to maximum uncertainty. The inside of the urn is so dark that you cannot have a look inside. In order to figure out which kind of urn it is, you have to *randomly* draw balls one by one. Let the two possible drawing events b and r be respectively a drawing of a black ball and a drawing of a red ball. Assuming that the number of balls in the urn is sufficiently large that when drawing a few balls, the proportion of colors does not significantly change, you know that at each drawing, the likelihoods are:

$$P(r \mid R) = \frac{2}{3} \quad P(b \mid R) = \frac{1}{3},$$
$$P(r \mid B) = \frac{1}{3} \quad P(b \mid B) = \frac{2}{3}.$$

Let's say that the first ball you draw is black. It corresponds to a Likelihood-Ratio (LR) $\frac{P(b|B)}{P(b|R)} = 2$. Multiplying this LR by your prior odds (Equation 4) gives your posterior odds representing your new belief about the urn in the light of the drawn ball. Your probability that the urn is B is now twice bigger than your probability for R. If you now draw a black ball again, your odds is again multiplied by 2 such that your posterior odds is now 4 and the urn appears to you four times more probable to be B than R.

corresponding to the Bayesian updating cannot move the distribution out of the simplex bounds. Therefore the magnitude of the translation is necessarily bounded and depends on the position of the prior, in particular on how close to the bounds the prior is. On the other hand, the logit transformation is kind of sending the simplex's vertices to infinity. The space on which the distributions are living is now unbounded. The Bayesian updating translation is here free to have any magnitude and does not depend on the prior distribution anymore.

After having drawn seven balls, you had observed a sequence S: bbrbrbb. When a black ball appears, the odds is multiplied by 2 and when a red ball appears, the odds is multiplied by $\frac{P(r|B)}{P(r|R)} = 0.5$. Your posterior odds is now:

$$\begin{aligned} o(B \mid S) &= 2 \times 2 \times 0.5 \times 2 \times 0.5 \times 2 \times 2 \times o(B) \\ &= 8 \end{aligned} \tag{6}$$

However, multiplication might be less intuitive than addition. In Equation 5, we have seen how the logarithm function transforms the odds form of the Bayes' rule as a sum. Equation 6 can therefore be rewritten as:

$$\begin{aligned} \log o(B \mid S) &= \log 2 + \log 2 + \log 0.5 + \log 2 + \log 0.5 + \log 2 + \log 2 + \log o(B) \\ &= \log 2 + \log 2 - \log 2 + \log 2 - \log 2 + \log 2 + \log 2 + \log o(B) \\ &= \log 2 + \log 2 + \log 2 + \log o(B) \\ &= 3 \log 2. \end{aligned} \tag{7}$$

There is now a symmetry: when a black ball is drawn, $\log 2$ is added to your log-odds and when a red ball is drawn $\log 2$ is subtracted from your log-odds. In this case, the balls have the same strength in changing the belief no matter the color. The color of the ball just tells in which direction the belief is changing.

To better visualize the symmetry induced by the logit transformation, Figure 2 shows the odds and log-odds as a function of the probability. They are respectively the function $p \mapsto \frac{p}{1-p}$ and $p \mapsto \log \frac{p}{1-p}$ for $p \in]0, 1[$. $\text{logit}(p + 0.5) = -\text{logit}(-p + 0.5)$ such that the log-odds are symmetric around $p = 0.5$.

A general formula of the successive Bayesian updating when observing a sequence $S = \{E_i\}_{1 \leqslant i \leqslant N_S}$ of size N_S is:

$$\log o(B \mid S) = \log o(B) + \sum_{i=1}^{N_S} \log \frac{P(E_i \mid B)}{P(E_i \mid R)}. \tag{8}$$

This expression shows a property known as *independent additivity* [78] where each coming evidence is adding a *weight* to the prior log-odds independently of the other evidences and independently of the prior log-odds.

The LLR is commonly used as a *weight-of-evidence* [69, 102] informing how an evidence is changing the prior belief into the posterior belief. In the above example, a positive LLR goes in favor of B against R while a negative LLR goes in favor of R against B.

2.2 APPLICATION OF BAYESIAN DECISION FRAMEWORK TO SPEAKER VERIFICATION

This section gives an application example of the Bayesian decision framework. The Bayesian decision framework has been supported in forensic science for matching fingerprints, DNA, or voice traces [104]. In this section, we discuss Automatic Speaker Verification (ASV) [15] as the task of deciding whether two speech utterances have been said by the same speaker or

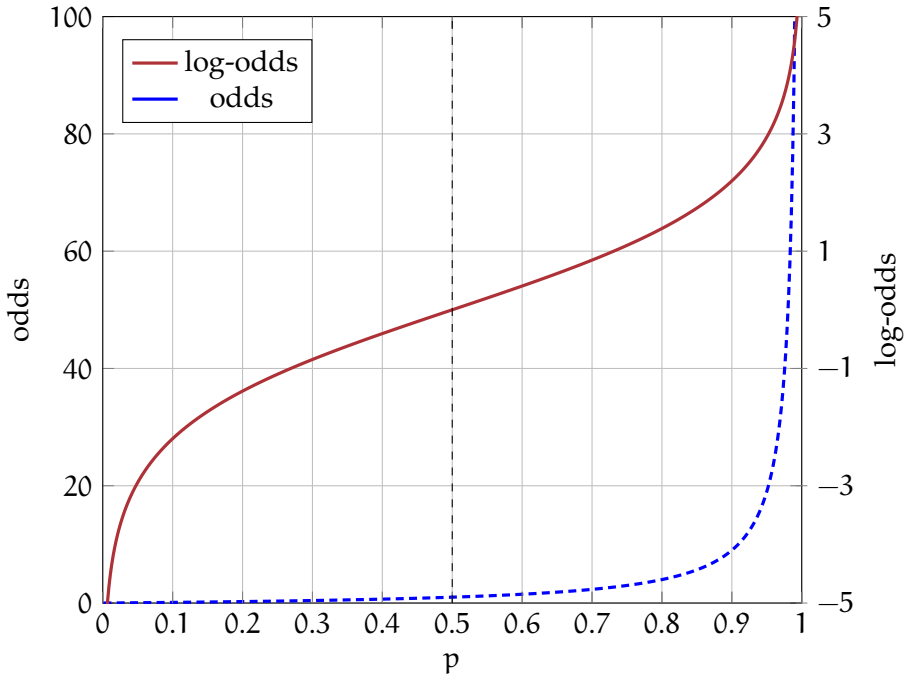


Figure 2: Odds and log-odds as a function of the probability p. The log-odds is the logit function and is symmetric around $p = 0.5$

by two different speakers. Forensic individualization [29, 31, 135] is not the only application of ASV, this can also be used for authentication purposes [61].

Let's consider this latter application as an example. A user requests an access by submitting a sample of their voice. The authentication side has to decide whether to accept or reject the request based on the comparison of the submitted speech utterance, called *test*, with one or several reference utterances called *enrolment(s)* given beforehand. A pair of test and enrolment is called a *trial*. The authentication side can provide access to the wrong person or deny access to the person in right. These cases correspond to the two types of *error* respectively called *false-alarm* and *miss*.

For a trial t, the authentication side uses a comparison system to produce a LLR:

$$l(t) = \log \frac{P(t \mid \mathrm{tar})}{P(t \mid \mathrm{nontar})}. \tag{9}$$

where tar refers to the *target* hypothesis: "the test and enrolment utterances have been produced by the same speaker"; and nontar refers to the *non-target* hypothesis: "they have been produced by two different speakers". A decision is taken by thresholding the LLR. Maximum a posteriori decision leads to the following threshold:

Access is given if:

$$\begin{aligned} P(\mathrm{tar} \mid t) \geqslant P(\mathrm{nontar} \mid t) &\iff \log \frac{P(\mathrm{tar} \mid t)}{P(\mathrm{nontar} \mid t)} \geqslant 0, \\ &\iff l(t) + \operatorname{logit} P(\mathrm{tar}) \geqslant 0, \\ &\iff l(t) \geqslant -\operatorname{logit} P(\mathrm{tar}). \end{aligned} \tag{10}$$

However, giving access to the wrong person is usually considered to have worst consequences than not giving access to the person in right. Different costs or penalties are therefore assigned to the different types of error. Let C_m and C_{fa} be respectively the costs assigned to the miss and the false-alarm errors. The decision rule is therefore rewritten to minimize the expected cost also known as the risk[6] [28]:

$$P(tar \mid t)C_m \geqslant P(nontar \mid t)C_{fa} \iff l(t) \geqslant \log \frac{C_{fa}}{C_m} - \operatorname{logit} P(tar), \tag{11}$$

where the costs are weighting the posterior probabilities.

Automatic speaker verification technologies:

Some of the recent technologies used for ASV are here briefly introduced. The reader not concerned with this can go to Section 2.3 but has to know that in practice, comparison systems for ASV produce scores that can not be directly interpreted as LLR. Additional processing of the scores, known as *calibration*, is therefore necessary to produce LLRs for minimum-risk decisions. Calibration will be discussed in more detail starting from Section 2.3.

For ASV and speech processing in general, Voice Activity Detection (VAD) is commonly applied to speech utterances and acoustic features are extracted. The most prominent acoustic features are the Mel-Frequency Cepstral Coefficients (MFCC) [39] and the Linear Predictive Coding (LPC)-based features [75]. These are short-term features extracted from a sliding window (of size commonly around 25ms with a 10ms sliding step) resulting in a sequence of acoustic feature vectors.

A few decades ago, ASV was done using a speaker-dependent Gaussian Mixture Model (GMM) [138, 142] trained on the acoustic feature vectors extracted from speech uttered by the speaker. The likelihood of a test feature vector or a set of feature vectors can be computed using the density function of the speaker-dependent GMM. For verification, this likelihood can be compared with the likelihood computed from a Universal Background Model (UBM) which is a GMM learned on a large cohort of speakers. Maximum A Posteriori (MAP) adaptation of the UBM was also proposed for obtaining a speaker or utterance-dependent GMM [62, 139].

Then, supervectors were introduced. Given a speech utterance, a GMM is learned with MAP adaptation of the UBM. The mean vectors of the GMM are concatenated to produce the so called *supervector*. In this way, a fixed-length vector is obtained for representing a variable-length speech utterance. Then, a kernel-based comparison of utterances is performed for verification [30].

Next, the research in speaker verification focused on modeling the supervector, using Factor Analysis (FA), trying to decompose the speaker-related and channel-related[a] variabilities into different latent variables of lower dimensionality [82, 83, 100]. FA-based approaches lead to the *i-vector* feature vector extraction. Instead of considering two separate subspaces, one for the speaker variabilities and one for the channel vari-

6 Here, not doing an error is considered to have no bad consequence i. e. no cost.

abilities, the idea is to have a single *total variability* space[b] [44, 45]. The estimated total variable can then be used as a new feature vector for representing a speech utterance. Probabilistic Linear Discriminant Analysis (PLDA) [76, 134] has been widely used for comparing two i-vectors for verification.

More recently, supervectors have been put aside in aid of Deep Neural Network (DNN)-based approaches for extracting, from speech utterances, fixed-length feature vectors. These are basically based on DNN that takes variable-length utterances as input and are trained on a speakers discrimination task. Then, the vector outputted by one layer of the DNN, also called *embedding*, can be used as a feature vector for representing the speaker-related information in the input utterance. This embedding is known as the *x-vector*. Many variations have been explored. They basically differ in the type of architecture used for the DNN. The original one was based on Time Delay Neural Network (TDNN) and statistical pooling [149]. For an overview of the DNN-based speaker embeddings, the reader can refer to [11]. For verification, the x-vectors are usually compared with PLDA or directly with cosine similarity.

We just mentioned a wide but non-exhaustive range of ASV technologies. Some approaches produce verification scores using kernel function or cosine similarity and therefore do not produce LLR. Approaches based on generative models, like GMM, FA, PLDA, produce LLR-like scores in the sense that they are computed from the ratio of probability density functions images. However, since these models are based on assumptions that may not be fulfilled, the computed LLRs are not necessarily *calibrated*. An additional calibration step is therefore required to transform the scores into calibrated LLRs necessary for minimum-risk decisions. Starting from the next section, the concept of calibration will be introduced.

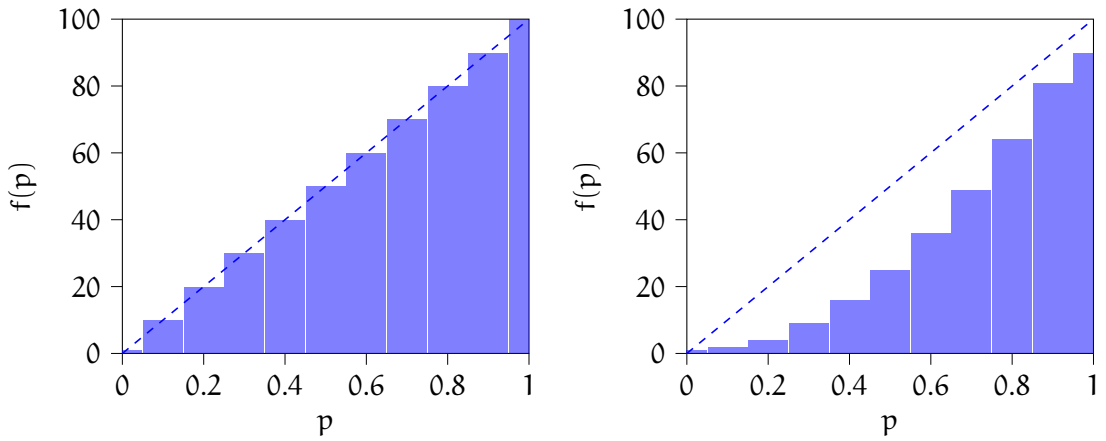
a The channel variabilities are induced by the different types of microphone, compression, the distance between the speaker and the microphone, etc. In other words, this is all the variability added to the signal from the moment the speech has been emitted by the speaker to the digitization of the signal.

b This has been motivated by the fact that the channel factor was still containing speaker-related information [46]. This could be due to the linearity of the approach which may not be flexible enough to disentangle speaker and channel contributions.

2.3 CALIBRATION AND DISCRIMINATION

In this section, the concept of calibration of probabilities is presented. It has been introduced in the context of weather forecasting [18, 43, 161] where the aim is to do a prevision by choosing a probability that it will rain on a given day. The previsions are said to be *calibrated* if they match the observed outcomes. To be more precise, over a long sequence of previsions and observed outcomes, the relative frequency $f(p)$ of days where it actually rained *and* on which the probability p has been assigned must be p [40, 43]. In other words, we want:

$$f(p) = \frac{N_r(p)}{N_r(p) + N_{\bar{r}}(p)} = p, \tag{12}$$



(a) Calibrated predictions

(b) Non-calibrated predictions

Figure 3: Artificial examples of empirical calibration plots. For good calibration, the bins' height have to be as close as possible to the dashed line $f(p) = p$. In (a), the previsions are well-calibrated while not calibrated in (b) where the forecaster tends to be over-confident.

where $N_r(p)$ is the number of days with the assigned probability p and where it has rained and $N_{\bar{r}}(p)$ is the number of days with the assigned probability p and where it has not rained.

Computing $f(p)$ requires that, in the set of previsions and observed outcomes, the number of elements with a probability p is not zero which is not necessarily the case for a continuous and infinite set of probabilities. To get around this issue, some applications—like weather forecasting—limit the set of allowable probability values to a finite and discrete set. To assess how well-calibrated the probabilities are, binning can also be used. The relative frequency is instead computed for a set of probabilities that are within an interval $p \pm \Delta p$. Empirical calibration plots allow us to visualize these frequencies as a function of the bin center probability p. Figure 3 shows examples of such plots. The height of a bin at p gives the relative frequency of days it actually rained and to which a probability within $p \pm \Delta p$ has been assigned. For good calibration, the height of each bin must be as close as possible to the dashed line $f(p) = p$.

Even if the concept of calibration has been introduced for probabilistic forecasts, it can be applied to the probability outputs of a classifier in pattern recognition. However, calibration is not the only desirable property: one also wants the probabilities to help in discriminating the different hypotheses (or classes) i. e. distinguishing days with rain and days without. This property will be here referred to as *discrimination* but is also known as *refinement* [43].

Many modern classification tasks in pattern recognition focus mainly on discrimination. However, calibration is important for making rational decisions [66], i. e. making decisions that minimize the expected cost, also referred to as *Bayes risk* [42], and which can be expressed in terms of the well-known *cross-entropy* as we will see in the next section.

2.4 MEASURING THE GOODNESS OF PROBABILITIES

The goodness of probabilities can be assessed with Proper Scoring Rule (PSR) [18, 43, 60, 63, 66, 143]. This section briefly defines what PSR are and how they relate to the cross-entropy, a well-known quantity in information theory [34] and pattern recognition.

2.4.1 *Proper scoring rules and cross-entropy*

Let $\mathcal{S}^N$ be a $(N-1)$-simplex[7], the sample space of probability distribution over the set of hypothesis $\mathcal{H} = \{H_1, H_2, \dots H_N\}$. A scoring rule S is a function that takes the assigned probability distribution $Q = [q_1, q_2, \dots q_N] \in \mathcal{S}^N$ and the event $h \in \mathcal{H}$ that *finally* occurred and gives a penalty $S(Q, h)$. Scoring rules can be defined in term of penalty like we are doing in this thesis and like it is done for instance in [22, 41] or in terms of reward [63] and utility [14]. The term *finally* is used here to highlight the fact that a prevision is done first, waiting for the event to occur, and once the event has been observed, the scoring rule assesses how good the prevision was[8]. When the assigned probability distribution is in accordance with an observed outcome, the score (the penalty) tends to be low, otherwise, it tends to be high. A scoring rule can therefore be seen as a cost in a decision problem.

Let's consider a reference probability distribution P that we, perhaps clumsily, interpret as the "true" distribution of the phenomenon we are interested in[9]. The expected score is:

$$S(Q, P) = \mathbb{E}_{H \sim P}\left[S(Q, h)\right], \tag{13}$$

where the subscript $H \sim P$ indicates that the expectation is done under the probability distribution P. A scoring rule S is *proper* if for all $P, Q \in \mathcal{S}^N$, $S(Q, Q) \leqslant S(P, Q)$ and is *strictly proper* if equality holds only when $P = Q$.

In the context of Bayesian decision, a PSR $S(Q, h)$ can be seen as a cost and its expectation $S(Q, P)$ and the infimum of its expectation $S(P, P) = \inf_{Q \in \mathcal{S}^N} S(Q, P)$ are referred to as *Bayes risks* [19, 42]. The function $H : P \mapsto H(P) = S(P, P)$ for $P \in \mathcal{S}^N$ is a concave function on the set $\mathcal{S}^N$ of discrete probability distributions and is an uncertainty or *entropy* function. It measures the level of uncertainty associated with a probability distribution [41, 42].

7 Simplex will be defined in more detail in the next chapter.

8 We present the concept of scoring rule in the context of weather forecast since this is how it was historically introduced [18, 43, 161]. However, as we will see later, scoring rules can also be used in pattern recognition to assess the goodness of the probabilities produced by a classifier. In this case, the word *finally* should not be seen in terms of time but rather in terms of available labels.

9 In my understanding of the subjective interpretation of probability [60], such a "true" distribution does not make sense since probability only exists subjectively. However, I do not want to go into this epistemological question here because no matter if such distribution exists or not, it is rarely available such that in practice, the expectations will be replaced by (empirical) averaging over a set of observed events and reference probability distribution will be chosen according to this set.

Depending on the PSR that is used, different entropy functions are obtained. This generalises the well-known Shannon's entropy [34, 145] obtained with the logarithmic scoring rule:

$$S(Q, h_i) = -\log q_i, \tag{14}$$

where q_i is the probability that has been assigned to the event h_i that occurred. This PSR leads to Shannon's entropy:

$$H_S(P) = -\sum_{i=1}^{N} p_i \log p_i. \tag{15}$$

A PSR also defines a discrepancy—also known as divergence—between two probability distributions $Q, P \in \mathcal{S}^N$:

$$D(Q, P) = S(Q, P) - H(P), \tag{16}$$

and $S(Q, P)$ is known as the cross-entropy. With the logarithmic scoring rule, the Kullback-Leibler divergence [87] is obtained:

$$D_{KL}(Q, P) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}. \tag{17}$$

We saw how a PSR defines an entropy function. Equivalently, any concave function defines a PSR. Figure 4 gives, in a two hypotheses case, a geometric intuition about the joint definition of a PSR and the corresponding information theoretic measures. Because this example is restricted to the binary case, we express the distributions $Q = [1 - q, q]$ and $P = [1 - p, p]$ by the probabilities q and p respectively. The concave blue curve represents any concave function. It defines an uncertainty or entropy function. $S(q, p)$ is the corresponding cross-entropy between the two distributions q and p, and the gap between $H(p)$ and $S(q, p)$, i. e. the gap between the blue and the red curves, gives the measure of divergence between the two distributions. The divergence can be seen as an additional loss of information when considering q instead of the reference probability distribution p. The images of the red tangent function $S(q, \cdot)$ at 0 and 1 define the scoring rule.

We have seen that scoring rules in probabilistic forecasting, Bayesian decision, and information theory are closely linked. For more details, the reader can refer for instance to [41, 42, 63] and Niko Brümmer's thesis [19] where he recalls the connection between information theory and the evaluation of the goodness of probability distribution in Bayesian decision using proper scoring rule.

2.4.2 *Cross-entropy and the calibration-discrimination decomposition*

Let's consider a binary classifier that takes an input e and outputs a posterior probability distribution over the set of hypotheses $\mathcal{H} = \{H_1, H_2\}$. Hypothesis H_1 states that e comes

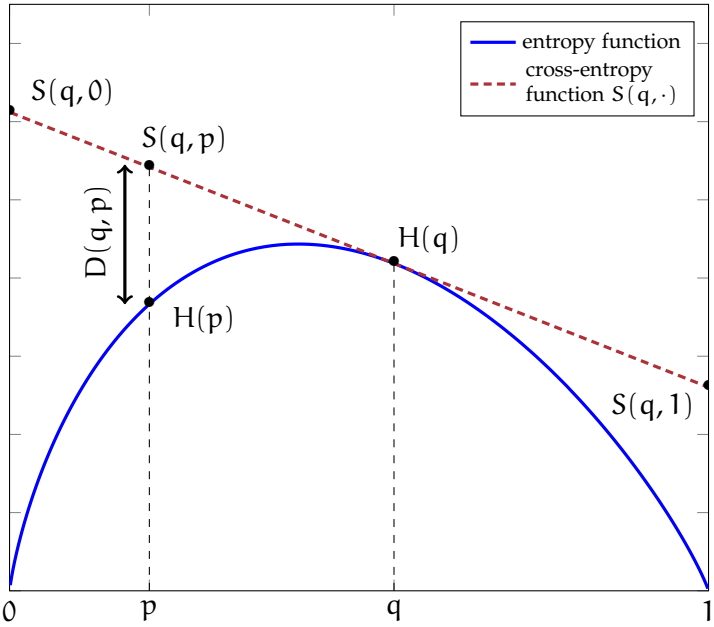


Figure 4: Entropy, divergence, cross-entropy, and proper scoring rule. The blue curve represents a concave function of the probability distribution. This defines an entropy function. Its tangent at q defines the scoring rule and a cross-entropy function $S(q, p)$. The gap between $S(q, p)$ and $H(p)$ gives the divergence $D(q, p)$ between q and p.

from the first class while H_2 states that e comes from the second class. Let $q_2 = 1 - q_1$ be the posterior probability assigned to H_2 by the classifier[10]. Considering the logarithmic scoring rule, the expected cross-entropy between the classifier's distribution and the reference "true" distribution p is:

$$\begin{aligned} \int_e S(q, p) P(e) de &= \int_e \left(-\sum_{i=1}^{2} P(H_i \mid e) \log q_i \right) P(e) de, \\ &= -\sum_{i=1}^{2} P(H_i) \int_e P(e \mid H_i) \log q_i de. \end{aligned} \tag{18}$$

This expression is sometimes simply called the cross-entropy [136]. The sampling probabilities, or likelihoods, $P(e \mid H_i)$ are usually not known. Having a set of observations with their corresponding class, an empirical approximation known as Empirical Cross-Entropy (ECE) is obtained [136]:

$$\text{ECE} = -\sum_{i=1}^{2} \frac{P(H_i)}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} \log q_i, \tag{19}$$

where $\mathcal{E}_i$ is the set of observations that belong to class i. The expectation has been replaced by empirical averaging assuming that for a given class, the observations are equally likely $P(e \mid H_i) \approx \frac{1}{|\mathcal{E}_i|}$.

10 q_1 and q_2 depend of course on the observation e and a prior but we do not explicitly write it to lighten the equations.

The ECE is a function of the prior probabilities $P(H_1)$ and $P(H_2)$. When chosen as $P(H_1) = \frac{|\mathcal{E}_1|}{|\mathcal{E}_1|+|\mathcal{E}_2|}$ and $P(H_2) = \frac{|\mathcal{E}_2|}{|\mathcal{E}_1|+|\mathcal{E}_2|}$ also known as the *empirical priors*, the resulting ECE is the well known binary cross-entropy loss commonly used in the training of classifier in pattern recognition.

Let $\pi = P(H_2) = 1 - p(H_1)$, and since $q_2 = 1 - q_1$, let's simply consider the probability q_2 that we will from now denote q. Let $\mathcal{Q}_1$ be the set of posterior probabilities assigned to the observations of class 1 and let $\mathcal{Q}_2$ be the set of posterior probabilities assigned to the observations of class 2, we rewrite the ECE as:

$$E_{CE}(\pi, \mathcal{Q}_1, \mathcal{Q}_2) = \frac{\pi - 1}{|\mathcal{Q}_1|} \sum_{q \in \mathcal{Q}_1} (\log(1-q)) - \frac{\pi}{|\mathcal{Q}_2|} \sum_{q \in \mathcal{Q}_2} \log q, \tag{20}$$

Let's now introduce a calibration mapping that aims in transforming the probabilities into "better" probabilities in the sense that they are better-calibrated and lower the ECE. It has been shown that the Pool Adjacent Violators Algorithm (PAVA)[11] gives a non-parametric mapping that perfectly calibrates the probabilities minimizing therefore the ECE (with fixed discrimination) [9, 24, 162]. However, these perfectly calibrated probabilities can be obtained only when the classes are available i. e. on a training set. PAVA should therefore only be used to create reference probabilities to evaluate how good were the predicted probabilities when the classes are available or once they have been obtained. The ECE can therefore be written in a decomposed form [136]:

$$E_{CE} = E_{CE}^{dis} + E_{CE}^{cal}, \tag{21}$$

where:

$$\begin{aligned} E_{CE}^{dis} &= E_{CE}(\pi, \gamma(\mathcal{Q}_1), \gamma(\mathcal{Q}_2)), \\ E_{CE}^{cal} &= E_{CE}(\pi, \mathcal{Q}_1, \mathcal{Q}_2) - E_{CE}^{dis}. \end{aligned} \tag{22}$$

where γ perfectly calibrates the set of probabilities with PAVA. These terms can respectively be seen as a *discrimination loss* and a *calibration loss* [22, 136]. See Appendix 10.1 for additional details on this decomposition in the context of both probabilistic forecasting [43] and binary classification [136].

2.5 CALIBRATION OF LOG-LIKELIHOOD-RATIOS

In pattern recognition and probabilistic forecasting, the concept of calibration has been usually formalized for probabilities [43, 71, 86, 115, 132]. It can however be extended to LLR as we briefly mentioned in Section 2.2. Driven by forensic science, research in ASV has been very active in the formalization of LLR calibration [19, 22, 28, 91, 96, 137].

11 For intuitive illustrations and examples on how PAVA is working, the reader can refer to Andreas Nautsch's PhD thesis Appendix B [111].

Given a prior π, there is a bijection between the posterior q and the LLR $l = \log \frac{P(e|H_1)}{P(e|H_2)}$ through the Bayes' rule. The ECE can therefore be rewritten as a function of the set of LLRs:

$$\begin{aligned} E_{CE}(\pi, \mathcal{L}_1, \mathcal{L}_2) = &\frac{1-\pi}{|\mathcal{L}_1|} \sum_{l \in \mathcal{L}_1} \log\left(1 + \frac{\pi}{1-\pi}\exp(-l)\right) \\ &+ \frac{\pi}{|\mathcal{L}_2|} \sum_{l \in \mathcal{L}_2} \log\left(1 + \frac{1-\pi}{\pi}\exp(l)\right), \end{aligned} \tag{23}$$

where $\mathcal{L}_1$ is the set of LLRs assigned to the observations of class 1 and $\mathcal{L}_2$ is the set of LLRs assigned to the observations of class 2. Calibrated LLRs result in calibrated posterior probabilities and minimize the ECE for a given prior and discrimination power. It has been shown that PAVA can be formalized as a LLR calibration rather than probabilities [24].

2.5.1 *Measuring the goodness of log-likelihood-ratios*

For a set of LLRs, empirical cross-entropy plots [136, 137] allow us to evaluate—for a wide range of prior—how good the LLRs are in terms of both discrimination and calibration. It represents two ECEs as a function of the prior: one with the actual set of LLRs and one with the LLRs calibrated using PAVA. The latter shows the discrimination capacity of the LLR set while the gap between the two curves gives the loss of information due to poor calibration. When the prior is $\pi = 0.5$ the metrics known as C_{llr} and C_{llr}^{min}—that have been used a lot in the NIST Speaker Recognition Evaluations—are recovered [19, 22]. Figure 5 shows examples of ECE plots[12]. The black dotted profile is the prior entropy (which is equivalent to the case where all LLRs are zero). We provide two examples: the blue and the dashdotted red profile. They correspond to two sets of LLRs. Both have the same discrimination capacity: when they are calibrated using PAVA, the minimum ECE profile (black dashed) is obtained. However, they both have different calibration qualities. Even if it is not perfectly calibrated, the blue profile is below the prior profile meaning that the set of LLRs provides useful information. However, being above the prior profile, the red one is an example of bad calibration corresponding to a loss of information. The ECE can be arbitrarily large because of the calibration loss. However, the minimum ECE, i. e. with perfect calibration, which can be interpreted as the posterior entropy, is always below the prior entropy [22, 136] which respects the information theory's inequality stating that, on average, observations reduce the uncertainty [34].

2.6 THE IDEMPOTENCE AND THE DISTRIBUTIONS OF THE LOG-LIKELIHOOD-RATIOS

This section presents some properties of the LLR that will be crucial for the rest of this thesis especially starting from Chapter 5. This section first presents the *idempotence* prop-

12 The natural logarithm is here used such that the entropy is measured in nat. Later in the thesis, the base 2 is used and the entropy is measured in bit.

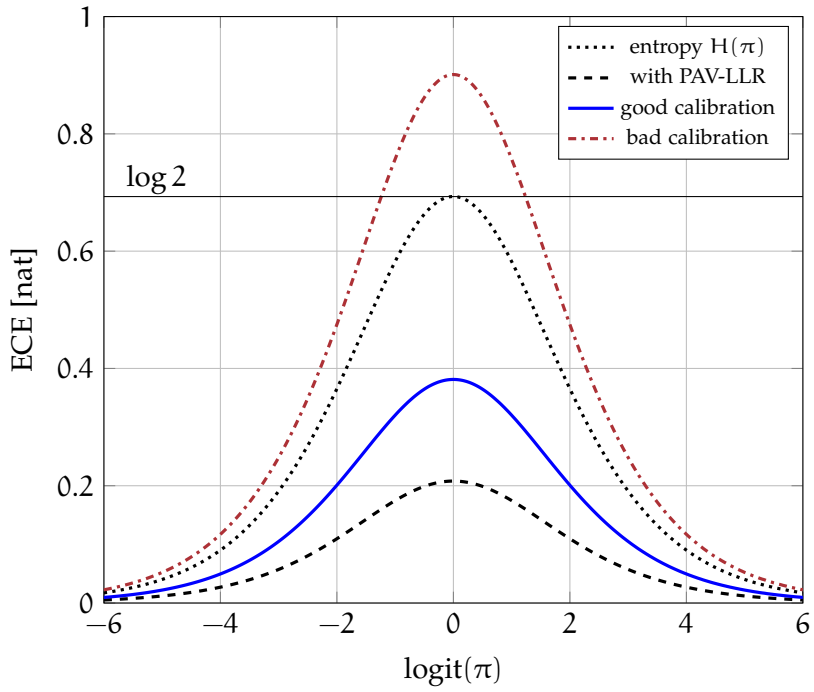


Figure 5: Artificial example of ECE plots.

erty of calibrated LLRs. This property leads to a constraint on the distributions of the LLR. These results have been reproofed and discussed in the context of speaker verification and forensic identification [91, 102]. However, they are known by statisticians since the works of Alan Turing and Irving John Good [67, 69].

2.6.1 *The idempotence property*

In the context of forensic speaker verification, a LLR $l = \log \frac{P(E|H_1)}{P(E|H_2)}$ is considered as calibrated if it results in the same posterior probabilities whether l or the data E is given [96, 97]:

$$
\begin{aligned}
&\forall i \in \{1,2\},\quad P(H_i \mid E) = P(H_i \mid l)\\
&\iff \frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(H_1 \mid l)}{P(H_2 \mid l)} \text{ since } P(H_2 \mid \cdot) = 1 - P(H_1 \mid \cdot),\\
&\iff \log \frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \log \frac{P(H_1 \mid l)}{P(H_2 \mid l)},\\
&\iff \log \frac{P(E \mid H_1)}{P(E \mid H_2)} + \operatorname{logit} P(H_1) = \log \frac{P(l \mid H_1)}{P(l \mid H_2)} + \operatorname{logit} P(H_1),\\
&\iff \log \frac{P(E \mid H_1)}{P(E \mid H_2)} = \log \frac{P(l \mid H_1)}{P(l \mid H_2)},\\
&\iff l = \log \frac{P(l \mid H_1)}{P(l \mid H_2)}.
\end{aligned}
\tag{24}
$$

The last line can be read as [91, 96, 97]:

"The LLR of the LLR is the LLR"

This is known as the idempotence property of the LLR. An alternative formulation and proof have been given in [102].

Having the posterior probabilities the same whether l or E is given ensures that l contains all the relevant information, about H_1 and H_2, contained in E. This has been expressed differently decades ago—but the idea is the same—in [69] as: *"[..] the weight of evidence tells us just as much as* E *does about the odd of [*H_1 *and* H_2*..]"*, stating directly that:

$$l = \log \frac{P(E \mid H_1)}{P(E \mid H_2)} = \log \frac{P(l \mid H_1)}{P(l \mid H_2)}. \tag{25}$$

In the next section, we will see how this property leads to a constraint on the distributions of the LLR.

2.6.2 *The LLR's conditional densities*

The idempotence of the LLR leads to a constraint on the conditional densities of the LLR.

Proposition 1. *If* $l \mid H_1 \sim \mathcal{N}(\mu, \sigma^2)$*, then* $l \mid H_2 \sim \mathcal{N}(-\mu, \sigma^2)$ *and* $\sigma^2 = 2\mu$.

In other words, if one conditional density of the LLR is Gaussian, the other is necessarily Gaussian, with an opposite mean, the variances are the same and are equal to two times the mean.

Proof. This is not a new result (proofs can be found in [69, 91, 102, 130]) but we report here a proof since it shares some elements with the proof of Proposition 2 presented in Chapter 6.

The idempotence can be written in term of the probability density functions:

$$l = \log \frac{f_{H_1}(l)}{f_{H_2}(l)} \iff f_{H_1}(l) = e^l f_{H_2}(l). \tag{26}$$

We have,

$$\begin{aligned} &l \mid H_1 \sim \mathcal{N}(\mu, \sigma^2), \text{ where } \mu \geqslant 0 \\ &f_{H_1}(l) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right), \end{aligned} \tag{27}$$

and because of Expression 26, we have:

$$\begin{aligned} f_{H_2}(l) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right) \exp(-l), \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(l-(\mu-\sigma^2)\right)^2}{2\sigma^2}\right) \exp\left(\frac{\sigma^2}{2} - \mu\right). \end{aligned} \tag{28}$$

Since f_{H_2} is a probability density function, its integral is one:

$$\begin{aligned}\int_{-\infty}^{+\infty} f_{H_2}(l)dl = 1 &\iff \exp\left(\frac{\sigma^2}{2} - \mu\right) = 1, \\ &\iff \sigma^2 = 2\mu.\end{aligned} \tag{29}$$

Therefore,

$$\begin{aligned}&f_{H_2}(l) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l-(-\mu))^2}{2\sigma^2}\right), \\ &l \mid H_2 \sim \mathcal{N}(-\mu, \sigma^2),\end{aligned} \tag{30}$$

and $\sigma^2 = 2\mu$. □

2.6.2.1 *The parameter and the separability*

The only parameter of the conditional densities is a scalar: the mean μ (or equivalently the variance σ^2 since $\sigma^2 = 2\mu$). In this section, we will see how this parameter can be expressed in terms of the separability between the two densities which can also be seen as the separabilities between the two classes in a pattern recognition context.

In [91], the author expressed the parameter in terms of the Equal Error Rate (EER). Here we express the parameter in terms of the Kullback-Leibler divergence (D_{KL}). Since the two densities are Gaussian with the same variance, the Kullback-Leibler divergence is symmetric. Since the LLR is a continuous variable, the D_{KL} between its conditional densities involves integrals rather than sums like in Equation 17 [34]:

$$D_{KL}(f_{H_1}, f_{H_2}) = \int_{-\infty}^{+\infty} f_{H_1}(l) \log \frac{f_{H_1}(l)}{f_{H_2}(l)} dl, \tag{31}$$

$$\begin{aligned}&= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right) \log \frac{\exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(l-(-\mu))^2}{2\sigma^2}\right)} dl, \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right) \frac{(l+\mu)^2 - (l-\mu)^2}{2\sigma^2} dl, \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right) \frac{2\mu l}{\sigma^2} dl,\end{aligned}$$

doing the substitution $x = \frac{l-\mu}{\sigma\sqrt{2}}$ we get:

$$\begin{aligned} D_{KL}(f_{H_1}, f_{H_2}) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-x^2\right) \left(4\mu x + \frac{2\sqrt{2}\mu^2}{\sigma}\right) dx, \\ &= \frac{4\mu}{\sigma\sqrt{2\pi}} \underbrace{\int_{-\infty}^{+\infty} x \exp\left(-x^2\right) dx}_{0} \\ &\quad + \frac{2\mu^2}{\sigma^2\sqrt{\pi}} \underbrace{\int_{-\infty}^{+\infty} \exp\left(-x^2\right) dx}_{\sqrt{\pi}}, \\ &= \frac{2\mu^2}{\sigma^2} = \frac{\sigma^2}{2} = \mu \text{ because } \sigma^2 = 2\mu. \end{aligned} \tag{32}$$

The D_{KL} is therefore equal to the mean μ. This result may first seem surprising as long as the idempotence property has not been digested. Indeed, from Equation 31, because $l = \log \frac{f_{H_1}(l)}{f_{H_2}(l)}$ thanks to the idempotence, we recognise that the D_{KL} is the expected LLR for the hypothesis H_1 and equivalently, minus the expected LLR for the other hypothesis:

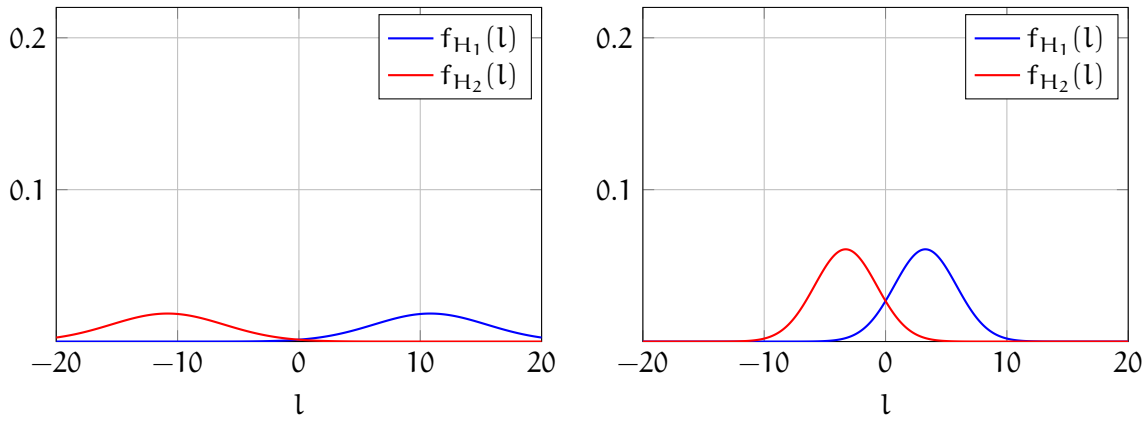
$$\begin{aligned} D_{KL}(f_{H_1}, f_{H_2}) &= \int_{-\infty}^{+\infty} f_{H_1}(l) l dl = E_{l|H_1}[l] = \mu, \\ D_{KL}(f_{H_2}, f_{H_1}) &= \int_{-\infty}^{+\infty} f_{H_2}(l) \log \frac{f_{H_2}(l)}{f_{H_1}(l)} dl \\ &= -\int_{-\infty}^{+\infty} f_{H_2}(l) \log \frac{f_{H_1}(l)}{f_{H_2}(l)} dl \\ &= -E_{l|H_2}[l] = \mu. \end{aligned} \tag{33}$$

This can also be expressed in terms of the EER as already mentioned: $D_{KL} = 2\left(\text{probit}(EER)\right)^2$ where the probit function is the inverse function of the cumulative normal distribution function.

Figure 6 shows examples of conditional densities of the LLR. When the mean increases, the variance and the separability increase. When the separability is 0, the two conditional densities are a Dirac delta function at 0.

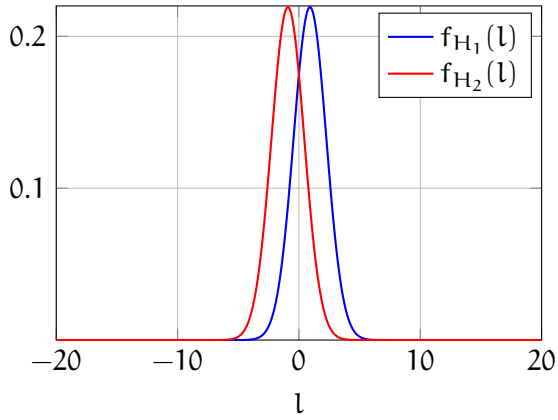
2.7 NON-UPDATING OF BELIEF: ZERO-EVIDENCE & PRIVACY

Let's do a slight digression about one use of the Bayesian view for privacy. We have seen so far that observed data, also called evidence, changes the observer's belief. In his 1949 paper "Communication theory of secrecy systems" [146], Claude Shannon defined the idea of *perfect secrecy* where the posterior probabilities remain equal to the prior ones for all observed evidence. When an attacker observes data to infer some sensitive information in it, his or her belief is not changing, and the data turns out to be useless for the attacker. In [113], Andreas Nautsch applied this concept to privacy and biometric recognition. When the attacker wants to infer which of two hypotheses is true, perfect privacy is reached when

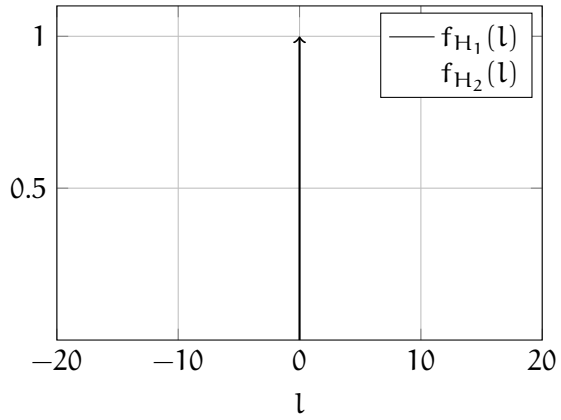


(a) LLR's conditional densities. EER = 0.01, $D_{KL} \approx 10.8$.

(b) LLR's conditional densities. EER = 0.1, $D_{KL} \approx 3.3$.



(c) LLR's conditional densities. EER = 0.25, $D_{KL} \approx 0.9$.



(d) LLR's conditional densities. EER = 0.5, $D_{KL} = 0$.

Figure 6: Examples of LLR's conditional densities. When the separability between the two conditional densities increases, the mean and variance increase. When there is no difference between the densities, i. e. when EER = 0.5 and $D_{KL} = 0$, the two densities are a Dirac delta function at 0.

the LLR is equal to zero for all evidences: this is zero-evidence. None of the observations is supporting a hypothesis against the other. These concepts will be discussed in more detail in Chapter 4.

2.8 SUMMARY

In this chapter, some basics of Bayesian decision theory have been introduced. The personal belief of an individual that is uncertain about which hypothesis is true over a set of possible ones can be represented by a discrete probability distribution. This personal belief can be revised using the well-known Bayes' rule when new data, also called evidence, is obtained.

When there are only two exhaustive and mutually exclusive hypotheses, the Bayes' rule can be written as a sum between the prior log-odds and the Log-Likelihood-Ratio (LLR). The former represents the initial knowledge of the individual while the latter represents the observed evidence and to which extent it supports a hypothesis against the other.

For doing cost-effective decisions, and to properly represent the information provided by the evidence, the probabilities and equivalently the likelihood functions—in the form of a LLR in a two hypotheses case—have to be well-calibrated. Proper Scoring Rule (PSR) and Empirical Cross-Entropy (ECE) can be used to assess how good probabilities and LLR are in terms of both calibration and discrimination.

We have seen a property of calibrated LLRs known as the *idempotence*. This ensures that all the relevant information contained in the data is contained in the LLR. This property leads to a constraint on the LLR's distribution. The parameter of the conditional densities are linked such that when one is Gaussian with a mean μ, the other is necessarily Gaussian with an opposite mean $-\mu$, the variances are the same and are equal to 2μ. The only parameter μ of these densities is the Kullback-Leibler divergence which can be seen as the separability between the two classes in a binary classification context. In Chapter 5, these properties will be used to design a new discriminant analysis in a two classes case. Chapter 5 will also show how this discriminant analysis approach can be used for privacy purposes when the attribute to hide is binary.

Perfect secrecy and zero-evidence as the idea of providing—to an attacker—no evidence about which hypothesis is true or false have been briefly introduced for privacy purposes. This will be discussed in more detail in Chapter 4.

Nevertheless, this chapter was focused on the two hypotheses case where the Bayes' rule can be written in an appealing way. However, the properties of linearity and independent additivity have been considered as not extensible to cases where more than two hypotheses are possible as for instance in Section 4.3 of E.T. Jaynes' book [78]. We agree with this but only when log-ratios are treated one by one, independently from one another. In the next chapter, we will see how treating probabilities and likelihood functions as compositional data provides an elegant manner for treating all log-ratios at once in a vector form and how linearity and independent additivity are recovered.

