The distribution of calibrated likelihood functions on the probability-likelihood Aitchison simplex

Paul-Gauthier Noé*

PAUL-GAUTHIER.NOE@CNRS.FR

Centre National de la Recherche Scientifique (CNRS) Laboratoire d'Informatique et des Systèmes (LIS) Aix-Marseille Université, France

Andreas Nautsch

Individual, Germany

Driss Matrouf,

Laboratoire d'Informatique d'Avignon (LIA) Avignon Université, France

Pierre-Michel Bousquet,

Laboratoire d'Informatique d'Avignon (LIA) Avignon Université, France

Jean-François Bonastre,

Laboratoire d'Informatique d'Avignon (LIA) Avignon Université, France

Abstract

While calibration of probabilistic predictions has been widely studied, this paper rather addresses calibration of likelihood functions. This has been discussed, especially in biometrics, in cases with only two exhaustive and mutually exclusive hypotheses (classes) where likelihood functions can be written as log-likelihood-ratios (LLRs). After defining calibration for LLRs and its connection with the concept of weight-of-evidence, we present the idempotence property and its associated constraint on the distribution of the LLRs. Although these results have been known for decades, they have been limited to the binary case. Here, we extend them to cases with more than two hypotheses by using the Aitchison geometry of the simplex, which allows us to recover, in a vector form, the additive form of the Bayes' rule; extending therefore the LLR and the weight-of-evidence to any number of hypotheses. Especially, we extend the definition of calibration, the idempotence, and the constraint on the distribution of likelihood functions to this multiple hypotheses and multiclass counterpart of the LLR: the isometric-log-ratio transformed likelihood function. This work is mainly conceptual, but we still provide one application to machine learning by presenting a non-linear discriminant analysis where the discriminant components form a calibrated likelihood function over the classes, improving therefore the interpretability and the reliability of the method.

Keywords: calibration, log-likelihood-ratio, weight-of-evidence, likelihood function, Bayes' rule, probability simplex, Aitchison geometry, multiple hypotheses & multiclass, discriminant analysis & generative classification

^{*.} Most of this work was done when Paul-Gauthier Noé was a PhD student at LIA, Avignon Université, supported by the VoicePersonae project ANR-18-JSTS-0001.

1 Introduction

Calibration has been introduced for probabilistic predictions in the context of weather forecasting, where a forecaster has to make previsions by assigning to each day a probability for rain (Brier, 1950; Winkler and Murphy, 1968; DeGroot and Fienberg, 1983). The predictions are said to be calibrated if the probabilities match the observed outcomes: over a long sequence of predictions, the relative frequency of days where it actually rained and on which the probability p has been assigned must be p (DeGroot, 1970; Dawid, 1982).

In machine learning, we are rather interested in data modeling and classification tasks. Like a weather forecaster, a classifier "is" uncertain, and then naturally outputs a probability distribution over the set of classes (we also call *hypotheses* here) rather than making a hard decision. However, for the uncertainty to be well-encoded in the classifier's probabilistic predictions, and for them to be used for cost-sensitive decisions, they have to be calibrated.

Considering a classifier $\mathbf{q}: \mathcal{X} \to \mathcal{S}^D$ that¹, given an input $x \in \mathcal{X}$, outputs a prediction in the form of a posterior probability distribution over the set of hypotheses (or classes): $\mathbf{q}(x) = [P_{\theta}(H_1 \mid x), \dots P_{\theta}(H_D \mid x)]^T \in \mathcal{S}^D$. A set \mathcal{Q} of probabilistic predictions is *perfectly* calibrated if (Bröcker, 2009),

$$\forall q(x) \in \mathcal{Q}, \quad \mathbb{P}(H_i \mid q(x)) = q_i(x) = P_{\theta}(H_i \mid x) \quad \forall i. \tag{1}$$

Meaning that, for all predictions, the conditional distribution over the set of hypotheses (or classes) given the prediction, is equal to the prediction. While calibration has been discussed in machine learning for decades (Zadrozny and Elkan, 2001, 2002), it has been the subject of renewed interest especially since Guo et al. (2017) discussed the tendency of modern neural networks to produce overconfident predictions.

In this paper, we will look at calibration from a different point of view than the one people in statistics and machine learning are generally used to. In our framework, a prediction will be in the form of a likelihood function over the set of possible hypotheses. This in no way reduces the relevance of our work for the statistical machine learning community since, as we will see, probabilistic predictions and corresponding likelihood functions are isomorphic given a prior and under some scale-invariance equivalence relation. To be more precise, we are rather interested in reporting statistical evidences rather than making predictions. A probabilistic prediction can then be made by combining a prior and a statistical evidence, represented by a calibrated likelihood function, through the Bayes' rule.

Reporting statistical evidence in the form of a likelihood function has been extensively discussed and promoted in forensic science (Aitken and Taroni, 2004; Meester and Slooten, 2021; Aitken et al., 2024) or in the context of medical diagnostic (Thornbury et al., 1975). In those cases, there are only two competiting hypotheses such that the likelihood function can be written in the form of a log-likelihood-ratio (LLR), or weight-of-evidence (WOE). The Bayes' rule can be written in its log-odds form: the posterior log-odds is the sum of the LLR and the prior log-odds. In this way, the LLR tells how new data is changing the personal belief from the prior to the posterior in an additive manner.