3

COMPOSITIONAL EVIDENCE

The previous chapter introduced some basics of Bayesian theory. We have seen how the logit transformation, and expressing the likelihood function in the form of a LLR, provide a good representation of the observed evidence in Bayesian updating when only two competing hypotheses are involved (Equation 5). The logit function transforms the one-dimensional simplex into the real line $\mathbb{R}$. When there are more than two possible hypotheses, the probability distribution lives in a higher-dimensional simplex. In this chapter, we will see how compositional data analysis and the Aitchison geometry of the simplex help in treating probability distributions and likelihood functions, and help in recovering the appealing linearity and independent additivity properties of the logit form of the Bayes' rule for more than two possible hypotheses.

This chapter is crucial for the rest of this thesis. Chapter 5 will propose a new approach to discriminant analysis for two classes where the data is mapped into a space where the discriminant component is the likelihood function in the form of a LLR. In order to extend this discriminant analysis to more than two classes, it is necessary to first generalize the concept of LLR to multiclass. The current chapter aims in doing so and proposes a way to represent an observed evidence in a multiple hypotheses case.

3.1 COMPOSITIONAL DATA ANALYSIS

A piece of basalt is made of several minerals. Let's say that one consists of 35% of pyroxene, 50% of plagioclase, 12% of olivine, and 3% of magnetite. These percentages represent the mineral composition of the stone. They sum to 100% such that knowing the three first is enough to know the last one. Vectors of such percentages are compositional data. Each element *describes a part of some whole* [128] like vectors of proportions, concentrations, and even probabilities. Compositional data analysis aims in treating such data by taking into account the compositional nature and structure of the data. For an overview of compositional data analysis, the reader can refer to the book *Modeling and Analysis of Compositional data* by Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado [128].

A N-part composition is a vector of N non-zero positive real numbers that sum to a constant k. Each element of the vector is a part of the *whole* k. The sample space of compositional data is known as the simplex:

$$\mathcal{S}^N = \left\{ \boldsymbol{x} = [x_1, x_2, \dots x_N]^T \in \mathbb{R}^N \mid \forall i \in [\![1, N]\!], x_i > 0 \text{ and } \sum_{i=1}^{N} x_i = k \right\}, \tag{34}$$

In a composition, the value of a part alone does not matter. Only the relative information between parts matters and John Aitchison introduced the use of log-ratios of components

to handle this [2]. He defined several operations on the simplex that leads to what is called the *Aitchison geometry of the simplex*.

3.2 THE AITCHISON GEOMETRY OF THE SIMPLEX

John Aitchison defined an internal operation called *perturbation*, an external one called *powering* and an inner product [4]:

- perturbation:

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left([x_1 y_1, \dots x_N y_N]\right), \tag{35}$$

- powering:

$$\alpha \odot \mathbf{x} = \mathcal{C}\left([x_1^\alpha, \dots x_N^\alpha]\right), \tag{36}$$

- inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j} \tag{37}$$

where $\mathbf{x}, \mathbf{y} \in \mathcal{S}^N$, $\alpha \in \mathbb{R}$ and $\mathcal{C}(\cdot)$ is the closure operator. Since only the relative information matter, scaling factors are irrelevant and a composition $\mathbf{x}$ is equivalent to $\lambda\mathbf{x} = [\lambda x_1, \lambda x_2, \dots \lambda x_N]$ for all $\lambda > 0$. This equivalence is materialized by the closure operator: $\mathbf{x} = \mathcal{C}(\lambda\mathbf{x})$. The closure is defined for $k > 0$ as:

$$\mathcal{C}(\mathbf{x}) = \left[\frac{k x_1}{\|\mathbf{x}\|_1}, \frac{k x_2}{\|\mathbf{x}\|_1}, \dots \frac{k x_N}{\|\mathbf{x}\|_1} \right]^T, \tag{38}$$

where $\mathbf{x} \in \mathbb{R}_+^{*N}$ and $\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|$. Therefore, any vector of positive real numbers can be projected onto the simplex using the closure.

Perturbation and powering give to the simplex a $(N-1)$-dimensional vector space structure and the inner product makes it Euclidean. The corresponding norm and distance are:

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\log \frac{x_i}{x_j} \right)^2}, \tag{39}$$

$$\begin{aligned} d_a(\mathbf{x}, \mathbf{y}) &= \|\mathbf{x} \ominus \mathbf{y}\|_a = \|\mathbf{x} \oplus ((-1) \odot \mathbf{y})\|_a \\ &= \sqrt{\frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2}, \end{aligned} \tag{40}$$

respectively called the *Aitchison norm* and the *Aitchison distance*. This Euclidean vector space structure of the simplex is called the *Aitchison geometry of the simplex*.

3.3 PROBABILITY DISTRIBUTION AND LIKELIHOOD FUNCTION AS COMPOSITIONAL DATA

Parameters of a discrete probability distribution live on a simplex ($k = 1$) called *probability simplex*. Indeed, probabilities are positive and sum to one. A vector of probabilities can therefore be treated as a composition. From now on, $\mathcal{S}^N$ will refer to the probability simplex.

Let's come back to our set of hypothesis $\mathcal{H} = \{H_1, H_2, \dots H_N\}$, to our evidence E, and to the Bayesian updating procedure. Let:

- $\boldsymbol{\pi} = [P(H_1), P(H_2), \dots P(H_N)]^T \in \mathcal{S}^N$ be the vector of prior probabilities assigned to each hypothesis,
- $\boldsymbol{w} = [P(E \mid H_1), P(E \mid H_2), \dots P(E \mid H_N)]^T \in \mathbb{R}_+^{*N}$ be the vector of likelihoods,
- $\mathbf{P} = [P(H_1 \mid E), P(H_2 \mid E) \dots P(H_N \mid E)]^T \in \mathcal{S}^N$ be the vector of posterior probabilities.

They respectively representent the prior probability distribution, the likelihood function, and the posterior probability distribution.

As already discussed in Section 2.1.1, likelihoods should not be seen as probabilities. Moreover, a vector of likelihoods does not sum to one. However, compositional data carries only relative information and are scale invariant. This defines *equivalent classes* [12]: all points in the positive orthant and on the line that passes through the origin are equivalent compositions. In this way, likelihood vectors can also be seen as compositions since they are scale-invariant and carry only relative information. Indeed, in Bayesian updating, it is known that no matter which scaling factor is applied to the likelihoods, the application of the Bayes' rule is the same:

$$\mathcal{B}(\boldsymbol{w}, \boldsymbol{\pi}) = \mathcal{B}(\lambda \boldsymbol{w}, \boldsymbol{\pi}) \tag{41}$$

where $\lambda > 0$ and $\mathcal{B}$ is the application of the Bayes' rule. In the Bayes' rule, the scaling factor λ factorizes in the law of total probability in the denominator and cancels out with the factor of the likelihood in the numerator.

Hence, From now on, when we discuss a likelihood vector $\boldsymbol{w}$ we refer to its scaled equivalence $\boldsymbol{w} \sim \mathcal{C}(\boldsymbol{w})$ that lives on the probability simplex. Figure 7 illustrates likelihood equivalent classes. Likelihood lines (in dashed blue) go through the probability simplex $\mathcal{S}^3$. Within a line, all likelihood functions are equivalent and are represented by the likelihood vector—given by the closure operator—that lives on the simplex. Therefore, likelihood functions live on the same space as probability vectors: the probability simplex[1].

1 In [19, 27], the authors already had the intuition that only the relative *values* of the likelihoods matter such that likelihood vectors live on a $(N-1)$-dimensional subspace of $\mathbb{R}^N$.

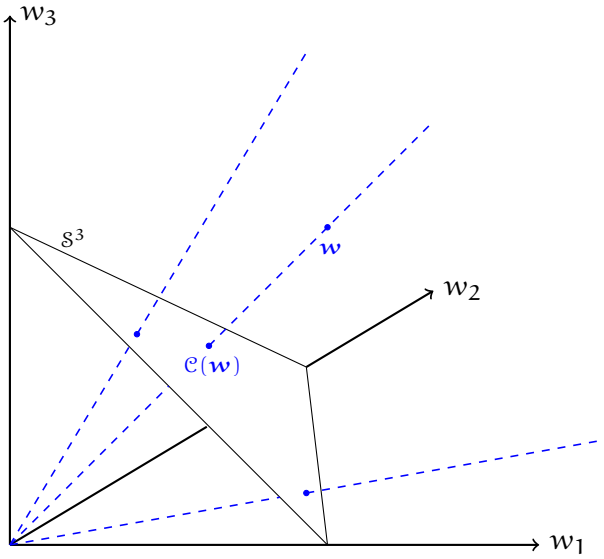


Figure 7: The probability simplex and likelihood lines as equivalent classes. All likelihood functions that live on the same blue dashed line are equivalent, and can be represented by, the likelihood vector that lives on the probability simplex.

The computation of the posterior probabilities through the Bayes' rule is the multiplication of the prior probabilities with the likelihoods normalized by $P(E)$ given by the law of total probability (see Section 2.1.1). This is exactly the perturbation of the prior probability vector by the likelihood vector, where the closure ensures the normalization:

$$\forall i,\ P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)} = \frac{w_i \pi_i}{\sum_{j=1}^{N} w_j \pi_j},$$

$$\begin{aligned} \mathbf{P} &= \left[\frac{w_1\pi_1}{\sum_{j=1}^{N} w_j\pi_j}, \frac{w_2\pi_2}{\sum_{j=1}^{N} w_j\pi_j}, \dots \frac{w_N\pi_N}{\sum_{j=1}^{N} w_j\pi_j} \right]^T, \\ &= \mathcal{C}\left([w_1\pi_1, w_2\pi_2, \dots w_N\pi_N]\right), \\ &= \boldsymbol{w} \oplus \boldsymbol{\pi}. \end{aligned} \tag{42}$$

With the Aitchison geometry of the simplex, the Bayes' rule is the perturbation of the prior distribution by the likelihood function [3, 55, 58]. The extension of the Aitchison geometry to the space of probability density functions has been studied in [53, 57, 157] but is out of the scope of this thesis.

3.3.1 *Zeros and the Cromwell's rule*

Since the Aitchison geometry is based on log-ratios of components, the components of a composition can not be equal to zero. The reader may have noticed that, in the definition of the sample space of compositions in Equation 34, the zeros are excluded. Dealing with zeros has been problematic in compositional data analysis [98, 128].

However, banning probabilities equal to zero is not an issue for us. We are indeed interested in the treatment of probabilities in the context of Bayesian updating, a realm where probabilities equal to zero are not desirable. Since a posterior probability is proportional to the product of the prior probability and the likelihood, if the prior probability is zero, the posterior is necessarily equal to zero no matter which evidence is observed. If you have a prior probability equal to zero, this means that this is already certain for you that the corresponding hypothesis is false, and no matter what evidence you observe or how someone is trying to convince you, your opinion about this hypothesis can not change. Doing so would be closed-minded. Any scientist must think that it is possible that he or she is wrong. The rule excluding certainty in the prior belief, i. e. banning prior probabilities equal to 0 or 1, has been proposed by Dennis Lindley and called *the Cromwell's rule* [93][2]. If you initially consider the set of possible hypotheses $\mathcal{H} = \{H_1, H_2, H_3\}$ and you finally proved *logically* that H_2 is wrong, you must not assign a probability 0 to H_2. You must instead redefine your decision problem and your range of possibility as $\mathcal{H} = \{H_1, H_3\}$ i. e. everything that is for you *neither certainly true nor certainly false* [60].

However, this argument holds for the probabilities only. There might be situations where the likelihood for a hypothesis of observing an evidence is zero. If such likelihood value is permitted in Bayesian updating, likelihood vectors as compositions can not contain zeros. In this case, the zeros have to be replaced. Zeros replacement strategies for compositional data are discussed in [98].

3.3.2 *The isometric log-ratio transformation for probability and likelihood*

In Section 3.2, the Euclidean vector space structure of the simplex has been defined. We are now interested in expressing the probability and likelihood vectors in a Cartesian coordinate system. This can be done using the Aitchison inner product by projecting the vectors on an orthonormal basis of the simplex. Let the set $\{\boldsymbol{e}^{(i)} \in \mathcal{S}^N, i = 1, \dots N-1\}$ be an *Aitchison* orthonormal basis of the simplex. The set $\{\text{ilr}(\boldsymbol{e}^{(i)}) \in \mathbb{R}^{N-1}, i = 1, \dots N-1\}$ forms a

2 The name of this rule comes from a quote Oliver Cromwell addressed to the Church of Scotland: *"I beseech you, in the bowels of Christ, think it possible that you may be mistaken."*.

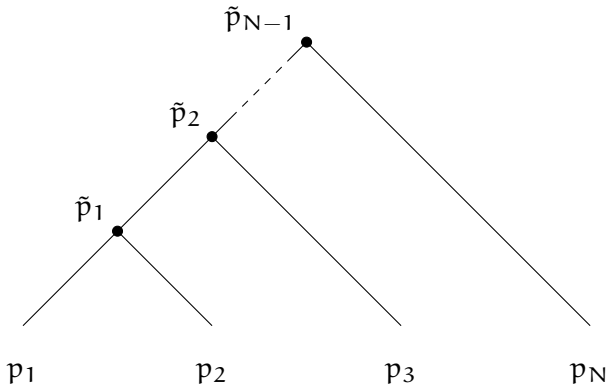


Figure 8: Bifurcating tree corresponding to the orthonormal basis obtained with the Gram-Schmidt procedure [56].

canonical basis in $\mathbb{R}^{N-1}$. The vectors of the Aitchison orthonormal basis obtained using the Gram-Schmidt procedure are defined for all $i \in [\![1, N-1]\!]$ as follows:

$$\boldsymbol{e}^{(i)} = \mathcal{C}\left(\left[\underbrace{\exp\left(\sqrt{\frac{1}{i(i+1)}}\right), \dots \exp\left(\sqrt{\frac{1}{i(i+1)}}\right)}_{\text{The first } i \text{ elements}}, \exp\left(-\sqrt{\frac{i}{i+1}}\right), 1, \dots 1\right]\right). \quad (43)$$

For details on the computation of this basis see [56].

The Isometric-Log-Ratio (ILR) transformation [56] allows a vector $\mathbf{p} \in \mathcal{S}^N$ to be expressed in a Cartesian coordinate system as follows[3]:

$$\tilde{\mathbf{p}} = \operatorname{ilr}(\mathbf{p}) = \left[\langle \mathbf{p}, \boldsymbol{e}^{(1)} \rangle_a, \dots \langle \mathbf{p}, \boldsymbol{e}^{(N-1)} \rangle_a\right]^T. \quad (44)$$

This defines an isometric isomorphism[4] between $\mathcal{S}^N$ and $\mathbb{R}^{N-1}$. Different bases could be used but the one presented above has a simple and intuitive recursive structure. The ILR transformation of the probability (or likelihood) vector results in a recursive grouping of the probabilities (or likelihoods) as illustrated by the bifurcation tree in Figure 8. Considering a vector $\mathbf{p} = [p_1, \dots p_N]^T \in \mathcal{S}^N$ and its ILR transformation $\tilde{\mathbf{p}} = \operatorname{ilr}(\mathbf{p}) = [\tilde{p}_1, \dots \tilde{p}_{N-1}]^T \in \mathbb{R}^{N-1}$, each node in the tree corresponds to a component $\tilde{p}_i$ of $\tilde{\mathbf{p}}$. The first component compares the probabilities for the two first hypotheses. Each next component then recursively compares the probabilities for the next hypothesis with the probabilities

3 We used here the definite article *the* to refer to the ILR transformation. However, this implies that there is only one ILR transformation while there is as many ILR transformations as they are Aitchison orthonormal bases on the simplex i. e. an infinite number. Along this thesis and without loss of generality, *the ILR transformation* will refer to the ILR transformation with the orthonormal basis obtained using the Gram-Schmidt procedure as defined in Equation 43. The use of this specific basis in no way excludes the general aspect of the following results since Aitchison orthonormal bases are related through unitary transformations [56].

4 An isometric isomorphism is an invertible mapping that preserves the distances: $\mathbf{x}, \mathbf{y} \in \mathcal{S}^N$, $d_a(\mathbf{x}, \mathbf{y}) = \|\operatorname{ilr}(\mathbf{x}) - \operatorname{ilr}(\mathbf{y})\|_2$.

for the previous ones. A general formula for the ith element $\tilde{p}_i$ of the ILR transformation of $\mathbf{p}$ is given by:

$$\tilde{p}_i = \langle \mathbf{p}, \boldsymbol{e}^{(i)} \rangle_a = \frac{1}{\sqrt{i\,(i+1)}} \log \left(\frac{\prod_{j=1}^{i} p_j}{(p_{i+1})^i} \right). \tag{45}$$

A proof of this result can be found in Appendix 10.2. An ILR component can be interpreted as a score. When the probability for the $(i+1)$th hypothesis increases and the probabilities for the hypotheses $H_{1 \leqslant j \leqslant i}$ decrease, the score $\tilde{p}_i$ decreases. Therefore, a low $\tilde{p}_i$ goes in favor of the $(i+1)$th hypothesis against the hypotheses $H_{1 \leqslant j \leqslant i}$ independently of the hypotheses $H_{i+2 \leqslant j \leqslant N}$.

A component of a composition carries relative information rather than absolute information. The treatment of compositional data is therefore based on ratios and in particular on log-ratios as already mentioned in Section 3.1. We can see the analogy with log-odds and Log-Likelihood-Ratio (LLR) presented in the previous chapter in the context of Bayesian updating.

3.3.3 *The Bayes' rule as an addition*

One benefit of the Aitchison geometry of the probability simplex is that it makes the Bayes' rule a perturbation: the posterior distribution is the perturbation of the prior distribution by the likelihood function. Within the Aitchison geometry of the simplex, the perturbation is an addition such that with the coordinate representation given by the ILR transformation, the Bayes' rule can be written as the translation of the prior by the likelihood vector in a Euclidean vector space [55, 58]:

$$\begin{aligned} \mathbf{P} &= \boldsymbol{w} \oplus \boldsymbol{\pi}, \\ \operatorname{ilr}(\mathbf{P}) &= \operatorname{ilr}(\boldsymbol{w}) + \operatorname{ilr}(\boldsymbol{\pi}), \\ \tilde{\mathbf{P}} &= \tilde{\boldsymbol{w}} + \tilde{\boldsymbol{\pi}}. \end{aligned} \tag{46}$$

Just like the logit transformation, the ILR transformation allows us to write the Bayes' rule as a sum between a term that depends only on the prior probabilities and a term that depends only on the likelihoods. In this way, the appealing additivity of Equation 5 is recovered.

To be more precise and to summarise, the likelihood vector $\tilde{\boldsymbol{w}}$ translates a prior probability distribution $\tilde{\boldsymbol{\pi}}$ into a posterior distribution $\tilde{\mathbf{P}}$. Moreover, the ILR transformation on a two hypotheses probability or likelihood vector is a one-element vector in which the element is the log-ratio of the components[5] which is consistent with Equation 5 and the LLR.

Figure 9 shows an example of Bayesian updating in a three hypotheses ILR space. The first component $\tilde{p}_1$ compares the probability for hypothesis H_1 with the probability for

5 To be more precise, with the given Aitchison basis, this is the log-ratio scaled by $\frac{1}{\sqrt{2}}$ (see Equation 45).

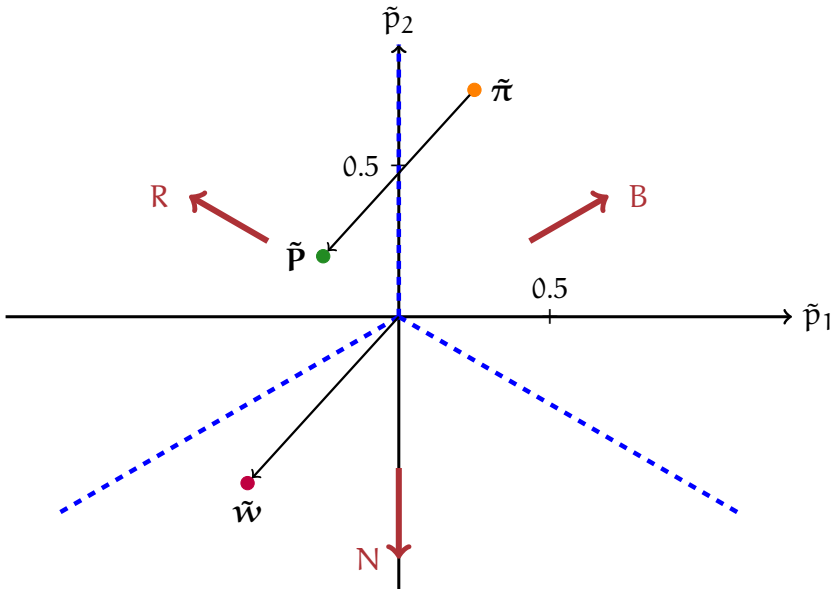


Figure 9: Bayesian updating in the three hypotheses ILR space. The posterior distribution $\tilde{P}$ is the translation of the prior distribution $\tilde{\pi}$ by the likelihood function $\tilde{w}$. The red arrows indicate the directions that go in favor of one hypothesis against the two others. The dashed blue rays mark out the maximum a posteriori decision regions.

hypothesis H_2 the second component $\tilde{p}_2$ compares the third probability against the two others as illustrated in the bifurcation tree in Figure 8. Each red arrow shows the direction that goes in favor of one hypothesis against the two others. These three directions are naturally separated by an angle of 120° i. e. one-third of 360°. The dashed blue rays mark out the maximum a posteriori decision regions (see Appendix 10.3 for their computation). Here, the simplex is 2-dimensional because we consider three possible hypotheses but keep in mind that for N hypotheses, the simplex is $(N-1)$-dimensional. When there are only two possible hypothesis, the simplex is one-dimensional such that if one goes against one hypothesis, it necessarily goes in favor of the other. With more hypotheses, the number of directions is now infinite.

Let's explicitly write in the notation of the likelihood vector the data or evidence E:

$$\tilde{w}(E) = \text{ilr}\left([P(E \mid H_1), \dots P(E \mid H_N)]\right), \tag{47}$$

and let $S = \{E_i\}_{1 \leqslant i \leqslant |S|}$ be a sequence of observed evidence. The posterior distribution over the set of hypothesis $\mathcal{H}$ given the sequence S is the successive (the order does not matter) translation of the probability distribution by the likelihood vectors:

$$\tilde{P} = \tilde{\pi} + \sum_{i=1}^{|S|} \tilde{w}(E_i). \tag{48}$$

We therefore recover the independent additivity as in the logit form of the Bayes' rule in the two hypotheses case (see Equation 8).

Example:

Coming back to the example we discussed in Section 2.1.1 with the urn that consists of either $\frac{2}{3}$ of black balls and $\frac{1}{3}$ of red balls or $\frac{2}{3}$ of red balls and $\frac{1}{3}$ of black balls. Let's consider a third possible hypothesis that states that none of the colors prevails. Let's summarise the set $\mathcal{H} = \{\mathrm{B}, \mathrm{R}, \mathrm{N}\}$ of possible hypotheses:

- The hypothesis B: there are $\frac{2}{3}$ of black balls,
- The hypothesis R: there are $\frac{2}{3}$ of red balls,
- The hypothesis N: none of the color prevails, there are as many black balls as red balls.

Either a black or a red ball can be drawn. Again, assuming that the number of balls in the urn is sufficiently large such that drawing a few balls does not significantly change the color proportions, the likelihood functions are:

$$\begin{aligned}
\boldsymbol{w}(\cdot) &= [\mathrm{P}(\cdot \mid \mathrm{B}), \mathrm{P}(\cdot \mid \mathrm{R}), \mathrm{P}(\cdot \mid \mathrm{N})]\,, \\
\tilde{\boldsymbol{w}}(\cdot) &= \left[\frac{1}{\sqrt{2}} \log \frac{\mathrm{P}(\cdot \mid \mathrm{B})}{\mathrm{P}(\cdot \mid \mathrm{R})}, \frac{1}{\sqrt{6}} \log \frac{\mathrm{P}(\cdot \mid \mathrm{B})\mathrm{P}(\cdot \mid \mathrm{R})}{\mathrm{P}(\cdot \mid \mathrm{N})^2}\right],
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{w}(\mathrm{b}) &= \left[\frac{2}{3}, \frac{1}{3}, \frac{1}{2}\right], \\
\tilde{\boldsymbol{w}}(\mathrm{b}) &= \left[\frac{1}{\sqrt{2}} \log 2, \frac{1}{\sqrt{6}} \log \frac{8}{9}\right],
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{w}(\mathrm{r}) &= \left[\frac{1}{3}, \frac{2}{3}, \frac{1}{2}\right], \\
\tilde{\boldsymbol{w}}(\mathrm{r}) &= \left[-\frac{1}{\sqrt{2}} \log 2, \frac{1}{\sqrt{6}} \log \frac{8}{9}\right].
\end{aligned}$$

Figure 10 shows the successive applications of the Bayes' rule when different sequences of draws are observed. The prior probability distribution $\tilde{\boldsymbol{\pi}}$ is at the origin i. e. the maximum uncertainty position. 10a shows a case where the observed sequence of draws is rbrbbr. There are as many observed black balls as red ones. The probability distribution moved toward the direction of hypothesis N against the two others that state that the urn is not balanced. In 10b, the observed sequence brrbrr contains $\frac{2}{3}$ of red balls. Although the resulting likelihood (sum of all likelihoods) goes in the opposite direction of B, the reader may first be surprised that it does not go straight to the direction of R. If we had had a likelihood going straight to R when drawing a red ball, it would have been against B and N equally. However, drawing a red ball goes indeed in favor of the falsity of B but not in favor of the falsity of N. This has to be regarded relatively: drawing further red balls would send the posterior deeper into the R region

> *and* further away from the N region. However, if the next drawn ball is instead black, the distribution falls into the N region. 10c shows the successive Bayesian updating when the observed sequence is bbrbrbb i. e. the one we studied in the two hypotheses case in Section 2.1.1. While in 2.1.1 we considered only two possible hypothesis, we have here considered a third possibility.

In the above example, all likelihoods go toward the lower half-part of the plan because there is no draw of a ball that would support the falsity of N. This particularity would not have been observed if for instance, we had considered instead an urn that contains three different colors and where the three possible hypotheses were about which color prevails.

3.3.4 *Multiple hypotheses evidence representation*

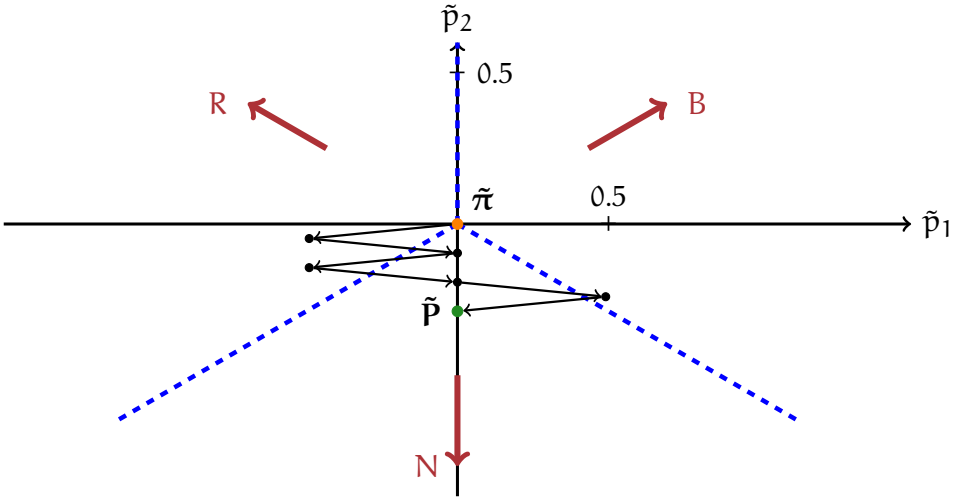
Recovering the additive form of the Bayes' rule—being the *basic property* of the weight-of-evidence [69]—the concept of LLR and weight-of-evidence can be now extended to cases with more than two hypotheses. The ILR transformation of the likelihood function, which we will now call Isometric-Log-Ratio-Likelihood (ILRL), can be seen as a multidimensional extension of the LLR making it a good candidate for representing the evidence when there are more than two possible hypotheses.

The direction of the ILRL informs which hypothesis or hypotheses the data may or may not support. The norm of the ILRL informs how strong. Like the absolute value of the LLR, the absolute value of each ILRL component gives the *strength-of-the-evidence* in the support of one hypothesis against some others as shown by the bifurcation tree (Figure 8). However, one basis does not provide all possible comparisons of hypotheses, this would have been redundant. If one wants to do a specific comparison, let's say for instance p_3 against p_1 only, he or she will have to use another basis resulting in a different bifurcation tree [54], which basically corresponds to a rotation and/or a permutation in the Euclidean space.

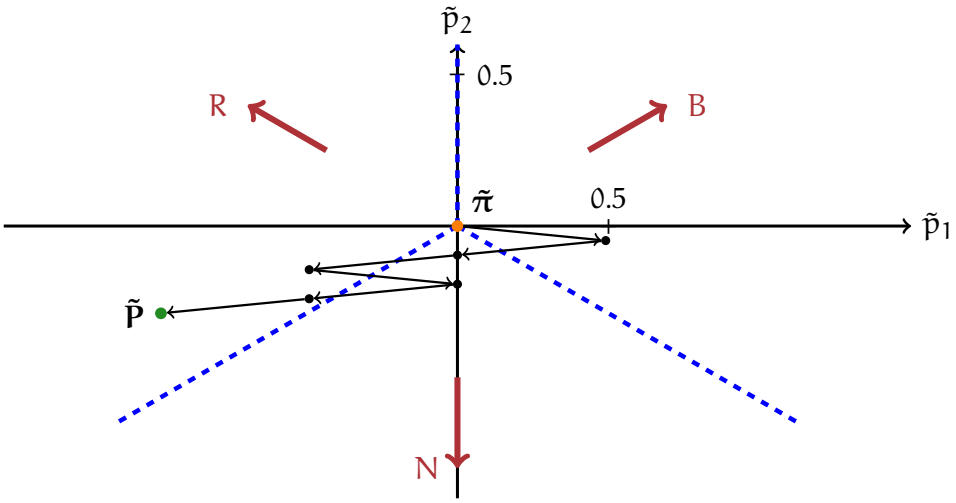
The Aitchison norm of the likelihood vector $\boldsymbol{w}$ (i. e. the Euclidean norm of its ILR transformation $\tilde{\boldsymbol{w}}$) can be regarded as a *global strength-of-evidence* and is given by:

$$\|\boldsymbol{w}(E)\|_a = \|\tilde{\boldsymbol{w}}(E)\|_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\log\frac{P(E \mid H_i)}{P(E \mid H_j)}\right)^2}. \tag{49}$$

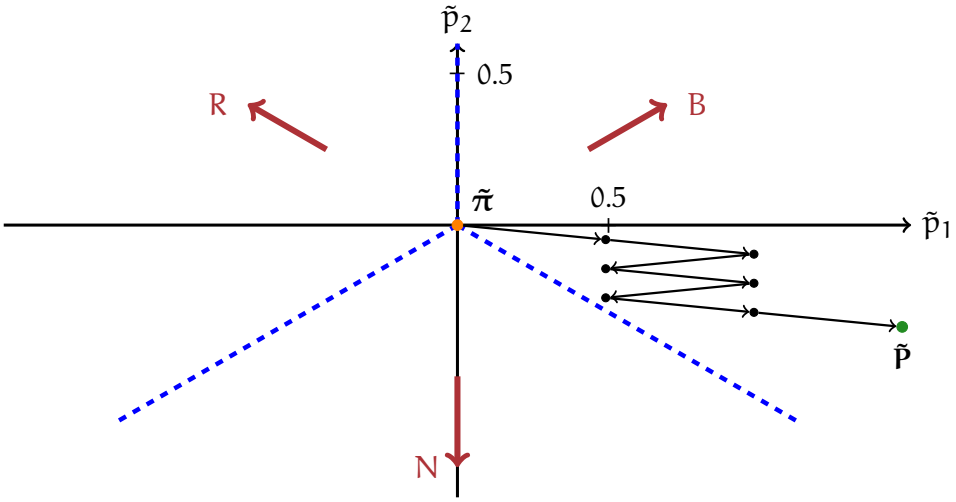
This is proportional to the square root of the sum of the square of all possible LLR. This informs how much the evidence E is changing the belief, i. e. how far the posterior distribution is from the prior distribution, regardless of any hypothesis. Since $\|\boldsymbol{w}\|_a = 0 \iff \tilde{\boldsymbol{w}} = \boldsymbol{0}$, a *global strength-of-evidence* equal to zero confirms there is no update of belief: the posterior remains equal to the prior. Such a situation is what is desirable in our conception of privacy where we want the belief of the attacker to not change when observing some data. The next chapter formally presents this conception of privacy.



(a) Sequence: rbrbbr



(b) Sequence: brrbrr



(c) Sequence: bbrbrbb

Figure 10: Bayesian updating in the three hypotheses ILR space when observing different sequences of ball draws. The prior probability vector $\tilde{\boldsymbol{\pi}}$ is successively translated by the likelihood vectors resulting in the posterior $\tilde{\mathbf{P}}$.

3.4 SUMMARY

This chapter briefly introduced compositional data analysis and the Aitchison geometry of the simplex. The latter gives to the probability simplex, i. e. the sample space of discrete probability distributions, a Euclidean vector space structure in which the Bayes' rule is the translation of the prior probability distribution by the likelihood function. The Isometric-Log-Ratio (ILR) transformation expresses the probability and the likelihood vector in this Euclidean vector space where the linearity and independent additivity of the Bayes' rule are recovered. The ILR transformation of the likelihood vector (ILRL) is a good candidate for representing the evidence and extending the idea of the LLR when more than two hypotheses are considered. For perfect privacy, the attacker's belief must not change. Therefore, we want the strength of the translation induced by the likelihood functions to be null i. e. we want the norm of the likelihood functions in the Aitchison geometry of the simplex to be zero: this is zero-evidence.

4

PERFECT SECRECY, PERFECT PRIVACY AND EVIDENCE INFORMATION

In 1948, Claude Shannon published his work "A Mathematical Theory of Communication" [145], a pioneer in the field of information theory. This work found numerous applications, especially in the development of telecommunications but also in cryptography and secrecy systems with the paper "Communication Theory of Secrecy Systems" that Shannon published a year after [146]. In this paper, Shannon studied the mathematical structure of general secrecy systems where a message is enciphered, before transmission, with a particular key known by the source of the message and by the receiver. During the transmission, an attacker may intercept the enciphered message—called cryptogram—and try to infer the message. The attacker's prior belief about the transmitted message is represented by a prior probability distribution over the set of all the possible messages. When observing the intercepted cryptogram, the belief of the attacker is changed into a posterior belief.

In order for the receiver to decipher the cryptogram, the enciphering transformation—with a given key—has to be invertible. In contrast, the concept of privacy we are interested in this thesis requires the sensitive information to be unrecoverable by any means. The study of secrecy systems in the context of cryptography is not the only contribution of Shannon in his 1949 paper [146]. He also discussed the theoretical concept of *perfect secrecy*. This concept can be used in the context of attribute privacy.

After presenting the idea behind *perfect secrecy*, this chapter presents the concept of *perfect privacy* and *zero-evidence* [113]. It then presents, when the sensitive information to protect is binary, an information-theoretic metric for assessing how far to *perfect privacy* a protection system is. Finally, the concept of *evidence information* as the amount of information disclosed by the data is discussed.

4.1 CLAUDE SHANNON'S PERFECT SECRECY

In this section, the definition of *perfect secrecy* developed by Shannon in [146] is presented. Let $\mathcal{M} = \{M_1, M_2, \dots M_N\}$ be the set of possible messages and $\mathcal{E} = \{E_1, E_2, \dots E_N\}$ be the set of possible cryptograms. A message $M \in \mathcal{M}$ is enciphered with the ith key which gives a cryptogram $E \in \mathcal{E}$: $E = f_i(M)$, where f_i is the enciphering transformation corresponding to the ith key. Let the attacker have a prior belief about the transmitted message represented by a prior probability distribution $\boldsymbol{\pi} \in \mathcal{S}^N$ over the set of possible messages:

$$\boldsymbol{\pi} = [P(M_1), \dots P(M_N)]^T = [\pi_1, \dots \pi_N]^T . \tag{50}$$

After intercepting the cryptogram and processing it by any means, the belief of the attacker is updated resulting in a posterior probability distribution:

$$\mathbf{P}(E, \boldsymbol{\pi}) = [P(M_1 \mid E), \dots P(M_N \mid E)]^T \in \mathcal{S}^N. \tag{51}$$

Perfect secrecy is simply defined as follow:

$$\forall E,\ \mathbf{P}(E, \boldsymbol{\pi}) = \boldsymbol{\pi}. \tag{52}$$

In this way, the posterior probabilities of the attacker remain equal to the prior probabilities and the attacker receives no information about the transmitted message.

Chapter 2 discussed the Bayes' rule as the natural way to update a prior belief when new data is obtained. The Bayes' rule is here rewritten in terms of the current cryptography problem:

$$P(M \mid E) = \frac{P(E \mid M)P(M)}{P(E)}, \tag{53}$$

where $P(E \mid M)$ is the likelihood of getting the cryptogram E when M appeared to be the transmitted message and $P(E)$ is the marginal. One can see from this equation that perfect secrecy is achieved when the ratio $\frac{P(E|M)}{P(E)}$ is one for all E and M. In other words, we want:

$$\forall E,\ \forall M,\ P(E \mid M) = P(E). \tag{54}$$

This has been presented in [146] as a necessary and sufficient condition for perfect secrecy. In most practical cryptography problems, perfect secrecy is not achievable since it requires the entropy of the key to be at least as great as the entropy of the message. Regardless, in this thesis, we are not concerned with this issue since we are interested in attribute privacy where no keys are involved and the concealed information does not have to be recoverable.

4.2 PERFECT PRIVACY & ZERO EVIDENCE

In the context of attribute privacy, the aim is to conceal in some data or observations, the information about an attribute of the person to whom the data belongs. The attribute is supposed to be discrete with a finite number of categories. In this context, perfect secrecy is renamed *perfect privacy* and is achieved when:

$$\forall E \in \mathcal{E}, \forall C \in \mathcal{C},\ P(C \mid E) = P(C), \tag{55}$$

where $\mathcal{E}$ and $\mathcal{C}$ are respectively the set of observations and the set of categories the attribute can take[1].

1 The notation C is here used in accordance with the concept of *class* in pattern recognition.

When the attribute to hide is binary with $\mathcal{C} = \{C_1, C_2\}$, we have seen in Section 2.1.2 that the Bayes' rule can be written as the sum of the Log-Likelihood-Ratio (LLR) and the prior log-odds:

$$\operatorname{logit} P(C_1 \mid E) = \log \frac{P(E \mid C_1)}{P(E \mid C_2)} + \operatorname{logit} P(C_1). \tag{56}$$

The LLR represents the evidence in the data, i. e. how much the data is supporting one hypothesis against the other and how the data is changing the belief of the observer. One can easily see from the above equation that perfect privacy is reached when:

$$\forall E \in \mathcal{E},\ \log \frac{P(E \mid C_1)}{P(E \mid C_2)} = 0. \tag{57}$$

In the two hypotheses or classes case, perfect privacy is reached when the LLRs are equal to zero: this is *zero-evidence*[2].

In the context of the VoicePrivacy initiative [155], the concept of *perfect secrecy* has been formulated as *perfect privacy* in Andreas Nautsch's paper [113]. The paper proposes an evaluation framework to empirically assess how close to perfect privacy a protection system is. This framework will be presented in more detail in the next section.

Equation 57 formulates the concept of zero-evidence and perfect privacy in the binary case. In Chapter 3, the Aitchison geometry of the simplex has been presented. Within this geometry, the Bayes' rule can be written as the *perturbation* of the prior probability distribution by the vector of likelihoods i. e. the likelihood function. With the coordinate system given by the Isometric-Log-Ratio (ILR) transformation, the Bayes' rule is the translation of the prior probability distribution by the likelihood function:

$$\forall E \in \mathcal{E}, \tilde{\mathbf{P}}(E, \boldsymbol{\pi}) = \tilde{\boldsymbol{w}}(E) + \tilde{\boldsymbol{\pi}}, \tag{58}$$

where $\tilde{\boldsymbol{w}}(E)$ is the ILRL i. e. the ILR transformation of the likelihood vector, and $\tilde{\boldsymbol{\pi}}$ and $\tilde{\mathbf{P}}(E, \boldsymbol{\pi})$ are respectively the ILR transformation of the prior distribution and the ILR transformation of the posterior distribution. Here, one can see that perfect privacy is reached if and only if:

$$\forall E \in \mathcal{E},\ \tilde{\boldsymbol{w}}(E) = \mathbf{0}, \tag{59}$$

where $\mathbf{0}$ is the $(N-1)$-dimensional zero vector: this is multiple hypotheses *zero-evidence*.

Chapter 3 already discussed the ILRL as the multiple hypotheses extension of the LLR for representing the evidence in Bayesian updating. The *strength-of-evidence* is given by the norm of the ILRL and informs how much the data is changing the observer's belief. Here, we naturally see that, for perfect privacy, the ILRL must be the zero vector corresponding to a null strength-of-evidence.

2 The expressions *perfect privacy* and *zero-evidence* will be used interchangeably. *Perfect privacy* has the advantage to follow Shannon's terminology while *zero-evidence* refers to the notion of evidence in biometry and Bayesian updating.

In his paper [146], Shannon presented the necessary and sufficient condition for perfect secrecy written in Equation 54 which of course implies that $\tilde{\boldsymbol{w}}(E) = \mathbf{0}$. However, as discussed in Section 3.3, likelihood vectors are scale-invariant such that the necessary and sufficient condition can therefore be rewritten as:

$$\textit{perfect privacy} \iff \forall E \in \mathcal{E}, \forall C \in \mathcal{C},\ P(E \mid C) = \alpha \iff \forall E \in \mathcal{E},\ \tilde{\boldsymbol{w}}(E) = \mathbf{0}, \tag{60}$$

where $\alpha \in \mathbb{R}_+^*$.

4.3 REGARDING THE AMOUNT OF INFORMATION DISCLOSED BY THE DATA

The concept of information is usually expressed in terms of *uncertainty* [146]. The information received by an observer is the difference between its prior uncertainty and its posterior uncertainty [92]. The entropy function H returns the uncertainty associated with a probability distribution such that the information—about a set of hypothesis $\mathcal{C} = \{C_1, C_2, \dots C_N\}$—received by an individual when observing e can be written as[3]:

$$I(e, \boldsymbol{\pi}) = H(\boldsymbol{\pi}) - H(\mathbf{P}(e, \boldsymbol{\pi})), \tag{61}$$

where $\boldsymbol{\pi}$ is the prior probability distribution of the individual and $\mathbf{P}(e, \boldsymbol{\pi})$ is the posterior probability distribution of the individual when observing e. Note that this quantity can be negative but is positive on average. Indeed the entropy of *one* posterior distribution should not be confused with what is called *conditional entropy* [34] or *posterior entropy* [136] defined as:

$$H(\mathcal{C} \mid E) = \int_e P(e) H(\mathbf{P}(e, \boldsymbol{\pi}))\, de, \tag{62}$$

which is the expected entropy of the posterior distributions. This quantity is always below the entropy of the prior distribution:

$$H(\mathcal{C} \mid E) \leqslant H(\boldsymbol{\pi}). \tag{63}$$

This inequality is well-known in information theory [34]. However, in the context of pattern recognition, the probabilities $P(e)$ are usually not known. This is where the cross-entropy and the Empirical Cross-Entropy (ECE) [136] presented in Section 2.4.2 reappear. When the scores are calibrated, we have seen that the ECE is below the prior entropy which is consistent with the above inequality. The gap between the prior entropy and the ECE can be seen as the amount of information an individual received from the scores given by the recognizer.

Nevertheless, whether in terms of posterior entropy or in terms of ECE, the measure of information received from an observation or from a set of observations still depends on the observer's prior. Let's consider the following example to better understand how measuring

3 From now on, the expression *entropy function* will refer to the Shannon's entropy as defined in Equation 15.

the amount of information received by an observed in terms of a drop of uncertainty necessarily depends on the prior. Let's say that the individual is almost certain about which hypothesis is true. When observing something that confirms his or her intuition, he or she is not learning something "new" since he or she was already almost certain before the observation.

In a privacy context, since he or she does not know the prior belief of the attacker, the privacy safeguard must think in terms of *"information disclosed"* rather than in terms of *"information received by the attacker"*. Indeed, as discussed above, the latter necessarily depends on the unknown initial knowledge of the attacker. It is in this context that the Zero Evidence Biometric Recognition Assessment (ZEBRA) framework [113] proposes a measure called the *expected privacy disclosure* independent of the attacker's prior.

4.3.1 *ZEBRA's expected privacy disclosure*

The ZEBRA framework [113] has been proposed to measure the amount of information disclosed by scores of biometric recognition or any binary attribute recognition systems.

Let's consider a set of two hypotheses or classes $\mathcal{C} = \{C_1, C_2\}$. These can be for instance $\{\mathtt{tar}, \mathtt{nontar}\}$ for biometric recognition based on *verification* as discussed for instance in the case of ASV in Section 2.2 and in the original ZEBRA paper [113]. In this case, the privacy objective is to hide the identity of the individual by making verification impossible. $\mathcal{C}$ can also be a set of categories or classes a binary attribute can take as for instance $\{\mathtt{male}, \mathtt{female}\}$. In any case, we want the data to provide no information i. e. no evidence about which hypothesis is true or false.

When the attacker observes a sample from which he or she wants to infer some information about the sensitive attribute, the evidence given by the sample to the attacker is represented by a LLR. In practice, this LLR is computed using a pattern recognizer. ZEBRA provides a measure called the *expected privacy disclosure* D_{ECE} which gives the expected private information disclosed by the recognizer independently of the attacker's prior knowledge.

D_{ECE} basically measures how far the *minimum* Empirical Cross-Entropy (ECE) curve is from the prior entropy profile. We recall here the expression of the ECE[4]:

$$\begin{aligned} E_{CE}(\pi, \mathcal{L}_1, \mathcal{L}_2) = &\frac{1-\pi}{|\mathcal{L}_1|} \sum_{l \in \mathcal{L}_1} \log_2 \left(1 + \frac{\pi}{1-\pi} \exp(-l)\right) \\ &+ \frac{\pi}{|\mathcal{L}_2|} \sum_{l \in \mathcal{L}_2} \log_2 \left(1 + \frac{1-\pi}{\pi} \exp(l)\right), \end{aligned} \tag{64}$$

where $\pi = P(C_2) = 1 - P(C_1)$ is the prior probability of the attacker, $\mathcal{L}_1$ is the set of LLRs for samples of class C_1, and $\mathcal{L}_2$ is the set of LLRs for samples of class C_2. When all LLR scores are zero, the ECE is equal to the prior entropy for all priors: this is zero-evidence. When

4 Base 2 is used here for the logarithmic scoring rule as in the original paper [113]. The information is therefore measured in bit.

the scores provide useful information to the attacker, the ECE is below the prior entropy. For perfect privacy, the aim is to reduce the area between the two profiles as illustrated in Figure 11. D_{ECE} is the measure of this area and is computed by integrating out the attacker's prior as follow:

$$D_{ECE}(\mathcal{L}_1, \mathcal{L}_2) = \int_0^1 (H(\pi) - E_{CE}(\pi, \mathcal{L}_1, \mathcal{L}_2))\, d\pi, \tag{65}$$

where $H(\pi) = -\pi \log_2 \pi - (1-\pi)\log_2(1-\pi)$ is the prior entropy (as defined in Equation 15). Computing the integral in Equation 65 gives the following expression for the D_{ECE}:

$$D_{ECE}(\mathcal{L}_1, \mathcal{L}_2) = \frac{1}{2\log 2}\left(\frac{1}{|\mathcal{L}_1|}\sum_{l\in\mathcal{L}_1}(Z(l)) + \frac{1}{|\mathcal{L}_2|}\sum_{l\in\mathcal{L}_2}(Z(-l))\right), \tag{66}$$

where:

$$Z(l) = \frac{1}{2} + \frac{l + (1-\exp(l))}{(1-\exp(l))^2}, \tag{67}$$

$$\lim_{l\to-\infty} Z(l) = -\infty,\ \lim_{l\to 0} Z(l) = 0,\ \lim_{l\to\infty} Z(l) = \frac{1}{2}.$$

The expression $H(\pi) - E_{CE}(\pi, \mathcal{L}_1, \mathcal{L}_2)$, in the integral of Equation 65, can be seen as an empirical formulation of $H(\pi) - H(\mathcal{C} \mid E)$. This tells, for a given prior π, how much the uncertainty of the attacker has decreased and therefore how much information the attacker has received. However, as discussed above, this depends on the attacker's prior and this is why the prior is integrated out[5].

The ECE discussed here is the *minimum* ECE where the LLRs in $\mathcal{L}_1 \cup \mathcal{L}_2$ are perfectly calibrated through PAVA. In practice, the scores of the attacker are not necessarily calibrated. D_{ECE} is therefore measuring an overestimated amount of useful information disclosed to the attacker. However, considering such a pessimistic scenario is a safe strategy for privacy design. The scenario where the attacker's scores are not calibrated has been briefly discussed as the *expected calibration distortion* in [121].

> **Remark:**
> In contrast to differential privacy [52] which is a theoretical guarantee of the amount of privacy ensured by a protection mechanism, the level of privacy given by a protection system is here empirically measured.

4.3.2 *Uncertainty and evidence information*

In this section, the conceptual differences between the *information received* expressed in terms of uncertainty and the *information disclosed* will be discussed in more detail.

5 Assuming that all prior values have the same probability of occurrence. In a situation where the privacy safeguard has some information about the attacker's prior, non-uniform probability distribution for the attacker's prior value could be used instead, on condition that the integral can be computed.

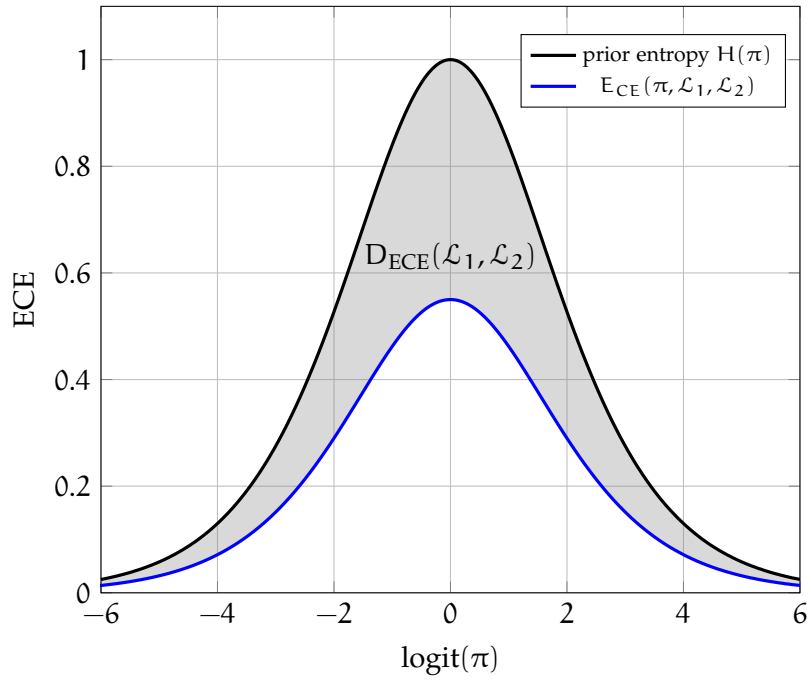


Figure 11: Area $D_{ECE}(\mathcal{L}_1, \mathcal{L}_2)$ between the prior entropy and the ECE curve as the expected amount of information disclosed by the set of scores $\mathcal{L}_1 \cup \mathcal{L}_2$.

Considering only two hypotheses, the absolute value of the logit of one of the probabilities is related to an uncertainty function. As discussed in Section 2.4.1, an uncertainty function is concave on the simplex. The absolute logit function is convex as one can see in Figure 12. A minimum absolute logit aligns with a maximum uncertainty while an infinite absolute logit aligns with no uncertainty. In the two hypotheses case, the absolute logit function is equivalent to the Aitchison distance between one distribution and the uniform distribution (up to a scaling factor $\frac{1}{\sqrt{2}}$). However, in a multiple hypotheses case, even if this distance informs how far a probability distribution is from the uniform distribution, i. e. the distribution corresponding to the maximum uncertainty, this can not be interpreted as an uncertainty like the Shannon's entropy. Figure 13a shows the entropy function in the three hypotheses ILR space of probability distributions and Figure 13b shows the distance from the uniform distribution[6]. $\tilde{p}_1$ and $\tilde{p}_2$ are respectively the first and the second ILR component[7]. Both the entropy and the distance reach their extremum at the uniform distribution ($\tilde{p}_1 = 0$ and $\tilde{p}_2 = 0$). However, the entropy is "aware" of the hypothesis directions while the distance is not. Indeed, contrary to the entropy, the distance evolves equally regardless of the distribution's direction. When the distribution goes in the direction of a hypothesis, the entropy drops to zero (see the blue regions in Figure 13a). When the distribution goes in the opposite direction of one hypothesis, the entropy diminishes but not to zero. Indeed some certainty has been obtained: it is more certain that one hypothesis is *not* true, but

6 This is the Aitchison norm of the probability vector but within the coordinate system given by the ILR transformation, this norm is simply the 2-norm.

7 Considering a set of three possible hypotheses and a probability distribution $\mathbf{p} = [p_1, p_2, p_3]$ over this set, $\tilde{p}_1 = \frac{1}{\sqrt{2}} \log \frac{p_1}{p_2}$ and $\tilde{p}_1 = \frac{1}{\sqrt{6}} \log \frac{p_1 p_2}{p_3^2}$.

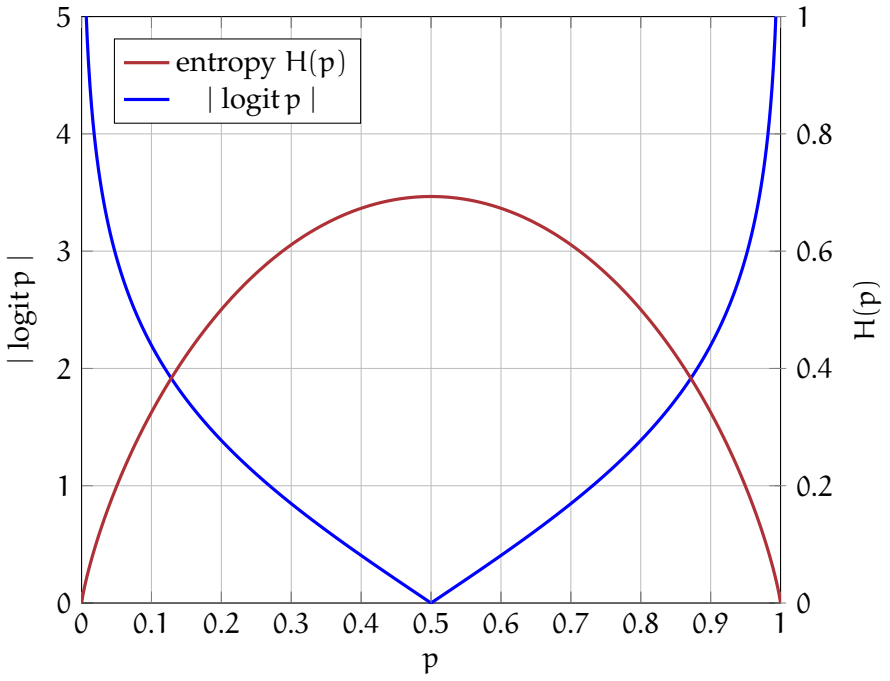


Figure 12: Shannon's entropy and absolute value of the logit function on the one dimensional simplex.

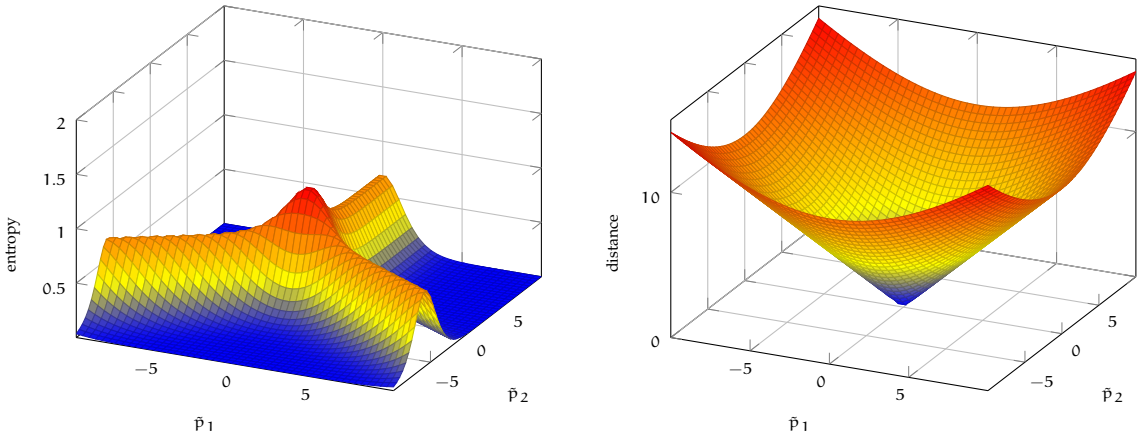
uncertainty about the two other hypotheses remains. Starting from the uniform distribution and going straightly to the opposite direction of a hypothesis, along one of the three ridges, reduces the entropy from $\log 3$ to $\log 2$ i. e. from the three hypotheses maximum uncertainty to the two hypotheses maximum uncertainty.

Even if the Aitchison distance from the uniform distribution, i. e. the Aitchison norm, can not be interpreted as an uncertainty function, this can be used on the vector of likelihoods as a measure of information.

As already discussed in Section 3.3.4, the Aitchison norm of the likelihood function is the *strength-of-evidence* telling how much an observation is changing the belief. This can be seen as a measure of information provided by the observation. The idea of using the Aitchison norm as a measure of information has been discussed in [55, 58].

Remark:

Here, a clarification of the terminology employed is necessary. In [55, 58], the authors used the expression *evidence information* to refer to the information given by an observation, or to refer to the information captured by a prior or a posterior. In this thesis, the word *evidence* refers to the observation (or the data, the experiment, whatever you call it depending on the context) and we are only concerned by the amount of information it provides. Following the terminology employed in [68], and considering the ILR transformation as the multidimensional extension of the logit transformation, the ILRL can be seen as the multiple hypotheses extension of the *weight-of-evidence* while the ILR transformation of a probability vector refers to the *plausibility*. This may



(a) Entropy in the ILR space.

(b) Euclidean distance from the uniform distribution in the ILR space.

Figure 13: Shannon's entropy and Euclidean distance (Aitchison distance on the simplex) from the uniform in the three hypotheses ILR space of probability distribution.

> somehow sound absurd since the terms *weight* and *plausibility* refer both to a scalar while in a multiple hypotheses context, the ILR transformation of likelihood or probability vector is multidimensional. Be that as it may, this thesis is concerned only with the likelihood function for representing the evidence and the expression *evidence information* will here refer to the Aitchison norm of the likelihood function as a measure of the information disclosed by an observation.

Within the Aitchison simplex, the likelihood function is translating the prior into the posterior. Its norm is the distance between the prior and the posterior and therefore tells how much the belief has changed. This change of belief is here considered as the amount of *evidence information* i. e. the amount of information—about the hypotheses—disclosed by the data. For perfect privacy, the aim is to provide no information in the sense that the attacker's belief does not change. In this context, the term information refers therefore to the evidence information.

Considering the Aitchison norm of the likelihood function as the amount of information disclosed by an observation has some benefits. First, this is prior independent contrary to the difference between the prior and the posterior uncertainty. Secondly, considering a single observation, measuring the information in terms of a uncertainty drop can be misleading. When the entropy of the posterior is greater than the entropy of the prior, the measure of information is negative. In the context of zero-evidence, what does "negative" information mean when the belief has changed? As another example, let's consider, in a three hypotheses case, a prior belief of $[0.8, 0.1, 0.1]$ and a posterior belief of $[0.1, 0.8, 0.1]$. These two beliefs are different: initially, the individual was going in favor of the first hypothesis and then, he or she is going in favor of the second hypothesis. And yet, they both correspond to the same uncertainty such that measuring the information in terms of uncertainty would give zero. With the evidence information, measured by the Aitchison norm of the likelihood function, the information is positive as soon as the belief has changed.

Even if the observation is misleading and the uncertainty of the attacker has increased, it is considered that a certain amount of information has been disclosed. Indeed, in the context of perfect privacy, the goal is not to mislead or to dupe the attacker but rather to provide no evidence that may change the attacker's belief. Perfect privacy is reached when the evidence information is zero corresponding to a zero strength-of-evidence and equal likelihoods as mentioned in Section 3.3.4.

Actually, one can not say that among these two ways of measuring information—measuring the information as a drop of uncertainty or as a change of belief—one is better. Comparing them is meaningless since they both refer to different notions of information:

- The *information received*, by an individual, that necessarily depends on the prior belief and is expressed as a difference between the prior and the posterior uncertainty of the individual,
- The *information disclosed* by the data tells how much the data is supporting the trueness or falseness of one hypothesis or a set of hypotheses and tells how much the belief is changing when observing the data. This is expressed by the Aitchison norm of the likelihood function but can not be interpreted in terms of the observer's uncertainty.[8].

While no information disclosed, i. e. zero-evidence, implies no information received by the attacker, no information received does not imply no information disclosed.

4.4 SUMMARY & DISCUSSION

Perfect secrecy is a theoretical concept developed by Shannon stating that the attacker's knowledge about a sensitive transmitted message must not change when observing the cryptogram. Formulated in terms of subjective probabilities, the posterior probability distribution of the attacker must remain equal to the prior one. This concept can be used in the context of privacy preservation where the posterior belief of the attacker about the sensitive attribute must be equal to the prior belief. In this case, the data provides no information about the attribute to the attacker. This is *perfect privacy* also referred to as *zero-evidence* and is reached when all likelihoods are equal i. e. when the ILRL is the null vector.

The *expected privacy disclosure* D_{ECE} measures—independently of the attacker's prior—how much a privacy system is far from perfect privacy. It measures the amount of information disclosed by the scores of a binary attribute recognizer that proceeds on protected data. However, the recognizer is actually the one used by the attacker which is unknown. Even if, for the computation of the D_{ECE}, the scores are perfectly calibrated, the discrimination component of the ECE depends on how flexible the recognizer is[9]. It is therefore implicitly assumed that the family of recognizers used by the privacy safeguard for the

8 The ZEBRA's expected privacy disclosure presented in Section 4.3.1 is a measure of the information disclosed. However, this can be interpreted in terms of uncertainty. This is the expected drop in the uncertainty of the attacker when the expectation is done over the attacker's prior.

9 By "flexible" we mean the ability for the recognizer to extract discriminant information from the data.

computation of D_{ECE} is wide enough to cover the attacker's recognizer. The expected privacy disclosure D_{ECE} is formalized for binary recognition problems. Its extension to the multiple hypotheses case has not been studied yet and remains an open question.

Section 4.3.2 detailed nuances between two different notions of information. One is the information *received* by an individual (the attacker in a privacy context) which is a drop of uncertainty and necessarily depends on the prior knowledge. The other is the information *disclosed* by one observation which tells how much the belief of the observer is changing and is expressed by the Aitchison norm of the likelihood function also called here the *strength-of-evidence*.

The next chapter presents a new approach to discriminant analysis for a binary classification. It is based on the indempotence property of the LLR and on Proposition 1. This approach allows the manipulation in the data of the evidence information represented by a LLR. These can be set to zero for privacy purposes, being therefore consistent with the zero-evidence formulation of privacy presented above. In order to extend—in Chapter 7—this discriminant analysis to applications with more than two classes, Chapter 6 extend the idempotence property and Proposition 1 to the ILRL.

5

A NON-LINEAR DISCRIMINANT ANALYSIS FOR BINARY ATTRIBUTE PRIVACY

This chapter presents a new discriminant analysis. In addition to recognition tasks, the proposed model can be used to remove any evidence about a set of two hypotheses or classes in some data. We will here adopt a pattern recognition terminology where $\mathbf{x}$ refers to an observed feature vector that belongs either to class C_1 or to class C_2. Along this thesis, the term *class* and *hypothesis* are used interchangeably depending on the context. In the current pattern recognition context, the probability for an element to belong to a class i is equivalent to the probability for the hypothesis "The element belongs to class i" to be true.

Discriminant analysis aims to map the data into a space that maximizes the separability between the classes. The discriminant direction on which the data is mapped can often be seen as a LLR-like score. When the mapping is invertible, the data can be mapped back into the original feature space after having annihilated the LLR. This idea will be used to divert the use of discriminant analysis for privacy purposes in Section 5.3. Let's first introduce in more detail the discriminant analysis by presenting the Linear Discriminant Analysis (LDA).

5.1 LINEAR DISCRIMINANT ANALYSIS AND OTHERS

With the LDA, the class-conditional densities in the feature space are assumed to be multivariate normal distributions with a shared covariance:

$$\begin{aligned} \mathbf{x} \mid C_1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \\ \mathbf{x} \mid C_2 &\sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \end{aligned} \tag{68}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors and $\boldsymbol{\Sigma}$ is the shared covariance matrix. A LLR-like score[1] for an observed vector $\mathbf{x}$ is computed as follows:

$$\log \frac{f(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{f(\mathbf{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})} = \boldsymbol{\nu}^\top \mathbf{x} + \frac{1}{2}\left(\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1\right) \tag{69}$$

where $f(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\nu} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the eigenvector with non-zero eigenvalue of the matrix $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_B$ where $\boldsymbol{\Sigma}_B$ is the between-class covariance matrix. The

1 We sometimes use the expression *LLR-like score* to refer to scores that are log-likelihood-ratio in the sense that they are computed from the ratio of probability density functions images but are not necessarily calibrated and do not necessarily respect the idempotence property. Keep in mind that the idempotence property is a property of *calibrated* LLR or a property that LLR *"should"* respect.

feature vector $\mathbf{x}$ is projected onto the discriminant direction given by $\mathbf{v}$ and shifted. This can be seen as the projection of the data that minimizes the within-class variability and maximizes the between-class variability. However, since the observations are not necessarily normally distributed, as assumed by the model, the computed scores are not necessarily calibrated LLR and therefore misrepresent the information related to the classes.

Other discriminant analysis approaches have softer assumptions on the distributions of the features. The QDA also assumes the features to be normally distributed for a given class, but the shared covariance assumption is dropped such that each class-conditional density has its own covariance and the LLR-like score becomes:

$$\begin{aligned}\log \frac{f(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{f(\mathbf{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} = &\frac{1}{2}\mathbf{x}^T \left(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x} + \mathbf{x}^T \left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \\ &+ \frac{1}{2}\left(\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\end{aligned} \tag{70}$$

This results in a quadratic form and a non-linear mapping of the features. However, this is still based on a Gaussian assumption and is not necessarily invertible contrary to the LDA's mapping [16, 73].

Some approaches make no assumption on how the feature vectors are distributed. In [50], the authors proposed what they called "DeepLDA" where the general LDA eigenvalue problem is solved on the top of an artificial neural network. However, the approach is fully discriminative and looses the generative nature of the LDA. In this sense, we do not consider this as a non-linear extension of the LDA. Generalized [152] and kernel-based [106] discriminant analysis are good candidates for generalizing the LDA in a non-linear manner with no assumptions on the distribution of the features. However, like the QDA, they do not have a trivial inverse mapping from the discriminant space back to the original feature space [159]. Works like [8, 77] propose generative classifiers by modeling the class-conditional densities of the features using invertible neural networks transformation, known as Normalizing Flow (NF), between the feature space and a *base*[2] space. In the base space, the class-conditional densities are chosen to be multivariate normal distributions. However, these approaches are not designed to have, in the base space, a discriminant direction interpretable as a calibrated LLR.

In the following section, we present a non-linear discriminant analysis with no explicit assumption on the distribution of the feature vectors. It allows mapping, in an invertible manner, input feature vectors into a space where the discriminant component is a calibrated LLR. This chapter is focused on the two classes case. Chapter 7 will extend the proposed discriminant analysis to any number of classes using the concept of Isometric-Log-Ratio-Likelihood (ILRL) presented in Chapter 3 and its properties presented in Chapter 6.

2 Some literature on Normalizing Flow (NF) uses the term *latent* rather than *base*. We were also doing so in [123], however, agreeing with the argument proposed in [127], we have adopted the term *base*. The mapping is indeed deterministic such that mapping the observed variable into the base space is just another way to represent the same random variable.

5.2 PROPOSED NON-LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

This section presents the new discriminant analysis that we introduced in [123]. The approach makes no explicit[3] assumption on how the feature vectors are distributed. It maps, in an invertible manner, the data into a space where one component is discriminant, respects the idempotence property, and forms a calibrated LLR. The main reasons why an invertible mapping is desirable are for preserving the sampling ability of generative models and for manipulating the data for privacy preservation. This latter point will be discussed in the next section.

Let $\{C_1, C_2\}$ be the set of classes and $\mathcal{X}$ be the d-dimensional feature space of the observed data. Let $l(\mathbf{x}) \in \mathcal{L} \subset \mathbb{R}$ be the LLR of an observation $\mathbf{x} \in \mathcal{X}$:

$$l(\mathbf{x}) = \log \frac{P(\mathbf{x} \mid C_1)}{P(\mathbf{x} \mid C_2)}, \tag{71}$$

which is, as we saw in Section 2.1.2, a good candidate for representing the evidence and which hypothesis the data is supporting and how strong. Let $\mathbf{r}(\mathbf{x}) = [r_1(\mathbf{x}), \dots r_{d-1}(\mathbf{x})]^T \in \mathcal{R} \subset \mathbb{R}^{d-1}$ be what we call the *residual* of $\mathbf{x}$. The residual is everything in the observation that is independent of the class variable. We want to map the feature vectors into a base space $\mathcal{Z} = \mathcal{L} \oplus \mathcal{R}$ where the first dimension represents the evidence represented by the LLR and the other dimensions form the residual and embed the variability not related to the classes. The first dimension is the discriminant one, while the other dimensions contain the remaining variability not related to the classes.

5.2.1 *Class-conditional densities in the base space*

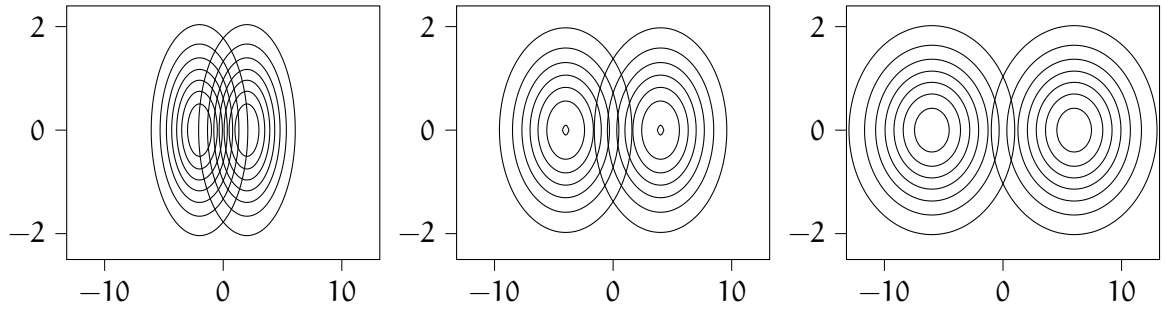
Since we want the first dimension of the base space to properly represent the evidence, we want it to be a calibrated LLR. Hence, its distribution has to respect the idempotence property constraint of Proposition 1. The class-conditional densities in the base space are therefore chosen accordingly:

$$\begin{aligned} \mathbf{z} \mid C_1 &\sim \mathcal{N}(\mu \mathbf{e}_1, \boldsymbol{\Sigma}), \\ \mathbf{z} \mid C_2 &\sim \mathcal{N}(-\mu \mathbf{e}_1, \boldsymbol{\Sigma}), \end{aligned} \tag{72}$$

where:

- $\mu \in \mathbb{R}^{*+}$ is the only parameter of the densities in the base space and is the Kullback-Leibler divergence between the two densities as discussed in Section 2.6.2.1,
- $\mathbf{e}_1 = [1, 0, \dots 0]^T$,
- $\boldsymbol{\Sigma} = \mathrm{diag}(2\mu, 1, \dots 1)$ is a diagonal covariance matrix with variances equal to one except the variance of the first component which is 2μ in order to respect Proposition 1.

3 To be more precise, the approach assumes there is a diffeomorphism that would transform the class-conditional densities of the feature vectors into the target densities. For more detail, see the remark at the end of Section 5.2.2.



(a) $\mu = 2$ (b) $\mu = 4$ (c) $\mu = 6$

Figure 14: Few contours of class-conditional densities in the base space. The first component is the LLR distributed according to Proposition 1. The other component is not discriminant and is normally distributed with a zero mean vector and an identity covariance matrix regardless of the classes.

In this way, the first component z_1 of $\boldsymbol{z}$ is the LLR of $\boldsymbol{z}$:

$$\begin{aligned} l(\boldsymbol{z}) &= \log \frac{(2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\mu\boldsymbol{e}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\mu\boldsymbol{e}_1)\right)}{(2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}+\mu\boldsymbol{e}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}+\mu\boldsymbol{e}_1)\right)}, \\ &= \log \frac{\exp\left(\mu \boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_1\right)}{\exp\left(-\mu \boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_1\right)}, \\ &= 2\mu \boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_1, \\ &= \frac{2\mu z_1}{2\mu} \text{ because } \boldsymbol{\Sigma}^{-1} = \operatorname{diag}\left(\frac{1}{2\mu}, 1, \dots 1\right), \\ &= z_1, \end{aligned} \tag{73}$$

and respects the idempotence constraint on the LLR's distribution. The other dimensions are normally distributed with a zero mean and an identity covariance matrix regardless of the class.

Figure 14 shows 2-dimensional examples of class-conditional densities in the base space. Only the first component is discriminant and is the LLR distributed according to Proposition 1. When the parameter μ increases, the separability between the two classes increases as discussed in Section 2.6.2.1.

5.2.2 *Mapping between the feature space and the base space*

Let's assume that there is a *diffeomorphism*—i. e. an invertible and differentiable mapping—g between $\mathcal{X}$ and $\mathcal{Z}$ such that $\boldsymbol{x} = g(\boldsymbol{z})$ and $\boldsymbol{z} = g^{-1}(\boldsymbol{x})$. Normalizing Flow (NF) is a family of differentiable and invertible neural network transformations that aims to transform a base probability distribution into an "observed" one. NF has been mostly developed in an unsupervised context but can be used with a supervised training for modeling class-conditional densities as briefly mentioned in Section 5.1.

Since the mapping is invertible, it allows both sampling and inference and if the determinant of the Jacobian can be computed, the likelihood of the data can be exactly computed. Let $\mathcal{D} = \{(\mathbf{x}^{(1)}, C^{(1)}), \dots (\mathbf{x}^{(|\mathcal{D}|)}, C^{(|\mathcal{D}|)})\}$ be a set of observed feature vectors with their corresponding class. The mapping g can be learned by maximizing the log-likelihood of the observed data:

$$\log f(\mathcal{D} \mid \theta_g, \mu) = \sum_{i=1}^{2} \left(\sum_{(\mathbf{x}, C_i) \in \mathcal{D}} \log f_{\mathcal{X}|C_i}(\mathbf{x} \mid \theta_g, \mu) \right), \tag{74}$$

where θ_g are the parameters of g and since the mapping is a diffeomorphism, the change of variable formula gives:

$$f_{\mathcal{X}|C_i}(\mathbf{x} \mid \theta_g, \mu) = f_{\mathcal{Z}|C_i}(\mathbf{z} \mid \mu) \left| \det \left(\frac{\partial g(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1}, \ \forall i \in \{1, 2\}. \tag{75}$$

In our work, g is a neural network with the Real NVP [48] architecture which allows easy inversion and computation of the Jacobian determinant.

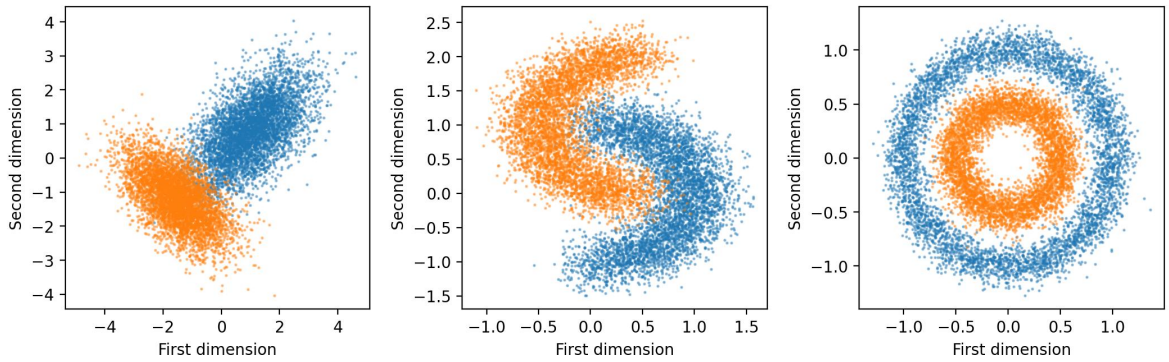
> **Remark on the implicit assumption made about the data:**
> While LDA and QDA assume the data to be normally distributed, the proposed approach does not have an explicit assumption on how the data is distributed. However, this does not mean there is no assumption at all. The proposed approach indeed assumes the existence of a diffeomorphism that would transform the features' class-conditional densities into the target Gaussian densities.

5.2.3 *Estimation of* μ

As one can see in Equation 72, μ is the only parameter of the densities in the base space. Like the parameters θ_g of the mapping, μ can be learned using automatic differentiation and gradient descent. In [123] we instead used a two steps optimization strategy. Indeed, μ appears only through the first dimension of $\mathbf{z}$, and for a batch $\mathcal{B}_\mathcal{Z} = \{\mathbf{z} = g^{-1}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{B}_\mathcal{X} \subset \mathcal{D}\}$ of samples in the base space, a maximum likelihood estimator is given by:

$$\hat{\mu}_{\text{MLE}}(\mathcal{B}_\mathcal{Z}) = -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_\mathcal{Z}|} \sum_{\mathbf{z} \in \mathcal{B}_\mathcal{Z}} z_1^2}, \tag{76}$$

where z_1 is the first component of $\mathbf{z}$. See Appendix 10.4 for a detailed derivation. The two steps optimization is summarised in the following algorithm where α is an adaptation



(a) Gaussians dataset. (b) Moons dataset. (c) Circles dataset.

Figure 15: Training sets for the binary discriminant analysis examples. The color indicates to which of the two classes a sample belongs: blue for C_1 and orange for C_2.

weight for the update of μ:

Choose $\alpha \in [0, 1]$,
Initialise θ_g and μ,
for *all batches* $\mathcal{B}_X$ **do**
  $\mathcal{B}_Z \leftarrow g^{-1}(\mathcal{B}_X)$
  $\theta_g \leftarrow \underset{\theta_g}{\operatorname{argmax}} \log f(\mathcal{B}_X \mid \theta_g, \mu)$
  $\mu \leftarrow \alpha\mu + (1-\alpha)\hat{\mu}_{\text{MLE}}(\mathcal{B}_Z)$
end

For each batch, the parameter θ_g of the mapping is first optimized using automatic differentiation and gradient descent and μ is then optimized using the maximum likelihood estimator.

5.2.4 *Toy examples*

In this section, we provide a few toy examples to illustrate the proposed discriminant analysis. We compare in terms of both discrimination and calibration our approach with both LDA and QDA on three generated datasets. The first dataset, we call *Gaussians*, consists of two multivariate Gaussians with their own mean and covariance matrix. The second one is known as the *Moons* dataset and consists of two interleaving noisy half-circles. The third one is the *Circles* dataset which consists of a large noisy circle containing a smaller one. These datasets are generated with scikit-learn [129]. For each set, 12000 samples are generated, 10000 are used for training and 2000 are used for testing the discriminant analysis. The results are assessed in terms of C_{llr} and $C_{\text{llr}}^{\text{min}}$ (see Section 2.5.1) and scatter plots visualization.

Table 1: C_{llr} measures of the discriminant analysis on the toy examples.

datasets	LDA		QDA		Proposed	
	C_{llr}^{min} [bit]	C_{llr} [bit]	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
Gaussians	0.125	0.198	0.115	0.126	0.117	0.155
Moons	0.387	0.432	0.387	0.432	0.105	0.118
Circles	0.839	1.000	0.023	0.491	0.040	0.054

Gaussians:

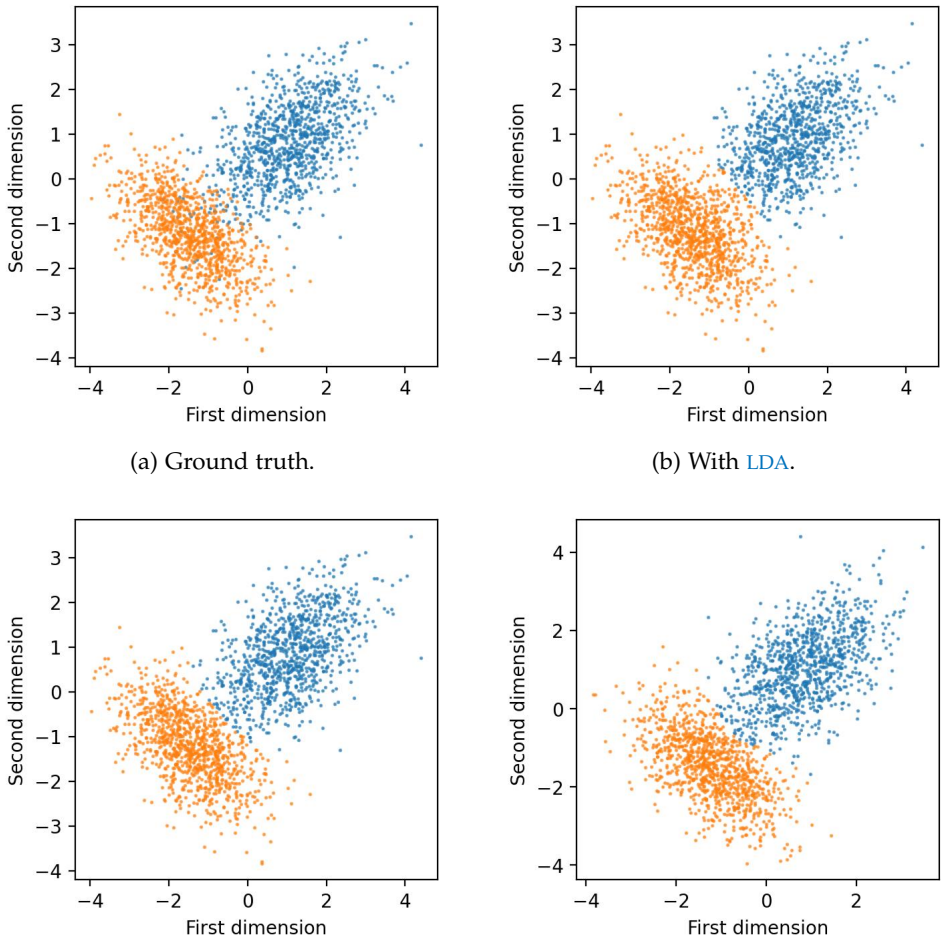
Figure 15a shows the training set for the Gaussians example. Figure 16 shows the results of the maximum likelihood classification using LDA, QDA, and the proposed discriminant analysis. C_{llr} measures are reported in Table 1. Both LDA and QDA are based on the Gaussian assumption. However, the LDA assumes that both classes share the same covariance which is not the case here. The LDA has therefore a discrimination and a calibration that are not as good as the QDA. The C_{llr}^{min} is 0.125 bit for the LDA while it is 0.115 bit for the QDA and the calibration cost $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$ is 0.073 bit for the LDA and 0.011 for the QDA. The QDA is even better than the proposed approach which has a C_{llr}^{min} of 0.117 bit and a C_{llr}^{cal} of 0.038 bit. The goodness of the QDA is here not surprising since it assumes the data to be normally distributed with different covariance matrices for each class which is actually how the data is distributed.

With interleaving moons.

Figure 15b shows the training set for the Moons example. Figure 17 shows the results of the maximum likelihood classification using LDA, QDA, and the proposed discriminant analysis. C_{llr} measures are reported in Table 1. Both LDA and QDA hardly separate the two classes while the proposed discriminant analysis does better in terms of both discrimination and calibration with $C_{llr} = 0.105$ bit and $C_{llr}^{cal} = C_{llr} - C_{llr}^{min} = 0.013$ bit.

With circles.

Figure 15c shows the training set for the Circles example. Figure 18 shows the results of the maximum likelihood classification using LDA, QDA, and the proposed discriminant analysis. C_{llr} measures are reported in Table 1. Being linear, the LDA cannot separate the two classes. The QDA has the best discrimination with a C_{llr}^{min} of 0.023. The proposed approach has a slightly bigger C_{llr}^{min} of 0.040. We can indeed see in Figure 18 a tiny slice of blue samples that are miss-classified as orange on the left-bottom part of the larger circle (this is more discernible on the training set we show in Figure 19). This is due to the fact that the family of mappings is restricted to a family of diffeomorphisms where none allows a "perfect" transformation of these interleaving circles into two distinct Gaussians. However, even if QDA has the best discrimination ability—thanks to the quadratic nature of the circle-shape boundary—it is still based on Gaussian assumptions while the data are definitely not normally distributed. This results in a calibration that is not as good as the



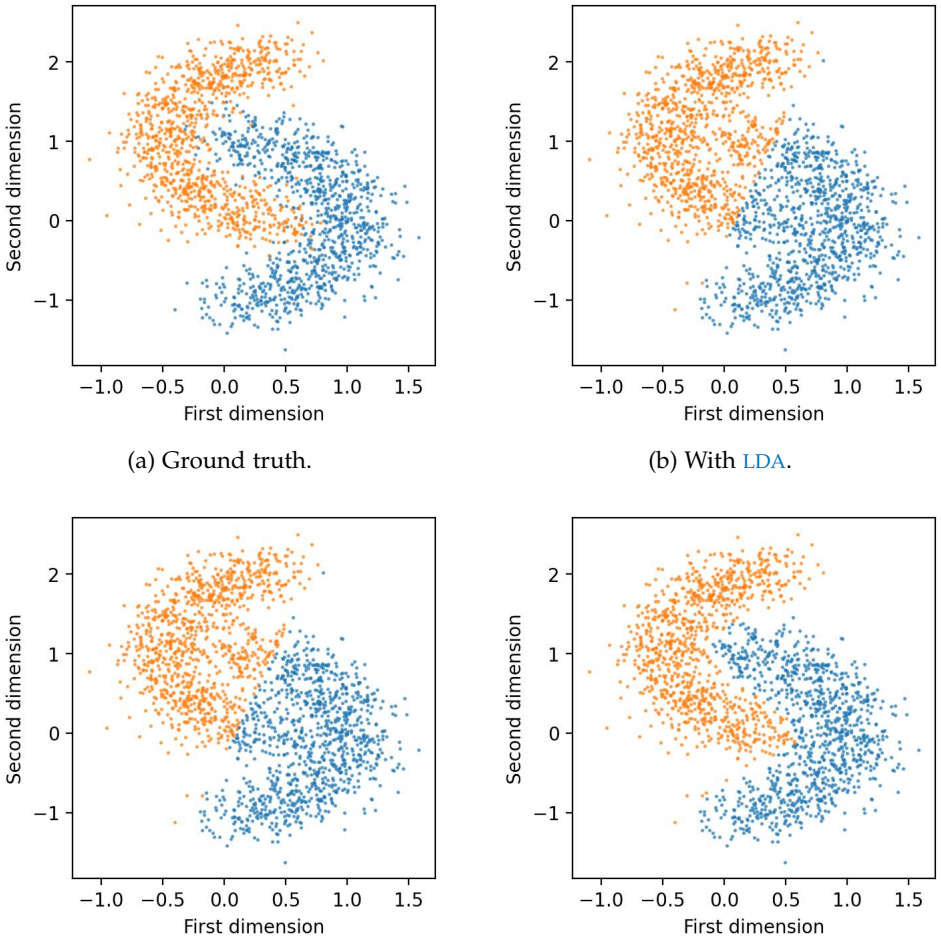
(a) Ground truth.

(b) With LDA.

(c) With QDA.

(d) With the proposed discriminant analysis.

Figure 16: Maximum likelihood classification on the Gaussians dataset. In (a), the colors indicate the true label while for the other figures, the colors indicate the predicted class.



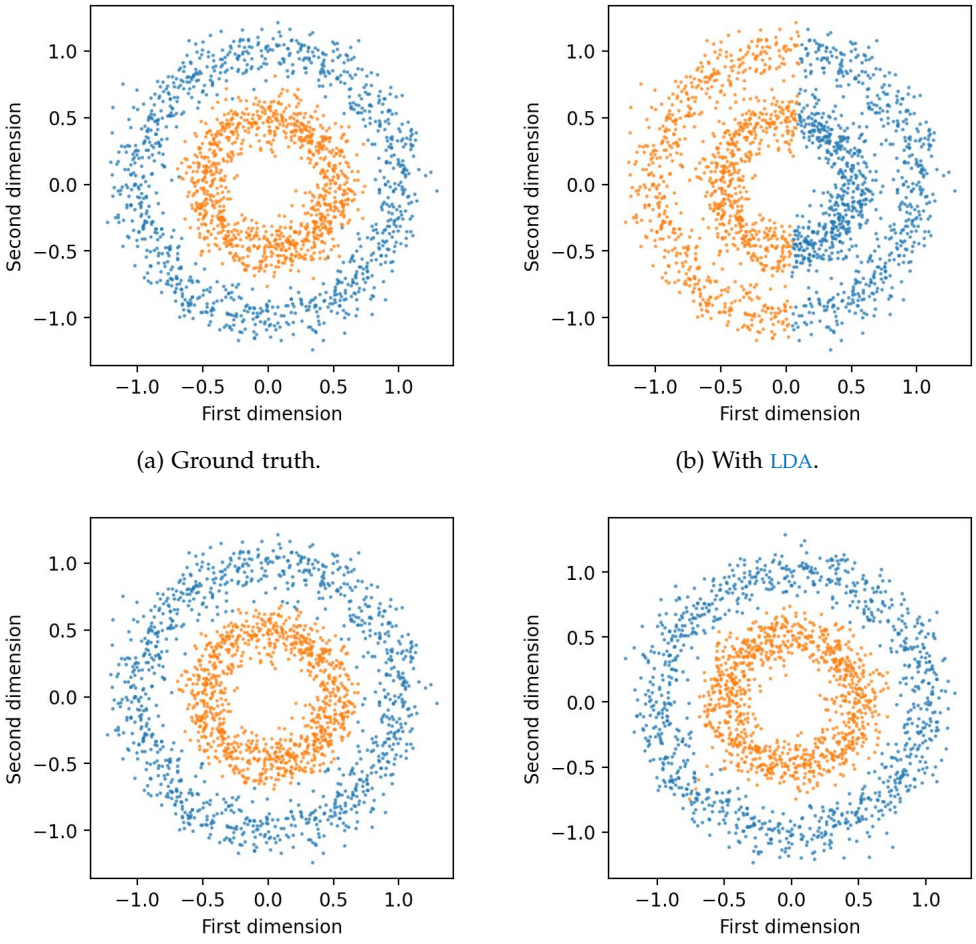
(a) Ground truth.

(b) With LDA.

(c) With QDA.

(d) With proposed.

Figure 17: Maximum likelihood classification on the Moons dataset. In (a), the colors indicate the true label while for the other figures, the colors indicate the predicted class.



(a) Ground truth.

(b) With LDA.

(c) With QDA.

(d) With proposed.

Figure 18: Maximum likelihood classification on the Circles dataset. In (a), the colors indicate the true label while for the other figures, the colors indicate the predicted class.