The calibration of LLRs has been specially developed in the context of speaker verification (Brümmer and du Preez, 2006; Brümmer, 2010; Ramos, 2007). In particular, the

^{1.} \mathcal{S}^D is the probability simplex. See Section 3 for more details.

idempotence property of calibrated LLRs and its associated constraint on their distribution are of great importance for understanding the calibration of LLRs. The idempotence tells that "the LLR of the LLR is the LLR" and that if calibrated LLRs are normally distributed under one hypothesis, it is also normally distributed under the other hypothesis, with an opposite mean, and a shared variance equal to twice the mean. These results have been proofed for calibrated LLRs in the context of speaker verification (van Leeuwen and Brümmer, 2013) but have been known for the WOE since at least the 40s (Good, 1979).

However, the LLR, its calibration, and its associated properties are defined only for the binary case, i.e. when there are only two exhaustive and mutually exclusive hypotheses. The main purpose of this work is to extend these concepts to the multiple hypotheses and multiclass case. Starting from the log-odds form of the Bayes' rule, Section 2 presents the WOE and recalls the definition of calibration for the LLR, the idempotence property, how it is related to the WOE, and the associated constraint on the distribution of calibrated LLRs. Section 3 presents the Aitchison geometry of the probability simplex (Aitchison, 1982). We will see how this allows us to extend the log-odds and additive form of the Bayes' rule generalizing therefore the concept of LLR, in a vector form, to a multiple hypotheses setting. This multiple hypotheses counterpart of the LLR is called the isometric-log-ratio transformed likelihood function (ILRL) and we will see in Section 4 how the idempotence property applies to it. We will also see how the constraint on the distribution generalizes to ILRs i.e. to likelihood functions over a set of more than two possible hypotheses. Finally, Section 5 presents one application of these results to machine learning, by proposing a non-linear discriminant analysis where the discriminant space is designed according to the idempotence property to form a space of calibrated likelihood functions.

Our contributions can be summarized as follows:

- By taking the work by Egozcue and Vera (2018) over, we extend the concept of LLR to any number of hypotheses thanks to the Aitchison geometry of the simplex. The resulting quantity is the isometric-log-ratio transformed likelihood function (ILRL);
- We extend the concept of calibration and the idempotence of the LLR to its multiple hypotheses counterpart: the ILRL;
- We prove a constraint on the distribution of calibrated ILRLs: if they are normally distributed under one hypothesis, they are also normally distributed for the other hypotheses with some additional constraints on their parameter. This result generalizes what has been known for the weight-of-evidence and calibrated LLRs for decades (Peterson et al., 1954; Good, 1979, 1985; van Leeuwen and Brümmer, 2013);
- We present, as an application of the above results, a non-linear discriminant analysis we call *Compositional discriminant analysis*, where the discriminant components form a calibrated likelihood function over the set of classes making this approach reliable and easy-to-interpret².

^{2.} This contribution has been presented as a poster at CoDaWork2024, the 10th International Workshop on Compositional Data Analysis (Noé et al., 2024a). The binary case has been presented earlier in the context of privacy preservation in speech technologies (Noé et al., 2022).

2 From the weight-of-evidence to calibrated log-likelihood-ratios

Let's consider a set of exhaustive and mutually exclusive *simple* hypotheses $\mathcal{H} = \{H_1, \dots H_D\}^3$. Let's consider an individual who wants to infer which hypothesis is true given the data or *evidence* x. Its posterior probabilities are given by the Bayes' rule:

$$\forall H \in \mathcal{H}, \ P(H \mid x) \propto P(x \mid H)P(H), \tag{2}$$

where $[P(x \mid H_i)]_{1 \leq i \leq D} \in \mathbb{R}_+^{*D}$ is the likelihood function over the set of hypotheses, and $[P(H_i)]_{1 \leq i \leq D} \in \mathcal{S}^D$ is the prior probability distribution representing the prior personal belief of the individual, i. e. its belief based on all the information available to him or her other than the evidence x. When there are only two competing hypotheses, i. e. $\mathcal{H} = \{H_1, H_2\}$, the Bayes' rule can be written in its log-odds or logit form:

$$\underbrace{\operatorname{logit} P(H_1 \mid x)}_{\text{posterior log-odds}} = \underbrace{\operatorname{log} \frac{P(x \mid H_1)}{P(x \mid H_2)}}_{\text{weight-of-evidence}} + \underbrace{\operatorname{logit} P(H_1)}_{\text{prior log-odds}} \tag{3}$$

where $logit(p) = log \frac{p}{1-p}$ for 0 .

The posterior is here the sum between a term that depends only on the prior probabilities and a term that depends only on the likelihoods. The latter is the weight-of-evidence—or log Bayes-factor—and informs about the contribution of the data x in the computation of the posterior. In Good (1985), the author wrote that "[...] the weight-of-evidence tells us just as much as |x| does about the odd of $|H_1|$ and $|H_2|$ " stating therefore that:

$$w(x) = \log \frac{P(x \mid H_1)}{P(x \mid H_2)} = \log \frac{P(w(x) \mid H_1)}{P(w(x) \mid H_2)}.$$
 (4)

This makes the weight-of-evidence w(x) a good candidate for representing the statistical evidence—in favor of H_1 and against H_2 —in the data x.

However, the hypotheses are here *simple* statistical hypotheses, such that Equation 4 is an intrinsic property of the weight-of-evidence as in Meester and Slooten (2021) and Good (1985). In machine learning, especially with generative classifiers, the likelihoods $P(x \mid H_1)$ and $P(x \mid H_2)$ are computed with respect to statistical models that may not reflect the "true" distribution of the data. This would result in an uncalibrated representation of the statistical evidence. Equation 4 becomes therefore a desired property for the log-ratio of the likelihoods—computed with respect to the models—to properly represent the statistical evidence, and to be interpreted as a weight-of-evidence.

2.1 Calibration for log-likelihood-ratios

In machine learning, especially in the context of generative classification, we do not have access to a weight-of-evidence. We rather compute a log-likelihood-ratio (LLR) as a log-ratio of probability density functions:

$$l_{\theta}(x) = \log \frac{f_{\theta_{\mathcal{X}_1}}(x)}{f_{\theta_{\mathcal{X}_2}}(x)} \tag{5}$$

^{3.} In a classification context, each hypothesis would correspond to a class such that for a given sample, the hypothesis H_i should be read as "the sample belongs to the *i*th class".

where $\theta_{\mathcal{X}_i}$ refers to a statistical model for the data under hypothesis H_i . The classifier produces here a LLR and the posterior is obtained as a function of the LLR and a prior⁴ through the Bayes' rule:

$$\log \operatorname{tr} P_{\theta}(H_{1} \mid x) = \log \frac{f_{\theta \chi_{1}}(x)}{f_{\theta \chi_{2}}(x)} + \operatorname{logit} P(H_{1})$$

$$\iff q_{1}(x) = P_{\theta}(H_{1} \mid x) = \operatorname{sigmoid} \left(\log \frac{f_{\theta \chi_{1}}(x)}{f_{\theta \chi_{2}}(x)} + \operatorname{logit} P(H_{1}) \right),$$
(6)

where sigmoid(l) = $1/(1 + \exp(-l))$, with $l \in \mathbb{R}$, and is the inverse of the logit.

From Equation 6, we can see that, for a given a prior, there is a bijection between the posterior and the LLR. In the definition of calibration in Equation 1, we can therefore interchange the set of probabilistic prediction Q with the set \mathcal{L} of corresponding LLRs:

$$\forall l_{\theta} \in \mathcal{L} = \{l_{\theta}(x) \mid x \in \mathcal{X}\},\,$$

$$\mathbb{P}(H_i \mid l_\theta) = q_i(x) = P_\theta(H_i \mid x) \quad \forall i \in \{1, 2\},\tag{7}$$

$$\iff \log \frac{\mathbb{P}(H_1 \mid l_{\theta})}{\mathbb{P}(H_2 \mid l_{\theta})} = \log \frac{P_{\theta}(H_1 \mid x)}{P_{\theta}(H_2 \mid x)}, \tag{8}$$

$$\iff \log \frac{f_{\mathcal{L}_1}(l)}{f_{\mathcal{L}_2}(l)} + \log \frac{P(H_1)}{P(H_2)} = \log \frac{f_{\theta_{\mathcal{X}_1}}(x)}{f_{\theta_{\mathcal{X}_2}}(x)} + \log \frac{P(H_1)}{P(H_2)}, \tag{9}$$

$$\iff \log \frac{f_{\mathcal{L}_1}(l_{\theta})}{f_{\mathcal{L}_2}(l_{\theta})} = l_{\theta}, \tag{10}$$

where $f_{\mathcal{L}_i}$ is the probability density function⁵ of the "true" distribution of the LLR under hypothesis H_i . The last line can be read as:

This expression was popularized in the context of calibrated LLRs for speaker verification systems (van Leeuwen and Brümmer, 2013) but can be traced back to the theory of signal detectability (Birdsall, 1966).

This is the *idempotence* property of calibrated LLRs and takes us back to Equation 4 where the weight-of-evidence of the weight-of-evidence is the weight-of-evidence itself. In a way, one intuition behind the calibration of log-likelihood-ratios is to make them interpretable as weights-of-evidence.

The equality in Equation 10 may not hold because the actual distribution of the LLR may not match the statistical models' assumed distribution. Hence the following definition of calibrated LLRs (van Leeuwen and Brümmer, 2013):

Definition 1 A set \mathcal{L} of log-likelihood-ratios is perfectly calibrated if they are idempotent:

$$\forall l_{\theta} \in \mathcal{L}, \quad \log \frac{f_{\mathcal{L}_{1}}(l_{\theta})}{f_{\mathcal{L}_{2}}(l_{\theta})} = l_{\theta}. \tag{11}$$

^{4.} Usually taken as the empirical class proportion in the training set.

^{5.} Assuming it exists.

The intuition is the same as the standard definition of calibration in Equation 1, where we want the conditional distribution over the hypotheses given the prediction to be equal to the prediction. Here, we want the log-likelihood-ratio of the prediction—where the prediction is here in a form of a LLR—to be equal to the prediction, i.e. the LLR.

In the following, we will see how the idempotence property leads to a constraint on the distribution of the LLRs.

2.2 The distribution of calibrated LLRs

Equation 11 can be rewritten as $f_{\mathcal{L}_1}(l) = e^l f_{\mathcal{L}_2}(l)$. This shows that if the distribution of the log-likelihood-ratio is known for one hypothesis, the distribution under the other hypothesis is completely determined. Therefore, the idempotence property leads to a constraint on the distribution of calibrated LLRs. The Gaussian case is illustrated by the following proposition.

Theorem 2 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 calibrated LLRs are normally distributed for one hypothesis, they are necessarily normally distributed for the other hypothesis, with an opposite mean, and the variances are the same and are equal to twice the mean.