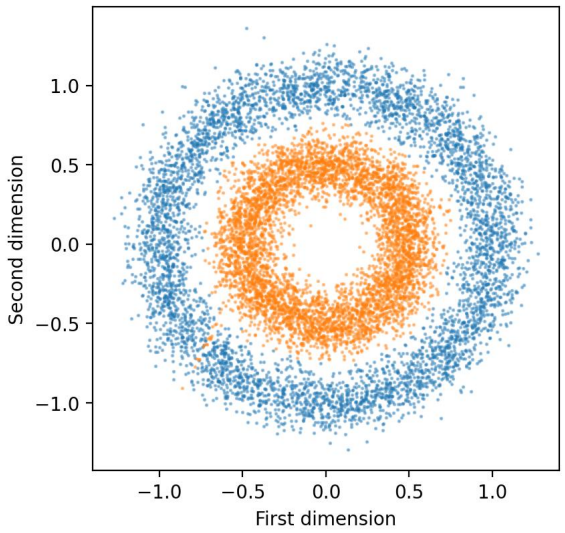
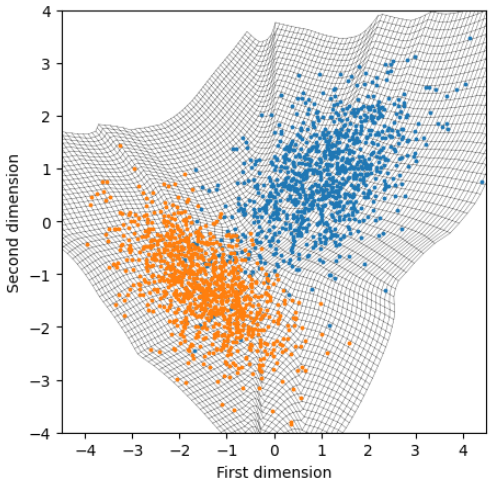


Figure 19: Proposed discriminant analysis on Circles training set (ground truth is given in Figure 15c). Since it is restricted to invertible and differentiable transformations, this discriminant analysis will never "perfectly" separate the two classes as the tiny slice of miss-classification illustrates.

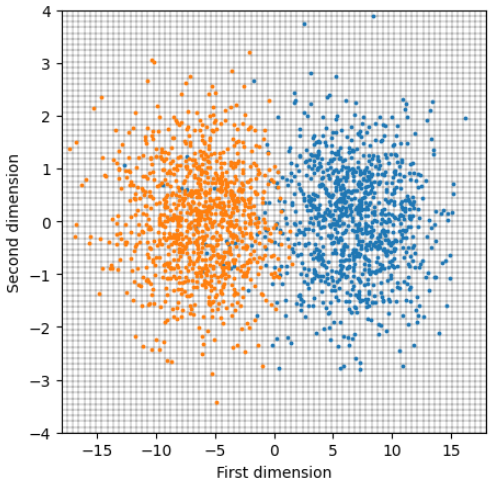
calibration of the proposed discriminant analysis. The QDA has indeed a C_{llr}^{cal} of 0.468 while the proposed discriminant analysis has a C_{llr}^{cal} of 0.037 on the testing set. This suggests that in this example—while the Gaussian assumption of the QDA leads to the worst calibration of the three methods—if the implicit assumption discussed in Section 5.2.2 is not fulfilled, the invertibility and differentiability constraints of the proposed approach limit the discrimination ability in the aid of a better calibration.

Grid visualisation of the learned transformation.

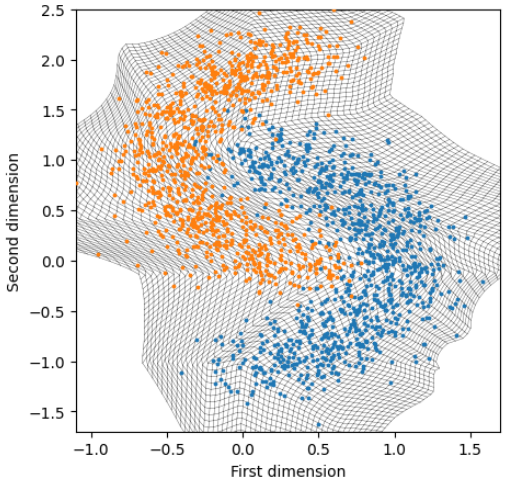
Since the above examples are two-dimensional, a learned diffeomorphism can be visualized as a transformation of a two-dimensional grid as shown in Figure 20. The right part of Figure 20 shows the testing data in the learned base space for the three examples. As expected, for each class, the data looks normally distributed and symmetric around the zero of the first dimension i. e. the zero-LLR line. For visualizing the learned diffeomorphism, a set of points is generated—in the form of a grid—homogeneously and regularly over the base space. The set of points are transformed through the learned diffeomorphism and the resulting deformed grid is visualized in the feature space on the left part of the figure. In the feature space, the transformed regular grid approximates the manifold on which the data lives. The testing sets are shown over the grids. The colors represent the true label of the samples. One can see on the bottom left of Figure 20e the slice of miss-classification of the blue circle. The orange circle's samples are "going through" the blue circle at the expense of the few blue samples that will be miss-classified and that can be seen in Figure 18d for the testing set and in Figure 19 for the training set. This confirms what we discussed in the above paragraph. On the circle example, the proposed approach favors the calibration quality at the expense of the discrimination quality.



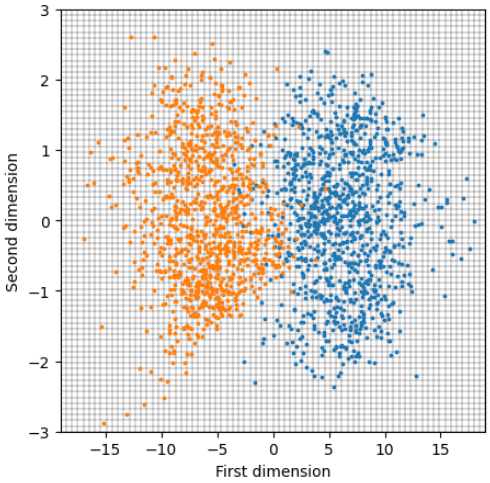
(a) Transformed grid in the feature space for the Gaussians examples.



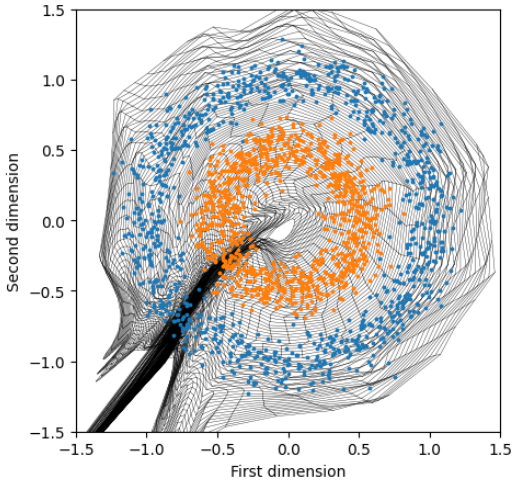
(b) Regular grid in the base space for the Gaussians example.



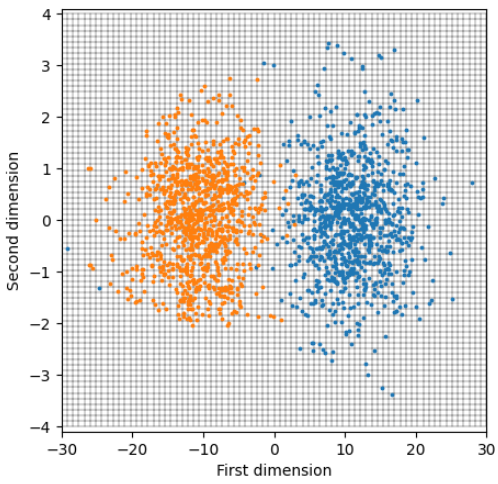
(c) Transformed grid in the feature space for the Moons example.



(d) Regular grid in the base space for the Moons examples.



(e) Transformed grid in the feature space for the Circles example.



(f) Regular grid in the base space for the Circles example.

Figure 20: Grid vizualisation of the learned diffeomorphisms on the toy examples. The samples from the testing sets are shown over the grid.

5.3 APPLICATION TO PRIVACY: ZERO-LLR SEX-RECOGNITION SPEAKER EMBEDDING

The approach presented above can be used for discriminant analysis. However, in this thesis, we are particularly interested in privacy preservation. We will see in this section how the discriminant analysis can be used for the concealment of a binary attribute in some data.

When the variable to hide is binary, the data can be mapped, using the proposed discriminant analysis, into a base space where the first dimension represents the LLR. In Section 2.7 and Chapter 4, we saw that perfect privacy is reached when the LLR is equal to zero for all samples. For privacy, the LLR is therefore set to zero in the base space before mapping the data back into the original feature space. Indeed, setting the LLR to zero in the base space $\mathcal{Z}$ is equivalent to setting the LLR to zero in $\mathcal{X}$ annihilating therefore the strength of the evidence in the observed data:

$$\frac{f_{\mathcal{Z}|C_1}(\boldsymbol{z})}{f_{\mathcal{Z}|C_2}(\boldsymbol{z})} = \frac{f_{\mathcal{X}|C_1}(\boldsymbol{x})\left|\det\left(\frac{\partial g(\boldsymbol{z})}{\partial \boldsymbol{z}}\right)\right|}{f_{\mathcal{X}|C_2}(\boldsymbol{x})\left|\det\left(\frac{\partial g(\boldsymbol{z})}{\partial \boldsymbol{z}}\right)\right|} = \frac{f_{\mathcal{X}|C_1}(\boldsymbol{x})}{f_{\mathcal{X}|C_2}(\boldsymbol{x})}. \tag{77}$$

In Expression 77, the change of variable formula is used and the Jacobian determinants cancel out such that the LR is the same for $\boldsymbol{x}$ and $\boldsymbol{z}$. This holds because the mapping is a diffeomorphism which is also required for mapping the data back into the original feature space $\mathcal{X}$.

The protection strategy is summarized as follows:

- Map each sample into the base space,
- For each sample, set the first dimension of the base space (i. e. the LLR) to zero,
- Map each sample back into the feature space using the inverse transformation.

In [123], we tested this approach for the concealment of the speaker's sex in x-vector speaker embeddings [149]. These are neural network embeddings obtained from a network trained using a speaker classification criterion. They are commonly used for speaker verification where two embeddings are compared using a PLDA [76]. Recently, different architectures have been explored for the embedding extraction and the PLDA is sometimes replaced by a similarity measure like the cosine similarity [47].

In [120], we proposed an approach based on adversarial disentanglement [88] for hiding the sex of the speaker in x-vectors. The vectors are fed into an autoencoder and a classifier tries to predict the sex of the speaker from the encoded representation. With an adversarial training where the encoder tries to fool the classifier while the latter tries to predict the correct sex, the discrimination between the sex can be reduced in the encoded representation. However, this approach has not been designed to represent the sex-related information into a LLR and is therefore not suitable for the zero-evidence formulation of privacy.

In [123], we used instead the new discrimination analysis-based protection presented in this chapter. We report the results of our experiments on kaldi's x-vector [149] on VoxCeleb

dataset [33, 110] (see [123] for more details). V2D and V2T respectively denote a subset of VoxCeleb2 development part and a subset of VoxCeleb2 test part. They are used to respectively train and test our protection approach. V2T is split into two parts V2T-train and V2T-test respectively used for the training and the testing of a sex classifier. The latter can be seen as the attacker's classifier. Therefore, for good protection, we want this classifier to perform poorly on protected data. The classifier is here trained on protected data which corresponds to an *informed* attacker which is strong in comparison to an *ignorant* attacker that trains its system on original non-protected data [150][4].

Our approach is compared with an LDA-based protection that can be expressed with the following whitening:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \frac{\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}}{\|\boldsymbol{v}\|^2}\boldsymbol{x} + \frac{1}{2\|\boldsymbol{v}\|^2}\left(\boldsymbol{\mu}_{\mathrm{F}}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{F}} - \boldsymbol{\mu}_{\mathrm{M}}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{M}}\right)\boldsymbol{v}, \tag{78}$$

where for each class the data is assumed to be normally distributed in $\mathcal{X}$ with means $\boldsymbol{\mu}_{\mathrm{F}}$ and $\boldsymbol{\mu}_{\mathrm{M}}$ (where F stands for Female and M stands for Male) and shared covariance $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}_{\mathrm{B}}$ is the between-class covariance matrix and $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_{\mathrm{F}} - \boldsymbol{\mu}_{\mathrm{M}}\right)$ is the non-zero eigenvalue eigenvector of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{B}}$. This whitening is equivalent to setting the LDA's LLR of Equation 69 to zero. We also compared the proposed protection system with the one based on adversarial disentanglement autoencoding (adv-AE) [120].

Remark:

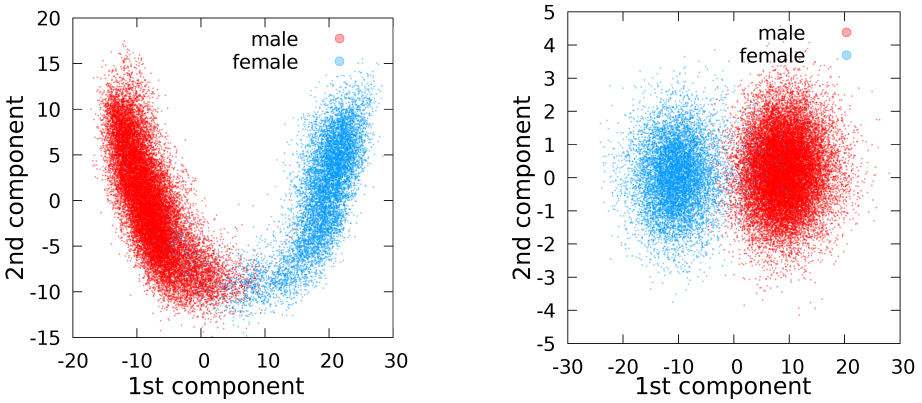
Randomly assigning a target sex to each speaker or x-vector could be a solution to conceal the true sex of the speaker. It would actually lead to better protection results when the performance is measured in terms of automatic detection of sex. However, we see several limitations to this approach:

- First, the assigned sex will sometimes be the actual true sex. Those samples will therefore not be protected;
- Secondly, the resulting voice would still be sexed, containing some evidence in favor of the sex male or female no matter the true sex of the speaker. However, as discussed in Chapter 4, our concept of privacy is not to fool the attacker but rather to not provide any evidence about the speaker's sex.

Detailed architecture of the proposed protection system:

The proposed protection system is based on the discriminant analysis proposed in Section 5.2. In our experiment, the invertible mapping is implemented through a NF architecture based on Real NVP [48]. It consists of six stacked coupling layers where the scale and translation functions are multilayer perceptrons. The scale function is made of three linear layers with two LeakyReLU activation functions [94] and an output hyperbolic tangent ac-

4 In our work on adversarial disentanglement [120], only ignorant attacks were considered. The protection ability findings of the system were therefore overestimated.



(a) First two principal components of the x-vectors

(b) First two components of the x-vectors in the base space

Figure 21: V2T's x-vectors visualization in the original feature space and in the base space. In the base space, only the first component is discriminant and represents the LLR. Its distribution tends to respect Proposition 1 imposed by the idempotence property.

tivation. The translation function is made of three linear layers with also two LeakyReLU activation functions but no output activation. Adam algorithm [84] with a 10^{-4} learning rate is used for optimization. At the time we did this experiment, the parameter μ was manually initialized, but it can be automatically initialized at the Kullback-Leibler divergence value computed in the $\mathcal{X}$ space assuming Gaussian class-conditional densities with shared covariance. In this way, we expect the initial μ value to be not too eccentric. The initialization of the base densities parameter will be discussed in more detail in Chapter 7 when we will extend the proposed discriminant analysis to more classes using the ILRL. We here report the results in [123] where μ were initialized manually at 10 and converged around 9 with $\alpha = 0.99$ (see Section 5.2.3).

5.3.1 *Protection and utility assessment of the zero-LLR speaker embeddings*

In this section, we test the protection ability of the system by checking whether a sex recognizer is able to perform on the protected embeddings. We also report ASV results to see if these embeddings can still be used for authentication purposes.

Figure 21 shows the x-vectors of the test set V2T in both the $\mathcal{X}$ space and the base space $\mathcal{Z}$. In 21b, we can see that in $\mathcal{Z}$, the samples are normally distributed as designed in Section 5.2.1 and respect the idempotence constraint of Proposition 1. Only the first component is discriminant and represents the LLR. The C_{llr}^{min} for this set of LLR is equal to 0.067 and the C_{llr}^{cal} is 0.008. The low value of the calibration cost shows that these LLRs are well-calibrated and therefore properly embed the information about the sex of the speaker.

For privacy, the LLR dimension must be set to zero. Then, the data is protected and is mapped back to the $\mathcal{X}$ space by using the inverse transformation.

In order to assess the protection ability of our system we train four two-layers perceptron sex classifiers. One on each version of V2T-train:

- without any protection (original data),
- protected with the LDA following the Equation 78,
- protected with the adversarial approach (adv-AE) [120],
- protected with the proposed strategy (zero-Log-Likelihood-Ratio (zLLR)).

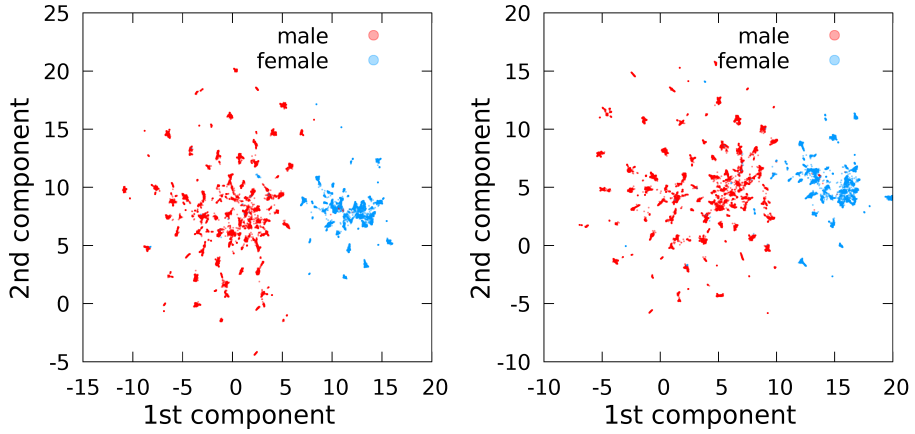
Each trained classifier is tested on the corresponding V2T-test. C_{llr}^{min} and D_{ECE} measures are given in the left part of Table 2 denoted by *Sex classification performance*. For a good protection, we want the C_{llr}^{min} to be close to one and the D_{ECE} to be close to zero. The classifiers hardly generalize to the test set when the data are protected, especially with the proposed approach (zLLR) where the C_{llr}^{min} is $95.75 10^{-2}$ and the D_{ECE} is 0.029. This illustrates the difficulty of an attacker to use such a classifier to detect the sex of the speaker from the x-vectors when the proposed approach is used[5].

We also report unsupervised visualization of the protected x-vectors. Figure 22 shows Uniform Manifold Approximation and Projection (UMAP) visualization [101] of the x-vectors. We can see from 22b and 22c that even with the protection based on the LDA and on the adversarial disentanglement, UMAP still allows distinguishing two classes while this is not the case with the proposed approach (22d).

The proposed approach reduces the amount of information related to the sex of the speaker in the x-vectors. We want now to see if the protected x-vectors can still be used for speaker verification. If yes, this would argue that some of the speaker variability not related to the sex has been preserved and that authentication by voice is still possible while allowing the user to not provide information about their sex. We use a PLDA trained on non-protected x-vectors for comparing the protected enrolment and test x-vectors. Indeed, we suppose that the authentication side is using a general ASV backend since it does not know that the user is protecting its data[6]. The EER and C_{llr}^{min} are reported in the right part of Table 2—denoted by *ASV performance*—for the original data, protected with the LDA, protected with adv-AE, and with the proposed approach (zLLR). Be aware that these results are for the binary classification *target/non-target* rather than *female/male* as in left part of the table. Therefore, we want a low EER and a low C_{llr}^{min} for good ASV performance. In addition to a good protection ability, zLLR results in better speaker verification in comparison to the adv-AE approach but the EER increases from 1.72% to 2.11% in comparison to the original

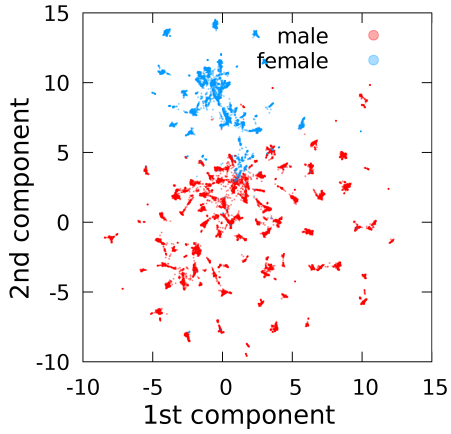
5 In our original paper [123], we also reported an estimated mutual information (MI) measures between the x-vector's dimensions and the sex variable. However, we do not report these results here and ask the reader who will refer to the original paper to interpret these results with caution. Indeed, we reported MI values averaged over the x-vector's dimensions but this is relevant only if the dimensions are independent, which is not guaranteed. Even if the target distribution is Gaussian with uncorrelated dimensions, this target is not necessarily reached to perfection during the training procedure.

6 In this example, only the x-vectors are given to the authentication side and not the speech utterances. This may not be practical in a real-life application as this requires the x-vectors to be extracted locally (on the user's device). However, we want to provide a toy example as a modest proof of concept rather than providing a direct solution usable in a real-life application.

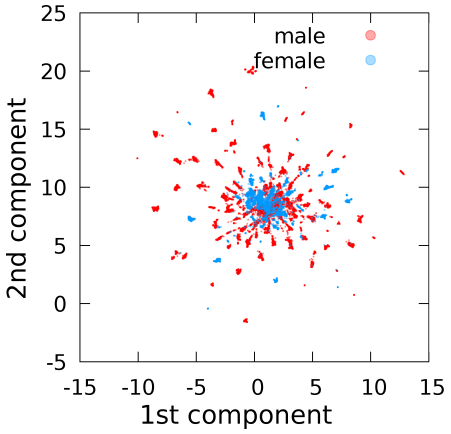


(a) Original data

(b) Data protected with the LDA



(c) Data protected with adv-AE



(d) Data protected with the proposed approach (zLLR)

Figure 22: UMAP vizualisation (with euclidean metric and 30 neighbors) [101] of the original and protected V2T's x-vectors. Even with the LDA protection and the adv-AE protection, unsupervised visualization allows visualizing two separated classes while with the proposed approach, male and female x-vectors are mixed in the visualization.

Table 2: Sex classification and ASV performance on non-protected and protected speaker embeddings. The first line is for non-protected data, the second is with the LDA-based protection, the third line is for the adv-AE protection and the last line is with the proposed system. The left part of the table reports the sex classifier performance on V2T. For good privacy, C_{llr}^{min} must be close to 1 and D_{ECE} must be as low as possible on V2T-test. The right part of the table reports the ASV performance. For good ASV ability, the EER and the C_{llr}^{min} must be as low as possible.

System	Sex classification performance				ASV performance	
	V2T-train		V2T-test			
	$C_{llr}^{min}\,10^{-2}$ [bit]	D_{ECE} [bit]	$C_{llr}^{min}\,10^{-2}$	D_{ECE}	EER [%]	C_{llr}^{min}
original data	8.39	0.658	2.12	0.703	1.72	0.067
LDA	12.20	0.628	57.75	0.295	1.57	0.065
adv-AE	30.43	0.493	74.21	0.179	2.36	0.097
zLLR	48.45	0.362	95.75	0.029	2.11	0.086

non-protected speech. However, this can be considered a low price to pay for privacy. In this experiment, the LDA-based protection results in even better ASV performance compared to the original data but fails in protecting the data.

5.3.2 *Summary*

In this section, the proposed non-linear discriminant analysis has been tested for the concealment of the sex-related information in speaker embeddings. For protectection, the data is mapped into the base space where the first dimension forms the LLR that represents the sex-related information contained in the data. The LLR can be set to zero and the resulting protected data mapped back into the feature space. This approach is theoretically consistent with the zero-evidence formulation of privacy of Chapter 4. The experiments conducted on speaker embeddings from the VoxCeleb datasets showed the ability of the approach to remove the sex information while preserving remaining speaker-related variability in the data.

5.4 VOICE CONVERSION WITH ZERO-LLR SEX-RECOGNITION SPEAKER EMBEDDING

In the previous section, we discussed how the discriminant analysis approach proposed in Section 5.2 can be used for the concealment of the sex of the speaker in speaker embeddings. These embeddings are usually used for ASV, however, their manipulation can also be explored for voice conversion. Voice Conversion (VC) [147] is the process of transforming a speech signal to make it sounds like another speaker by changing the perceived voice while keeping linguistic content unchanged. This can be used for privacy purposes [150, 155] by avoiding the listener to perceive the original voice and therefore hiding the identity of the speaker (assuming that the prosody and the linguistic content do not contain information about the speaker's identity).

Most of the research in voice conversion for privacy focuses on hiding the full identity of the speaker while there might be situations where *attribute privacy* is preferable. Like in the previous section, we focus here on the sex as an attribute to hide and want to transform a voice into a voice that provides no evidence about the sex of the speaker.

Recently, methods have been proposed for hiding the sex of the speaker with voice conversion. In [140], in order to avoid sex-related bias in speech model training, the authors proposed making speech "sex-neutral" beforehand by automatically searching for pitch and formants shifting that would lead to the maximum uncertainty sex classifier score, i. e. 50%. However, in their paper, there is no guarantee that the classifier they used is well-calibrated, and a search for shifting parameters must be done for each utterance. Here, we prefer to have a single transformation that can be applied regardless of the input utterance, which appears to us more suitable for a real-life application of privacy systems. In [151], the authors proposed removing the speaker's sex using an adversarial approach. Their approach also aims to protect the speaker's identity instead of leaving the speaker's other information unchanged.

Here, we report our work presented in [119] that aims in altering only the speaker's sex while preserving the other speaker-related variabilities. We consider the explicit "disentanglement" of the sex information in the speech as a desirable step. The proposed discriminant analysis can be used but it has been designed for vector inputs only and extending this approach to waveforms is challenging. Hence, we instead include it in an analysis/synthesis framework. We use the HiFiGAN[7] vocoder [85] fed by the f_0 trajectory, a HuBERT soft content representation [116], and an ECAPA-TDNN [47] speaker embedding. Once the analysis/synthesis pipeline is trained, we apply the protection proposed in Section 5.3 to the speaker embedding and an affine transformation to the f_0 to remove sex-related information. In our experiments, we test the protection ability of our approach in terms of automatic sex recognition performance with both *ignorant* and *informed* attacks [150], respectively regarded as *weak* and *strong* attacks. We evaluate the performance of Automatic Speech Recognition (ASR) and Automatic Speaker Verification (ASV) as downstream tasks. Listening tests are also done in order to assess the human ear perception of protected speech.

5.4.1 *Detailed architecture of the voice conversion-based protection*

Analysis/synthesis is the process of extracting features from which the original speech signal can be recovered using a vocoder. In speech technology, this approach has been widely used. The intermediate characterization of speech can be used for speech transmission, voice conversion, speech anonymization, etc. In speech anonymization, we want a part of the intermediate features to represent speaker-related information that can be manipulated for privacy. In [105], the authors proposed to update the first VoicePrivacy's

7 Recently, some speech synthesis approaches have been inspired by the adversarial learning procedure of the Generative Adversarial Network (GAN) [70]. However, these systems, like the HiFiGAN we are using and contrary to the GAN, are not generative models in the statistical sense even if they are named as such. They transform input features into an output waveform in a deterministic manner.

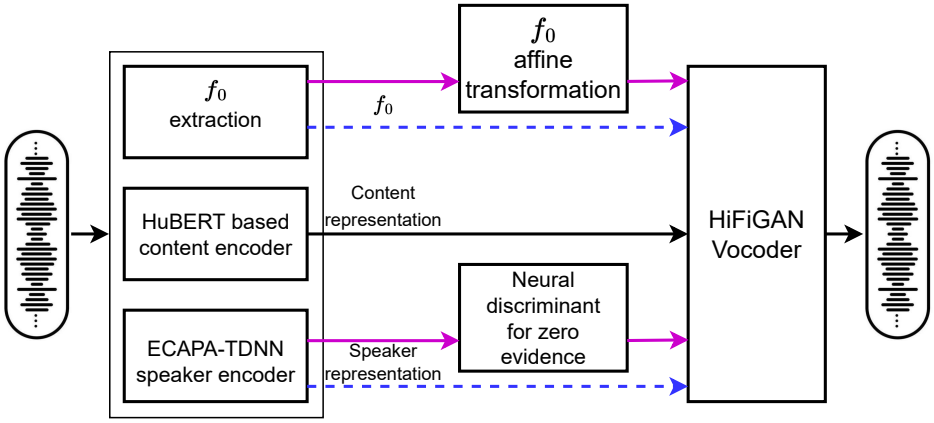


Figure 23: Architecture of the voice conversion system for speaker's sex protection. The blue-dashed path is used during training and the purple path during protection. Neural discriminant for zero evidence refers to the zLLR approach of Section 5.3.

baseline [155] by replacing the neural source-filter vocoder [158] with a HiFiGAN vocoder [85]. They also replaced the x-vector Kaldi TDNN speaker embedding [149] with a x-vector ECAPA-TDNN speaker embedding [47]. The authors also got rid of the acoustic model by using instead a HuBERT-based soft content representation [116]. In [105], the authors used this system for the VoicePrivacy task and studied its application to unseen languages. Here, we use this system but replace the speaker embedding averaging used for voice anonymization with the discriminant analysis-based protection of Section 5.3. In addition, since the fundamental frequency (f_0) is known to contain information about the sex, we apply an affine transformation to the f_0 in order to force the vocoder to output a fixed target f_0 trajectory's mean and standard deviation that we expect to be sex-neutral. The target f_0 trajectory's mean and standard deviation are computed from a training set where the means and standard deviations from f_0 trajectories are first averaged at the speaker level and are then averaged over all males and all females resulting in two means and two standard deviations f_0 (one for each sex). Then, the target mean for f_0 is obtained by taking the average between the male and the female mean for f_0, and the target standard deviation is obtained by taking the average between the male and the female standard deviation. These careful ways of averaging are done in order to avoid bias due to an unbalanced number of utterances per speaker and speakers per sex in the training set.

Figure 23 shows an outline of our system. The blue-dashed arrows show the training path of the HiFiGAN. The feature extractors are pretrained and fixed. Once the vocoder has been trained, the purple path is used. Both the f_0 and the speaker representation are transformed to reduce the sex-related information they contain. The content representation is assumed to not contain sex-related information. However, in real applications, the speaker might explicitly reveal their sex but we do not consider this scenario and instead focus on hiding the sex information in the acoustic features.

5.4.2 *Training, testing sets, and baselines*

The sets used in our experiment for training and testing the proposed system are presented above.

- **Vocoder's training:**

  LibriTTS train-clean-100 [163] was used to train the HiFiGAN as in [105]. The feature extraction modules were pretrained and fixed. ECAPA-TDNN [47] with 80-coefficient FBank features [105] was used for the 192-dimensional speaker representation extraction and was trained on VoxCeleb2 development set [33]. The 200-dimensional content representation was extracted by a HuBERT soft content encoder [116] fine-tuned from a pretrained HuBERT base model[8] on LibriTTS train-clean-100. Its training procedure is detailed in [105]. We used YAAPT [81] for the extraction of f_0, which does not require any training.

- **Training of the protection modules:**

  Once the analysis/synthesis system is trained, it can be used for voice conversion and privacy by manipulating the speaker representation and the f_0 trajectory. In our experiment, the protection of the speaker representation is done using the discriminant analysis for zero-evidence sex recognition presented in Section 5.3. The discriminant analysis model was trained on ECAPA-TDNN speaker embeddings from LibriTTS train-other-500. The target f_0 mean and standard deviation were computed, with the averaging strategy described in the previous paragraph, from LibriTTS train-other-500 also.

- **Testing sets:**

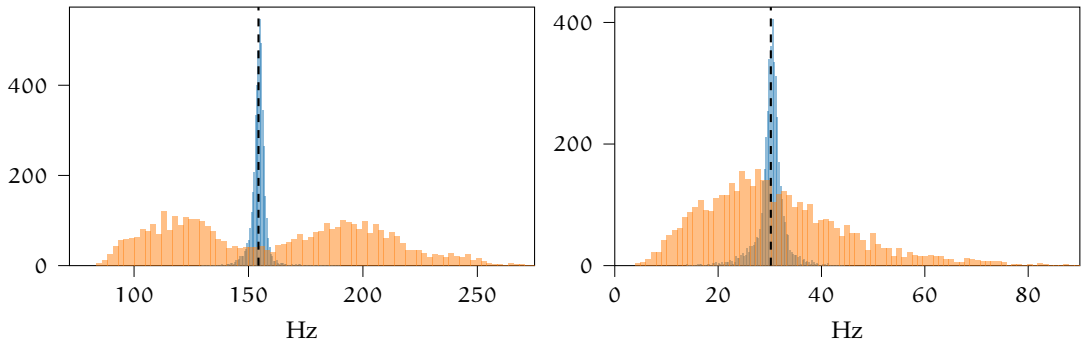
  The VoicePrivacy challenge provides a complete evaluation protocol. For conciseness, we merged its libri_dev and libri_test sets—subsets of LibriSpeech [126]— for the evaluation of our system, resulting in 35 females with a total of 1185 utterances and 34 males with a total of 1136 utterances.

We compared the proposed approach with two baselines:

- The first one, we call *global*, is the same as our proposed approach but instead of using the discriminant analysis-based protection of the speaker representation, we simply fed the HiFiGAN with one global averaged x-vector for all utterances. The averaging was done in such a way as to avoid bias as it was done for the computation of the target f_0 trajectory moments. In this case, we expect that the sex of the original speaker will be hidden but that all the remaining speaker variability will be altered such that the resulting voices all look the same.

- The second baseline simply transforms the f_0 using Time Domain Pitch Synchronous Overlap Add (TDPSOLA) [108]. Since we know that sex information is also contained

8 https://github.com/pytorch/fairseq/tree/main/examples/hubert



(a) Mean f_0 histograms.

(b) Standard deviation f_0 histograms.

Figure 24: Mean f_0 and standard deviation f_0 histograms for original speech (orange) and for generated protected speech (blue) with target f_0 mean ≈ 154 and standard deviation ≈ 30 (dashed lines).

in the spectral envelope, we expect that the sex of the speaker will not be satisfactorily hidden.

Is the HifiGAN vocoder following the target f_0 moments?

Before going further, we want to make sure that the HifiGAN vocoder is following the target f_0 mean and standard deviation. Figure 24 shows histograms of generated f_0 trajectory mean and standard deviation when the protection proposed in Section 5.4.1 is applied. The target moments are shown by the dashed vertical lines: ≈ 154 for the mean and ≈ 30 for the standard deviation. The histograms of the mean and standard deviation of the generated f_0 are in blue. We can see that they are narrow around the target values shown by the dashed lines which suggest that the moments of the outputted f_0 trajectory follow the target ones.

5.4.3 *Protection results with automatic sex classification*

We report the results for:

- *original* speech i. e. natural non processed speech,
- *synthesised* speech i. e. speech utterances that have been processed by the analysis/synthesis system of Section 5.4.1 but the protection of the x-vector and f_0 have not been applied,
- protected speech with the proposed approach of Section 5.4.1 (zLLR-VC, for zero-LLR-voice-conversion),
- protected speech with the *global* baseline,
- protected speech with TDPSOLA-based baseline.

In order to assess the protection ability of the systems, we report how much an automatic sex classifier is able to detect the sex of the speaker from the speech. This paragraph is

concerned with automatic attacks only. An attacker may try to infer the sex of the speaker by listening manually to the data. This point will be discussed later when reporting the listening test results. We study two kinds of attack:

- *Ignorant attack* where the classifier is trained on original speech data. In this scenario, the attacker does not have access to the protection system or may not be aware that the data has been protected. Therefore, he or she uses a sex classifier trained on natural non-protected speech;
- *Informed attack* where the classifier is trained on protected speech data. This is the strongest attack we consider. In this scenario, the attacker has access to the protection system. He or she can apply the protection to the data used for training of the automatic sex classifier. The training of the classifier can therefore benefit from the sex-related information that could remain in the protected data.

The classifier we used in our experiment is based on fine-tuned HuBERT base features extraction (with frozen convolution), statistical pooling, and multilayer perceptron. Ideally, an objective measure of information leakage should not depend on any classifier architecture. We however consider, as an approximation, that this classifier is flexible enough for the attacker to extract the information about the speaker's sex that potentially remains in the speech. Table 3 reports the results in terms of two metrics: the EER and the D_{ECE}. For privacy, we want a low D_{ECE} and an EER close to 50%. The first line shows the initial ability to distinguish the sex of the speakers. We can see from the second line that this is slightly altered when the speech has been processed even without protection applied. The next two lines show how the classification performance drops when the *global* or the zLLR-VC protection is applied. For the ignorant attack, we have a drop in D_{ECE} of 78% with the zLLR-VC approach and 66% with *global*. The methods have also a positive effect against the informed attack with a drop in D_{ECE} of 65% and 60%. The TDPSOLA baseline is not competitive, which is not a surprise since it alters only the f_0 while it is known that differences in vocal tract shape between males and females are significantly related to the spectral envelope. However, we do not have a clear explanation as to why the zLLR-VC approach protects better than the *global* baseline does. This could be due to uncontrolled bias in the data or to the fact that the *global* x-vector—which is a linear interpolation—may go out of the manifold on which the x-vectors are living, however, this requires further study.

5.4.4 *Automatic speech recognition and speaker verification results*

We want now to make sure that Automatic Speech Recognition (ASR) and Automatic Speaker Verification (ASV) can still be performed as downstream tasks.

ASR aims in recognizing what has been said in a speech utterance. It transcripts a speech utterance into a text sentence. The produced text can be compared with the true transcription by computing a Word Error Rate (WER). This can be used to assess how much a voice conversion system is altering the speech content.

Table 3: Sex classification results for protection assessment of the voice conversion systems.

	ignorant		informed	
system	EER [%]	D_{ECE} [bit]	EER	D_{ECE}
original	3.67	0.578		
synthesised	4.32	0.542	4.01	0.593
global	24.95	0.198	20.60	0.233
TDPSOLA	6.30	0.504	4.36	0.542
zLLR-VC	28.99	0.128	24.13	0.200

Table 4: Automatic speech recognition results in term of WER.

	WER [%]
original	4.02
synthesised	4.79
global	4.92
TDPSOLA	4.43
zLLR-VC	4.81

We use here the same ASR evaluation as in the VoicePrivacy challenge [155]. The WER is reported in Table 4. Processing the speech slightly increases the WER, but among the two systems that provide good protection, the proposed approach seems to alter the ASR performance less than the *global* approach (at least on this dataset). At worst, 0.9% is added to the WER which is a relatively low price to pay for privacy.

We also report ASV results. We used the same ASV system used for evaluation in the VoicePrivacy challenge. It consists of a Kaldi TDNN speaker embedding extractor [149] with a PLDA backend. Both enrolment and test utterances were processed by the system. The EER and C_{llr}^{min} are reported in Table 5. We can see that processing the data without applying protection already slightly reduces the ASV performance. This suggests that the HiFiGAN vocoder results in a small distortion or domain shift for the ASV backend. Applying the protection further reduces the ASV performance. For *global*, all the speaker variability in the x-vector is annihilated by the global averaging, increasing therefore the confusion between voices. With the zLLR-VC approach, the x-vector is transformed in order to alter only the speaker's sex. In this way, other speaker variabilities must be preserved and, as expected, the resulting protected voices remain consistent to some extent. Indeed, the proposed approach does far better than the global one although—compared with original data—significant ASV ability is lost with an increase in C_{llr}^{min} from 0.278 to 0.445 and from 0.040 to 0.345 for female and male respectively. In addition to the domain shift induced by the HiFiGAN synthesis, this drop in performance could be explained by both the reduction in sex information as a component of the speaker variability that significantly helps in distinguishing speakers, not only a male speaker from a female one but also between speakers

Table 5: Automatic speaker verification results. F and M refer respectively to in-between female and in-between male trials, while FM refers to cross-sex trials.

system	EER [%]			C_{llr}^{min} [bit]		
	F	M	FM	F	M	FM
original	8.15	1.13	5.77	0.278	0.040	0.204
synthesised	9.42	7.18	6.86	0.325	0.245	0.240
global	39.88	39.81	35.86	0.931	0.943	0.903
TDPSOLA	9.36	1.26	6.38	0.332	0.046	0.237
zLLR-VC	13.22	9.85	11.55	0.445	0.345	0.407

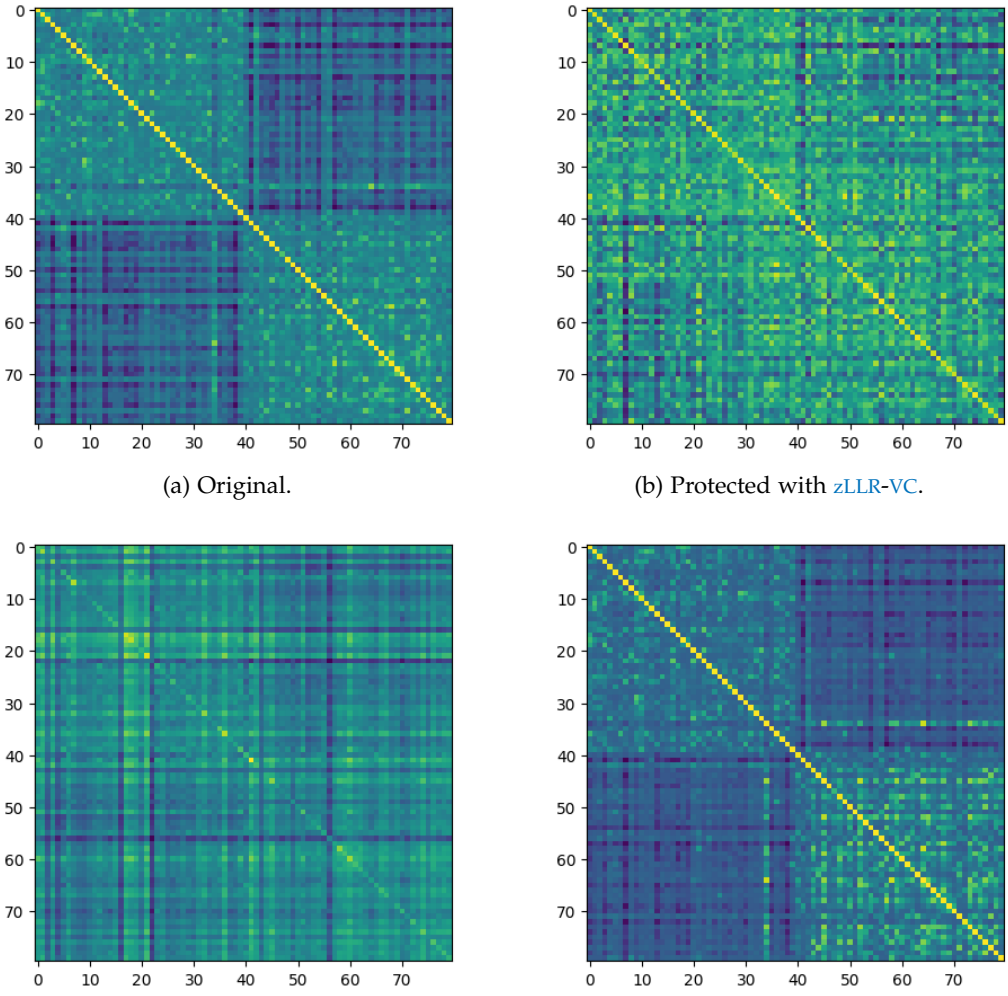
with the same sex, and also by imperfect "disentanglement" of the sex component from other speaker-related information.

In [118, 121], we proposed, in the context of the VoicePrivacy initiative, to visualize speaker voice similarity matrices to investigate the behavior of a protection system at both a speaker and global level (see Appendix 10.5 for more details). Here, our task is different than the VoicePrivacy challenge, but we can still visualize voice similarity matrices to assess the consistency of the protected voices and to pay attention to any sex-related patterns that could appear in the matrices. We report four of these matrices in Figure 25. Speakers were grouped by sex such that squares appear in the matrix 25a. Indeed, male speakers generally look more like other males than females and vice versa. When we have good sex protection in 25b and 25c, we can see that these squares tend to disappear. The near disappearance of the diagonal in 25c confirms that *global* is not suitable enough to preserve other speaker variabilities compared with zLLR-VC.

5.4.5 *Listening tests*

The results presented so far show the machine's perception. In this section, we discuss how the human ear may perceive the protected speech by reporting listening test results. In addition to using automatic regonizers, an attacker may try to infer the sex of the speaker by listening to the speech utterance. Listening tests are therefore crucial to assess the robustness of a protection system to this kind of attack.

In our test, 19 listeners—all native English speakers—were asked to assess the naturalness of speech on a discrete scale from 1 (unnatural) to 10 (natural) and whether the speech sounds like a male (1), a female (5), or neutral (3), also allowing for some nuance with scores of 2 and 4. Figures 26a and 26b show respectively the distributions of the naturalness score and distributions of the sex-perception score for each system. The blue and red dots show the median of each distribution and the cyan and orange dots show the means of each distribution. The speech processed by the analysis/synthesis even without protection does not sound as natural as the original speech. The mass of the naturalness score distribution *synthesised* is indeed moved to lower values (Figure 26a). However, applying the *global* or the zLLR-VC protection does not further decrease, on average, the naturalness.



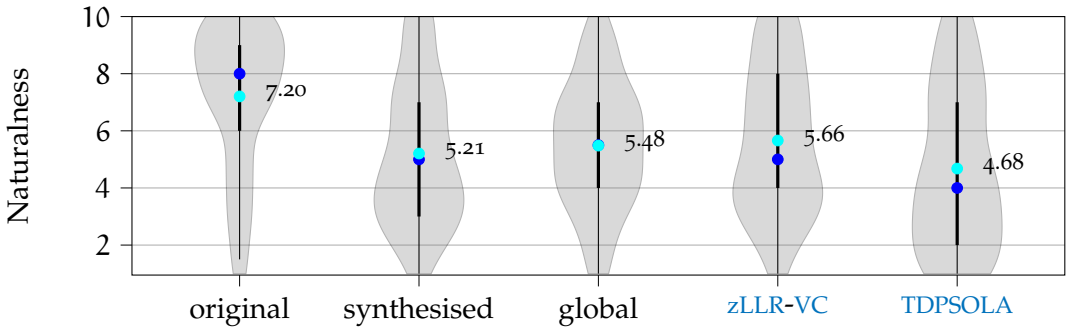
(a) Original.

(b) Protected with zLLR-VC.