Because of Equation 4, this result holds for the weight-of-evidence and is known by statisticians since at least I.J. Good and A.M. Turing's work (Good, 1979, 1985), and has been used in detection theory for radars like in (Peterson et al., 1954). It has then been reproofed and discussed later in the context of calibrated LLRs with applications on speaker verification (van Leeuwen and Brümmer, 2013) and in the context of forensic identification (Meester and Slooten, 2021). We give a detailed proof in Appendix A.

Example: In order to illustrate Definition 1 and Theorem 2, let's consider the linear discriminant analysis as a simple example. Let's consider two classes H_1 and H_2 for which the data is assumed to be normally distributed with the same covariance matrix:

$$x \mid H_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}),$$

 $x \mid H_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$ (12)

The LLR is given by:

$$l(\mathbf{x}) = \log \frac{f_{\theta_{\mathcal{X}_1}}(\mathbf{x})}{f_{\theta_{\mathcal{X}_2}}(\mathbf{x})} = \mathbf{x}^T \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \left(\boldsymbol{\mu}_2^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_1 \right).$$
(13)

Let's consider that the Gaussian assumptions are correct and that the data is indeed distributed according to Equation 12 (this is like having *simple* hypotheses where each hypothesis specify the distribution of the data). Then, for each class, l(x) is normally distributed (because this is a projection of a normally distributed random variable). It can be shown that for the *i*th class, the mean and variance of the LLR are:

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{i}, \boldsymbol{\Sigma})} [l(\boldsymbol{x})] = \frac{(-1)^{i-1}}{2} (\boldsymbol{\mu}_{1} - \boldsymbol{\mu}_{2})^{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{1} - \boldsymbol{\mu}_{2}),$$

$$= (-1)^{i-1} \boldsymbol{\mu}, \quad \text{where } \boldsymbol{\mu} = \frac{1}{2} (\boldsymbol{\mu}_{1} - \boldsymbol{\mu}_{2})^{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{1} - \boldsymbol{\mu}_{2})$$

$$\mathbb{V}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{i}, \boldsymbol{\Sigma})} [l(\boldsymbol{x})] = (\boldsymbol{\mu}_{1} - \boldsymbol{\mu}_{2})^{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{1} - \boldsymbol{\mu}_{2}) = 2\boldsymbol{\mu},$$

$$(14)$$

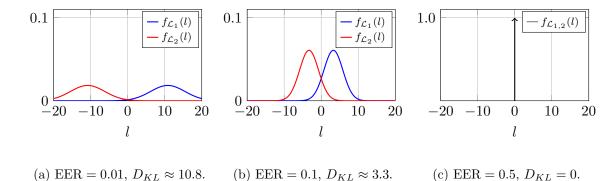


Figure 1: Examples of Gaussian densities of calibrated LLRs under each hypothesis.

respecting therefore Theorem 2. With the normally distributed LLRs as described just above, it can also be shown that $\log \frac{f_{\mathcal{L}_1}(l)}{f_{\mathcal{L}_2}(l)} = l$ recovering the idempotence property. However, note that the idempotence and Theorem 2 are here valid because we have considered the data as distributed according to the Gaussian assumptions. Otherwise, this would not have been the case, resulting in uncalibrated LLRs.

The parameter and the separability. The only parameter of the distributions in Theorem 2 is a scalar: the mean μ (or equivalently the variance $\sigma^2 = 2\mu$). 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 classification context. In van Leeuwen and Brümmer (2013), the authors 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}) :

$$D_{KL}(f_{\mathcal{L}_{1}}||f_{\mathcal{L}_{2}}) = \int_{-\infty}^{+\infty} f_{\mathcal{L}_{1}}(l) \log \frac{f_{\mathcal{L}_{1}}(l)}{f_{\mathcal{L}_{2}}(l)} dl,$$

$$= \int_{-\infty}^{+\infty} f_{\mathcal{L}_{1}}(l) l dl \text{ because of the idempotence: } \log \frac{f_{\mathcal{L}_{1}}(l)}{f_{\mathcal{L}_{2}}(l)} = l,$$

$$= \mathbb{E}_{l \sim \mathcal{N}(\mu, 2\mu)}[l] = \mu.$$
(15)

The mean is therefore equal to the Kullback-Leibler divergence. Note that since the two densities are Gaussian with the same variance, the D_{KL} is symmetric. Figure 1 shows examples of Gaussian densities of the LLR under each hypothesis. When the mean increases, the variance and the separability increase. When the separability is 0, i. e. when EER = 0.5 and $D_{KL} = 0$, both densities are a Dirac delta function at 0.

Theorem 2 gives a reference distribution for the LLRs to be calibrated, which has been applied especially for the calibration and the evaluation of of speaker verification systems. However, all the results presented so far are for the two hypotheses case: we have presented the concept of calibration for LLRs. The additivity of the Bayes' rule in its log-odds form, and the concepts of LLR and WOE, have been considered as not extensible to cases where more than two hypotheses are possible, see for instance Section 4.3 in Jaynes (2003). We agree with this but only when log-ratios are treated one by one, independently from one another. In the next section, we will see how treating probability distributions and likelihood functions as compositional data provides an elegant manner for treating all log-ratios, at

once, in a vector form; and how the additivity of the Bayes' rule is recovered generalizing therefore the concept of LLR.

3 The Aitchison geometry of the probability-likelihood simplex

In this section, we will see how treating discrete probability distributions and discrete likelihood functions as compositional data allows us to recover the additive form of the Bayes' rule, extending therefore the concept of LLR, in a vector form, to any number of hypotheses.