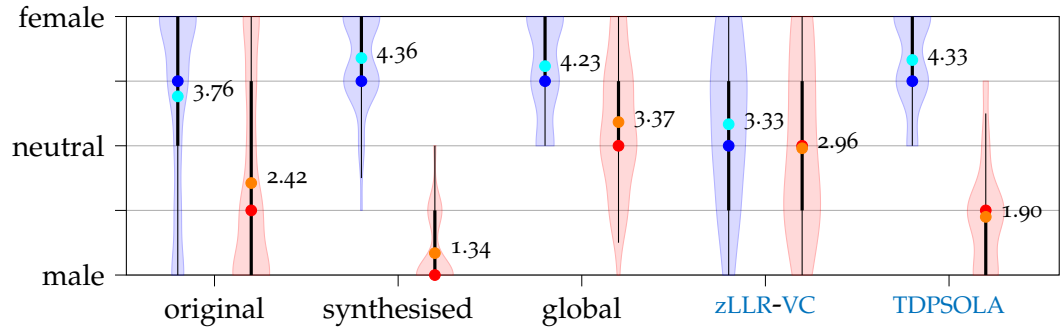
(c) Protected with *global*

(d) Protected with TDPSOLA

Figure 25: Voice log-similarity matrices (see Appendix 10.5 for more details).



(a) Distributions of the perceived naturalness scores. The blue and the cyan points respectively show the medians and the means.



(b) Distributions of the sex-perception scores. The blue and red points show the medians and the cyan and orange points show the means.

Figure 26: Listening test results. Violin plots of perceived speech naturalness (top). Violin plots of perceived speaker's sex (bottom), blue for female and red for male; blue and red dots show medians, cyan and orange dots show means.

For assessing the robustness to listening attacks, Figure 26b shows the distributions of the sex-perception scores. As expected, TDPSOLA does not sufficiently change the perception of the sex. The scores are on average low for male and high for female. The *global* approach seems to change the perception of the speech from males: the mass of the distribution is around 3. However, it does not have the expected behavior for females[9]. The zLLR-VC approach works for both male and female with a good average score close to 3. This suggests that the proposed protection tends to make attacks by listening inefficient. However, for better zero-evidence protection of each utterance, it would have been better to have narrower distributions around the neutral score[10].

5.5 SUMMARY

In this chapter, a new discriminant analysis for a binary attribute has been presented. It allows mapping the data into a space where only the first dimension is discriminant and forms a calibrated LLR representing the attribute-related information contained in the data.

9 We do not have an explanation as to why *global* does not protect the data as well as zLLR-VC.

10 Daniel Ramos has rightly pointed out that if these distributions are not narrow around the neutral score, it is because the scores are not calibrated.

The mapping is invertible such that the LLR can be set to zero before mapping the data back into the feature space. This allows to reduce in the data the strength of evidence about the attribute and is consistent with the zero-evidence formulation of privacy.

This approach to privacy has been tested on the concealment of the speaker's sex in speaker's embeddings. Our experiments on VoxCeleb2 showed the ability of the approach to hide the sex of the speaker into the speaker embeddings while keeping remaining speaker variabilities.

Such protected speaker embeddings have also been used for voice conversion for reducing the sex-related information in speech utterances. Our experiments on LibriSpeech, based on automatic sex classification, and listening tests showed the ability of the approach to reduce, in the voice, the information related to the sex of speaker while preserving reasonable intelligibility according to the ASR, ASV, and listening test results.

The idempotence property of the LLRs and its constraint on their distributions is a crucial aspect for the design of our approach. The next chapter extends the idempotence property on ILRL that also leads to a constraint on the ILRL's conditional densities. This will be used in Chapter 7 to extend the discriminant analysis presented above to multiclass cases i. e. where more than two classes are involved.

6

THE IDEMPOTENCE AND DISTRIBUTIONS OF THE LIKELIHOOD VECTORS ON THE AITCHISON SIMPLEX

In the previous chapter, we presented a discriminant analysis that can be used for the concealment of a binary variable by setting the LLR to zero. The first step to extend this work to non-binary variables is to extend the logit form of the Bayes' rule to cases where more than two hypotheses or categories are involved. This has been discussed in Chapter 3 where we saw that the Isometric-Log-Ratio (ILR) transformation of the likelihood function (ILRL) is a good candidate for going beyond the LLR and summarising the evidence in a multiple hypotheses case.

The design of the two classes discriminant analysis presented in the previous chapter is based on the properties of the LLR that have been discussed in Section 2.6. In this chapter, these properties are generalized to the ILRL.

6.1 THE IDEMPOTENCE PROPERTY

As we mentioned in Section 2.6.1, in the two hypotheses case, we want the posterior distribution to be the same whether the LLR or the data E are given. Equivalently, we want here the posterior distribution to be the same whether the ILRL or the data E are given meaning that the ILRL contains all the relevant information, about the hypotheses, contained in E:

$$\forall i \in [\![1, N]\!],\ P(H_i \mid E) = P(H_i \mid \mathbf{l}), \tag{79}$$

Let $\tilde{\mathbf{P}}(\cdot) = \operatorname{ilr}\left([P(H_i \mid \cdot)]_{1 \leqslant i \leqslant N}\right)$ be the ILRL transformation of the posterior probability vector and $\mathbf{l} = \tilde{\boldsymbol{w}}(E) = \operatorname{ilr}(\boldsymbol{w}(E))$ be the ILRL vector. Since, within the Aitchison geometry of the simplex, the Bayes' rule is the translation of the prior probability vector by the likelihood vector, Expression 79 can be rewritten as:

$$\begin{aligned} \tilde{\mathbf{P}}(\mathbf{l}) = \tilde{\mathbf{P}}(E) &\iff \tilde{\boldsymbol{w}}(\mathbf{l}) + \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{w}}(E) + \tilde{\boldsymbol{\pi}}, \\ &\iff \tilde{\boldsymbol{w}}(\mathbf{l}) = \tilde{\boldsymbol{w}}(E), \\ &\iff \tilde{\boldsymbol{w}}(\mathbf{l}) = \mathbf{l}, \end{aligned} \tag{80}$$

where $\tilde{\boldsymbol{\pi}}$ is the ILR transformation of the prior probability vector and by definition, $\mathbf{l} = \tilde{\boldsymbol{w}}(E)$. The result $\tilde{\boldsymbol{w}}(\mathbf{l}) = \mathbf{l}$ is the *idempotence* of the ILRL that can be read as:

"The ILRL of the ILRL is the ILRL itself".

6.2 THE CONDITIONAL DENSITIES OF THE ISOMETRIC-LOG-RATIO-LIKELIHOOD VECTOR

Exactly like in Section 2.6.2 with the LLR, we will see how the idempotence property of the ILRL leads to a constraint on their distributions. Let $\mathbf{A} \in \mathcal{M}_{N-1,N-1}(\mathbb{R})$ be a real square matrix defined as:

$$\begin{aligned} &\mathbf{A} = \{\alpha_{ij}\}_{1 \leqslant i,j \leqslant N-1} \\ &\alpha_{ij} = \begin{cases} 2\sqrt{\frac{i+1}{i}}, & \text{if } i = j \\ \frac{2}{\sqrt{j(j+1)}}, & \text{if } j < i \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{81}$$

and let $\mathbf{B} \in \mathcal{M}_{N-1,(N-1)^2}(\mathbb{R})$ be a block matrix:

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}^{(1)} & \mathbf{B}^{(2)} & \dots & \mathbf{B}^{(N-1)} \end{bmatrix} \tag{82}$$

where $\mathbf{B}^{(b)} \in \mathcal{M}_{N-1,N-1}(\mathbb{R})$ is the bth block and is defined as:

$$\begin{aligned} &\mathbf{B}^{(b)} = \{\beta_{ij}^{(b)}\}_{1 \leqslant i,j \leqslant N-1} \\ &\beta_{ij}^{(b)} = \begin{cases} \frac{b+1}{b}, & \text{if } i = j = b \\ 2\sqrt{\frac{i+1}{ib(b+1)}}, & \text{if } (i = j) \wedge (b < i) \\ \frac{1}{jb\sqrt{(j+1)(b+1)}}, & \text{if } (b < i) \wedge (j < i) \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{83}$$

The idempotence property of ILRL leads to the following property on their distributions:

Proposition 2. *If* $\mathbf{l} \mid H_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, *then* $\forall i \in [\![2, N]\!]$, $\mathbf{l} \mid H_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, *where*

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}\mathbf{a}_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1}\boldsymbol{\Sigma}\mathbf{a}_j,$$

and $\boldsymbol{\mu}_1 = \mathbf{A}^{-1}\mathbf{B}\,\mathrm{vec}(\boldsymbol{\Sigma})$, *where the* $(N-1)^2$*-dimensional vector* $\mathrm{vec}(\boldsymbol{\Sigma})$ *is the vectorization of the covariance matrix* $\boldsymbol{\Sigma}$ *and* $\forall i \in [\![1, N-1]\!]$, $\mathbf{a}_i = \sqrt{\frac{i+1}{i}}\mathbf{e}_i$ *where* $\mathbf{e}_i$ *is the* i*th vector of the standard canonical basis of* $\mathbb{R}^{N-1}$.

In other words, if one of the conditional densities of the likelihood vector on the Aitchison simplex is a multivariate Gaussian[1], then the others are also Gaussian with the same covariance matrix and the means are entirely defined by the covariance matrix. A proof of this result is given in Appendix 10.6.

1 The multivariate normal distribution that appears with the ILR coordinate representation is also known as the *additive logistic-normal distribution*, the *logistic-normal distribution* [5], or simply the *normal distribution on the simplex* [128].

Since $\boldsymbol{\mu}_1 = \mathbf{A}^{-1}\mathbf{B}\,\mathrm{vec}(\boldsymbol{\Sigma})$, the only parameter of the distributions is $\boldsymbol{\Sigma}$ which is a $(N-1)\times(N-1)$ symmetric positive definite matrix[2]. Therefore, it corresponds to $\frac{N(N-1)}{2} = \binom{N}{2}$ scalar parameters which is equal to the number of pairs of hypotheses. In the next paragraph, we will see how these parameters can be expressed in terms of the Kullback-Leibler divergences between each conditional density. This relation between the mean vector and the covariance matrix, and how it is related to the divergences, extends what we discussed in the two hypotheses case in Section 2.6.2.1 where $\mu = \frac{\sigma^2}{2}$ and is equal to the D_{KL} between the conditional densities of the LLR.

6.2.1 *The covariance matrix and the divergences*

We will see in this section how the covariance matrix $\boldsymbol{\Sigma}$, i.e. the parameter of the ILRL distributions, can be expressed in terms of the Kullback-Leibler divergences (D_{KL}) between each conditional density. In a pattern recognition context, these divergences can be seen as the between class separabilities.

Let $\boldsymbol{\Delta} = \{d_{i,j}\}_{1\leqslant i,j\leqslant N} \in \mathcal{M}_{N\times N}(\mathbb{R}_+)$, where $d_{i,j} = D_{KL}\left(f_{H_i}, f_{H_j}\right) = D_{KL}\left(f_{H_j}, f_{H_i}\right)$, be the matrix of Kullback-Leibler divergences between each conditional density. Since the densities are multivariate Gaussian with the same covariance matrix, the divergences are symmetric and the $\boldsymbol{\Delta}$ is therefore a symmetric matrix. Since $d_{i,j} = 0$ for $i = j$, the N diagonal elements are 0. There remain therefore $\frac{N(N-1)}{2}$ degrees of freedom for the matrix of divergences, corresponding to the number of variances and covariances in $\boldsymbol{\Sigma}$.

The divergence between the densities for the hypotheses or classes i and j can be written as:

$$\begin{aligned} d_{i,j} &= \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ &= \frac{1}{2}\left(\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right), \end{aligned} \tag{84}$$

and since $\forall i \in [\![1, N-1]\!]$, $\boldsymbol{\mu}_{i+1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$ (see Appendix 10.6 for a proof),

$$d_{i,j} = \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j. \tag{85}$$

Replacing $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ with the expression given in Proposition 2 we get:

$$d_{i,j} = \boldsymbol{\zeta}_{i,j}^T\boldsymbol{\mu}_1 - \boldsymbol{\eta}_{i,j}^T\,\mathrm{vec}(\boldsymbol{\Sigma}) \tag{86}$$

2 In general, covariance matrices are symmetric positive semi-definite but here, we restrict to symmetric positive definite because we need $\boldsymbol{\Sigma}^{-1}$.

where

$$\begin{aligned}
\zeta_{i,j} &= \left(\alpha_{i-1} + \alpha_{j-1} + \sum_{k=1}^{i-2} \frac{1}{k+1} \alpha_k + \sum_{k=1}^{j-2} \frac{1}{k+1} \alpha_k \right), \\
\eta_{i,j} &= \left(\operatorname{vec}(\alpha_{i-1} \alpha_{j-1}^\top) + \sum_{k=1}^{j-2} \frac{1}{k+1} \operatorname{vec}(\alpha_{i-1} \alpha_k^\top) + \sum_{k=1}^{i-2} \frac{1}{k+1} \operatorname{vec}(\alpha_{j-1} \alpha_k^\top) \right. \\
&\quad \left. + \sum_{k=1}^{i-2} \sum_{k=1}^{j-2} \frac{1}{(k+1)(l+1)} \operatorname{vec}(\alpha_k \alpha_l^\top) \right).
\end{aligned} \tag{87}$$

When $2 \leqslant i, j \leqslant N$ and $i = j$, these vectors are respectively the $(i-1)$th row of $\mathbf{A}$ and the $(i-1)$th row of $\mathbf{B}$. Since $\mu_1 = \mathbf{A}^{-1}\mathbf{B} \operatorname{vec}(\boldsymbol{\Sigma})$, the divergences can be written as follow:

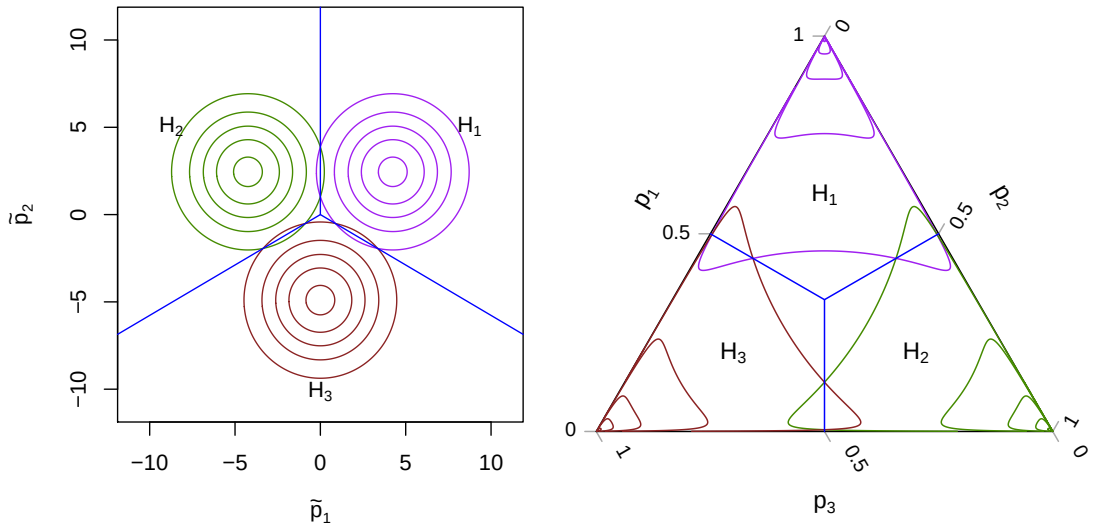
$$\begin{aligned}
&\forall i \in [\![1, N-1]\!], \\
&\forall j \in [\![i+1, N]\!], \\
&d_{i,j} = \left(\zeta_{i,j}^\top \mathbf{A}^{-1}\mathbf{B} - \eta_{i,j}^\top\right) \operatorname{vec}(\boldsymbol{\Sigma}).
\end{aligned} \tag{88}$$

Let vech be the half-vectorization of a matrix and $\operatorname{vech}_{\neg\backslash}$ be the half-vectorization without the diagonal elements. The above set of equations can therefore be written in the following matrix form:

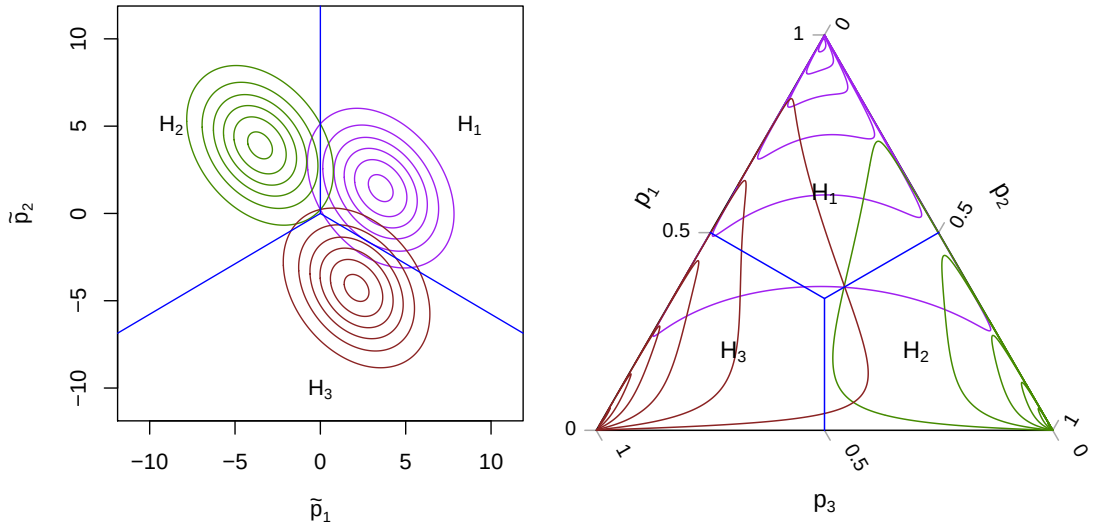
$$\begin{aligned}
\operatorname{vech}_{\neg\backslash}(\boldsymbol{\Delta}) &= \underbrace{\begin{bmatrix} \zeta_{1,2}^\top \mathbf{A}^{-1}\mathbf{B} - \eta_{1,2}^\top \\ \zeta_{1,3}^\top \mathbf{A}^{-1}\mathbf{B} - \eta_{1,3}^\top \\ \vdots \\ \zeta_{N-1,N}^\top \mathbf{A}^{-1}\mathbf{B} - \eta_{N-1,N}^\top \end{bmatrix}}_{\mathbf{M}} \mathbf{D}_{N-1} \operatorname{vech}(\boldsymbol{\Sigma}), \\
\operatorname{vech}_{\neg\backslash}(\boldsymbol{\Delta}) &= \mathbf{M} \operatorname{vech}(\boldsymbol{\Sigma}),
\end{aligned} \tag{89}$$

where $\mathbf{D}_{N-1}$ is the duplication matrix [95] such that $\operatorname{vec}(\boldsymbol{\Sigma}) = \mathbf{D}_{N-1} \operatorname{vech}(\boldsymbol{\Sigma})$ and $\mathbf{M} \in \mathcal{M}_{\frac{N(N-1)}{2} \times \frac{N(N-1)}{2}}(\mathbb{R})$ is a real square matrix. The divergences can therefore be computed from the parameter $\boldsymbol{\Sigma}$. Unfortunately, we did not prove that $\mathbf{M}$ is invertible even if in our experiments we assumed $\mathbf{M}$ to be so, such that the covariances can be expressed in terms of the divergences as: $\operatorname{vech}(\boldsymbol{\Sigma}) = \mathbf{M}^{-1} \operatorname{vech}_{\neg\backslash}(\boldsymbol{\Delta})$.

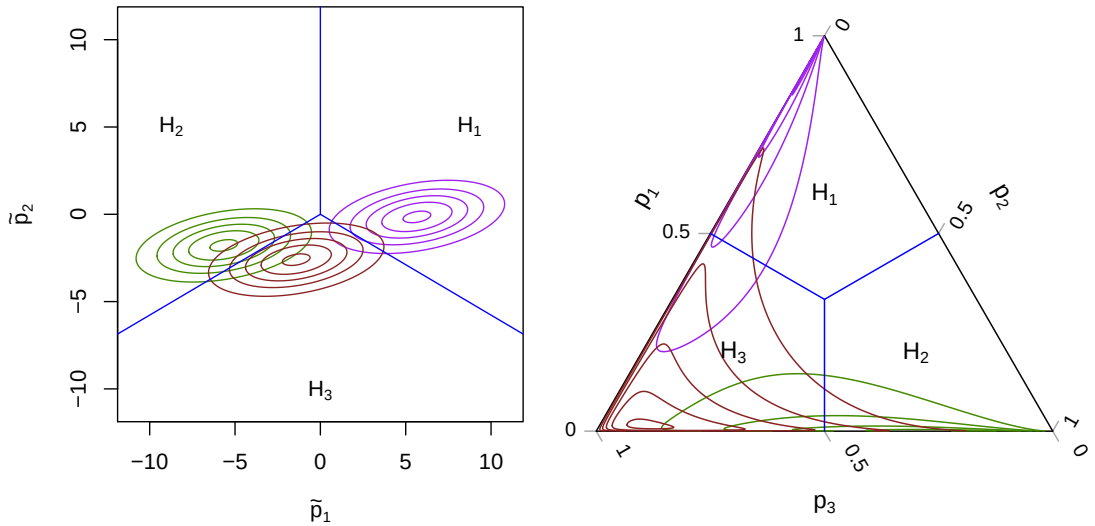
Figure 27 shows a few examples of conditional densities of the ILRL with three hypotheses or classes H_1, H_2, and H_3. The blue lines mark out the maximum probability decision regions. The figures on the right side show the densities on a ternary plot and the figures on the left side show the densities with the ILR coordinate representation. The densities are multivariate normal and their parameters are linked and constrained according to Proposition 2. Their parameters can be expressed in terms of the three divergences. When the separabilities between each class are all the same, the covariance matrix is isotropic and the means are equidistant. When they are different, the densities stretch accordingly. When two classes get closer, the densities crush on the corresponding decision boundary and when



(a) $\mathrm{EER}_{1,2} = \mathrm{EER}_{2,3} = \mathrm{EER}_{1,3} \approx 4.16\%$, $\mathrm{d}_{1,2} = \mathrm{d}_{2,3} = \mathrm{d}_{1,3} = 6$.



(b) $\mathrm{EER}_{1,2} \approx 5.69\%$, $\mathrm{d}_{1,2} = 5$; $\mathrm{EER}_{2,3} \approx 3.07\%$, $\mathrm{d}_{2,3} = 7$; $\mathrm{EER}_{1,3} \approx 7.86\%$, $\mathrm{d}_{1,3} = 4$.



(c) $\mathrm{EER}_{1,2} \approx 2.28\%$, $\mathrm{d}_{1,2} = 8$; $\mathrm{EER}_{2,3} \approx 15.87\%$, $\mathrm{d}_{2,3} = 2$; $\mathrm{EER}_{1,3} \approx 7.86\%$, $\mathrm{d}_{1,3} = 4$.

Figure 27: Few contours of Gaussian conditional densities of ILRL in a three hypotheses case. They are here parameterized by a covariance matrix that can be expressed in terms of the three separabilities between each class. The densities on the ILR space (left) are with respect to the Lebesgue measure while the densities on the simplex (right) are with respect to the Aitchison measure [99].

the separability between all classes goes to 0, the densities collapse at $\mathbf{0}$ and tend to be a Dirac delta function.

6.3 SUMMARY

This chapter presented properties of the isometric log-ratio transformation of the likelihood function (ILRL) which can be seen as a multidimensional and multiple hypotheses extension of the LLR. Just like the LLR, the ILRL is idempotent which leads to a constraint on its conditional densities. If one of the conditional densities is a multivariate Gaussian, the others are also a multivariate Gaussian with the same covariance matrix and the means are entirely defined by the covariance matrix. The latter can be expressed in terms of the separability between each class for a better interpretation of the densities' parameter.

In the next chapter, these properties on the distribution of ILRL are used to design the multiclass generalization of the discriminant analysis presented in Chapter 5.

7

COMPOSITIONAL DISCRIMINANT ANALYSIS

In Chapter 5 we saw how the idempotence property and its constraint on the distributions of the LLR can be used to design a new discriminant analysis for the two classes case. Chapter 6 showed how these properties generalize to the isometric log-ratio transformation of the likelihood vector (ILRL) that we consider as a multidimensional extension of the LLR for representing the evidence in a multiple hypotheses or classes case. The current chapter presents how these properties can be used to extend the discriminant analysis of Chapter 5 to any number of classes.

In Chapter 5, the idea was to map the data into a space where the discriminant dimension is designed to be a calibrated LLR, the idea is here to map the data into a space where the discriminant dimensions form a calibrated ILRL.

We call this approach Compositional Discriminant Analysis (CDA) not to be confused with discriminant analysis that aims in modeling compositional data as discussed for instance in [59] or in section 8.4 of [128]. In these cases, the data is compositional, its sample space is the simplex defined in Equation 34. Since standard statistical analysis methods are designed for data that lives on a Euclidean space, the data is beforehand transformed using for instance the isometric log-ratio transformation. Our approach is different. The data is not necessarily compositional and the compositional nature of our approach is on the treatment of the likelihoods and probabilities computed through the discriminant analysis. The discriminant space is designed to be the space of ILRL having therefore the intuitive structure of the Aitchison geometry of the simplex. Moreover, for better calibration, the distributions of the likelihood vectors are designed to respect the idempotence property constraint of Proposition 2.

7.1 PROPOSED COMPOSITIONAL DISCRIMINANT ANALYSIS FOR MULTICLASS

Like in Section 5.2, the proposed discriminant analysis makes no explicit assumption on the distribution of the observed data contrary to the LDA or the QDA[1]. Let's consider the set $\mathcal{C} = \{C_1, C_2, \dots C_N\}$ of N classes and an observed vector x which belongs to one of these classes. The proposed discriminant analysis, named CDA, is a generative model which models the class-conditional densities in the feature space by learning an invertible and differentiable mapping between the feature space and a base space in which the class-conditional densities are known. The class-conditional densities in the base space are de-

1 This does not mean there is no assumption at all. The same remark of Section 5.2.2 holds here. Even if there is no explicit assumption on how the feature vectors are distributed, the use of the CDA assumes the existence of a diffeomorphism that transforms the class-conditional densities into Gaussian ILRL and Gaussian residual densities. As we saw in the two classes example on the Circles dataset in Section 5.2.4, when this assumption is not fulfilled, this tends to affect the discrimination quality rather than the calibration quality.

signed according to the idempotence property constraint of Proposition 2. In this way, the mapping transforms the observed vectors into a same-dimensional base space where the first $N-1$ dimensions form the ILRL and the others form the residual which can be seen as "everything in the observed variable that is independent of the class variable".

Let $\mathcal{X}$ be the d-dimensional feature space and let $\mathbf{l}(\mathbf{x}) \in \mathcal{L} \subset \mathbb{R}^{N-1}$ be the ILRL of an observation $\mathbf{x} \in \mathcal{X}$ where the ith component is given by (see Equation 45):

$$\forall i \in [\![1, N-1]\!], \qquad l_i(\mathbf{x}) = \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod_{j=1}^{i} P(\mathbf{x} \mid C_j)}{P(\mathbf{x} \mid C_{i+1})^i} \right). \tag{90}$$

Let $\mathbf{r}(\mathbf{x}) = [r_1(\mathbf{x}), r_2(\mathbf{x}), \dots r_{d-N+1}(\mathbf{x})]^T \in \mathcal{R} \subset \mathbb{R}^{d-N+1}$ be the residual of $\mathbf{x}$. We want the data to be mapped from the feature space to a base space $\mathcal{Z} = \mathcal{L} \oplus \mathcal{R}$ in which the first $N-1$ dimensions form the ILRL, representing therefore the evidence, and the other dimensions form the residual.

7.1.1 *Class-conditional densities in the base space*

In order to properly represent the evidence, the first $N-1$ dimensions of the base space have to form a calibrated ILRL. The class-conditional densities in the base space have to be chosen accordingly to respect the idempotence property constraint of Proposition 2:

$$\forall i \in [\![1, N]\!], \qquad \mathbf{z} \mid C_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{C}), \tag{91}$$

where:

- $\mathbf{\Sigma}$ is a $(N-1) \times (N-1)$ symmetric positive definite matrix and the only parameter of the distributions in the base space,
- the means $\mathbf{m}_i \in \mathcal{Z} \subset \mathbb{R}^d$ are the concatenation of $\boldsymbol{\mu}_i \in \mathcal{L}$ and the $(d-N+1)$-dimensional zero vector. $\boldsymbol{\mu}_i$ is defined as in Proposition 2 and its expression is given by (see Appendix 10.6 for the proof):

$$\forall i \in [\![1, N]\!], \qquad \boldsymbol{\mu}_i = \mathbf{A}^{-1}\mathbf{B}\,\mathrm{vec}(\mathbf{\Sigma}) - \mathbf{\Sigma}\mathbf{a}_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1}\mathbf{\Sigma}\mathbf{a}_j, \tag{92}$$

  where $\mathbf{A} \in \mathcal{M}_{N-1,N-1}(\mathbb{R})$ and $\mathbf{B} \in \mathcal{M}_{N-1,(N-1)^2}(\mathbb{R})$ are defined as in Equations 81, 82 and 83, $\mathbf{a}_i = \sqrt{\frac{i+1}{i}}\mathbf{e}_i$ where $\mathbf{e}_i$ is the ith vector of the standard canonical basis of $\mathbb{R}^{N-1}$ and $\mathbf{a}_0 = \mathbf{0}$ is the $(N-1)$-dimensional zero vector,

- the covariance matrix $\mathbf{C}$ is the following block matrix:

$$\mathbf{C} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0}_{N-1,d-N+1} \\ \mathbf{0}_{d-N+1,N-1} & \mathbf{I}_{d-N+1} \end{bmatrix}, \tag{93}$$

  where $\mathbf{I}_K$ is the $K \times K$ identity matrix and $\mathbf{0}_{K,L}$ is the $K \times L$ zero matrix.

In this way, the first $N-1$ dimensions of z form the ILRL (see Appendix 10.7 for a proof) and the others form the residual normally distributed with a zero mean vector and an identity covariance matrix regardless of the class.

Note that, as shown in the case of the LLR in Equation 77, since the mapping g between $\mathcal{X}$ and $\mathcal{Z}$ is invertible, the ILRL of z and the ILRL of x are the same: $l(x) = l(z)$ where $z = g^{-1}(x)$. Indeed, for a given observation x, the likelihood vector is scaled by the Jacobian determinant of the mapping to get the likelihood vector of z and, as we discussed in Section 3.3, likelihood vectors are scale invariant.

We will not discuss in detail how the mapping g is learned as this is basically the same as in Section 5.2.2 but to summarise, g is based on a composition of differentiable and invertible neural networks, named Normalizing Flow (NF) [127], learned through negative log-likelihood minimization with automatic differentiation and gradient descent. Nevertheless, the next section will briefly discuss the initialization and optimization of the base space class-conditional densities' parameter $\boldsymbol{\Sigma}$.

7.1.2 *Regarding the initialisation and estimation of* $\boldsymbol{\Sigma}$

In our experiments, the training of the proposed discriminant analysis turned out to be very sensitive to the initialization of $\boldsymbol{\Sigma}$. In this section, we propose an initialization method for starting the optimization with a $\boldsymbol{\Sigma}$ that we expect to be not too eccentric.

Section 6.2.1 showed how the covariance matrix $\boldsymbol{\Sigma}$ can be expressed in terms of the Kullback-Leibler divergences (D_{KL}) within each pair of classes[2]. We propose here to initialize the mapping g as the identity function and to initialize $\boldsymbol{\Sigma}$ with the D_{KL} measured in the feature space assuming that each class-conditional densities are multivariate Gaussians with shared covariance (this is the standard LDA assumption). Even if there is no strong theoretical foundation for this choice of the initial $\boldsymbol{\Sigma}$ and g, these initializations appeared to be effective in our experiments. Note that this does not mean that an assumption on how the feature vectors are distributed is made. This is only for the initialization of the optimization. The parameters of g and $\boldsymbol{\Sigma}$ are then free to take any value under the constraints of differentiability and invertibility for g and symmetric positive definiteness for $\boldsymbol{\Sigma}$.

The symmetric positive definiteness of $\boldsymbol{\Sigma}$ is insured by optimizing instead the lower triangular matrix $\mathbf{L}$ from the Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ and since the diagonal elements of $\mathbf{L}$ must be positive, the log-Cholesky parametrization [131] is used. In our experiments, the estimation of $\mathbf{L}$ and the parameters of g is done with automatic differentiation and

2 Keep in mind that since the class-conditional densities in the base space are Gaussian with the same covariance matrix, the Kullback-Leibler divergences are symmetric.

gradient descent. We did not find a straightforward computation of a maximum likelihood estimator of $\boldsymbol{\Sigma}$ contrary to $\boldsymbol{\mu}$ in Section 5.2.3.

7.1.3 *Regarding the interpretability of the compositional discriminant analysis*

The proposed discriminant analysis maps the data into a space where only the first $N-1$ dimensions are discriminant and form the ILRL of the observation. With the standard LDA, the $N-1$ dimensions given by the non-zero eigenvalue eigenvectors of the matrix $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$, where $\boldsymbol{\Sigma}_W$ is the shared within-class covariance matrix and $\boldsymbol{\Sigma}_B$ is the between-class covariance matrix, are also the discriminant ones. They are usually sorted in the descending order of the eigenvalues which inform how much each direction is discriminant.

In the proposed approach, the discriminant dimensions are not sorted according to their discriminant power. However, thanks to the compositional nature of the base space, each dimension is instead opposing a class with a group of classes in an intuitive recursive manner as illustrated by the bifurcation tree in Figure 8. The discriminations between the classes are given by the parameter $\boldsymbol{\Sigma}$ as discussed in Section 6.2.1.

Moreover, since the densities of the ILRL in the base space are designed to respect the idempotence constraint of Proposition 2, the approach tends to produce a set of ILRLs that is well-calibrated. The resulting classifier can therefore be used for uncertainty-aware decisions avoiding under or overconfident predictions. In addition, the ILRL nature of the first $N-1$ dimensions of the base space allows straightforward computation of the posterior probability distribution over the set of classes by simply shifting the prior distribution by the likelihood on the Aitchison simplex.

7.1.4 *Toy examples*

This section discusses two toy examples to illustrate the use of the CDA. The first example is the classification of multivariate normal distributions. The second example is the recognition of hand-written digits with the well-known MNIST database [89].

With Gaussians

Let's consider here a three classes example for visualization convenience. Indeed, for three classes, the probability simplex is 2-dimensional and the isometric-log-ratio coordinate system can be visualized into a 2-dimensional figure. Let the class-conditional densities be 4-dimensional Gaussians. In this way, the residual space is also 2-dimensional.

In this example, each class is generated by a multivariate normal distribution with its own mean and covariance matrix. Figure 28 shows the training set (10000 samples per class). Figure 29 shows the base space on which the testing set (2000 samples per class) is mapped using the learned transformation. Within the ILRL space 29a (i. e. the first two dimensions of the base space) the first dimension discriminates the two first classes (blue and orange) while the second dimension discriminates the third class (green) from the two others without discriminating the first two. This is ensured by the chosen Aitchison basis

corresponding to the bifurcation tree of Figure 8. The class-conditional densities in the base space tend to respect the distribution of calibrated ILRL. The residual components are not discriminant and are normally distributed with zero mean and identity covariance matrix.

The learned parameter $\boldsymbol{\Sigma}$ of the ILRL's class-conditional density is:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 35.50 & -0.87 \\ -0.87 & 8.71 \end{bmatrix},$$

and can be interpreted in terms of the following divergences, or separability between the classes, using Equation 89:

$$\begin{aligned} d_{1,2} &= 35.50, \\ d_{1,3} &= 14.65, \\ d_{2,3} &= 16.16. \end{aligned}$$

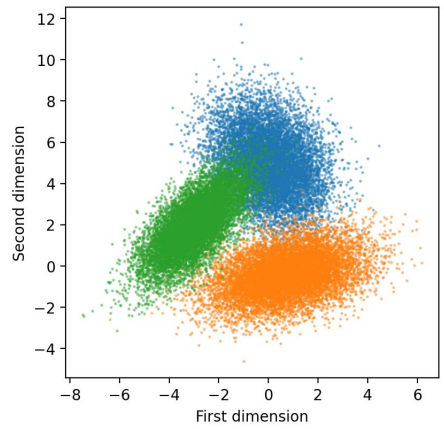
We compare the proposed CDA with the LDA and the QDA. Figures 29c and 29d show the projection of the testing set using the LDA. The first two components are discriminant but are hardly interpretable by other means than the discriminative power given by the eigenvalues. The other dimensions are less discriminant but are not identically distributed contrary to the CDA's residual space (Figure 29b).

QDA is not designed to have an information-preserving mapping of the data into a same-dimensional space. This is why there are no results for the QDA in Figure 29. QDA can however be used to compute LLR scores as given by Equation 70. Figure 30 shows the histograms of the scores obtained with LDA (Figures 30a,30b and 30c), QDA (Figures 30d,30e and 30f), and CDA (Figures 30g, 30h and 30i). For the latter, the LLR in favor of a class against another is obtained by projecting the ILRL vector on the orthogonal direction of the maximum probability decision boundaries (given in Appendix 10.3)[3]. For the LDA and the CDA, the class-conditional densities of the scores look Gaussian as expected. However, for the LDA, they are not symmetric as required by the idempotence property. We therefore expect the scores of the LDA to have less good calibration. For the QDA the histograms are not symmetric but this does not suggest that the scores are not calibrated. Indeed, the idempotence constraint of Propositions 1 and 2 is formalized for normally distributed scores while the scores are here not Gaussian[4].

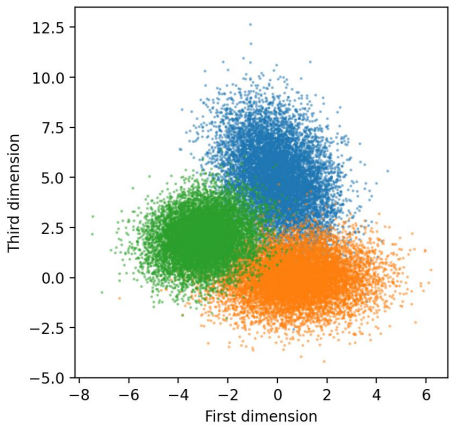
To better assess the discrimination and calibration of the scores, Table 6 provides C_{llr} measures. The LDA has the worst discrimination and calibration which is not surprising as it is based on the shared covariance assumption. QDA models the best the data and the resulting scores have the best discrimination and calibration which is again not surprising since the QDA's assumption is actually how the data are really distributed. However, as we already

3 To be more precise, projecting the data into the unit vector orthogonal to the decision boundaries gives the LLR up to a scaling factor $\frac{1}{\sqrt{2}}$. See the definition of the ILR transformation in Equation 45: its first component is $\frac{1}{\sqrt{2}}$ times the log-ratio.

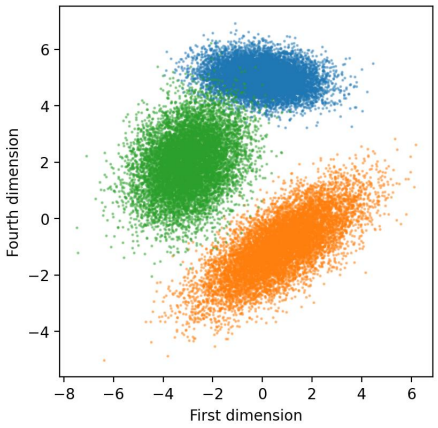
4 To be more precise, since the data is here normally distributed and the mapping is quadratic, the scores are distributed according to a generalised chi-squared distribution.



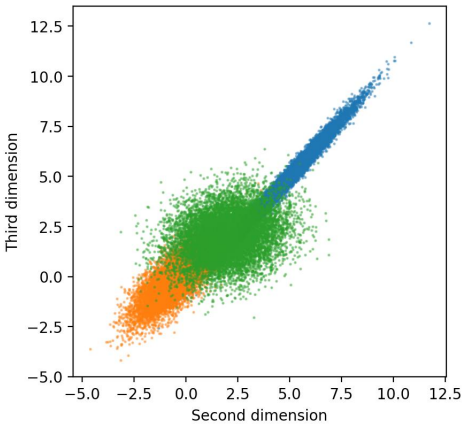
(a) The first and second dimensions



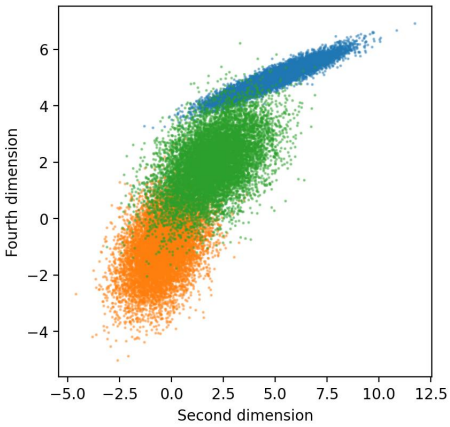
(b) The first and third dimensions.



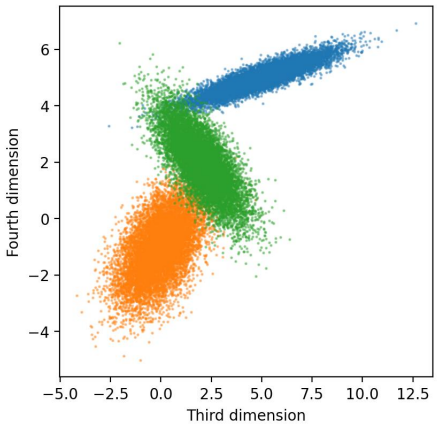
(c) The first and fourth dimensions



(d) The second and third dimensions.

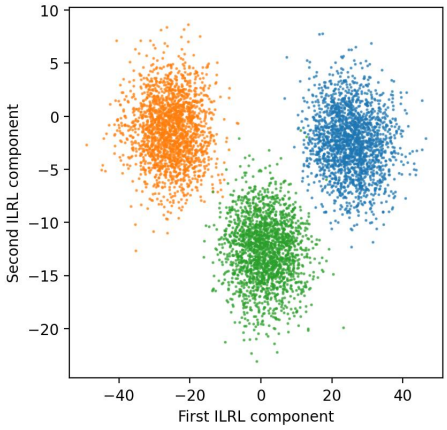


(e) The second and fourth dimensions

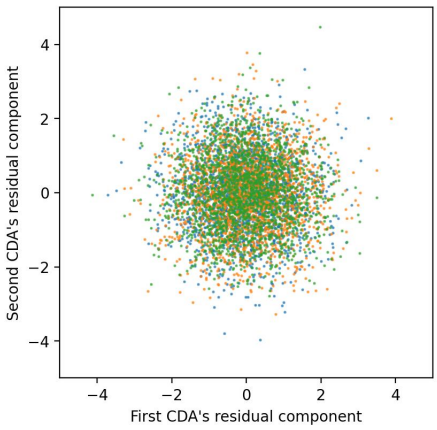


(f) The third and fourth dimensions.

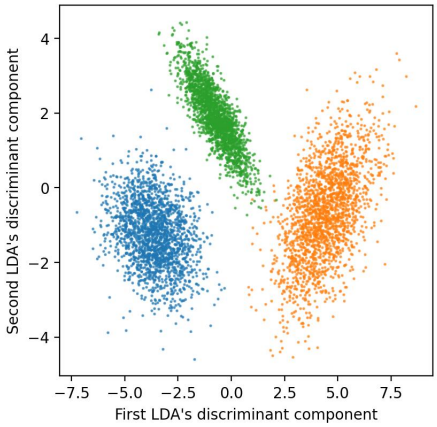
Figure 28: Training set for the three classes CDA example with non-shared covariance Gaussian. The colors indicate to which of the three classes a sample belongs: blue for C_1, orange for C_2 and green for C_3.



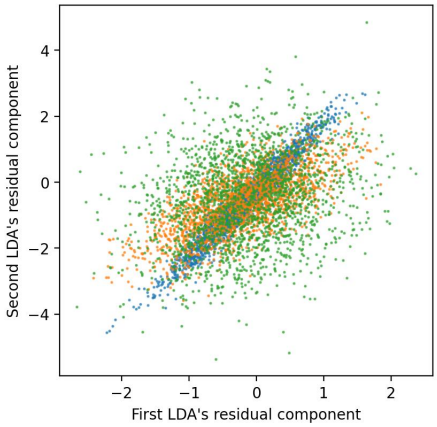
(a) The CDA's ILRL components



(b) The CDA's residual components.

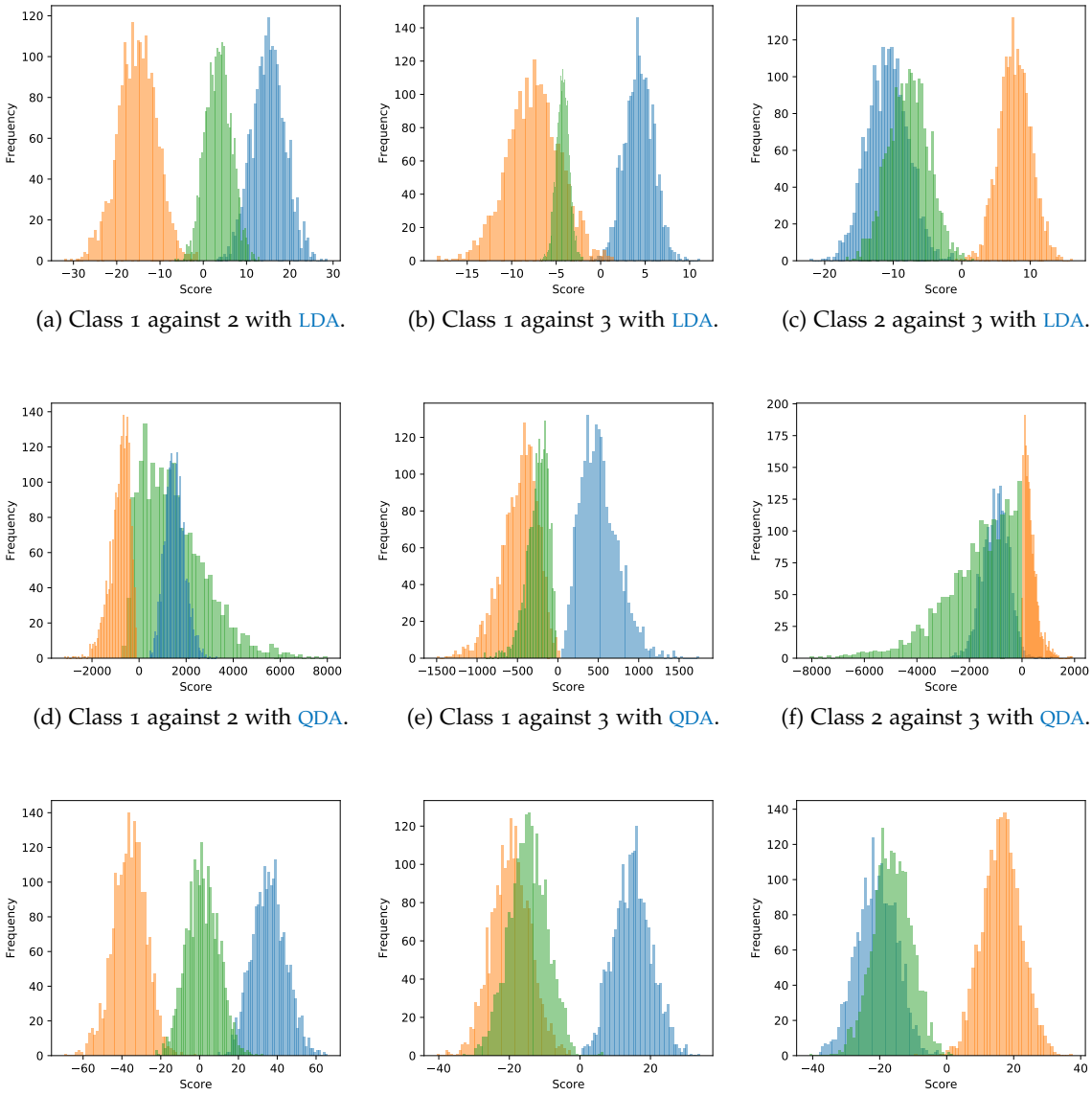


(c) The LDA's discriminant components



(d) The LDA's residual components.