Compositional data carries relative information. A composition is a vector where each element describes a part of some whole (Pawlowsky-Glahn et al., 2015) like vectors of proportions, concentrations, and probabilities. Compositional data analysis aims in treating such data by taking into account the compositional nature and structure of the data⁶. A D-part composition is a vector of D 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:

$$S^{D} = \left\{ \boldsymbol{x} = [x_{1}, x_{2}, \dots x_{D}]^{T} \in \mathbb{R}_{+}^{*D} \middle| \sum_{i=1}^{D} x_{i} = k \right\}.$$
 (16)

This is how the simplex is usually defined. However, having the sum of each parts equal to a constant k is not what really matter. Only the relative information between parts is important. We therefore introduce the following equivalence relation:

$$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_{+}^{*D}, \quad \boldsymbol{y} \sim \boldsymbol{x} \iff \exists c > 0, \text{ such that } \boldsymbol{y} = c\boldsymbol{x}.$$
 (17)

The simplex is then defined as the set of equivalent classes, i.e. as the quotient space:

$$S^D = \mathbb{R}_+^{*D} / \sim . \tag{18}$$

This formulation allows us to see both the probability distributions and the likelihood functions as living in the same space: the probability simplex as the set of equivalent classes (where k=1). Indeed, while the sum of the probabilities is equal to one, the likelihoods do not sum to a constant. However, since multiplying all the likelihoods by the same constant carries the same information⁷, likelihood functions can be seen as compositional data too. Hence, from now on, when we discuss a likelihood function, as a vector $\mathbf{w} \in \mathbb{R}_+^{*D}$ of likelihoods, we refer to its equivalent that lives on the probability simplex.

We refer to this simplex as the *probability-likelihood* simplex. Figure 2 illustrates likelihood equivalent classes. Likelihood lines (in dashed blue) go through the simplex S^3 . Within a line, all likelihood functions are equivalent and we take the likelihood function C(w) that lives on the simplex as the representative of this equivalent class.

This equivalence is materialized by the closure operator \mathcal{C} . Since only the relative information matter, scaling factors are irrelevant and a composition x is equivalent to its

^{6.} For an overview of compositional data analysis, the reader can refer to Pawlowsky-Glahn et al. (2015).

^{7.} See the likelihood princible (Berger and Wolpert, 1988).

normalized version that lives on the simplex. The closure is defined for k=1 as:

$$C\left(\boldsymbol{x}\right) = \left[\frac{x_1}{\|\boldsymbol{x}\|_1}, \frac{x_2}{\|\boldsymbol{x}\|_1}, \dots \frac{x_D}{\|\boldsymbol{x}\|_1}\right]^T, \tag{19}$$

where $\boldsymbol{x} \in \mathbb{R}_+^{*D}$ and $\|\boldsymbol{x}\|_1 = \sum_{i=1}^D |x_i|$. Therefore, any vector of positive real numbers can be mapped to its equivalent on the simplex using the closure.

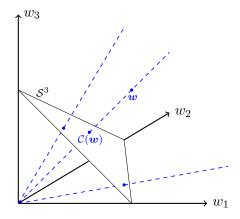


Figure 2: The probability-likelihood simplex and likelihood lines as equivalent classes. All likelihood functions that live on the same blue dashed ray are equivalent, and can be represented by the likelihood function that lives on the probability simplex.

To handle the scale-invariance nature of compositional data, John Aitchison introduced the use of log-ratios of components (Aitchison, 1982). He defined several operations on the simplex leading to the *Aitchison geometry of the simplex*.

3.1 The Aitchison geometry of the simplex

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

• perturbation⁸:

$$\boldsymbol{x} \oplus \boldsymbol{y} =_{\sim} [x_1 y_1, \dots x_D y_D]^T, \tag{20}$$

• powering:

$$\alpha \odot \boldsymbol{x} =_{\sim} [x_1^{\alpha}, \dots x_D^{\alpha}]^T, \tag{21}$$

• inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}$$
 (22)

^{8.} Where "= \sim " is "= $\mathcal{C}(\cdot)$ " or also " \propto ". Any scaling factor is indeed irrelevant under the equivalent relation \sim .

where $x, y \in \mathbb{R}_+^{*D}$ and $\alpha \in \mathbb{R}$. The perturbation and powering give to the simplex a (D-1)-dimensional vector space structure and the inner product makes it Euclidean. The corresponding norm and distance are:

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \sqrt{\frac{1}{2N} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j}\right)^2},$$
 (23)

$$d_{\mathcal{A}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_{\mathcal{A}} = \|\boldsymbol{x} \oplus ((-1) \odot \boldsymbol{y})\|_{\mathcal{A}}$$
$$= \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j}\right)^2},$$
(24)

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*.

One can already notice the extensive use of log-ratios of parts. Hence the analogy with the log-odds of Section 2.

3.2 The isometric-log-ratio transformation

Thanks to the Euclidean vector space structure of the simplex, the probability distributions and likelihood functions can be expressed in a Cartesian coordinate system using the Aitchison inner product and an orthonormal basis of the simplex. Let the set $\{e^{(i)} \in \mathcal{S}^D, i \in \{1, \dots D-1\}\}$ be such an Aitchison orthonormal basis. The elements of one basis obtained using the Gram-Schmidt procedure as in Egozcue et al. (2003) are defined for all $i \in \{1, \dots D-1\}$ as follows:

$$e^{(i)} = \mathcal{C}\left(\left[\underbrace{\exp\left(\sqrt{\frac{1}{i(i+1)}}\right), \dots \exp\left(\sqrt{\frac{1}{i(i+1)}}\right)}_{\text{The first } i \text{ elements}}, \exp\left(-\sqrt{\frac{i}{i+1}}\right), 1, \dots 1\right]\right). \quad (25)$$

The Isometric-Log-Ratio (ILR) transformation (Egozcue et al., 2003) allows to express a composition $p \in \mathcal{S}^D$ in a Cartesian coordinate system by projecting it onto the basis as follows⁹:

$$\tilde{\boldsymbol{p}} = ilr(\boldsymbol{p}) = \left[\langle \boldsymbol{p}, \boldsymbol{e}^{(1)} \rangle_{\mathcal{A}}, \dots \langle \boldsymbol{p}, \boldsymbol{e}^{(D-1)} \rangle_{\mathcal{A}} \right]^{T}.$$
 (26)

This defines an isometric isomorphism¹⁰ between S^D and \mathbb{R}^{D-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 3. Considering

^{9.} We use the definite article the to refer to the ILR transformation. This may suggest that there is only one ILR transformation, while there are as many ILR transformations as they are Aitchison orthonormal bases on the simplex i.e. an uncountable number. Along this article and without loss of generality, the expression "the ILR transformation" will refer to the one with the orthonormal basis defined in Equation 25. 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 (Egozcue et al., 2003).

^{10.} An isometric isomorphism is an invertible mapping that preserves the distances.

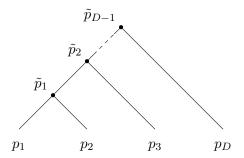


Figure 3: Bifurcating tree corresponding to the orthonormal basis of Equation 25 obtained with the Gram-Schmidt procedure (Egozcue et al., 2003).

a vector $\boldsymbol{p} = [p_1, \dots p_D]^T \in \mathcal{S}^D$ and its ILR transformation $\tilde{\boldsymbol{p}} = \operatorname{ilr}(\boldsymbol{p}) = [\tilde{p}_1, \dots \tilde{p}_{D-1}]^T \in \mathbb{R}^{D-1}$, each node of the tree corresponds to a component \tilde{p}_i of $\tilde{\boldsymbol{p}}$. The first component compares the probabilities (or likelihoods) for the two first hypotheses. Each next component then recursively compares the probability (likelihood) for the next hypothesis with the probabilities (likelihoods) for the previous ones. The *i*th element \tilde{p}_i of the ILR transformation of a composition \boldsymbol{p} can be obtained with the following formula:

$$\tilde{p}_i = \langle \boldsymbol{p}, \boldsymbol{e}^{(i)} \rangle_{\mathcal{A}} = \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod\limits_{j=1}^{i} p_j}{(p_{i+1})^i} \right).$$
 (27)

An ILR component can therefore be interpreted as a weight (like a weight-of-evidence) comparing a probability (likelihood) with a group of other probabilities (likelihoods). When the probability (likelihood) for the (i + 1)th hypothesis increases and the probabilities (likelihoods) for the hypotheses $H_{1 \leq j \leq i}$ 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 $H_{1 \leq j \leq i}$ independently of the hypotheses $H_{i+2 \leq j \leq D}$.

We saw that a composition carries relative rather than absolute information. The treatment of compositional data is therefore based on ratios and in particular on log-ratios. It is worth noting the natural analogy with log-odds and log-likelihood-ratios as presented in Section 2 with the Bayes' rule. In the next section, we take a deeper dive into this analogy.

3.3 The Bayes' rule as a vector translation

The computation of the posterior probabilities through the Bayes' rule is the product of the prior probabilities with the likelihoods, normalized by P(x) given by the law of total probability. This is exactly the perturbation (Equation 20) of the prior probability vector by the likelihood vector (where the closure ensures the normalization). Let¹¹:

^{11.} Note that the Aitchison geometry is based on log-ratios such that a composition can not contain zeros. In the definition of the simplex in Equations 16 and 18, the zeros are indeed excluded. Dealing with zeros has been problematic in compositional data analysis (Martín-Fernández et al., 2003; Pawlowsky-Glahn et al., 2015). However, banning probabilities equal to zero is not an issue for us. In the context of

- $\pi = [P(H_1), P(H_2), \dots P(H_D)]^T \in \mathcal{S}^D$ be the vector of prior probabilities assigned to each hypothesis, i.e. the prior probability distribution;
- $\boldsymbol{w} = [P(x \mid H_1), P(x \mid H_2), \dots P(x \mid H_D)]^T \in \mathbb{R}_+^{*D}$ be the vector of likelihoods, i. e. the likelihood function;
- $\mathbf{P} = [P(H_1 \mid x), P(H_2 \mid x) \dots P(H_D \mid x)]^T \in \mathcal{S}^D$ be the posterior probability distribution.

The Bayes' rule is:

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

$$\iff \mathbf{P} = \begin{bmatrix} \frac{w_1 \pi_1}{\sum_{j=1}^{D} w_j \pi_j}, \frac{w_2 \pi_2}{\sum_{j=1}^{D} w_j \pi_j}, \dots, \frac{w_D \pi_D}{\sum_{j=1}^{D} w_j \pi_j} \end{bmatrix}^T,$$

$$\iff \mathbf{P} = \mathcal{C}\left([w_1 \pi_1, w_2 \pi_2, \dots w_D \pi_D]\right) = \mathbf{w} \oplus \mathbf{\pi}.$$

$$(28)$$

The Bayes' rule is the perturbation of the prior distribution by the likelihood function.