Figure 29: Testing set in the LDA and CDA base spaces for the non-shared covariance Gaussian example.



(a) Class 1 against 2 with LDA. (b) Class 1 against 3 with LDA. (c) Class 2 against 3 with LDA.

(d) Class 1 against 2 with QDA. (e) Class 1 against 3 with QDA. (f) Class 2 against 3 with QDA.

(g) Class 1 against 2 with CDA. (h) Class 1 against 3 with CDA. (i) Class 2 against 3 with CDA.

Figure 30: LLR score histograms of one class against another for the non-shared covariance Gaussian example given by LDA, QDA, and CDA. C_1, C_2, and C_3 are respectively blue, orange, and green.

Table 6: C_{llr} measures for the non-shared covariance example. Samples from the non-concerned class are discarded.

compared classes	LDA		QDA		CDA	
	C_{llr} [bit]	C_{llr}^{min} [bit]	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}
1 vs 2	$1.72_{10^{-3}}$	0.0	0.0	0.0	$4.85_{10^{-5}}$	0.0
1 vs 3	1.98	$1.43_{10^{-1}}$	$1.72_{10^{-9}}$	0.0	$9.46_{10^{-3}}$	$5.04_{10^{-3}}$
2 vs 3	$2.00_{10^{-1}}$	$1.76_{10^{-2}}$	$6.37_{10^{-4}}$	0.0	$8.46_{10^{-3}}$	$5.31_{10^{-3}}$

discussed, the QDA does not provide an information-preserving transformation necessary for the "disentanglement" of the discriminant components from the non-discriminant ones contrary to the CDA, which, still provides good discrimination and calibration.

With the MNIST database

The MNIST database consists of grayscale images of size 28×28. Each image is a handwritten digit between 0 and 9. The training set is made of 60000 samples and the testing set is made of 10000 samples. The task is here to extract from an image the information about which digit is written. Figure 31 shows one randomly selected example for each class.

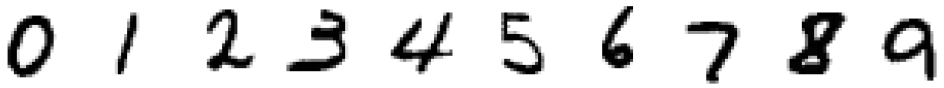


Figure 31: Examples from the MNIST database.

In the following experiment, each image is flattened and normalized into a 784-dimensional feature vector[5]. One can see from Figure 31 that the pixels on the edges of the images tend to all have the same low intensity. This leads to collinearities between some of the features such that methods that require the inversion of covariance matrices in the feature space (like the LDA and the QDA) can not be directly used. The dimensionality of the feature vector is therefore reduced to 40 using a Principal Component Analysis (PCA).

In the above Gaussian example, where only three classes were considered, C_{llr} measures were reported for each of the three pairs of classes. In the current experiments, there are 10 classes corresponding to 45 pairs of classes. Thus, we instead report the Empirical Cross-Entropy (ECE) C_{mc} of Equation 95 as a summary measure.

5 State-of-the-art discriminative approaches for classification on MNIST are based on Convolutional Neural Network (CNN). Even if coupling layers [48] (used in our implementation) can be made of convolutions, since our experiments are just toy examples to illustrate the use of CDA, we do not consider this way of processing the images.

About the cross-entropy measure in the multiclass case:

The Empirical Cross-Entropy (ECE) expression of Equation 19 can be written for any number N of hypotheses or classes as follow:

$$\mathrm{ECE} = -\sum_{i=1}^{N} \frac{P(C_i)}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} \log q_i, \tag{94}$$

where $q_i = P(C_i \mid e, \boldsymbol{\pi})$ and $\boldsymbol{\pi} = [P(C_1), \dots P(C_N)]^T$ are respectively the posterior and the prior probability distribution.

In the two classes case, the Pool Adjacent Violators Algorithm (PAVA) calibrates a set of scores into a reference set of probabilities or likelihoods that minimizes the ECE with a fixed discrimination allowing therefore the decomposition of the ECE into a calibration term and a discrimination term (see Section 2.4.2).

When more than two classes are involved, there is no available method to obtain perfectly calibrated probabilities or likelihoods as reference. Thus, the ECE can not be decomposed into a calibration term and a discrimination term. However, the multiclass ECE of Equation 94 can still be used to summarise the amount of useful information given by the pattern recognizer as done for instance in the context of language recognition [141]. Having a multiclass ECE lower than the entropy of the prior probability distribution informs that the system has extracted useful information.

In our experiments, we report the multiclass ECE at a uniform prior (i. e. $P(C_i) = \frac{1}{N}$ for all $i \in [\![1, N]\!]$). We call this cost C_{mc} and its expression is given by:

$$C_{mc} = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} s_i, \tag{95}$$

where s_i is the log-likelihood score, for the class C_i, given by the pattern recognizer.

In addition, since we do not have a minimum ECE as a discrimination measure, we report the accuracy of the system with a maximum likelihood decision[a]. However, be aware that the cross-entropy-based measures and the accuracy differ by nature. The accuracy measures the goodness of hard decisions while the ECE measures the goodness of probabilities or likelihoods regardless of the operating point or decision boundaries. In this way, the accuracy can not substitute a minimum ECE measure.

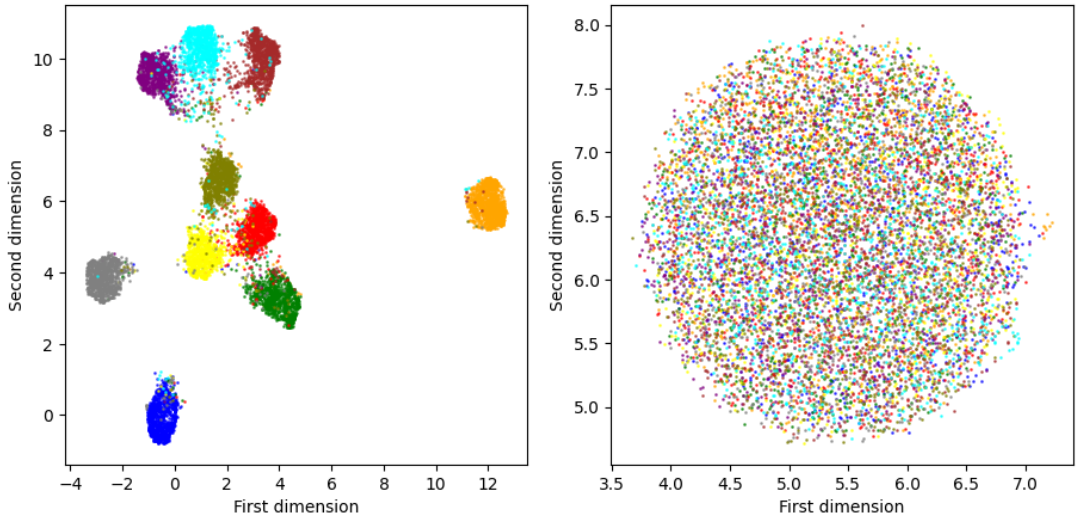
a The accuracy is given by the ratio of well recognized examples over the total number of examples.

Table 7 gives the C_{mc} in nat[6] and the accuracy on the testing set for the LDA, the QDA, and CDA. All the systems result in a cross-entropy lower than the entropy of the uniform prior distribution: $C_{mc} < \log 10 \approx 2.30$, meaning that all the systems extract useful information from the images. Interestingly, QDA has the best accuracy but the worst cross-entropy. This confirms that having a good accuracy does not mean that a system is good for making rational decisions minimizing the expected cost. The CDA results in the lowest cross-entropy

6 A nat is a unit of information when the natural logarithm is used while a bit is a unit of information with the base two.

Table 7: Cross-entropy and accuracy measures on the testing set for the MNIST's digit recognition task with LDA, QDA, and CDA.

system	C_{mc} [nat]	accuracy [%]
LDA	$5.44 10^{-1}$	87.67
QDA	$8.56 10^{-1}$	96.24
CDA	$2.23 10^{-1}$	94.43



(a) UMAP visualization of the estimated ILRLs.

(b) UMAP visualization of the residual.

Figure 32: UMAP visualization of the MNIST testing data in the CDA's base space. The color indicates to which class a sample belongs: blue for 0, orange for 1, green for 2, red for 3, purple for 4, yellow for 5, gray for 6, brown for 7, olive for 8, and cyan for 9.

which shows that it extracts the most useful information from the images. In addition, Figure 32 shows UMAP visualization of the testing data in the base space. Figure 32a shows an UMAP visualization of the nine first dimensions i. e. of the estimated ILRL. One can see that, the visualization results in clusters corresponding to the ten digits while in Figure 32b—showing an UMAP visualization of the residual—no clusters are produced. This suggests, as expected, that the information related to the digit that has been written is concentrated in the ILRL components. Table 8 shows the estimated Kullback-Leibler divergences between the digit's class-conditional densities in the base space. These divergences are computed from the estimated $\boldsymbol{\Sigma}$ using Equation 89. This informs us about the separability between the classes. Even if the observed distance between the clusters in Figure 32a can not be read as they are, they share some overall tendencies with the estimated divergences.

The CDA can be used without dimensionality reduction beforehand. This allows the learning of an information-preserving transformation between the space of images and a base space. In this case, CDA can be used for generating images. The interpretability of the base space allows manipulation or generation of images. As an example, Appendix 10.8 shows interpolation between digits. The parameter $\boldsymbol{\Sigma}$ of the class-conditional densities

Table 8: Estimated Kullback-Leibler divergences (rounded with one decimal) between the digit's conditional densities in the base space.

digit	0	1	2	3	4	5	6	7	8	9
0	0	-	-	-	-	-	-	-	-	-
1	38.3	0	-	-	-	-	-	-	-	-
2	18.5	15.6	0	-	-	-	-	-	-	-
3	16.2	16.0	8.8	0	-	-	-	-	-	-
4	24.8	25.0	17.8	17.4	0	-	-	-	-	-
5	12.4	19.2	14.1	7.2	13.3	0	-	-	-	-
6	18.4	25.9	13.3	19.6	13.0	13.5	0	-	-	-
7	21.2	21.6	17.5	15.0	14.2	16.9	27.1	0	-	-
8	18.0	16.6	8.3	7.2	13.8	6.7	14.5	17.5	0	-
9	21.2	21.3	17.9	12.8	5.1	11.9	19.0	7.3	10.9	0

in the base space defines the centroid for each digit (see Section 7.1.1) and the Euclidean structure allows linear interpolation for generating "digits in between"[7].

7.2 REGARDING THE USE OF THE COMPOSITIONAL DISCRIMINANT ANALYSIS FOR PRIVACY

The discriminant analysis presented in Chapter 5 is a particular case of the CDA: this is the two hypotheses or classes formulation of the CDA. Section 5.3 discussed its use for privacy where the information to protect is about a binary attribute. The multiclass extension of the discriminant analysis proposed in the current chapter can similarly be used for the concealment of attributes with more than two categories. The information about the sensitive attribute is represented by the ILRL components in the CDA's base space. For protection, the ILRL components of all observations can be set to zero in the base space being consistent with the *zero-evidence* formulation of privacy, while keeping the residual information unchanged. The protected observations can then be mapped back into the feature space.

Unfortunately, there is no experiments in this thesis to illustrated the use of CDA for multiclass attribute privacy. We hope this will be tested in future works.

7.3 SUMMARY

This chapter extended the discriminant analysis proposed in Chapter 5 to any number of classes. This discriminant analysis we called CDA models the class-conditional densities of the data by learning a diffeomorphism between the feature space and a base space in which the densities are designed according to the distribution of calibrated likelihood vectors. In

7 We trained the CDA without pre-dimensionality reduction (i. e. on the 784-dimensional features vectors) resulting in a C_{mc} of $3.81_{10^{-1}}$ nats and an accuracy of 89.88%.

this way, the data can be mapped into a space where the discriminant components form the likelihood function on the probability simplex with the Euclidean vector space structure known as the Aitchison geometry of the simplex. This compositional structure of the base space allows easy interpretation of the discriminant components. Each component is opposing a class with a group of classes. The calibration cost of standard discriminant analysis methods usually comes from assumptions—about the distribution of the data—which are not fulfilled. The proposed CDA makes no explicit assumptions on the distribution of the data and is designed to produce calibrated likelihood functions. For this reason, CDA is a good candidate for application where rational decisions in presence of uncertainty is crucial.

8

FLOW-BASED CALIBRATION

The concept of calibration of LLRs has been discussed in Section 2.5. Some pattern recognizers produce scores[1] with good discrimination but bad calibration making them not suitable for rational decision-making. Calibration methods can be used on top of a recognizer to transform the scores into calibrated LLRs or *better* calibrated LLRs[2].

Considering a binary classification problem and a set of scores, a calibration aims in transforming these scores into a calibrated set of LLRs. This chapter focuses on *pure* calibration only. This refers to a calibration that does not change the discrimination power of the scores and seeks to reduce the calibration cost only in opposition to *hybrid discrimination-calibration* which seeks to reduce both the discrimination and the calibration cost [135]. Therefore, in order to avoid "swappings" between scores and to keep the same discrimination power, pure calibration methods are restricted to invertible and monotonically increasing mappings.

Chapters 5 and 7 introduced a new approach to discriminant analysis. The idea is to learn of a diffeomorphism between the feature space and a space where the first dimensions contain the class-related information in the form of a calibrated likelihood function and the other dimensions form the residual i. e. the variability not related to the class. This transformation is learned through Normalizing Flow (NF) which is a cascade of learnable diffeomorphisms. For calibration purpose, this chapter introduces the use of NF to transform class-conditional densities of scores into densities that respect the constraint of Proposition 1. The aim is not to propose a calibration approach that works better than others but rather to share a new idea that we hope could be helpful for those who want to go beyond standard calibration methods. First, let's discuss some of the LLR calibration methods usually studied in the context of Automatic Speaker Verification (ASV).

8.1 CALIBRATION METHODS FOR LOG-LIKELIHOOD-RATIOS

Automatic Speaker Verification (ASV) research has been very active in the design of calibration methods for LLRs. Indeed, scores produced by *target/non-target* classifiers can rarely be directly treated as calibrated LLRs. Calibration methods are therefore required to transform these scores into calibrated LLRs. There are several approaches to calibration:

- *Discriminative* approaches learn the parameter of a monotonically increasing mapping by minimizing an objective function (usually based on proper scoring rules, it can be for instance the C_{llr}). The most common is the linear calibration learned

1 The scores are not necessarily LLR-like scores. They can be more general following the property of having a large value for one class and a low value for the other class.

2 We already discussed calibration mapping in Section 2.4.2 in the context of probability calibration. Here, we discuss the calibration of LLRs rather than the calibration of probabilities.

through logistic regression [20, 21, 28, 148]. Be aware that even if these approaches are called *discriminative* this does not mean they improve the discrimination power of the scores. Since the mapping is subject to a monotonicity constraint, it can only improve the calibration. The word *discrimination* refers here to the training procedure in opposition to the *generative* approaches.

- *Generative* approaches for calibration aim in modeling the class-conditional densities of the scores. In [91], the class-conditional densities of the scores are assumed to be Gaussian with shared variance. Scores are affinely transformed to respect the idempotence constraint of Proposition 1 (which is actually equivalent to the discriminative trained logistic regression discussed above). Non-linear generative calibrations have been proposed in [26] by dropping the shared covariance assumption, resulting in a quadratic transformation, or assuming more general distributions like Student's t, normal-inverse-Gaussian distributions [26], or normal variance-mean mixture distributions [35, 37]. In [36, 38], the authors showed that variance-gamma and the wider family of generalized hyperbolic distributions better model the ASV scores produced by the Probabilistic Linear Discriminant Analysis (PLDA). With the generative approach to calibration, unsupervised and semi-supervised[3] training can also be investigated but are out of the scope of this thesis [23].

- The two calibration kinds presented above are parametric. The PAVA-based calibration [24], discussed in Section 2.4.2, is *non-parametric*. While PAVA results in perfectly calibrated LLRs on the set used for its training, this does not generalize to unseen scores. PAVA should be used only for the computation of the minimum ECE as discussed in Sections 2.4.2 and 2.5.1 and not for the calibration of unseen scores for rational decision-making.

The next section introduces a new idea for generative calibration by modeling the class-conditional densities of the scores using NF.

8.2 PROPOSED FLOW-BASED CALIBRATION OF LOG-LIKELIHOOD-RATIOS

In this section, a new generative calibration method is proposed. Let's consider a set of classes $\{C_1, C_2\}$ and a pattern recognizer that outputs a score in $\mathbb{R}$ which is low for class C_1 and large for class C_2. The aim of the calibration is to transform the scores produced by the recognizer into calibrated LLR. NF is used to model and transform the class-conditional densities of the scores into Gaussian densities respecting the constraint of Proposition 1 . In order to do so, a composition of mixtures of Cumulative Distribution Functions (CDF) and logit transformations is used to learn the monotonically increasing calibration function[4].

3 *Semi-supervised* refers to a set where the labels are not known for all samples.

4 The use of mixtures of CDF is highly inspired by Lecture 3 of the Spring 2020 University of California Berkeley course on Deep Unsupervised Learning: https://sites.google.com/view/berkeley-cs294-158-sp20/ (last visit in February 2023).

The CDF for a normal distribution (that we will call normal CDF) with mean μ and standard deviation σ is given by:

$$F_{CDF}(x \mid \mu, \sigma) = \int_{-\infty}^{x} f(t \mid \mu, \sigma)dt = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right), \tag{96}$$

where $f(\cdot \mid \mu, \sigma)$ is the probability density function of the normal distribution with mean μ and standard deviation σ, and erf is the error function: $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. Figure 33 shows the normal CDF for some couples of parameters. A normal CDF looks like a sigmoid function: it is a bijection between $\mathbb{R}$ and $]0,1[$. A mixture of $K \in \mathbb{N}^*$ CDFs is defined as follows:

$$F_{MCDF}(x \mid M, \Sigma, W) = \sum_{i=1}^{K} w_i F_{CDF}(x \mid \mu_i, \sigma_i), \tag{97}$$

where $M = \{\mu_i \in \mathbb{R}\}_{i\in[\![1,K]\!]}$ is the set of means, $\Sigma = \{\sigma_i \in \mathbb{R}^{*+}\}_{i\in[\![1,K]\!]}$ is the set of standard deviations, and $W = \{w_i \in \mathbb{R}^{*+}; \sum_{j=1}^{K} w_j = 1\}_{i\in[\![1,K]\!]}$ is the set of weights.

A mixture of CDFs defines a wider family of bijections between $\mathbb{R}$ and $]0,1[$. An arbitrary example of a mixture of three normal CDFs is shown in Figure 34.

For the calibration of scores into LLRs, the transformation has to be a bijection between $\mathbb{R}$ and $\mathbb{R}$. In order to do so, the mixture of CDFs is composed with a logit function to obtain the following bijection:

$$B(x \mid M, \Sigma, W) = \operatorname{logit}(F_{MCDF}(x \mid M, \Sigma, W)). \tag{98}$$

Several of these bijections can be composed to build an even wider family of transformations as follows[5]:

$$F_{FLOWCAL}(x \mid \theta) = \bigcirc_{i=1}^{N} B(x \mid M_i, \Sigma_i, W_i), \tag{99}$$

where $\theta = \{M_i, \Sigma_i, W_i\}_{i\in[\![1,N]\!]}$ is the set of all the parameters. Given the following target class-conditional densities of calibrated LLR (see Proposition 1):

$$\begin{aligned} l \mid C_1 &\sim \mathcal{N}(-\mu, 2\mu), \\ l \mid C_2 &\sim \mathcal{N}(\mu, 2\mu), \end{aligned} \tag{100}$$

where $l = F_{FLOWCAL}(x \mid \theta)$, the parameters θ, and μ can be learned through negative log-likelihood minimization using automatic differentiation and gradient descent. Since

5 We define the notation $\bigcirc_{i=1}^{N} f_i(x)$ as the composition $f_N(f_{N-1}(\dots f_1(x)))$.

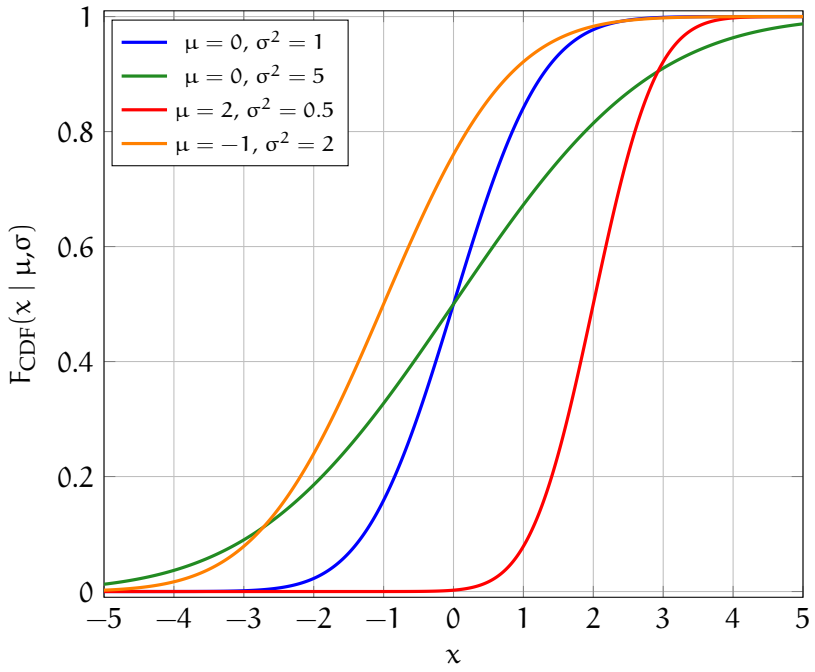


Figure 33: Normal cumulative distribution functions for some values of the parameters μ and σ.

the mapping is invertible, the exact log-likelihood of the data can be computed using the change of variable formula. The log-likelihood of a score x belonging to the class C_i is:

$$\log f_{X|C_i}(x \mid \theta, \mu) = \log f_{\mathcal{L}|C_i}(l \mid \mu) + \log \left| \frac{d}{dx} \left(F_{\text{FLOWCAL}}(x \mid \theta) \right) \right|, \tag{101}$$

see Appendix 10.9 for the computation of the derivative of $F_{\text{FLOWCAL}}(\cdot \mid \theta)$.

8.2.1 *A toy experiment*

Let's come back to the Circles example of Section 5.2.4. The good discrimination of the QDA's scores suggests that the QDA is extracting useful information about the classes. However, the ECE measure is not as low as the two classes CDA. In order to improve the representation quality of the information extracted by the QDA, the scores can be further calibrated. The flow-based calibration presented above will be here use to improve the calibration of the QDA's scores and will be briefly compared with the standard linear and quadratic calibrations. In order to do so, an additional set of 5000 samples sampled from the Circles distributions is used to train the calibration methods.

Figure 35 shows the histograms of the testing set of scores when the different calibrations are applied. The histograms of the original scores are showed in Figure 35a. They look almost Gaussian but are not symmetric as one would expect from the idempotence property constraint. As it is restricted only to scaling and shifting, the linear method is not enough to produce densities that would respect the idempotence constraint and the asymmetry of the histograms is not fixed (Figure 35b). The quadratic calibration does not result in Gaussian

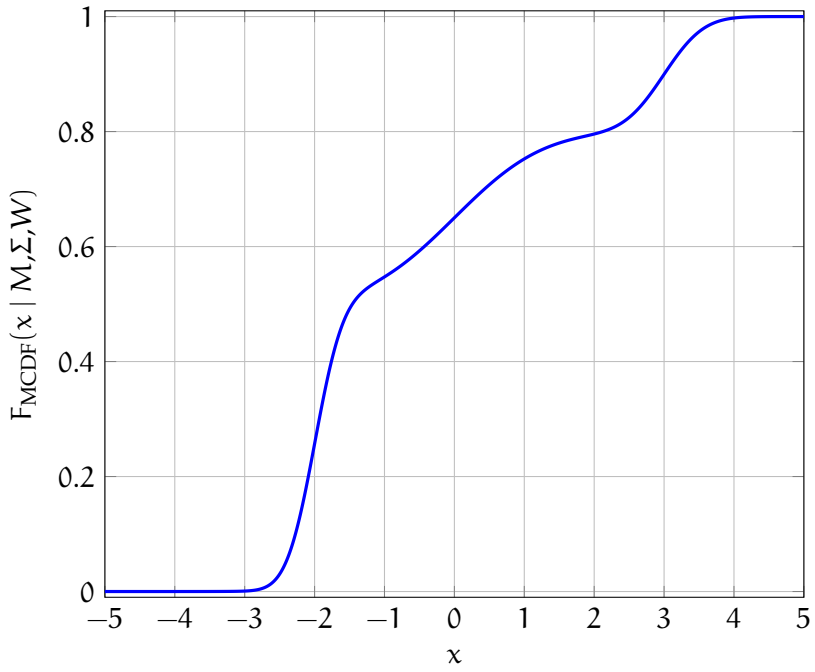
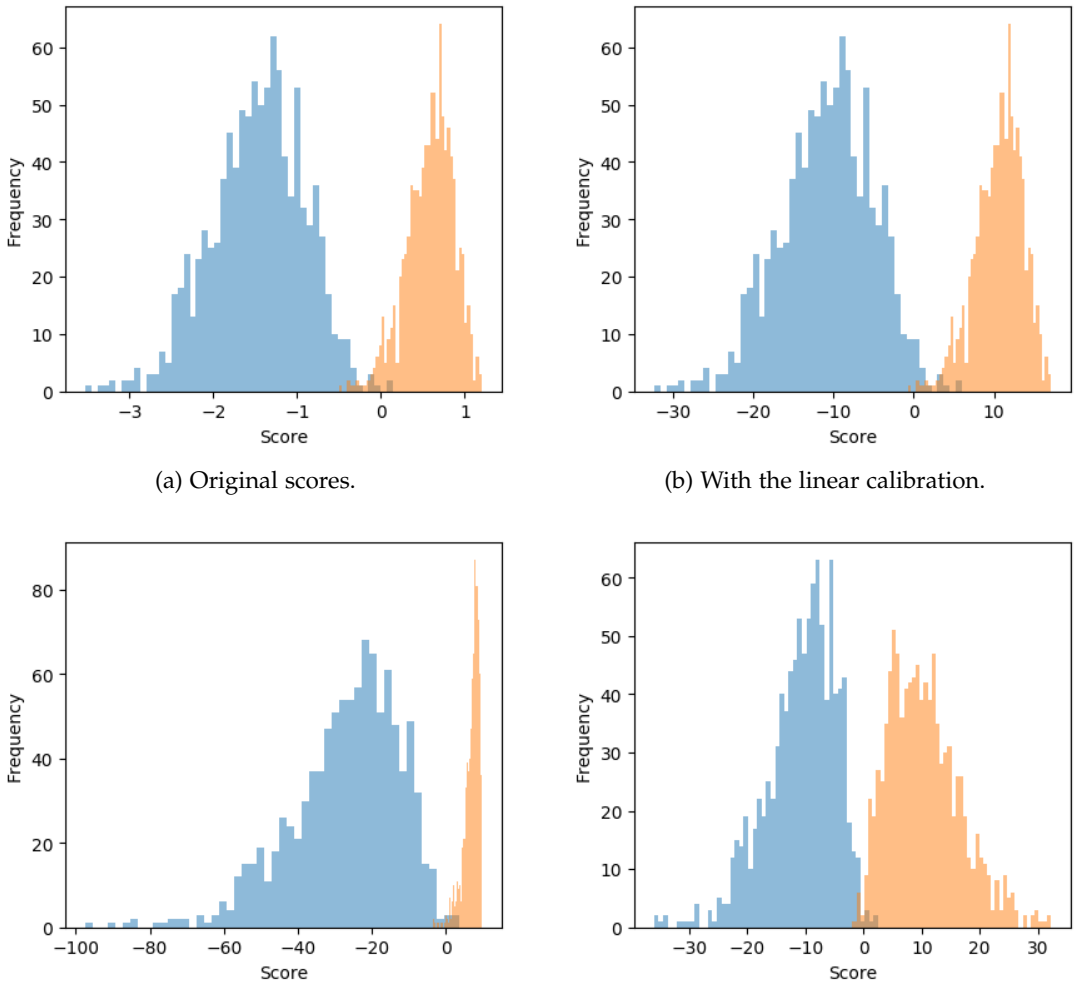


Figure 34: Example of a mixture of three normal CDFs with $M = \{0, 3, -2\}$, $\Sigma = \{1, \sqrt{0.2}, \sqrt{0.1}\}$ and $W = \{0.3, 0.2, 0.5\}$.

scores (Figure 35c). Since a quadratic transformation of normally distributed scores results in a generalized chi-squared distribution and that Proposition 1 concerns only normally distributed scores, these histograms can not be interpreted in terms of the idempotence property. The histograms of the flow-based calibrated scores look almost Gaussian and are symmetric around zero (Figure 35d). This symmetry is at least one requirement of Proposition 1.

Figure 36 shows the ECE values on the testing set for a wide range of prior π and for the different calibration methods. All the profiles with calibration (purple, red, green, and blue) are below the profile without (orange). All calibrations are therefore improving the quality of the LLR scores. The quadratic and flow-based calibrations perform similarly and both produce better scores in comparison to the linear calibration. Nonetheless, we can see that they differ significantly for some priors: for a logit prior between 0 and 3, the flow-based produces slightly better-calibrated scores. For other prior values, the quadratic calibration seems to perform better. The purple profile shows the ECE for a perfectly calibrated set of LLR given by the PAVA.

Figure 37 shows the learned calibration functions. Even if the resulting calibration performance in terms of ECE of the quadratic and the flow-based approaches are close, the calibration functions are significantly different. Indeed, the flow-based calibration transforms the scores into normally distributed calibrated scores while the quadratic approach does not have Gaussian target distributions. One drawback of the flow-based calibration is that, in practice, outputs of the mixtures of CDF very close to zero or one will cause numerical issues since the logit function goes to infinity at zero and one. A small value $\epsilon \approx 2.22_{10^{-16}}$ is therefore added or subtracted from the output values rounded to zero or one. This however lim-



(a) Original scores.

(b) With the linear calibration.

(c) With the quadratic calibration.

(d) with the flow-based calibration.

Figure 35: Effect of calibration on the QDA's LLR scores histograms from the Circles test set. 35a shows the scores without calibration, 35b, 35c, and 35d respectively show the scores' histograms with linear, quadratic, and flow-based calibration.

its the image domain of the flow-based calibration to $]-\text{logit}\,\epsilon, +\text{logit}\,\epsilon[\approx]-36.04, 36.04[$. This saturation can be seen on the left part of the blue curve in Figure 37. This is problematic because the tails of the Gaussian can saturate and it is even worst for scores with a low discrimination cost which corresponds to a large absolute value of the mean and standard deviation. Actually, the class-conditional densities of the scores in Figure 35d do not look so Gaussian, each class-conditional density is not symmetric and seems to be "pressed" on zero. We suspect that this is due to the bounded domain, especially with this example where the discriminant power of the scores is large. Therefore, we expect that the performance of the flow-based calibration can be improved by extending the bounds of the domain. We tried to add to the flow of transformations an affine transformation for this purpose. However, our experiments are for now not conclusive, and reporting these results in this thesis would be too hasty without further investigation.

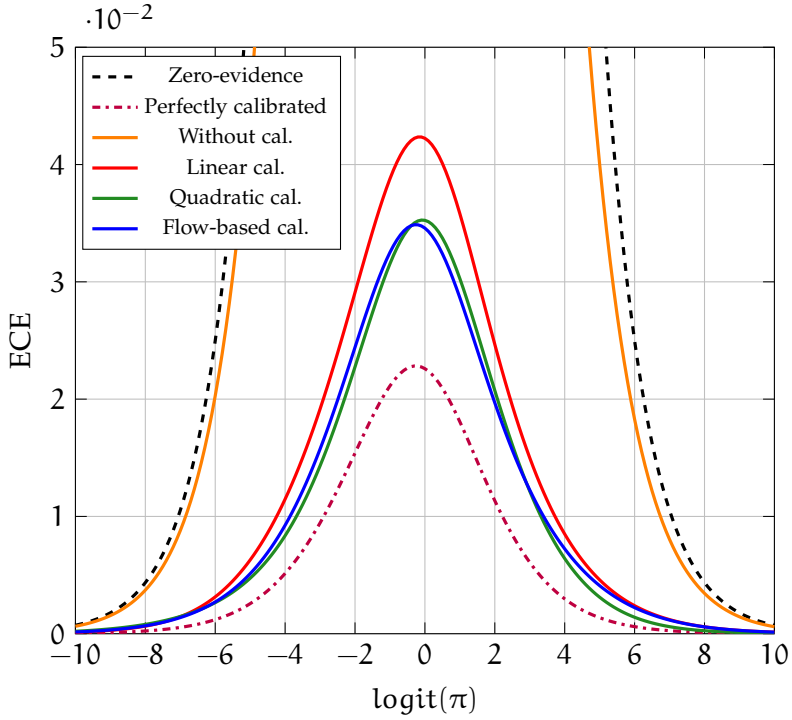


Figure 36: ECE plots with different calibrations on the QDA's scores from the test set of the Circles example.

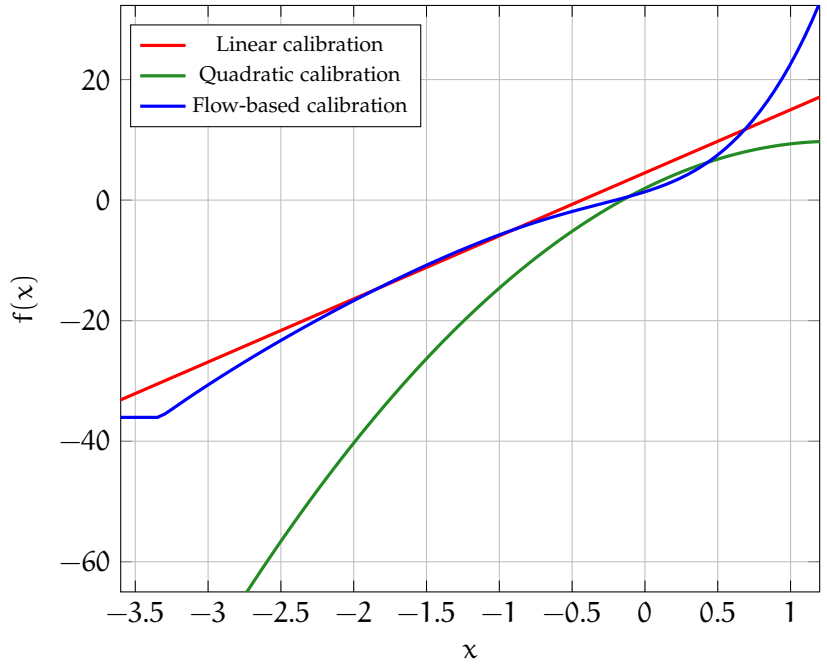


Figure 37: Linear, quadratic and flow-based calibration function trained on QDA's scores from the Circles dataset.

8.3 REGARDING THE CALIBRATION OF MULTICLASS LIKELIHOOD FUNCTIONS

The calibration of multiclass classifier scores has been mostly studied to produce probabilities rather than likelihoods [71, 86, 162]. Calibration of multiclass likelihood vectors and its difficulty has however been discussed in the context of language identification [27]. Chapter 3 discussed some advantages of expressing, in a multiclass context, the likelihood functions with the coordinate system given by the ILR transformation. This way of expressing a likelihood function is named Isometric-Log-Ratio-Likelihood (ILRL) and is considered in this thesis as the multiclass extension of the LLR. Chapter 6 discussed how the idempotence property of calibrated ILRL leads to a constraint on their distribution. Proposition 2 states that if one class-condition density of calibrated ILRL is Gaussian, the others are also Gaussian, and their parameters are all interdependent and can be expressed in terms of the Kullback-Leibler divergence within each pair of Gaussians. This is the multiclass formulation of the idempotence constraint on the LLRs' distributions (Proposition 1) used for the design of the above flow-based calibration. It is therefore tempting to simply extend the flow-based calibration to multiclass by learning a multidimensional flow-based calibration of ILRLs with the target densities respecting Proposition 2. This could be done, for instance, by training the CDA on ILRLs with a residual of dimension zero. However, the family of transformations learnable through NF is only restricted to diffeomorphisms and this is not sufficient for a *pure* calibration. Indeed, while in the two classes case the invertibility condition was enough to avoid "swappings" between scores, this is not the case anymore in the multiclass case. The concept of "swapping" is even not defined in the multiclass case. In the two classes case, the scores live on a one-dimensional line where the concept of order is well-defined and where a monotonically increasing function preserves the order of the scores. In the multiclass case, a "score" is not a scalar anymore but is a vector that lives in an infinite directional space where the concept of order is not defined. Using NF for the calibration of multiclass likelihood vectors remains therefore an open question.

8.4 SUMMARY

This chapter briefly discussed a new generative approach to LLR calibration. A 1-dimensional Normalizing Flow (NF), based on compositions of mixtures of normal CDFs and logit transformations, is used to transform the densities of the non-calibrated scores into Gaussian densities of calibrated LLR given by Proposition 1. However, in its current implementation, this flow-based calibration is limited due to some numerical issues and needs to be improved in the future.

It would be natural to extend this calibration to ILRL for multiclass cases. However, even if the invertibility constraint of 1-dimensional NF is enough to ensure—in the two classes case—that the calibration is *pure*, this is not the case anymore in the multiclass case. The application of flow-based calibration to ILRL remains therefore an open question.

Actually, NF could be used differently for calibrating LLRs. Here, we proposed to train a single NF model for the two classes. However, two models could be trained instead, one for

each class. No matter the shape of the base distributions, the role is here to fit each class-conditional density. The likelihood of a score for each class can still be computed through the change of variable formula. We hope to explore such an approach to calibration in the future. The reason why we preferred to first study the approach presented in this chapter is that it is conceptually the same approach as the discriminant analysis at the core of this thesis: it is the discriminant analysis with a residual dimension of zero.

9

CONCLUSION & DISCUSSION

9.1 SUMMARY AND FINDINGS

Starting from the idea of attribute privacy, i. e. the protection of a personal attribute in some data, this thesis addressed two main questions: how to properly represent the attribute-related information in an observation, and how to manipulate this information for privacy purposes?

Being restricted to attributes taking one value over a finite set of values, the knowledge of an attacker about the attribute is represented by a discrete probability distribution. Indeed, in the Bayesian interpretation of probability, a probability distribution represents a personal knowledge in presence of uncertainty. The Bayes' rule is the paradigm of information acquisition and updates a prior belief—i. e. the knowledge of the attacker before observing the data—into a posterior belief—i. e. the knowledge of the attacker considering all the available information, including the observed data— using the *likelihood function*.

This thesis proposed to represent the attribute-related information contained in the data—i. e. the *evidence*— by a *likelihood function*.

A likelihood function lives in a finite-dimensional vector space and since it is scale-invariant, it can be treated as compositional data. Thanks to the Aitchison geometry of the simplex, the probability distributions and the likelihood functions can be expressed in a Euclidean vector space where the Bayes' rule is a sum between the prior probability distribution and the likelihood function. Within this space and with the coordinate representation given by the Isometric-Log-Ratio (ILR) approach, the likelihood function is called Isometric-Log-Ratio-Likelihood (ILRL).

This thesis proposed to treat likelihood functions as compositional data and to use the ILRL as the multiple hypotheses and multidimensional extension of the Log-Likelihood-Ratio (LLR).

The LLR is also known in Bayesian updating as the weight-of-evidence. It has been widely used in forensic science and more generally in any Bayesian updating process with only two competing hypotheses. It expresses how the data is supporting one hypothesis against another and how the prior belief of the observer is changed into the posterior belief. The absolute value of the LLR is known as the *strength-of-evidence* and informs about the distance between the posterior and the prior. This can be seen as a measure of information disclosed by an observation. The multiple hypotheses extension of this concept is naturally given by

the Aitchison norm of the likelihood function. This notion of information is called *evidence information*.

For *perfect privacy*—the application of Claude Shannon's definition of *perfect secrecy* to privacy—the posterior belief of the attacker must remain equal to the prior belief. Hence, the evidence information must be equal to zero which corresponds to a zero ILRL.

> *Perfect privacy* is reached when the ILRL is zero for all observations: this is *zero-evidence*.

For the information to be well represented, the likelihood functions have to be calibrated. The notion of calibration is usually applied to probabilities but this thesis has been focused on the calibration of likelihood functions. The idempotence property of the ILRL ensures that all the sensitive information in the data is contained in the likelihood function. This ensures that the likelihood function is calibrated and leads to a constraint on the distributions of the ILRLs. This property has been known for LLR for decades but has been here extended to the ILRL for multiple hypotheses and multiclass applications.

> This thesis showed that the *idempotence* property, known for calibrated LLRs, holds naturally for ILRLs thanks to the compositional nature of the likelihood functions.

> As with the LLRs, the idempotence property of ILRLs leads to a constraint on their distributions. When, for one hypothesis or class, the ILRL is normally distributed, the ILRL is also normally distributed for each of the other classes with the same covariance matrix. The means are defined by the covariance matrix such that the latter is the only parameter of these distributions and can be expressed in terms of the separability between each Gaussian density.

Once it has been stated that the information related to an attribute can be represented by a likelihood function in the form of a calibrated ILRL respecting the idempotence property constaint, this thesis has been focused on how to model the data in order to manipulate these quantities and, consequently, to manipulate the information in the data. For this purpose:

> A new approach to discriminant analysis has been proposed. This is based on a generative model that models the data by mapping—in an invertible manner—the input vectors into a base space where the discriminant components are the likelihood function in the form of a calibrated ILRL. This approach is called Compositional Discriminant Analysis (CDA) since it benefits from the Euclidean vector space structure—given by the Aitchison geometry of the simplex—to express the likelihood function.

For privacy, since the mapping is invertible, the samples can be mapped back into the original feature space after setting the ILRL in the base space to zero which is consistent with the perfect privacy definition.

The CDA can be used for privacy by setting the estimated ILRLs to zero.

This has been tested for the concealment of the sex of the speaker in DNN-based speaker embeddings resulting in speaker representations that provides no evidence about the speaker's sex.

This transformation of speaker embeddings—combined with an affine transformation of the pitch trajectory—can be used in an analysis/synthesis pipeline for voice conversion for the concealment of the speaker's sex. The experiments based on automatic recognitions and on listening tests shows the ability of the approach to significantly reduce the sex-related information in the speech.

However, privacy is not the only possible application of this discriminant analysis. This can be used for any problem where the accurate modeling and/or manipulation of the information contained in the data is desirable. Many pattern recognition problems are concerned only with the discrimination power at the expense of the quality of the information representation i. e. the calibration. Even if the approach proposed in this thesis may have a comparatively low discrimination power, its focus on calibration and information quality makes it a good candidate for applications where humans have to objectively interpret and explain the model.

9.2 DISCUSSION

In contrast to the randomized mechanisms that have a theoretical protection guarantee formalized by the differential privacy framework and its variations, the protection given by the discriminant analysis-based approach presented in this thesis must be assessed empirically. Indeed, the protection capacity of the approach depends on the underlying ability of the generative model to model the data. However, our approach significantly differs from randomized mechanisms. For some kinds of complex data like speech, it is not obvious how and where to incorporate noise to alter only one attribute[1]. Hence, we focused instead on the accurate disentanglement of the attribute-related information in embeddings.

The thesis proposed to express the information *contained* in the data by a likelihood function. However, the information expressed by the likelihood function should rather be

1 A recent work [144] explored the use of differential private feature extractors in an analysis/synthesis pipeline for speaker anonymization purposes.

seen as the information *extracted* by the learned probabilistic model. This has been ignored in this thesis on the implicit assumption that the Normalizing Flow (NF) approach we used is sufficient to model the data properly.

Another limitation of the thesis is that it is restricted to attributes with a finite and relatively low number of possible values. The number of dimensions of the space of likelihood functions linearly increases with the number of possible values the attribute can take. Therefore, this number has to remain sensibly lower than the dimensionality of the data. Moreover, if the attribute of interest is continuous like many natural quantities like for instance, a height, a distance, an age, and so on, the likelihood function may live in an infinite dimensional space which is technically and conceptually challenging for traditional machine learning approaches.

Despite the above points, we consider that this thesis opens several research directions and opportunities:

- Representation learning [13] is the field of extracting from the data meaningful vectors of features, sometimes called *embeddings*, containing useful information for a downstream task like classification. To some extent, we believe that the idea of learning a representation in the form of a calibrated likelihood function[2] as presented in this thesis could be an interesting research direction for the field of representation learning. Indeed, likelihoods and probabilities have a statistical meaning and, when they are calibrated, are understandable by humans in contrast to many modern machine learning models that suffer from a lack of explainability of their outputs.

- Calibration methods for multiclass applications have been ignoring the compositional nature of probability distributions and likelihood functions. In this thesis, the idempotence property of multiple hypotheses likelihood functions—that naturally appears thanks to the Aitchison geometry of the simplex—has been studied and used for the design of a discrimination analysis that produces well-calibrated likelihood functions. We believe that, in the future, these results could help in the design of new calibration methods for multiclass applications.