In the Isometric-Log-Ratio (ILR) space, i.e. the space \mathbb{R}^{D-1} isometrically isomorphic to the simplex through the ILR transformation, a perturbation is a vector translation. Therefore, in the coordinate representation given by the ILR transformation, the Bayes' rule can be written as a vector translation of the prior by the likelihood function (Egozcue and Vera, 2018):

$$P = \boldsymbol{w} \oplus \boldsymbol{\pi},$$

$$ilr(\boldsymbol{P}) = ilr(\boldsymbol{w}) + ilr(\boldsymbol{\pi}),$$

$$\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{w}} + \tilde{\boldsymbol{\pi}}.$$
(29)

Just like the logit transformation in Equation 3, the ILR transformation allows us to write the Bayes' rule as a sum between a term that depends only on the prior probabilities

Bayesian updating, probabilities equal to zero are indeed 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, it means that this is already certain for you that the corresponding hypothesis is false. No matter what evidence you observe or how someone is trying to convince you, your opinion about this hypothesis can not change. 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 is called the Cromwell's rule (Lindley, 2006). If you initially consider a 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" (de Finetti, 1975). However, the Cromwell's rule holds for 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 treated as compositions can not contain zeros. In this case, the zeros have to be replaced. Zeros replacement strategies for compositional data are discussed by Martín-Fernández et al. (2003).

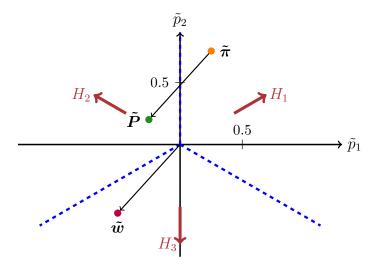


Figure 4: Bayesian updating in the three-hypotheses ILR space. The posterior distribution $\tilde{\boldsymbol{P}}$ is the translation of the prior distribution $\tilde{\boldsymbol{\pi}}$ by the likelihood function $\tilde{\boldsymbol{w}}$. The red arrows indicate the directions that go in favor of one hypothesis against the two others. The dashed blue rays mark out the maximum probability (or likelihood) decision regions.

and a term that depends only on the likelihoods. The ILR transformation is therefore the multidimensional, multiple hypotheses, or multiclass, extension of the logit transformation.

In this way, the appealing additivity of the Bayes' rule is recovered. To be more precise, the likelihood function $\tilde{\boldsymbol{w}}$ translates a prior probability distribution $\tilde{\boldsymbol{\pi}}$ into a posterior distribution $\tilde{\boldsymbol{P}}$. Moreover, the ILR transformation in a two hypotheses case results in a one-element vector: the log-ratio of the probabilities (or likelihoods)¹²; which is consistent with Equation 3.

Figure 4 shows an example of a Bayesian updating in a three-hypotheses ILR space. The first component \tilde{p}_1 compares the probability (likelihood) for hypothesis H_1 with the probability (likelihood) for hypothesis H_2 , and the second component \tilde{p}_2 compares the third probability (likelihood) against the two others as illustrated by the bifurcation tree in Figure 3. 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 probability (likelihood) decision regions. Here, the simplex is 2-dimensional because we consider three possible hypotheses but keep in mind that for D hypotheses, the simplex is (D-1)-dimensional. When there are only two possible hypotheses, the simplex is one-dimensional such that if one goes against one hypothesis, it necessarily goes in favor of the other, and we recover the situation of Section 2. With more hypotheses, the number of directions is now uncountable.

^{12.} To be more precise, with the basis of Equation 25, there is a scaling factor $\frac{1}{\sqrt{2}}$.

3.4 Statistical evidence representation for multiple hypotheses: from the LLR to the Isometric-Log-Ratio Likelihood function

Recovering the additive form of the Bayes' rule—being the "basic property" of the weight-of-evidence (Good, 1985)—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—that we will now call ILRL for Isometric-Log-Ratio transformed Likelihood function—can be seen as a multidimensional extension of the LLR making it a good candidate for representing the statistical evidence when there are more than two possible hypotheses.

The direction of the ILRL informs which 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 3). 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 (Egozcue and Pawlowsky-Glahn, 2005), or alternatively, to project the ILRL vector on the corresponding direction. The Aitchison norm of the likelihood function (i. e. the Euclidean norm of its ILR transformation) can be regarded as a global strength-of-evidence and is given by:

$$\|\boldsymbol{w}(x)\|_{\mathcal{A}} = \|\tilde{\boldsymbol{w}}(x)\|_{2} = \sqrt{\frac{1}{D} \sum_{i=1}^{D} \sum_{j=i+1}^{D} \left(\log \frac{P(x \mid H_{i})}{P(x \mid H_{j})}\right)^{2}}.$$
 (30)

This is proportional to the square root of the sum of the square of all possible LLR. This informs how much the evidence x is changing the belief, i. e. how far the posterior distribution is from the prior distribution, regardless of any direction: this is the Aitchison distance between the posterior distribution and the prior distribution.

4 Calibrated likelihood functions on the simplex

In Section 2.1, we presented the idempotence property of calibrated LLRs and the constraint on their distribution. In Section 3, by introducing and completing elements from Egozcue et al. (2003), and Egozcue and Vera (2018), we saw how the Aitchison geometry of the probability simplex allows us to extend the concept of LLR and weight-of-evidence, in a vector form, to any number of hypotheses. In the current section, we extend the definition of calibration and the idempotence property to the multiple hypotheses extension of the LLR: the Isometric-Log-Ratio Likelihood function (ILRL). We also show how the constraint on the distribution of calibrated LLRs generalizes to the ILRLs. This gives a reference distribution for the likelihood function to be calibrated and to properly represent the statistical evidence in a multiple hypotheses and multiclass context.