- Calibrated probabilities and likelihoods are crucial for rational decision making and we have seen in this thesis how the Aitchison geometry of the simplex helps in treating them. However, costs are just as much crucial but have been ignored in this thesis since we were concerned only with the extraction and the modeling of the information in the data rather than the decision stage. Exploring if the Aitchison geometry of

2 This idea of representing the information of interest in the form of a likelihood function has already been encountered. In [25], *meta-embeddings* as a probabilistic generalization of embeddings have been proposed. Meta-embeddings are likelihood functions for a hidden variable of interest. However, the authors were concerned with a *continuous* speaker variable in the context of ASV in contrast to an attribute with a finite number of possible values. In their case—depending on how the variable is modeled—the likelihood function may live in a high, or even infinite dimensional space making it computationally challenging. To be practical, the infinite family of meta-embeddings has to be restricted. In our case, the variable of interest is not hidden and the likelihood function can be represented in a finite-dimensional vector space. This can therefore be used with standard supervised machine learning approaches designed so that the likelihood function is a part of the learned representation.

the simplex can help in the treatment of the costs is also a research direction we wish to explore in the future.

- While this thesis has been limited to supervised learning of information representations, unsupervised and semi-supervised training of the normalizing flow-based mapping between the data and the space of likelihood functions and residuals could be explored. Indeed, supervised training requires the data to be labeled while labels are not available in many applications.
- The use of the proposed discriminant analysis for attribute-privacy has been tested on real speech data only for the concealment of the speaker's sex. In the future, we wish to test this approach for the concealment of other attributes like the native language or the health state for example.
- Finally, we are interested in studying Bayes Hilbert spaces [157] which are generalizing the Aitchison geometry of the simplex to continuous probability distributions and continuous likelihood functions. This has been out of the scope of this thesis, but we wish to explore whether it can help to go beyond discrete attributes.

10

APPENDICES

10.1 SOME ADDITIONAL REMARKS ABOUT THE CALIBRATION-DISCRIMINATION DECOMPOSITION

The concept of calibration has been introduced for probabilistic forecasts. In [43], Morris H. DeGroot showed how the average PSR can be decomposed into a discrimination and a calibration term. The concept of calibration and the decomposition of the cross-entropy have been then formalized in the context of pattern recognition like for instance in Niko Brümmer's [19, 22] and Daniel Ramos' [135, 136] works. This has already been discussed in Section 2.4.2.

In order to help the unfamiliar reader in understanding the connection between the concept of calibration in probabilistic forecasting and in pattern recognition, this appendix gives additional details.

Once it has been observed whether it has rained or not, the forecaster's probabilities can be compared with the actual outcomes using PSR. In [43], the forecaster is allowed to choose a probability p within a finite set of allowable values. In this way, the relative frequency $f(p)$ of rainy days on which the probability p has been assigned is defined (this is like considering bins as discussed in Section 2.3). We here reformulate—in terms of the logarithmic scoring rule—what Morris H. DeGroot wrote with the Brier score in Section 4 of his paper [43].

Among those days on which the forecaster's probability of rain is p, the score will be $-\log p$ with relative frequency $f(p)$ and will be $-\log(1-p)$ with relative frequency $1-f(p)$. Let $\nu(p)$ be the relative frequency that the forecaster's probability is p for a day. The averaged scoring rule can be written as:

$$\begin{aligned}
&\sum_p \nu(p)\left[-f(p)\left(\log p\right)-\left(1-f(p)\right)\log(1-p)\right] \\
&\quad=\sum_p \nu(p)\left[-f(p)\log f(p)-\left(1-f(p)\right)\log(1-f(p))\right] \\
&\qquad+\sum_p \nu(p)\left[f(p)\log\frac{f(p)}{p}+\left(1-f(p)\right)\log\frac{1-f(p)}{1-p}\right] \\
&\quad=\sum_p \nu(p)H(f(p))+\sum_p \nu(p)D_{KL}(f(p),p).
\end{aligned} \tag{102}$$

The averaged Kullback-Leibler divergence (D_{KL}) term is the calibration loss. It measures how far the probabilities p are from $f(p)$, i. e. how far the bins' height is from $f(p)=p$ in Figure 3. The other term, the averaged entropy of $f(p)$, is the discrimination or refinement loss. Here, $f(p)$ plays the role of the *reference* probability distribution. In the context of

pattern recognition, the set of available values for p is not restricted to a finite discrete set. The set of possible values is continuous such that $f(p)$ can not be computed. The reference probability distribution has to be chosen differently: this is where PAVA steps in.

Let's summarise the nuances between the formulation just above and the formulation in the context of binary classification as discussed in Section 2.4.2. In both cases, an Empirical Cross-Entropy (ECE) is computed for assessing how good the set of probabilities is. $v(p)$ corresponds to $P(e)$ in Equation 18 and $f(p)$ corresponds to reference probability $P(H_i \mid e)$. In DeGroot's formulation, the decomposition appears because $f(p)$ can be computed and corresponds to the reference calibrated probabilities. In the pattern recognition formulation like in [136] and in Section 2.4.2, $P(e)$ and $P(H_i \mid e)$ are combined through the Bayes' rule resulting in the prior probability $P(H_i)$ and the likelihood $P(e \mid H_i)$. The latter is approximated by $\frac{1}{|\mathcal{E}_i|}$ where $|\mathcal{E}_i|$ is the number of observations belonging to class i. The reference probability $P(H_i \mid e)$ does not appear anymore in this formulation of the ECE (Equations 19 and 20). The probabilities q are then chosen to be the reference probabilities that minimize the ECE without changing the discrimination ability. These reference probabilities are given by the Pool Adjacent Violators Algorithm (PAVA).

10.2 GENERAL FORMULA FOR THE ILR COMPONENTS

The set $\{\boldsymbol{e}^{(i)} \in \mathcal{S}^N, i = 1, \dots N-1\}$ is the Aitchison orthonormal basis obtained by the Gram-Schmidt procedure [56] where for $k = 1, \dots N-1$:

$$\boldsymbol{e}^{(k)} = \mathcal{C}\left[\exp\left(\underbrace{\sqrt{\frac{1}{k(k+1)}}, \dots \sqrt{\frac{1}{k(k+1)}}}_{k \text{ elements}}, -\sqrt{\frac{k}{k+1}}, 0, \dots 0\right)\right] \tag{103}$$

forms an Aitchison orthonormal basis on the $(N-1)$-simplex $\mathcal{S}^N$ [56].

The isometric log-ratio transformation (ILR) of a vector $\mathbf{p} \in \mathcal{S}^N$ is defined as $\tilde{\mathbf{p}} = \text{ilr}(\mathbf{p}) = \left[\langle \mathbf{p}, \boldsymbol{e}^{(1)} \rangle_a, \dots \langle \mathbf{p}, \boldsymbol{e}^{(N-1)} \rangle_a\right]$. This is an isometric isomorphism between $\mathcal{S}^N$ and $\mathbb{R}^{N-1}$. The kth component of the ILR image of the vector $\mathbf{p}$ is given by the following formula:

$$\tilde{p}_k = \langle \mathbf{p}, \boldsymbol{e}^{(k)} \rangle_a = \frac{1}{\sqrt{k(k+1)}} \log\left(\frac{\prod_{i=1}^{k} p_i}{(p_{k+1})^k}\right) \tag{104}$$

Proof.

$$\begin{aligned}
\langle \mathbf{p}, \boldsymbol{e}^{(k)} \rangle_a &= \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \log\frac{p_i}{p_j} \log\frac{e_i^{(k)}}{e_j^{(k)}} \qquad \text{(application of the inner product)}, \\
&= \frac{1}{2N} \sum_{i=1}^{N} \left(\sum_{j=1}^{k} \log\frac{p_i}{p_j} \log\frac{e_i^{(k)}}{\exp\left(\sqrt{\frac{1}{k(k+1)}}\right)} + \log\frac{p_i}{p_{k+1}} \log\frac{e_i^{(k)}}{\exp\left(-\sqrt{\frac{k}{k+1}}\right)} \right. \\
&\quad \left. + \sum_{j=k+2}^{N} \log\frac{p_i}{p_j} \log\frac{e_i^{(k)}}{1} \right), \\
&= \frac{1}{2N} \left(\sum_{i=1}^{k} \left(\sum_{j=1}^{k} \log\frac{p_i}{p_j} \log 1 + \log\frac{p_i}{p_{k+1}} \log\frac{\exp\left(\sqrt{\frac{1}{k(k+1)}}\right)}{\exp\left(-\sqrt{\frac{k}{k+1}}\right)} \right. \right. \\
&\quad \left. + \sum_{j=k+2}^{N} \log\frac{p_i}{p_j} \log\exp\left(\sqrt{\frac{1}{k(k+1)}}\right) \right) + \sum_{j=1}^{k} \log\frac{p_{k+1}}{p_j} \log\frac{\exp\left(-\sqrt{\frac{k}{k+1}}\right)}{\exp\left(\sqrt{\frac{1}{k(k+1)}}\right)} \\
&\quad + \log 1 \log 1 + \sum_{j=k+2}^{N} \log\frac{p_{k+1}}{p_j} \log\exp\left(-\sqrt{\frac{k}{k+1}}\right) \\
&\quad + \sum_{i=k+2}^{N} \left(\sum_{j=1}^{k} \log\frac{p_i}{p_j} \log\frac{1}{\exp\left(\sqrt{\frac{1}{k(k+1)}}\right)} + \log\frac{p_i}{p_{k+1}} \log\frac{1}{\exp\left(-\sqrt{\frac{k}{k+1}}\right)} \right. \\
&\quad \left. \left. + \sum_{j=k+2}^{N} \log\frac{p_i}{p_j} \log 1 \right) \right),
\end{aligned}$$

(105)

$$
\begin{aligned}
\langle \mathbf{p}, \boldsymbol{e}^{(k)} \rangle_a &= \frac{1}{2N}\Bigg(\sum_{i=1}^{k} \Bigg(\sqrt{\frac{k+1}{k}} \log \frac{p_i}{p_{k+1}} + \sum_{j=k+2}^{N} \sqrt{\frac{1}{k(k+1)}} \log \frac{p_i}{p_j} \Bigg) \\
&\quad + \sum_{j=1}^{k} \sqrt{\frac{k+1}{k}} \log \frac{p_j}{p_{k+1}} + \sum_{j=k+2}^{N} \sqrt{\frac{k}{k+1}} \log \frac{p_j}{p_{k+1}} \\
&\quad + \sum_{i=k+2}^{N} \Bigg(\sum_{j=1}^{k} \sqrt{\frac{1}{k(k+1)}} \log \frac{p_j}{p_i} + \sqrt{\frac{k}{k+1}} \log \frac{p_i}{p_{k+1}} \Bigg) \Bigg), \\
&= \frac{1}{2N}\Bigg(2\sqrt{\frac{k+1}{k}} \sum_{i=1}^{k} \log \frac{p_i}{p_{k+1}} + 2\sqrt{\frac{1}{k(k+1)}} \sum_{i=1}^{k} \sum_{j=k+2}^{N} \log \frac{p_i}{p_j} \\
&\quad + 2\sqrt{\frac{k}{k+1}} \sum_{i=k+2}^{N} \log \frac{p_i}{p_{k+1}} \Bigg), \\
&= \frac{1}{N}\Bigg(\sqrt{\frac{k+1}{k}} \Big(\sum_{i=1}^{k} \log p_i - k \log p_{k+1} \Big) \\
&\quad + \frac{1}{\sqrt{k(k+1)}} \Big((N-k-1) \sum_{i=1}^{k} \log p_i - k \sum_{j=k+2}^{N} \log p_j \Big) \\
&\quad + \sqrt{\frac{k}{k+1}} \Big(\sum_{i=k+2}^{N} \log p_i - (N-k-1) \log p_{k+1} \Big) \Bigg), \\
&= \frac{1}{N}\Bigg(\Bigg(\sqrt{\frac{k+1}{k}} + \frac{N-k-1}{\sqrt{k(k+1)}} \Bigg) \sum_{i=1}^{k} \log p_i \\
&\quad - \Bigg(k\sqrt{\frac{k+1}{k}} + (N-k-1)\sqrt{\frac{k}{k+1}} \Bigg) \log p_{k+1} \Bigg), \\
&= \frac{1}{N}\Bigg(\frac{N}{\sqrt{k(k+1)}} \sum_{i=1}^{k} \log p_i - N\sqrt{\frac{k}{k+1}} \log p_{k+1} \Bigg), \\
&= \frac{1}{\sqrt{k(k+1)}} \Bigg(\log \prod_{i=1}^{k} p_i - k \log p_{k+1} \Bigg), \\
&= \frac{1}{\sqrt{k(k+1)}} \log \Bigg(\frac{\prod_{i=1}^{k} p_i}{(p_{k+1})^k} \Bigg).
\end{aligned}
\tag{106}
$$

□

10.3 THREE HYPOTHESIS MAXIMUM PROBABILITY DECISION REGIONS ON THE AITCHISON SIMPLEX

In this section, we will see how the maximum probability decision boundaries are derived. Let's consider the set of three possible hypotheses $\{H_1, H_2, H_3\}$, the vector of probabilities over the set of hypotheses $\mathbf{P} = [p_1, p_2, p_3]$ and its isometric log-ratio transformation:

$$\tilde{\mathbf{P}} = \left[\frac{1}{\sqrt{2}} \log \frac{p_1}{p_2}, \frac{1}{\sqrt{6}} \log \frac{p_1 p_2}{p_3^2}\right] = [\tilde{p}_1, \tilde{p}_2]\,. \tag{107}$$

Decision region for the first hypothesis

The probability for the hypothesis H_1 is the maximum one if:

$$\begin{aligned}
&(p_1 > p_2) \wedge (p_1 > p_3) \\
&\iff \left(\log p_1 > \log p_2\right) \wedge \left(\left(\log (p_1 p_2) > \log (p_2 p_2)\right) \wedge \left(\log p_3^2 < \log p_1^2\right)\right), \\
&\iff \left(\log \frac{p_1}{p_2} > 0\right) \wedge \left(\left(\log (p_1 p_2) > \log (p_2 p_2)\right) \wedge \left(-\log p_3^2 > -\log p_1^2\right)\right), \\
&\iff \left(\tilde{p}_1 > 0\right) \wedge \left(\log \frac{p_1 p_2}{p_3^2} > \log \frac{p_2}{p_1}\right), \\
&\iff \left(\tilde{p}_1 > 0\right) \wedge \left(\tilde{p}_2 > -\frac{1}{\sqrt{3}} \tilde{p}_1\right).
\end{aligned} \tag{108}$$

Figure 38a shows in red the region where the proposition $\tilde{p}_1 > 0$ is true and in orange the region where $\tilde{p}_2 > -\frac{1}{\sqrt{3}}\tilde{p}_1$ is true. The intersection, marked out by the dashed blue lines, is the maximum probability decision region for hypothesis H_1.

Decision region for the second hypothesis

The probability for the hypothesis H_2 is the maximum one if:

$$\begin{aligned}
&(p_2 > p_1) \wedge (p_2 > p_3) \\
&\iff \left(\log p_2 > \log p_1\right) \wedge \left(\left(\log(p_1 p_2) > \log(p_1 p_1)\right) \wedge \left(\log p_3^2 < \log p_2^2\right)\right), \\
&\iff \left(\log \frac{p_2}{p_1} > 0\right) \wedge \left(\left(\log(p_1 p_2) > \log(p_1 p_1)\right) \wedge \left(-\log p_3^2 > -\log p_2^2\right)\right), \\
&\iff \left(\tilde{p}_1 < 0\right) \wedge \left(\log \frac{p_1 p_2}{p_3^2} > \log \frac{p_1}{p_2}\right), \\
&\iff \left(\tilde{p}_1 < 0\right) \wedge \left(\tilde{p}_2 > \frac{1}{\sqrt{3}} \tilde{p}_1\right).
\end{aligned} \tag{109}$$

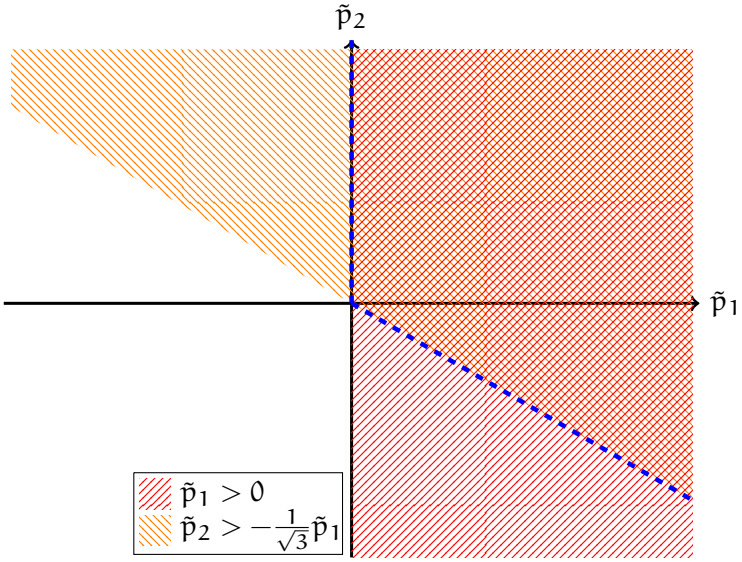
Figure 38b shows in red the region where the proposition $\tilde{p}_1 < 0$ is true and in orange the region where $\tilde{p}_2 > \frac{1}{\sqrt{3}}\tilde{p}_1$ is true. The intersection, marked out by the dashed blue lines, is the maximum probability decision region for hypothesis H_2.

Decision region for the third hypothesis

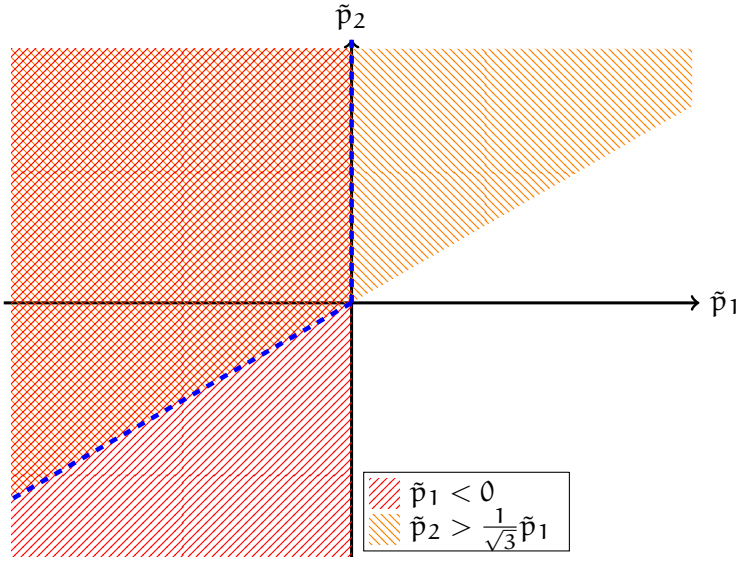
The probability for the hypothesis H_3 is the maximum one if:

$$\begin{aligned}
&(p_3 > p_1) \wedge (p_3 > p_2) \\
&\iff \left(-\log p_3^2 < -\log p_1^2\right) \wedge \left(-\log p_3^2 < -\log p_2^2\right), \\
&\iff \left(\log \frac{p_1 p_2}{p_3^2} < \log \frac{p_1 p_2}{p_1^2}\right) \wedge \left(\log \frac{p_1 p_2}{p_3^2} < \log \frac{p_1 p_2}{p_2^2}\right), \\
&\iff \left(\tilde{p}_2 < -\frac{1}{\sqrt{3}} \tilde{p}_1\right) \wedge \left(\tilde{p}_2 < \frac{1}{\sqrt{3}} \tilde{p}_1\right).
\end{aligned} \tag{110}$$

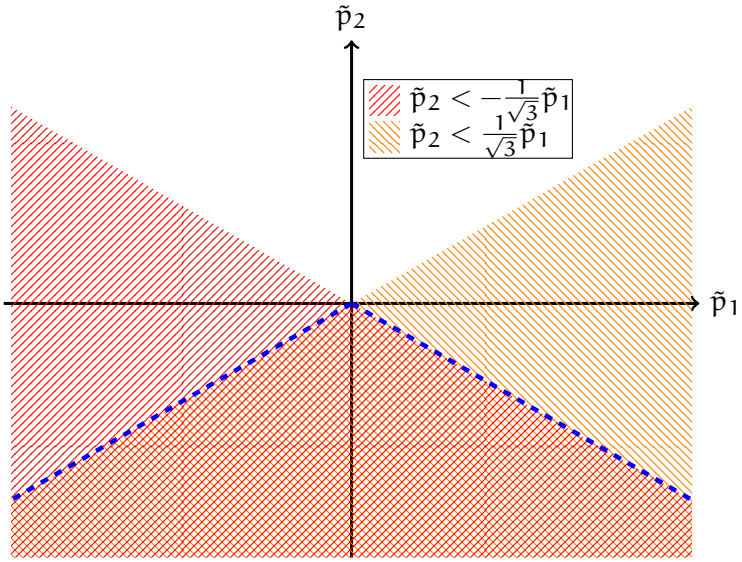
Figure 38c shows in red the region where the proposition $\tilde{p}_2 < -\frac{1}{\sqrt{3}}\tilde{p}_1$ is true and in orange the region where $\tilde{p}_2 < \frac{1}{\sqrt{3}}\tilde{p}_1$ is true. The intersection, marked out by the dashed blue lines, is the maximum probability decision region for hypothesis H_3.



(a) Maximum probability decision region for hypothesis H_1.



(b) Maximum probability decision region for hypothesis H_2.



(c) Maximum probability decision region for hypothesis H_3.

Figure 38: Maximum probability decision regions in a three hypotheses case.

10.4 MAXIMUM LIKELIHOOD ESTIMATOR OF μ

In this section, we derive the maximum likelihood estimator presented in Equation 76. The log-likelihood of a batch $\mathcal{B}_\mathcal{Z} = \{(z^{(1)}, C^{(1)}), \dots (z^{(|\mathcal{B}_\mathcal{Z}|)}, C^{(|\mathcal{B}_\mathcal{Z}|)})\}$ in the base space is:

$$
\begin{aligned}
&\log f(\mathcal{B}_\mathcal{Z} \mid \mu) \\
&= \sum_{i=1}^{2} \left(\sum_{\substack{(z,C)\in\mathcal{B}_\mathcal{Z} \\ C=C_i}} \log \left(\frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \left(z + (-1)^i \mu \boldsymbol{e}_1 \right)^\top \boldsymbol{\Sigma}^{-1} \left(z + (-1)^i \mu \boldsymbol{e}_1 \right) \right) \right) \right), \\
&= \sum_{i=1}^{2} \left(\sum_{\substack{(z,C)\in\mathcal{B}_\mathcal{Z} \\ C=C_i}} \left(-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \left(z + (-1)^i \mu \boldsymbol{e}_1 \right)^\top \boldsymbol{\Sigma}^{-1} \left(z + (-1)^i \mu \boldsymbol{e}_1 \right) \right) \right), \\
&= -\frac{d\,|\mathcal{B}_\mathcal{Z}|}{2} \log 2\pi - \frac{|\mathcal{B}_\mathcal{Z}|}{2} \log 2\mu \\
&\quad - \frac{1}{2} \sum_{i=1}^{2} \sum_{\substack{(z,C)\in\mathcal{B}_\mathcal{Z} \\ C=C_i}} \left(z^\top \boldsymbol{\Sigma}^{-1} z + 2(-1)^i \mu z^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_1 + (-1)^{2i} \mu^2 \boldsymbol{e}_1 \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_1 \right), \\
&= -\frac{d\,|\mathcal{B}_\mathcal{Z}|}{2} \log 2\pi - \frac{|\mathcal{B}_\mathcal{Z}|}{2} \log 2\mu - \frac{1}{2} \sum_{i=1}^{2} \sum_{\substack{(z,C)\in\mathcal{B}_\mathcal{Z} \\ C=C_i}} \left(\frac{z_1^2}{2\mu} + (-1)^i z_1 + \frac{\mu}{2} \right).
\end{aligned}
\tag{111}
$$

Setting the first derivative to zero:

$$
\begin{aligned}
\frac{\partial \log f(\mathcal{B}_\mathcal{Z} \mid \mu)}{\partial \mu} = 0 &\iff -\frac{|\mathcal{B}_\mathcal{Z}|}{2\mu} - \frac{|\mathcal{B}_\mathcal{Z}|}{4} + \frac{1}{4\mu^2} \sum_{z\in\mathcal{B}_\mathcal{Z}} z_1^2 = 0, \\
&\iff |\mathcal{B}_\mathcal{Z}|\,\mu^2 + 2\,|\mathcal{B}_\mathcal{Z}|\,\mu - \sum_{z\in\mathcal{B}_\mathcal{Z}} z_1^2 = 0, \\
&\iff \mu = -1 \pm \sqrt{1 + \frac{1}{|\mathcal{B}_\mathcal{Z}|} \sum_{z\in\mathcal{B}_\mathcal{Z}} z_1^2}.
\end{aligned}
\tag{112}
$$

Since μ is positive

$$
\frac{\partial \log f(\mathcal{B}_\mathcal{Z} \mid \mu)}{\partial \mu} = 0 \iff \mu = -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_\mathcal{Z}|} \sum_{z\in\mathcal{B}_\mathcal{Z}} z_1^2}. \tag{113}
$$

This extremum is a maximum if the second derivative of the log-likelihood function at this point is negative:

$$\begin{aligned}
\frac{\partial^2 \log f(\mathcal{B}_Z \mid \mu)}{\partial \mu^2} < 0 &\iff \frac{|\mathcal{B}_Z|}{2\mu^2} - \frac{1}{2\mu^3} \sum_{z \in \mathcal{B}_Z} z_1^2 < 0, \\
&\iff \mu < \frac{1}{|\mathcal{B}_Z|} \sum_{z \in \mathcal{B}_Z} z_1^2, \\
&\iff -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_Z|} \sum_{z \in \mathcal{B}_Z} z_1^2} < \frac{1}{|\mathcal{B}_Z|} \sum_{z \in \mathcal{B}_Z} z_1^2 \\
&\qquad \text{at the extremum } \mu = -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_Z|} \sum_{z \in \mathcal{B}_Z} z_1^2},
\end{aligned} \tag{114}$$

which is true since $x + 1 - \sqrt{1+x} > 0$ for $x > 0$. Therefore:

$$\hat{\mu}_{\text{MLE}}(\mathcal{B}_Z) = -1 + \sqrt{1 + \frac{1}{|\mathcal{B}_Z|} \sum_{z \in \mathcal{B}_Z} z_1^2}. \tag{115}$$

10.5 VOICE SIMILARITY MATRICES

In the context of the VoicePrivacy Challenge [154], we proposed in [118, 121] the use of voice similarity matrices to visualize potential differences in protection performance between speakers. In addition, this work proposed two summarizing metrics to assess the global *de-identification* and *voice distinctiveness* preservation of a protection system. In this section, we review some basic elements of these publications.

De-identification and voice distinctiveness preservation

This work has been done in the context of the VoicePrivacy Challenge where *pseudonymization* is the idea to use a voice conversion system in order to protect the identity of the speakers while keeping the linguistic content unchanged. Both *de-identification* and *voice-distinctiveness* are desirable criteria. *De-identification* [10, 79, 80] refers here to the concealment of the speaker's identity so that it is not possible to recover who is speaking in protected speech utterances. This term is also referred to as voice-disguise [74, 164] or identity masking [133]. However, it is likely that the full concealment of identity is impossible in practice. One reason is that identity information can not be explicitly disentangled from other speech attributes that must be preserved like the speech content. Accordingly, de-identification must be viewed as a process that increases the uncertainty in the linkability between a speech utterance and the corresponding speaker identity.

Depending upon the application, it can be desirable that protected voices remain consistent and distinguishable, i. e. in the protected domain, segments uttered by the same speaker are mutually linkable but distinct from protected segments uttered by another speaker. This requirement is named *voice distinctiveness* preservation.

Figure 39 illustrates a scenario in which the voice distinctiveness preservation requirement is paramount. Consider three speakers who communicate using a teleconferencing system and wish to not disclose their identities. De-identification systems can be used to conceal identity but might produce three almost identical or indistinguishable voices as illustrated by the upper-right situation in Figure 39. This may result in a confusing and unnatural conversation. Voice distinctiveness preservation allows for a comfortable, natural conversation by ensuring that the three protected voices remain distinguishable as per the bottom-right situation in Figure 39. In this case, a *pseudo-voice* is assigned to each speaker. It is distinct from the others and from the original speaker's voice.

Voice similarity matrix

An element of a voice similarity matrix gives a measure of similarity between two speakers. The similarity is computed with LLRs obtained from perfectly calibrated scores[1] given by the one-by-one comparisons of the speech segments using a biometric classifier. Let `tar` refer to the *target* hypothesis and let `nontar` refer to the *non-target* hypothesis (see Section

1 Using PAVA-based calibration.

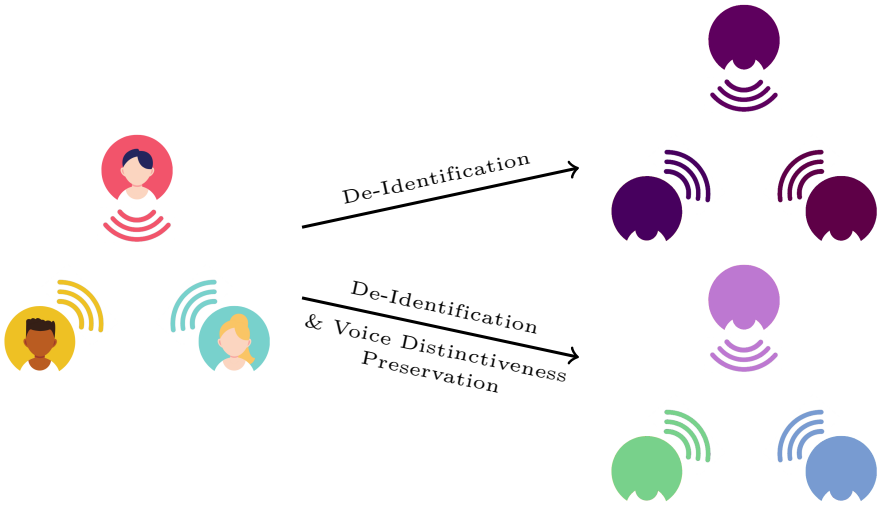


Figure 39: Illustration of the *de-identification* alone and together with the *voice distinctiveness* preservation. Using both results in the concealment of the speakers' identity while avoiding confusion between the protected voices.

2.2). To be more precise, a voice similarity matrix M is defined as $M = \{\mathrm{Sim}(i,j)\}_{1 \leqslant i,j \leqslant N}$ where N is the number of speakers in the set to protect and the similarity $\mathrm{Sim}(i,j)$ is the geometric mean of the posteriors, assuming a uniform prior $P(tar) = P(nontar) = 0.5$, and is computed as:

$$\mathrm{Sim}(i,j) = \prod_{l \in \mathcal{L}_{i,j}} \left(\sigma(l)^{\frac{1}{|\mathcal{L}_{i,j}|}} \right), \tag{116}$$

where $\sigma(l) = \frac{1}{1+\exp(-l)}$ and $\mathcal{L}_{i,j}$ is the set of LLRs from the biometric comparisons of all segments from speaker i with all segments from speaker j. When computing the similarity of a speaker with itself (self-similarity), scores from the comparison of identical segments are removed from the mean in order to avoid overestimated similarities. When computing the voice similarity matrix *within a* set, it might be surprising to not have all self-similarities equal to a maximum value one like in a standard confusion matrix. However, a self-similarity has to be seen as the reflection of the ASV's behavior on a speaker rather than a measure compatible with a distance between a speaker and itself: if a speaker has a small self-similarity, it suggests that she or he behaves like a goat facing the ASV system [49]. Thus, a diagonal value is related to the strength of the intra-speaker variability [6]. Actually, applying $-\log_2$ to Equation 116 gives:

$$-\log_2(\mathrm{Sim}(i,j)) = -\frac{1}{|\mathcal{L}_{i,j}|} \sum_{l \in \mathcal{L}_{i,j}} \log_2 \sigma(l). \tag{117}$$

One may notice the resemblance with the Empirical Cross-Entropy (ECE) of Equation 19. Despite that, $-\log_2$ is not applied to the similarities in order to keep the values bounded between zero and one which facilitates the visualization of the matrices.

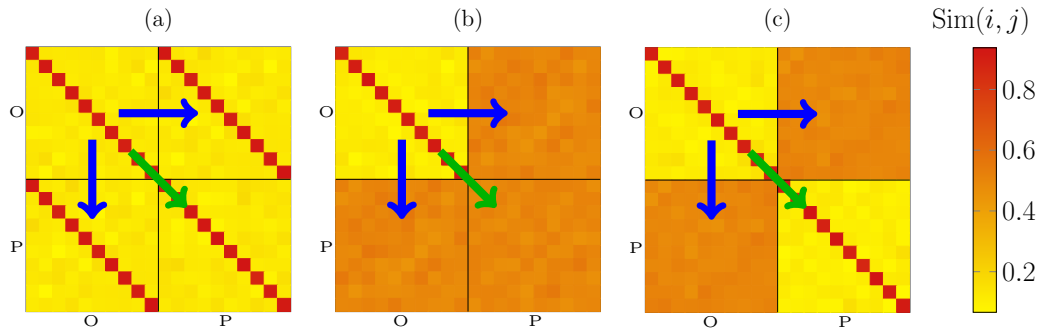


Figure 40: Three artificial examples of similarity matrices for pseudonymization assessment. For each case (a), (b), and (c), the upper-left matrix is M_{OO}, the upper-right and lower-left are M_{OP} (actually the lower-left matrix is $M_{PO} = M_{OP}^{\top}$ but as the biometric classifier is symmetric, the similarity is symmetric, and $M_{PO} = M_{OP}$) whereas the lower-right is M_{PP}.

Applying a pseudonymization system on a set of speech utterances, we call *original* set (O), results in a *protected* set (P). Three different voice similarity matrices can therefore be computed:

- M_{OO} which returns the speakers' similarities *within* the original set of speech utterances,
- M_{OP} which returns the speakers' similarities *between* the original and protected sets,
- M_{PP} which returns the speakers' similarities *within* the protected set.

The next section presents how these matrices are used to measure the level of de-identification and voice distinctiveness preservation of a pseudonymization system.

Assessing pseudonymization with the voice similarity matrices

In a voice similarity matrix, the emergence of the diagonal elements reflects how well an ASV system behaves when the two inputs of the biometric classifier may or may not be protected. Especially, in the case of M_{OP}, it shows if and which speaker could be identified from its protected segments. On the other hand, the diagonal elements of M_{PP} show if the speakers have low intra-speaker variability in the protected space, and the off-diagonal elements show similarities between the different pseudo-voices. Therefore, the presence of a diagonal in M_{PP} suggests that each speaker has, in the protected domain, a consistent pseudo-voice that does not confuse with the others. In Figure 40, cases (b) and (c) are examples of good de-identification (good protection) as the diagonal disappears in M_{OP} while case (a) is an example of poor de-identification. Cases (a) and (c) are examples where the voice distinctiveness is preserved as the diagonal remains in M_{PP}. Hence, case (c) respects two of the pseudonymization's requirements: both de-identification and voice distinctiveness preservation are fulfilled. Thus, these pseudonymization's requirements can be

measured by comparing the *dominance of the diagonal* between the matrices. The diagonal's dominance of a matrix M is defined as:

$$D_{diag}(M) = \left| \left(\sum_{1 \leqslant i \leqslant N} \frac{Sim(i,i)}{N} \right) - \left(\sum_{\substack{1 \leqslant j \leqslant N \\ 1 \leqslant k \leqslant N \\ j \neq k}} \frac{Sim(j,k)}{N(N-1)} \right) \right|. \tag{118}$$

In this way, the diagonal's dominance is equal to zero for a uniform matrix and one for the identity matrix, and for a matrix where all diagonal elements are zero and off-diagonal elements are one. Indeed, we consider that a scenario where a small group of segments is gathered as being from the same speaker if corresponding target trials result in high scores and non-target trials result in low scores and the scenario where a small group of segments is gathered as being from the same speaker if the target trials result in low scores and non-target trials all result in high scores are equivalent to the same risk of identification. However, if this property is appropriate for M_{OP} it might be an inconvenience for M_{PP} if a privacy safeguard results in a M_{PP} where diagonal values are low and off-diagonal values are high and close together.

The de-identification DeID measures how much the diagonal disappears from M_{OO} to M_{OP} as illustrated by the blue arrows in Figure 40 and is computed, assuming that $D_{diag}(M_{OO}) \neq 0$, as follows:

$$DeID = 1 - \frac{D_{diag}(M_{OP})}{D_{diag}(M_{OO})}. \tag{119}$$

DeID = 100% is the perfect de-identification while DeID = 0% corresponds to a system that achieves no de-identification or at least that the resulting privacy is not better than the initial one. Indeed, the initial privacy corresponds here to the limited capacity of an ASV system to discriminate the original voices. To help illustrate this, we consider a pair of twins who have exactly the same voice such that the ASV system can not distinguish between them. In this closed case, privacy is already high. Therefore, we are here only interested in a *relative* measure of privacy improvement rather than an absolute level of privacy.

The voice distinctiveness preservation is how much the diagonal remains from M_{OO} to M_{PP} as illustrated by the green arrows in Figure 40. Because a privacy safeguard can result in either a loss or an increase of voice distinctiveness, the voice distinctiveness preservation is computed as a gain of diagonal dominance:

$$G_{VD} = 10 \log_{10} \left(\frac{D_{diag}(M_{PP})}{D_{diag}(M_{OO})} \right), \tag{120}$$

such that a gain equal to zero means that the voice distinctiveness remains the same on average, and a gain above or below zero corresponds respectively to an increase or a loss of the global voice distinctiveness.

As we assumed that a privacy safeguard will never result in less privacy, a percentage of de-identification in Equation 119 is more suitable than a gain like G_{VD}[2]. Both are global metrics that summarize the performance over the set of speakers. The next section motivates how the matrices can be used for assessing the system speaker by speaker.

A zoo tour: an investigation at the speaker level

An ASV system does not perform equally on all speakers. To better interpret the ASV behavior across different speakers, George Doddington proposed to categorize speakers into four different "animals" [49]. This idea has been extended to other categories and the zoo plot has been proposed to visualize the performance of a biometric system across different users [51]. This section deals with the use of a voice similarity matrix to visualize, like in the zoo plots, the performance of a system at the speaker level. Figure 41 shows a few examples of voice similarity matrix with the corresponding zoo plot below[3]. The latter shows speakers in an averaged target/non-target score plane. In these examples, one can notice that different regions of the plane are covered. For the matrix in column (a), there are mostly two kinds of resulting pseudo-voices: those which do not confuse with others but have a self-similarity not so high (1, 2, 7, 8, 11, 12, and 14) and those which confuse with some others but have a higher self-similarity (3, 4, 5, 6, 9, 10, 13, and 15). These two kinds correspond to the two clusters in the zoo plot (bottom figure) in column (a) making them close to goats and wolves respectively. Column (b) shows an example where the ASV performs almost equally well on all speakers which corresponds to high averaged target and low averaged non-target scores (sheep). Lastly, column (c) shows an example of a good privacy preservation, corresponding to poor ASV performance with an OP setting. Speakers are thus placed in the zoo plot with zero scores making them close to worms. These examples show how the voice similarity matrices can be used to visualize heterogeneous performance across set of speakers.

Going further

Equation 118 has been written in a way to summarize the appearance of a matrix, especially for its diagonal. It does not have the information theoretical interpretation of cross-entropy-based measures like the C_{llr}. Actually, a de-identification and a voice distinctiveness preservation measure could have been based on empirical cross-entropy. A brief comparison of these formulations and the use of the voice similarity matrices in a practical privacy system assessment can be found in our paper [121]. For more examples of the use of these matrices

2 In electrical engineering, telecommunication and acoustics, it is common to express a gain in decibel.
3 For more details on which data these matrices are computed refer to our original paper [121]

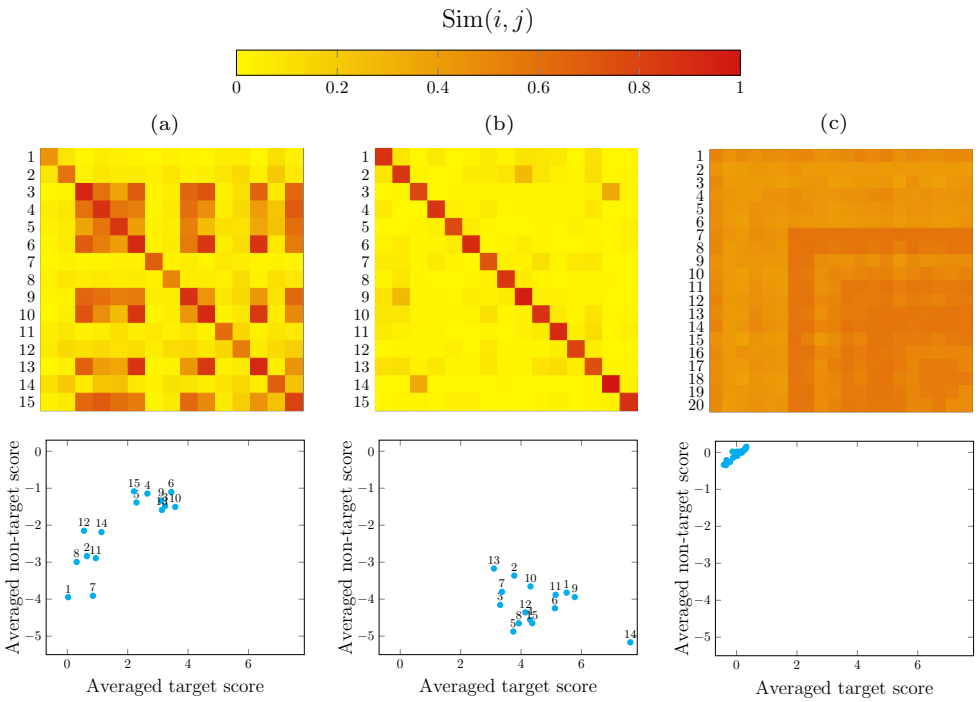


Figure 41: Examples of Voice Similarity Matrix with the corresponding zoo plot [51].

for assessing pseudonymization systems, the reader can also refer to the VoicePrivacy 2020 Challenge's results and findings [156].

10.6 PROOF OF PROPOSITION 2

In this section, we provide a proof for Proposition 2 given in Chapter 6.

Proof. Let's recall some notations. Let $\boldsymbol{w}(E) = [P(E \mid H_i)]_{1 \leqslant i \leqslant N}$ be the likelihood vector of E and $\tilde{\boldsymbol{w}}(E) = \operatorname{ilr}(\boldsymbol{w}(E)) = \boldsymbol{l}$ its ILR transformation.

Equation 45 gives a general expression for the ILR components of a likelihood vector:

$$l_i = \tilde{w}(E)_i = \frac{1}{\sqrt{i\,(i+1)}} \log \left(\frac{\prod_{j=1}^{i} w(E)_j}{(w(E)_{i+1})^i} \right) \tag{121}$$

where l_i is the ith component of $\boldsymbol{l} = \tilde{\boldsymbol{w}}(E)$. Using the idempotence property we can replace E by $\boldsymbol{l}$:

$$l_i = \frac{1}{\sqrt{i\,(i+1)}} \log \left(\frac{\prod_{j=1}^{i} w(\boldsymbol{l})_j}{(w(\boldsymbol{l})_{i+1})^i} \right) \tag{122}$$

After rewriting this expression and setting $\boldsymbol{a}_i = \sqrt{\frac{i+1}{i}}\boldsymbol{e}_i$ where $\boldsymbol{e}_i$ is the ith vector of the standard canonical basis for $\mathbb{R}^{N-1}$ i.e. with zero everywhere except with 1 at the ith position, we get:

$$w(\boldsymbol{l})_{i+1} = \exp\left(-\boldsymbol{a}_i^T \boldsymbol{l}\right) \sqrt[i]{\prod_{j=1}^{i} w(\boldsymbol{l})_j}. \tag{123}$$

We thus have a recursive way to get any $w(\boldsymbol{l})_i$ from $w(\boldsymbol{l})_1$. Expressing them in terms of probability density functions:

$$\boldsymbol{w}(\boldsymbol{l}) = [P(\boldsymbol{l} \mid H_i)]_{1 \leqslant i \leqslant N} = [f_{H_i}(\boldsymbol{l})]_{1 \leqslant i \leqslant N}\,, \tag{124}$$

and since $\boldsymbol{l} \sim \mathcal{N}(\boldsymbol{\mu}_1 \mid \boldsymbol{\Sigma})$, we can use 123 to recursively compute the conditional densities $f_{H_i}(\boldsymbol{l})$ for $i \in [\![2, N]\!]$ from:

$$f_{H_1}(\boldsymbol{l}) = \frac{1}{(2\pi)^{\frac{N-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{l} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{l} - \boldsymbol{\mu}_1) \right). \tag{125}$$

The main idea of the proof is to show—by induction and using the recursive relation of Expression 123—that:

$$
\begin{aligned}
&\text{for all integer } N \geqslant 2 \text{ we have:}\\
&\forall i \in [\![1, N-1]\!],\\
&f_{H_{i+1}}(l) = \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(l-\mu_{i+1})^T\Sigma^{-1}(l-\mu_{i+1})\right),\\
&\mu_{i+1} = \frac{1}{i}\sum_{j=1}^{i}\mu_j - \Sigma a_i,\\
&\mu_{i+1}^T\Sigma^{-1}\mu_{i+1} = \mu_1^T\Sigma^{-1}\mu_1.
\end{aligned}
\tag{126}
$$

The base case:

From Equation 123 and Equation 125 we get:

$$
\begin{aligned}
f_{H_2}(l) &= \exp\left(-a_1^T l\right) f_{H_1}(l),\\
&= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(l-\mu_1)^T\Sigma^{-1}(l-\mu_1) - a_1^T l\right),\\
&= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}l^T\Sigma^{-1}l + l^T\Sigma^{-1}\left(\mu_1 - \Sigma a_1\right) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right),\\
&= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(l-\mu_2)^T\Sigma^{-1}(l-\mu_2)\right) \exp\left(\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right),
\end{aligned}
\tag{127}
$$

where $\mu_2 = \mu_1 - \Sigma a_1$. Since f_{H_2} is a probability density function, its integral is one:

$$
\begin{aligned}
&\int_{l\in\mathbb{R}^{N-1}} f_{H_2}(l)\,dl = 1\\
&\iff \exp\left(\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) = 1,\\
&\iff \mu_2^T\Sigma^{-1}\mu_2 = \mu_1^T\Sigma^{-1}\mu_1.
\end{aligned}
\tag{128}
$$

We just showed that $l \mid H_2 \sim \mathcal{N}(\mu_2, \Sigma)$ where $\mu_2 = \mu_1 - \Sigma a_1$ and $\mu_2^T\Sigma^{-1}\mu_2 = \mu_1^T\Sigma^{-1}\mu_1$.

The induction step:

Let's assume that for an integer K we have:

$$\begin{aligned}
&\forall i \in [\![1, K-1]\!], \\
&f_{H_{i+1}}(l) = \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(l-\mu_{i+1})^T\Sigma^{-1}(l-\mu_{i+1})\right) \\
&\mu_{i+1} = \frac{1}{i}\sum_{j=1}^{i}\mu_j - \Sigma a_i, \\
&\mu_{i+1}^T\Sigma^{-1}\mu_{i+1} = \mu_1^T\Sigma^{-1}\mu_1.
\end{aligned} \tag{129}$$

Let's show that this still holds for $K+1$. Using Equation 123 we can write:

$$\begin{aligned}
f_{H_{K+1}}(l) &= \exp\left(-a_K^T l\right) \sqrt[K]{\prod_{j=1}^{K} f_{H_j}(l)}, \\
&= \exp\left(-a_K^T l\right) \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \sqrt[K]{\prod_{j=1}^{K} \exp\left(-\frac{1}{2}(l-\mu_j)^T\Sigma^{-1}(l-\mu_j)\right)}, \\
&= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-a_K^T l - \frac{1}{2K}\sum_{j=1}^{K}(l-\mu_j)^T\Sigma^{-1}(l-\mu_j)\right), \\
&= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-a_K^T l - \frac{1}{2K}\sum_{j=1}^{K}\left(l^T\Sigma^{-1}l + \mu_j^T\Sigma^{-1}\mu_j - 2\mu_j^T\Sigma^{-1}l\right)\right).
\end{aligned} \tag{130}$$

Since for all $j \in [\![1, K]\!]$, $\mu_j^T\Sigma^{-1}\mu_j = \mu_1^T\Sigma^{-1}\mu_1$, we have:

$$\begin{aligned}
f_{H_{K+1}}(l) &= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-a_K^T l - \frac{1}{2}l^T\Sigma^{-1}l - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{K}\left(\sum_{j=1}^{K}\mu_j\right)^T\Sigma^{-1}l\right), \\
&= \frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}l^T\Sigma^{-1}l - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + l^T\Sigma^{-1}\left(\frac{1}{K}\left(\sum_{j=1}^{K}\mu_j\right) - \Sigma a_K\right)\right).
\end{aligned} \tag{131}$$

Setting $\mu_{K+1} = \frac{1}{K}\sum_{j=1}^{K}\mu_j - \Sigma a_K$ we get:

$$\begin{aligned}
f_{H_{K+1}}(l) = &\frac{1}{(2\pi)^{\frac{N-1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(l-\mu_{K+1})^T\Sigma^{-1}(l-\mu_{K+1})\right) \\
&\times \exp\left(\frac{1}{2}\mu_{K+1}^T\Sigma^{-1}\mu_{K+1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right),
\end{aligned} \tag{132}$$

Since $f_{H_{K+1}}$ is a probability density function, its integral is one, which leads to:

$$\mu_{K+1}^T \Sigma^{-1} \mu_{K+1} = \mu_1^T \Sigma^{-1} \mu_1. \tag{133}$$

Therefore, 129 holds also for $K+1$. We hence have proved by induction the expressions 126.

A general formula for the means:

We will here proof by induction the following expression for the derivation of the mean vectors:

$$\begin{aligned} &\text{for all integer } N \geqslant 2: \\ &\forall i \in [\![2, N]\!], \\ &\mu_i = \mu_1 - \Sigma a_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1} \Sigma a_j, \text{ where } a_0 = 0 \text{ the zero vector.} \end{aligned} \tag{134}$$

The base case is straightforward so we provide only the induction step. Let's assume that the expression is true for an integer K:

$$\begin{aligned} &\forall i \in [\![2, K]\!], \\ &\mu_i = \mu_1 - \Sigma a_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1} \Sigma a_j. \end{aligned} \tag{135}$$

Let's show that this holds also for $K+1$. From Expression 126 that we proofed above, we know that:

$$\mu_{K+1} = -\Sigma a_K + \frac{1}{K} \sum_{j=1}^{K} \mu_j, \tag{136}$$

we replace μ_j according to Expression 135:

$$\begin{aligned} \mu_{K+1} &= -\Sigma a_K + \frac{1}{K} \sum_{j=1}^{K} \left(\mu_1 - \Sigma a_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \Sigma a_k \right), \\ &= \mu_1 - \Sigma a_K - \frac{1}{K} \sum_{j=2}^{K} \Sigma a_{j-1} - \frac{1}{K} \sum_{j=3}^{K} \sum_{k=1}^{j-2} \frac{1}{k+1} \Sigma a_k, \\ &= \mu_1 - \Sigma a_K - \frac{1}{K} \sum_{j=1}^{K-1} \Sigma a_j - \frac{1}{K} \sum_{j=1}^{K-2} \frac{n-1-j}{j+1} \Sigma a_j, \\ &= \mu_1 - \Sigma a_K - \frac{1}{K} \Sigma a_{K-1} - \frac{1}{K} \sum_{j=1}^{K-2} \left(\Sigma a_j + \frac{n-1-j}{j+1} \Sigma a_j \right), \end{aligned} \tag{137}$$

$$
\begin{aligned}
&= \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}\mathbf{a}_K - \frac{1}{K}\boldsymbol{\Sigma}\mathbf{a}_{K-1} - \sum_{j=1}^{K-2}\frac{1}{j+1}\boldsymbol{\Sigma}\mathbf{a}_j, \\
&= \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}\mathbf{a}_K - \sum_{j=1}^{K-1}\frac{1}{j+1}\boldsymbol{\Sigma}\mathbf{a}_j,
\end{aligned}
\tag{138}
$$

Hence,

$$
\begin{aligned}
&\forall i \in [\![2, K+1]\!], \\
&\boldsymbol{\mu}_i = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}\mathbf{a}_{i-1} - \sum_{j=1}^{i-2}\frac{1}{j+1}\boldsymbol{\Sigma}\mathbf{a}_j,
\end{aligned}
\tag{139}
$$

The general expression 134 has therefore been proved by induction.

About matrices A and B

In Proposition 2, the mean vector $\boldsymbol{\mu}_1$ is expressed in terms of the covariance matrix $\boldsymbol{\Sigma}$ and two constant matrices $\mathbf{A}$ and $\mathbf{B}$ as follow:

$$
\boldsymbol{\mu}_1 = \mathbf{A}^{-1}\mathbf{B}\,\mathrm{vec}(\boldsymbol{\Sigma}),
\tag{140}
$$

where $\mathbf{A} \in \mathcal{M}_{N-1,N-1}(\mathbb{R})$ and is defined as:

$$
\begin{aligned}
&\mathbf{A} = \{\alpha_{ij}\}_{1 \leqslant i,j \leqslant N-1} \\
&\alpha_{ij} = \begin{cases} 2\sqrt{\frac{i+1}{i}}, & \text{if } i = j \\ \frac{2}{\sqrt{j(j+1)}}, & \text{if } j < i \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
\tag{141}
$$

and where $\mathbf{B} \in \mathcal{M}_{N-1,(N-1)^2}(\mathbb{R})$ is a block matrix:

$$
\mathbf{B} = \begin{bmatrix} \mathbf{B}^{(1)} & \mathbf{B}^{(2)} & \dots & \mathbf{B}^{(N-1)} \end{bmatrix}
\tag{142}
$$

where $\mathbf{B}^{(b)} \in \mathcal{M}_{N-1,N-1}(\mathbb{R})$ is the bth block and is defined as:

$$
\begin{aligned}
&\mathbf{B}^{(b)} = \{\beta_{ij}^{(b)}\}_{1 \leqslant i,j \leqslant N-1} \\
&\beta_{ij}^{(b)} = \begin{cases} \frac{b+1}{b}, & \text{if } i = j = b \\ 2\sqrt{\frac{i+1}{ib(b+1)}}, & \text{if } (i = j) \wedge (b < i) \\ \frac{1}{jb\sqrt{(j+1)(b+1)}}, & \text{if } (b < i) \wedge (j < i) \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
\tag{143}
$$

In this paragraph, we show how these matrices are derived. The following system of equations, that comes from the expressions 126, can be written in a matrix form:

$$\begin{aligned} &\forall\, i \in [\![1, N-1]\!], \\ &\mu_{i+1}^T \Sigma^{-1} \mu_{i+1} = \mu_1^T \Sigma^{-1} \mu_1. \end{aligned} \tag{144}$$

Using the general expression 134 for the means, the system of equations becomes:

$$\begin{aligned} &\forall i \in [\![1, N-1]\!], \\ &2a_i^T \mu_1 + 2\mu_1^T \Sigma^{-1} \sum_{j=1}^{i-1} \frac{1}{j+1} \Sigma a_j = a_i^T \Sigma a_i + 2a_i^T \sum_{j=1}^{i-1} \frac{1}{j+1} \Sigma a_j + \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} \frac{1}{(j+1)(k+1)} a_j^T \Sigma a_k, \end{aligned} \tag{145}$$

Since $x^T \Sigma y = \text{vec}(\Sigma)^T \text{vec}(xy^T)$ and by setting $\theta_\Sigma = \text{vec}(\Sigma)$ we get:

$$\begin{aligned} &\forall i \in [\![1, N-1]\!], \\ &\left(2a_i + 2\sum_{j=1}^{i-1} \frac{1}{j+1} a_j\right)^T \mu_1 \\ &\qquad = \left(\text{vec}(a_i a_i^T) + 2\sum_{j=1}^{i-1} \frac{1}{j+1} \text{vec}(a_i a_j^T) + \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} \frac{1}{(j+1)(k+1)} \text{vec}(a_j a_k^T)\right)^T \theta_\Sigma. \end{aligned} \tag{146}$$

$a_i a_j^T$ is a $(N-1) \times (N-1)$ matrix with zero everywhere except the element at the ith row and jth column which is $\sqrt{\frac{(i+1)(j+1)}{ij}}$. Its vectorization is therefore the $(N-1)^2$-dimensional vector with zero everywhere except the $((j-1)(N-1)+i)$th element which is $\sqrt{\frac{(i+1)(j+1)}{ij}}$. Let's now rewrite this system in a matrix form:

$$A\mu_1 = B\theta_\Sigma, \tag{147}$$

where $A \in M_{N-1,N-1}(\mathbb{R})$, $B \in M_{N-1,(N-1)^2}(\mathbb{R})$. In 146, the vector on the left side of μ is the ith row of the matrix A and the vector on the left side of θ_Σ is the ith row of B. This is straightforward that A is triangular with diagonal elements $2\sqrt{\frac{i+1}{i}}$ for $i \in [\![1, N-1]\!]$. Consequently, its determinant is non zero, and therefore A is invertible. The mean vector μ_1 can therefore be written in terms of the variances and covariances as follow:

$$\mu_1 = A^{-1} B \theta_\Sigma = A^{-1} B\, \text{vec}(\Sigma). \tag{148}$$

□

10.7 PROOF THAT THE BASE SPACE'S FIRST DIMENSIONS FORM THE ILRL

This section shows that with the class-conditional densities in the base space as defined in Equation 91, the first $N-1$ dimensions of $z \in \mathcal{Z}$ form the ILRL.

Proof. Expressing Equation 90 in terms of the probability density functions for z, the ith component of the ILRL vector is given by:

$$\begin{aligned} \forall i \in [\![1, N-1]\!], \quad l_i(z) &= \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod_{j=1}^{i} f_{C_j}(z)}{f_{C_{i+1}}(z)^i} \right) \\ &= \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod_{j=1}^{i} \exp\left(-\frac{1}{2}(z-m_j)^T C^{-1}(z-m_j)\right)}{\exp\left(-\frac{i}{2}(z-m_{i+1})^T C^{-1}(z-m_{i+1})\right)} \right), \end{aligned} \tag{149}$$

where $m_i = m_i(\Sigma)$ and $C = C(\Sigma)$ are respectively the mean vector and the covariance matrix as defined in what follows Equation 91,

$$\begin{aligned} l_i(z) &= \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(-\frac{1}{2}(z-m_j)^T C^{-1}(z-m_j) \right) + \frac{i}{2}(z-m_{i+1})^T C^{-1}(z-m_{i+1}) \right), \\ &= \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(m_j^T C^{-1} z - \frac{1}{2} m_j^T C^{-1} m_j \right) + \frac{i}{2} m_{i+1}^T C^{-1} m_{i+1} - i m_{i+1}^T C^{-1} z \right), \\ &= \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(\mu_j^T \Sigma^{-1} z_{1:N-1} - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j \right) + \frac{i}{2} \mu_{i+1}^T \Sigma^{-1} \mu_{i+1} - i \mu_{i+1}^T \Sigma^{-1} z_{1:N-1} \right), \end{aligned} \tag{150}$$

where $z_{1:N-1} = [z_1, z_2, \dots z_{N-1}]^T$ is the vector of the first $N-1$ components of z. Since $\forall i \in [\![1, N]\!]$, $\mu_i^T \Sigma \mu_i = \mu_1^T \Sigma \mu_1$ (see Appendix 10.6), we have:

$$l_i(z) = \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(\mu_j^T \Sigma^{-1} z_{1:N-1} \right) - i \mu_{i+1}^T \Sigma^{-1} z_{1:N-1} \right), \tag{151}$$

using Equation 92, we get:

$$\begin{aligned} l_i(z) &= \frac{1}{\sqrt{i(i+1)}}\left(\sum_{j=1}^{i}\left(\mathbf{A}^{-1}\mathbf{B}\,\mathrm{vec}(\boldsymbol{\Sigma}) - \boldsymbol{\Sigma}\mathbf{a}_{j-1} - \sum_{k=1}^{j-2}\frac{1}{k+1}\boldsymbol{\Sigma}\mathbf{a}_k\right)^T \boldsymbol{\Sigma}^{-1}z_{1:N-1}\right. \\ &\qquad \left. - i\left(\mathbf{A}^{-1}\mathbf{B}\,\mathrm{vec}(\boldsymbol{\Sigma}) - \boldsymbol{\Sigma}\mathbf{a}_i - \sum_{k=1}^{i-1}\frac{1}{k+1}\boldsymbol{\Sigma}\mathbf{a}_k\right)^T \boldsymbol{\Sigma}^{-1}z_{1:N-1}\right), \\ &= \frac{1}{\sqrt{i(i+1)}}\left(\sum_{j=1}^{i}\left(-\mathbf{a}_{j-1}^T z_{1:N-1} - \sum_{k=1}^{j-2}\frac{1}{k+1}\mathbf{a}_k^T z_{1:N-1}\right)\right. \\ &\qquad \left. + i\sum_{k=1}^{i-1}\left(\frac{1}{k+1}\mathbf{a}_k^T z_{1:N-1}\right) + i\mathbf{a}_i^T z_{1:N-1}\right). \end{aligned} \tag{152}$$

In the next paragraph, we will see that:

$$\sum_{j=1}^{i}\left(-\mathbf{a}_{j-1}^T z_{1:N-1} - \sum_{k=1}^{j-2}\frac{1}{k+1}\mathbf{a}_k^T z_{1:N-1}\right) + i\sum_{k=1}^{i-1}\left(\frac{1}{k+1}\mathbf{a}_k^T z_{1:N-1}\right) = 0. \tag{153}$$

We therefore have:

$$\begin{aligned} l_i(z) &= \frac{i}{\sqrt{i(i+1)}}\mathbf{a}_i^T z_{1:N-1}, \\ &= \frac{i}{\sqrt{i(i+1)}}\sqrt{\frac{i+1}{i}}\mathbf{e}_i^T z_{1:N-1} = \mathbf{e}_i^T z_{1:N-1}, \\ &= z_i, \end{aligned} \tag{154}$$

the ith component of the ILRL is therefore the ith component of z for all $i \in [\![1, N-1]\!]$.

Proof of Expression 153:

In the following, we will show by induction that Expression 153 is true for all $i \in [\![1, N-1]\!]$ which is equivalent to show that:

$$\forall i \in [\![1, N-1]\!], \quad \sum_{j=1}^{i}\left(-\mathbf{a}_{j-1} - \sum_{k=1}^{j-2}\frac{1}{k+1}\mathbf{a}_k\right) + i\sum_{k=1}^{i-1}\left(\frac{1}{k+1}\mathbf{a}_k\right) = 0. \tag{155}$$

The base case of the proof by induction is straightforward, thus we will focus on the induction step. We assume that the expression is true for a $i = n$ where $n \in \mathbb{N}$:

$$\sum_{j=1}^{n}\left(-\mathbf{a}_{j-1} - \sum_{k=1}^{j-2}\frac{1}{k+1}\mathbf{a}_k\right) + n\sum_{k=1}^{n-1}\left(\frac{1}{k+1}\mathbf{a}_k\right) = 0. \tag{156}$$

Let's show this is still true for $i = n+1$:

$$
\begin{aligned}
&\sum_{j=1}^{n+1}\left(-a_{j-1}-\sum_{k=1}^{j-2}\frac{1}{k+1}a_k\right)+(n+1)\sum_{k=1}^{n}\left(\frac{1}{k+1}a_k\right)\\
&\quad=\sum_{j=1}^{n}\left(-a_{j-1}-\sum_{k=1}^{j-2}\frac{1}{k+1}a_k\right)+n\sum_{k=1}^{n-1}\left(\frac{1}{k+1}a_k\right)\\
&\qquad-a_n-\sum_{k=1}^{n-1}\left(\frac{1}{k+1}a_k\right)+\sum_{k=1}^{n-1}\left(\frac{1}{k+1}a_k\right)+\frac{n+1}{n+1}a_n\\
&\quad=\sum_{j=1}^{n}\left(-a_{j-1}-\sum_{k=1}^{j-2}\frac{1}{k+1}a_k\right)+n\sum_{k=1}^{n-1}\left(\frac{1}{k+1}a_k\right)\\
&\quad=0 \text{ according to Equation 156.}
\end{aligned}
\tag{157}
$$

□

10.8 INTERPOLATING BETWEEN MNIST'S NUMBERS

Section 7.1.4 discussed how the Compositional Discriminant Analysis (CDA) can be used to model MNIST's digits. The Euclidean base space allows easy interpolation between digits. This can be done with linear interpolation between the digits' centroid. The interpolation between the digit i and the digit j in the base space is given by:

$$z_{i,j}(\alpha) = \alpha m_i + (1 - \alpha) m_j, \tag{158}$$

where $\alpha \in [0, 1]$ and m_i and m_j are the learned digits' centroid as defined in Section 7.1.1. The image can then be constructed by mapping $z_{i,j}(\alpha)$ into the feature space using the learned diffeomorphism g: $x_{i,j}(\alpha) = g(z_{i,j}(\alpha))$. The feature vector is then unflattened to produce the image. The following set of images shows interpolation between each pair of digits for $\alpha = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

From 0 to 1:

From 0 to 2:

From 0 to 3:

From 0 to 4:

From 0 to 5:

From 0 to 6:

From 0 to 7:

From 0 to 8:

From 0 to 9:

From 1 to 2:

From 1 to 3:

From 1 to 4:

From 1 to 5:

From 1 to 6:

From 1 to 7:

From 1 to 8:

From 1 to 9:

From 2 to 3:

From 2 to 4:

From 2 to 5:

From 2 to 6:

From 2 to 7:

From 2 to 8:

From 2 to 9:

From 3 to 4:

From 3 to 5:

From 3 to 6:

From 3 to 7:

From 3 to 8:

From 3 to 9:

From 4 to 5:

From 4 to 6:

From 4 to 7:

From 4 to 8:

From 4 to 9:

From 5 to 6:

From 5 to 7:

From 5 to 8:

From 5 to 9:

From 6 to 7:

From 6 to 8:

From 6 to 9:

From 7 to 8:

From 7 to 9:

From 8 to 9:

10.9 DERIVATIVE OF THE FLOW-BASED CALIBRATION FUNCTION

This section discusses the derivation of the flow-based calibration function's derivative required by the change of variable in Equation 101. The calibration function $F_{\text{FLOWCAL}}(\cdot \mid \theta)$ as defined in Equation 99 is a composition of N logit transformed mixtures of normal CDFs. Let p_i be the output of the ith mixture of normal CDFs and l_i be the output of the ith logit. In this way, we can write:

$$\log \left| \frac{d}{dx} F_{\text{FLOWCAL}}(x \mid \theta) \right| = \sum_{i=1}^{N} \left(\log \left(\left| \frac{dl_i}{dp_i} \right| \left| \frac{dp_i}{dl_{i-1}} \right| \right) \right), \tag{159}$$

where l_0 is simply the input x of the calibration mapping. Since $l_i = \operatorname{logit} p_i$:

$$\frac{dl_i}{dp_i} = \frac{d}{dp_i} \operatorname{logit} p_i = \frac{1}{p_i(1-p_i)}, \tag{160}$$

and since $p_i = F_{\text{MCDF}}(l_{i-1} \mid M_i, \Sigma_i, W_i) = \sum_{i=1}^{K} w_i F(l_{i-1} \mid \mu_i, \sigma_i)$:

$$\begin{aligned} \frac{dp_i}{dl_{i-1}} &= \sum_{i=1}^{K} w_i \frac{d}{dl_{i-1}} F(l_{i-1} \mid \mu_i, \sigma_i) \\ &= \sum_{i=1}^{K} w_i f(l_{i-1} \mid \mu_i, \sigma_i), \end{aligned} \tag{161}$$

where $f(l_{i-1} \mid \mu_i, \sigma_i)$ is the probability density function of the normal distribution with mean μ_i and standard deviation σ_i.

BIBLIOGRAPHY

[1] Alberto Abad, Eugénio Ribeiro, Fábio N Kepler, Ramón Fernández Astudillo, and Isabel Trancoso. "Exploiting Phone Log-Likelihood Ratio Features for the Detection of the Native Language of Non-Native English Speakers." In: *Proc. ISCA Interspeech*. 2016, pp. 2413–2417.

[2] John Aitchison. "The Statistical Analysis of Compositional Data." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 139–177.

[3] John Aitchison. "The Statistical Analysis of Compositional Data." In: *Monographs on Statistics and Applied Probability* 25 (1986).

[4] John Aitchison. "Simplicial inference." In: *Algebraic Methods in Statistics and Probability*. Ed. by Donald St. P. Richards (eds.) Ams Special Session on Algebraic Methods in Statistics Marlos A. G. Viana. Contemporary Mathematics 287. American Mathematical Society, 2001.

[5] John Aitchison and S. M. Shen. "Logistic-Normal Distributions: Some Properties and Uses." In: *Biometrika* 67.2 (1980), pp. 261–272.

[6] Moez Ajili. "Reliability of voice comparison for forensic applications." PhD thesis. Avignon Université, 2017.

[7] Ranya Aloufi, Hamed Haddadi, and David Boyle. "Privacy-preserving Voice Analysis via Disentangled Representations." In: *Proc. SIGSAC Conference on Cloud Computing Security Workshop*. ACM. 2020, pp. 1–14.

[8] Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. "Training Normalizing Flows with the Information Bottleneck for Competitive Generative Classification." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 7828–7840.

[9] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. "An Empirical Distribution Function for Sampling with Incomplete Information." In: *The Annals of Mathematical Statistics* 26.4 (1955), pp. 641–647.

[10] Fahimeh Bahmaninezhad, Chunlei Zhang, and John Hansen. "Convolutional Neural Network Based Speaker De-Identification." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2018, pp. 255–260.

[11] Zhongxin Bai and Xiao-Lei Zhang. "Speaker recognition based on deep learning: An overview." In: *Neural Networks* 140 (2021), pp. 65–99.

[12] Carles Barceló-Vidal and Vera Pawlowsky-Glahn. "Mathematical foundations of compositional data analysis." In: *Proc. IAMG*. Vol. 1. 2001.

[13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[14] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 1994.

[15] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds. "A tutorial on text-independent speaker verification." In: *EURASIP Journal on Advances in Signal Processing* 4 (2004), pp. 1–22.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[17] Jean-François Bonastre et al. "Benchmarking and challenges in security and privacy for voice biometrics." In: *Proc. ISCA Symposium on Security and Privacy in Speech Communication*. 2021, pp. 52–56.

[18] Glenn W. Brier. "Verification of forecasts expressed in terms of probability." In: *Monthly Weather Review* 78.1 (1950), pp. 1–3.

[19] Niko Brümmer. "Measuring, refining and calibrating speaker and language information extracted from speech." PhD thesis. Stellenbosch, University of Stellenbosch, 2010.

[20] Niko Brümmer, Lukas Burget, Jan Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karafiat, David A Van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim. "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006." In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.7 (2007), pp. 2072–2084.

[21] Niko Brümmer and George R. Doddington. "Likelihood-ratio calibration using prior-weighted proper scoring rules." In: *Proc. ISCA Interspeech*. 2013, pp. 1976–1980.

[22] Niko Brümmer and Johan du Preez. "Application-independent evaluation of speaker detection." In: *Computer Speech & Language* 20.2 (2006), pp. 230–275.

[23] Niko Brümmer and Daniel Garcia-Romero. "Generative modelling for unsupervised score calibration." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 1680–1684.

[24] Niko Brümmer and J.A. Preez. "The PAV algorithm optimizes binary proper scoring rules." 2013.

[25] Niko Brümmer, Anna Silnova, Lukas Burget, and Themos Stafylakis. "Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2018, pp. 349–356.

[26] Niko Brümmer, Albert Swart, and David van Leeuwen. "A comparison of linear and non-linear calibrations for speaker recognition." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2014, pp. 14–18.

[27] Niko Brümmer and David A. Van Leeuwen. "On calibration of language recognition scores." In: *Proc. IEEE Odyssey - The Speaker and Language Recognition Workshop*. 2006, pp. 1–8.

[28] Niko Brümmer and Edward de Villiers. *The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF*. 2013.

[29] Joseph P Campbell, Wade Shen, William M Campbell, Reva Schwartz, Jean-Francois Bonastre, and Driss Matrouf. "Forensic speaker recognition." In: *IEEE Signal Processing Magazine* 26.2 (2009), pp. 95–103.

[30] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. "Support vector machines using GMM supervectors for speaker verification." In: *IEEE Signal Processing Letters* 13.5 (2006), pp. 308–311.

[31] Christophe Champod and Didier Meuwly. "The inference of identity in forensic speaker recognition." In: *Speech Communication* 31.2 (2000), pp. 193–203.

[32] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. "Broadening the Scope of Differential Privacy Using Metrics." In: *Privacy Enhancing Technologies*. Ed. by Emiliano De Cristofaro and Matthew Wright. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 82–102.

[33] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition." In: *Proc. ISCA Interspeech*. 2018, pp. 1086–1090.

[34] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[35] Sandro Cumani. "Normal Variance-Mean Mixtures for Unsupervised Score Calibration." In: *Proc. ISCA Interspeech*. 2019, pp. 401–405.

[36] Sandro Cumani. "On the Distribution of Speaker Verification Scores: Generative Models for Unsupervised Calibration." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 547–562.

[37] Sandro Cumani and Pietro Laface. "Tied Normal Variance–Mean Mixtures for Linear Score Calibration." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2019, pp. 6121–6125.

[38] Sandro Cumani and Salvatore Sarni. "A Generative Model for Duration-Dependent Score Calibration." In: *Proc. ISCA Interspeech*. 2021, pp. 4598–4602.

[39] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366.

[40] A. Philip Dawid. "The well-calibrated Bayesian." In: *Journal of the American Statistical Association* 77.379 (1982), pp. 605–610.

[41] A. Philip Dawid. *Coherent Measures of Discrepancy, Uncertainty and Dependence, with Applications to Bayesian Predictive Experimental Design*. 1998.

[42] Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, Inc., 1970.

[43] Morris H. DeGroot and Stephen E. Fienberg. "The Comparison and Evaluation of Forecasters." In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 32.1/2 (1983), pp. 12–22.

[44] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification." In: *Proc. ISCA Interspeech*. 2009, pp. 1559–1562.

[45] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. "Front-End Factor Analysis for Speaker Verification." In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.

[46] Najim Dehak, Patrick Kenny, Reda Dehak, Ondrej Glembek, Pierre Dumouchel, Lukas Burget, Valiantsina Hubeika, and Fabio Castaldo. "Support vector machines and Joint Factor Analysis for speaker verification." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 4237–4240.

[47] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." In: *Proc. ISCA Interspeech*. 2020, pp. 3830–3834.

[48] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP." In: *Proc. ICLR - International Conference on Learning Representations*. 2017.

[49] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. *Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation*. Tech. rep. National Inst of Standards and Technology Gaithersburg Md, 1998.

[50] Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. "Deep Linear Discriminant Analysis." In: *Proc. ICLR - International Conference on Learning Representations*. 2016.

[51] Ted Dunstone and Neil Yager. *Biometric System and Data Analysis: Design, Evaluation, and Data Mining*. eng. 1. Aufl. New York, NY: Springer-Verlag, 2009.

[52] Cynthia Dwork, Aaron Roth, et al. "The algorithmic foundations of differential privacy." In: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[53] Juan José Egozcue, José Luis Díaz-Barrero, and Vera Pawlowsky-Glahn. "Hilbert space of probability density functions based on Aitchison geometry." In: *Acta Mathematica Sinica* 22.4 (2006), pp. 1175–1182.

[54] Juan José Egozcue and Vera Pawlowsky-Glahn. "Groups of parts and their balances in compositional data analysis." In: *Mathematical Geology* 37.7 (2005), pp. 795–828.

[55] Juan José Egozcue and Vera Pawlowsky-Glahn. "Evidence information in Bayesian updating." In: *Proc. International Workshop on Compositional Data Analysis* (May 2011).

[56] Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. "Isometric logratio transformations for compositional data analysis." In: *Mathematical geology* 35.3 (2003), pp. 279–300.

[57] Juan José Egozcue, Vera Pawlowsky-Glahn, Raimon Tolosana-Delgado, MI Ortego, and Karl Gerald van den Boogaart. "Bayes spaces: use of improper distributions and exponential families." In: *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* 107.2 (2013), pp. 475–486.

[58] Juan José Egozcue and Pawlowsky-Glahn Vera. "Evidence functions: a compositional approach to information." In: *SORT-Statistics and Operations Research Transactions* 1.2 (Dec. 2018), pp. 101–124.

[59] Peter Filzmoser, Karel Hron, and Matthias Templ. "Discriminant Analysis for Compositional Data and Robust Parameter Estimation." In: vol. 27. Dec. 2011, p. 17.

[60] Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, 1975.

[61] Sadaoki Furui and Aaron Rosenberg. *Speaker verification*. 1999.

[62] Jean-Luc. Gauvain and Chin-Hui Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." In: *IEEE Transactions on Speech and Audio Processing* 2.2 (1994), pp. 291–298.

[63] Tilmann Gneiting and Adrian E Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation." In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[64] Jorge Andrés Gómez García, Laureano Moro Velázquez, Juan Ignacio Godino Llorente, and Germán Castellanos Domínguez. "Automatic age detection in normal and pathological voice." In: *Proc. ISCA Interspeech*. 2015, pp. 3739–3743.

[65] Pedro Gómez-Vilda, Roberto Fernández-Baillo, Victoria Rodellar-Biarge, Víctor Nieto Lluis, Agustín Álvarez-Marquina, Luis Miguel Mazaira-Fernández, Rafael Martínez-Olalla, and Juan Ignacio Godino-Llorente. "Glottal source biometrical signature for voice pathology detection." In: *Speech Communication* 51.9 (2009), pp. 759–781.

[66] I. J. Good. "Rational Decisions." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 14.1 (1952), pp. 107–114.

[67] I. J. Good. "Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II." In: *Biometrika* 66.2 (1979), pp. 393–396.

[68] I.J. Good. *Probability and the Weighing of Evidence*. 1950.

[69] I.J. Good. "Weight of evidence: A brief survey." In: *Bayesian statistics* 2 (1985), pp. 249–270.

[70] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[71] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." In: *Proc. ICML - International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 1321–1330.

[72] Hadi Harb and Liming Chen. "Voice-based gender identification in multimedia applications." In: *Journal of intelligent information systems* 24 (2005), pp. 179–198.

[73] Trevor Hastie and M Zhu. "Dimension reduction and visualization in discriminant analysis - Discussion." In: *Australian & New Zealand Journal of Statistics* 43 (June 2001), pp. 179–185.

[74] Rosa Gonzalez Hautamäki, Anssi Kanervisto, Ville Hautamaki, and Tomi Kinnunen. "Perceptual Evaluation of the Effectiveness of Voice Disguise by Age Modification." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2018, pp. 320–326.

[75] Hynek Hermansky. "Perceptual linear predictive (PLP) analysis of speech." In: *the Journal of the Acoustical Society of America* 87.4 (1990), pp. 1738–1752.

[76] Sergey Ioffe. "Probabilistic Linear Discriminant Analysis." In: *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, 2006, pp. 531–542.

[77] Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. "Semi-supervised learning with normalizing flows." In: *Proc. ICML - International Conference on Machine Learning*. PMLR. 2020, pp. 4615–4630.

[78] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge: Cambridge University Press, 2003.

[79] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black. "Speaker de-identification via voice transformation." In: *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. 2009, pp. 529–533.

[80] T. Justin, V. Štruc, S. Dobrišek, B. Vesnicer, I. Ipšić, and F. Mihelič. "Speaker de-identification using diphone recognition and speech synthesis." In: *Proc, IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. Vol. 04. 2015, pp. 1–7.

[81] Kavita Kasi and Stephen A. Zahorian. "Yet Another Algorithm for Pitch Tracking." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 2002, pp. I-361-I-364.

[82] Patrick Kenny. "Joint factor analysis of speaker and session variability: Theory and algorithms." In: *CRIM, Montreal,(Report) CRIM-06/08-13* 14.28-29 (2005), p. 2.

[83] Patrick Kenny and Pierre Dumouchel. "Disentangling speaker and channel effects in speaker verification." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 2004, pp. I–37.

[84] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *Proc. ICLR - International Conference on Learning Representations*. 2015, pp. 1–13.

[85] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 17022–17033.

[86] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[87] S. Kullback and R. A. Leibler. "On Information and Sufficiency." In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[88] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. "Fader networks: Manipulating images by sliding attributes." In: *Advances in Neural Information Processing Systems*. 2017, pp. 5967–5976.

[89] Yann LeCun. "The MNIST database of handwritten digits." 1998.

[90] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *nature* 521.7553 (2015), pp. 436–444.

[91] David van Leeuwen and Niko Brümmer. "The distribution of calibrated likelihood-ratios in speaker recognition." In: *Proc. ISCA Interspeech*. 2013, pp. 1619–1623.

[92] D. V. Lindley. "On a Measure of the Information Provided by an Experiment." In: *The Annals of Mathematical Statistics* 27.4 (1956), pp. 986–1005.

[93] Dennis V. Lindley. *Understanding Uncertainty*. Wiley-Interscience, 2006.

[94] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models." In: *Proc. ICML - International Conference on Machine Learning*. Vol. 30. 2013.

[95] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Second. John Wiley, 1999.

[96] Miranti Indar Mandasari. "Speaker Recognition System in Forensic Conditions: The Calibration and Evaluation of the Likelihood Ratio." PhD thesis. Nijmegen, Radboud University, 2018.

[97] Miranti Indar Mandasari, Rahim Saeidi, Mitchell McLaren, and David A. van Leeuwen. "Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions." In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.11 (2013), pp. 2425–2438.

[98] Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. "Dealing with zeros and missing values in compositional data sets using nonparametric imputation." In: *Mathematical Geology* 35.3 (2003), pp. 253–278.

[99] Glòria Mateu-Figueras, Vera Pawlowsky-Glahn, and Juan José Egozcue. "The principle of working on coordinates." In: *Compositional data analysis: Theory and applications* (2011), pp. 29–42.

[100] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, and Jean-François Bonastre. "A straightforward and efficient implementation of the factor analysis model for speaker verification." In: *Proc. ISCA Interspeech*. 2007, pp. 1242–1245.

[101] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform manifold approximation and projection for dimension reduction." 2018.

[102] Ronald Meester and Klaas Slooten. *Probability and Forensic Evidence: Theory, Philosophy, and Applications*. Cambridge University Press, 2021.

[103] Arianna Mencattini, Eugenio Martinelli, Giovanni Costantini, Massimiliano Todisco, Barbara Basile, Marco Bozzali, and Corrado Di Natale. "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure." In: *Knowledge-Based Systems* 63 (2014), pp. 68–81.

[104] Didier Meuwly. "Forensic Individualisation from Biometric Data." In: *Science & Justice* 46.4 (2006), pp. 205–213.

[105] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko. "Language-Independent Speaker Anonymization Approach Using Self-Supervised Pre-Trained Models." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2022, pp. 279–286.

[106] Sebastian. Mika, Gunnar. Rätsch, Jason. Weston, Bernhard. Schölkopf, and Klaus-Robert Müller. "Fisher discriminant analysis with kernels." In: *Proc. IEEE Workshop on Neural Networks for Signal Processing*. 1999, pp. 41–48.

[107] Mohammad Mohammadamini, Driss Matrouf, and Paul-Gauthier Noé. "Denoising x-vectors for Robust Speaker Recognition." In: *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*. 2020, pp. 75–80.

[108] Eric Moulines and Francis Charpentier. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." In: *Speech Communication* 9.5 (1990). Neuropeech '89, pp. 453–467.

[109] Kevin P. Murphy. *Machine learning: A Probabilistic Perspective*. 2012.

[110] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "VoxCeleb: A Large-Scale Speaker Identification Dataset." In: *Proc. ISCA Interspeech*. 2017, pp. 2616–2620.

[111] Andreas Nautsch. "Speaker recognition in unconstrained environments." PhD thesis. Darmstadt, Technische Universität, 2019.

[112] Andreas Nautsch, Jose Patino, N. Tomashenko, Junichi Yamagishi, Paul-Gauthier Noé, Jean-François Bonastre, Massimiliano Todisco, and Nicholas Evans. "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment." In: *Proc. Interspeech*. 2020, pp. 1698–1702.

[113] Andreas Nautsch, Jose Patino, Natalia Tomashenko, Junichi Yamagishi, Paul-Gauthier Noé, Jean-François Bonastre, Massimiliano Todisco, and Nicholas Evans. "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment." In: *Proc. ISCA Interspeech*. 2020, pp. 1698–1702.

[114] Andreas Nautsch et al. "Preserving privacy in speaker and speech characterisation." In: *Computer Speech & Language* 58 (2019), pp. 441–480.

[115] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting good probabilities with supervised learning." In: *Proc. ICML - International Conference on Machine Learning*. 2005, pp. 625–632.

[116] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. "A comparison of discrete and soft speech units for improved voice conversion." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2022, pp. 6562–6566.

[117] Paul-Gauthier Noé, Jean-François Bonastre, Driss Matrouf, N. Tomashenko, Andreas Nautsch, and Nicholas Evans. "Speech Pseudonymisation Assessment Using Voice Similarity Matrices." In: *Proc. Interspeech*. 2020, pp. 1718–1722.

[118] Paul-Gauthier Noé, Jean-François Bonastre, Driss Matrouf, Natalia Tomashenko, Andreas Nautsch, and Nicholas Evans. "Speech Pseudonymisation Assessment Using Voice Similarity Matrices." In: *Proc. ISCA Interspeech*. 2020, pp. 1718–1722.

[119] Paul-Gauthier Noé, Xiaoxiao Miao, Xin Wang, Junichi Yamagishi, Jean-François Bonastre, and Driss Matrouf. "Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2023.

[120] Paul-Gauthier Noé, Mohammad Mohammadamini, Driss Matrouf, Titouan Parcollet, Andreas Nautsch, and Jean-François Bonastre. "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation." In: *Proc. ISCA Interspeech*. 2021, pp. 1902–1906.

[121] Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino, Jean-François Bonastre, Natalia Tomashenko, and Driss Matrouf. "Towards a unified assessment framework of speech pseudonymisation." In: *Computer Speech & Language* 72 (2022), p. 101299.

[122] Paul-Gauthier Noé, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, and Jean-François Bonastre. "A Bridge between Features and Evidence for Binary Attribute-Driven Perfect Privacy." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2022, pp. 3094–3098.

[123] Paul-Gauthier Noé, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, and Jean-François Bonastre. "A bridge between features and evidence for binary attribute-driven perfect privacy." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2022, pp. 3094–3098.

[124] Paul-Gauthier Noé, Titouan Parcollet, and Mohamed Morchid. "CGCNN: Complex Gabor Convolutional Neural Network on Raw Speech." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2020, pp. 7724–7728.

[125] Tony O'Hagan. "Dicing with the unknown." In: *Significance* 1 (2004).

[126] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: An ASR corpus based on public domain audio books." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2015, pp. 5206–5210.

[127] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. "Normalizing Flows for Probabilistic Modeling and Inference." In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64.

[128] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

[129] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[130] W. Peterson, T. Birdsall, and W. Fox. "The theory of signal detectability." In: *Transactions of the IRE Professional Group on Information Theory* 4.4 (1954), pp. 171–212.

[131] José C. Pinheiro and Douglas M. Bates. "Unconstrained Parameterizations for Variance-Covariance Matrices." In: *Statistics and Computing* 6 (1996), pp. 289–296.

[132] John Platt. "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods." In: *Advances in Large Margin Classiers* (1999).

[133] M. Pobar and I. Ipšić. "Online speaker de-identification using voice transformation." In: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. May 2014, pp. 1264–1267.

[134] Simon J.D. Prince and James H. Elder. "Probabilistic Linear Discriminant Analysis for Inferences About Identity." In: *2007 IEEE 11th International Conference on Computer Vision*. 2007, pp. 1–8.

[135] Daniel Ramos. "Forensic evaluation of the evidence using automatic speaker recognition systems." PhD thesis. Madrid, Universidad Politécnica de Madrid, 2007.

[136] Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez. "Deconstructing Cross-Entropy for Probabilistic Binary Classifiers." In: *Entropy* 20.3 (2018).

[137] Daniel Ramos and Joaquin Gonzalez-Rodriguez. "Reliable support: Measuring calibration of likelihood ratios." In: *Forensic Science International* 230.1 (2013). EAFS 2012 6th European Academy of Forensic Science Conference The Hague, 20-24 August 2012, pp. 156–169.

[138] D.A. Reynolds and R.C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models." In: *IEEE Transactions on Speech and Audio Processing* 3.1 (1995), pp. 72–83.

[139] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models." In: *Digital Signal Processing* 10.1 (2000), pp. 19–41.

[140] Davit Rizhinashvili, Abdallah Hussein Sham, and Gholamreza Anbarjafari. "Gender Neutralisation for Unbiased Speech Synthesising." In: *Electronics* 11.10 (2022).

[141] Luis Javier Rodríguez-Fuentes, Niko Brümmer, Mikel Penagarikano, Amparo Varona, Germán Bordel, and Mireia Diez. "The Albayzin 2012 Language Recognition Evaluation." In: (2013), pp. 1497–1501.

[142] Richard C. Rose and Douglas A. Reynolds. "Text independent speaker identification using automatic acoustic segmentation." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech, and Signal Processing*. 1990, pp. 293–296.

[143] Leonard J. Savage. "Elicitation of Personal Probabilities and Expectations." In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801.

[144] Ali Shahin Shamsabadi, Brij Mohan Lal Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot. "Differentially Private Speaker Anonymization." In: *Proc. Privacy Enhancing Technologies Symposium* 2023.1 (2023), pp. 98–114.

[145] C. E. Shannon. "A mathematical theory of communication." In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[146] C. E. Shannon. "Communication theory of secrecy systems." In: *The Bell System Technical Journal* 28.4 (1949), pp. 656–715.

[147] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 132–157.

[148] Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans. "Probabilities for SV Machines." In: *Advances in Large-Margin Classifiers*. 2000, pp. 61–73.

[149] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "X-vectors: Robust DNN embeddings for speaker recognition." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2018, pp. 5329–5333.

[150] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. "Evaluating Voice Conversion-based Privacy Protection against Informed Attackers." In: *Proc. IEEE ICASSP - International Conference on Acoustics, Speech and Signal Processing*. 2020.

[151] Dimitrios Stoidis and Andrea Cavallaro. "Generating gender-ambiguous voices for privacy-preserving speech recognition." In: *Proc. ISCA Interspeech*. 2022, pp. 4237–4241.

[152] André Stuhlsatz, Jens Lippel, and Thomas Zielke. "Feature Extraction With Deep Neural Networks by a Generalized Discriminant Analysis." In: *IEEE Transactions on Neural Networks and Learning Systems* 23.4 (2012), pp. 596–608.

[153] Latanya Sweeney. "K-Anonymity: A Model for Protecting Privacy." In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 557–570.

[154] Natalia Tomashenko et al. "Introducing the VoicePrivacy Initiative." In: *Proc. ISCA Interspeech*. 2020, pp. 1693–1697.

[155] Natalia Tomashenko et al. "Introducing the VoicePrivacy Initiative." In: *Proc. ISCA Interspeech*. 2020, pp. 1693–1697.

[156] Natalia Tomashenko et al. "The VoicePrivacy 2020 Challenge: Results and findings." In: *Computer Speech & Language* 74 (2022), p. 101362.

[157] Karl Gerald Van den Boogaart, Juan José Egozcue, and Vera Pawlowsky-Glahn. "Bayes hilbert spaces." In: *Australian & New Zealand Journal of Statistics* 56.2 (2014), pp. 171–194.

[158] Xin Wang and Junichi Yamagishi. "Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis." In: *Proc. ISCA Workshop on Speech Synthesis*. 2019, pp. 1–6.

[159] Jason Weston, Bernhard Schölkopf, and Gökhan Bakir. "Learning to Find Pre-Images." In: *Advances in Neural Information Processing Systems*. Ed. by S. Thrun, L. Saul, and B. Schölkopf. Vol. 16. MIT Press, 2003.

[160] Jennifer Williams, Junichi Yamagishi, Paul-Gauthier, Cassia Valentini-Botinhao, and Jean-François Bonastre. "Revisiting Speech Content Privacy." In: *Proc. ISCA Symposium on Security and Privacy in Speech Communication*. 2021, pp. 42–46.

[161] Robert L. Winkler and Allan H. Murphy. ""Good" Probability Assessors." In: *Journal of Applied Meteorology* 7.5 (1968), pp. 751–758.

[162] Bianca Zadrozny and Charles Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates." In: *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2002).

[163] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. *LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech*. 2019.

[164] C. Zhang. "Acoustic analysis of disguised voices with raised and lowered pitch." In: *Proc. International Symposium on Chinese Spoken Language Processing*. 2012, pp. 353–357.

[165] Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. "Attribute Privacy: Framework and Mechanisms." In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 757–766.