Let's consider the prior π , the likelihood function w_{θ} , and the posterior P_{θ} as compositions:

$$\boldsymbol{\pi} = [P(H_1), \dots P(H_D)]^T \in \mathcal{S}^D,$$

$$\boldsymbol{w}_{\theta}(\cdot) = \left[f_{\theta_{\mathcal{X}_1}}(\cdot), \dots f_{\theta_{\mathcal{X}_D}}(\cdot) \right]^T \in \mathbb{R}_+^{*D},$$

$$\boldsymbol{P}_{\theta}(\cdot) = \left[P_{\theta}(H_1 \mid \cdot), \dots P_{\theta}(H_D \mid \cdot) \right]^T \in \mathcal{S}^D,$$
(31)

and their isometric-log-ratio transform:

$$\tilde{\boldsymbol{\pi}} = \operatorname{ilr}(\boldsymbol{\pi}) \in \mathbb{R}^{D-1},$$

$$\boldsymbol{l}_{\theta}(\cdot) = \tilde{\boldsymbol{w}}_{\theta}(\cdot) = \operatorname{ilr}(\boldsymbol{w}_{\theta}(\cdot)) \in \mathbb{R}^{D-1},$$

$$\tilde{\boldsymbol{P}}_{\theta}(\cdot) = \operatorname{ilr}(\boldsymbol{P}_{\theta}(\cdot)) \in \mathbb{R}^{D-1},$$
(32)

where $f_{\theta_{\mathcal{X}_i}}$ is the probability density function of the statistical model for the data under hypothesis H_i .

Let \mathcal{L} be a set of (isometric-log-ratio transformed) likelihood functions (ILRLs). Starting from the definition of calibration like in Equation 7 and applying the isometric-log-ratio transformation we get:

$$\forall \boldsymbol{l}_{\theta} \in \mathcal{L} = \{ \boldsymbol{l}_{\theta}(x) \mid x \in \mathcal{X} \},$$

$$\mathbb{P}(H_{i} \mid \boldsymbol{l}_{\theta}) = q_{i}(x) = P_{\theta}(H_{i} \mid x) \ \forall i \in \{1, \dots D\} \iff [\mathbb{P}(H_{1} \mid \boldsymbol{l}_{\theta}), \dots \mathbb{P}(H_{D} \mid \boldsymbol{l}_{\theta})] \\
= [P_{\theta}(H_{1} \mid x), \dots P_{\theta}(H_{D} \mid x)], \\
\iff \boldsymbol{l}_{\mathcal{L}}(\boldsymbol{l}_{\theta}) + \tilde{\boldsymbol{\pi}} = \boldsymbol{l}_{\theta}(x) + \tilde{\boldsymbol{\pi}} \\
\iff \boldsymbol{l}_{\mathcal{L}}(\boldsymbol{l}_{\theta}) = \boldsymbol{l}_{\theta}.$$
(33)

where $\boldsymbol{l}_{\mathcal{L}}(\boldsymbol{l}_{\theta}) = \operatorname{ilr}\left([f_{\mathcal{L}_{1}}(\boldsymbol{l}_{\theta}), \dots f_{\mathcal{L}_{D}}(\boldsymbol{l}_{\theta})]^{T}\right)$, and where $f_{\mathcal{L}_{i}}$ refers to the probability density function of the distribution, over \mathbb{R}^{D-1} , of \boldsymbol{l}_{θ} under hypothesis H_{i} . This is the same reasoning as in Equations 7-10 but in a multiple hypothesis and multidimensional setting. Here LLRs are replaced by ILRLs, and the logit transformation is the isometric-log-ratio transformation.

Hence the extension of the idempotence property for calibrated likelihood functions that can be read as:

or simply:

"the likelihood function of the likelihood function is the likelihood function"

since the ILRL and the likelihood function are isomorphic through the ILR transformation. Hence the following definition of calibrated likelihood functions:

Definition 3 A set \mathcal{L} of (isometric-log-ratio) likelihood functions is perfectly calibrated if they are idempotent:

$$\forall l_{\theta} \in \mathcal{L}, \quad l_{\mathcal{L}}(l_{\theta}) = l_{\theta}. \tag{34}$$

Exactly like in Theorem 2 with the LLR, we will see how the idempotence property leads to a constraint on the distributions of the likelihood functions.

4.1 The distribution of calibrated ILRLs

Let $A \in \mathcal{M}_{D-1,D-1}(\mathbb{R})$ be a real square matrix and $B \in \mathcal{M}_{D-1,(D-1)^2}(\mathbb{R})$ be a real block matrix. These matrices are fixed and defined by the used Aitchison basis. See Appendix B for more details

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

Theorem 4 If $l \mid H_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, then $\forall i \in \{2, \dots D\}$, $l \mid H_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where

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

and $\boldsymbol{\mu}_1 = \boldsymbol{A}^{-1}\boldsymbol{B}\operatorname{vec}(\boldsymbol{\Sigma})$, where the $(D-1)^2$ -dimensional vector $\operatorname{vec}(\boldsymbol{\Sigma})$ is the vectorization of the covariance matrix $\boldsymbol{\Sigma}$, and $\forall i \in \{1, \dots D-1\}$, $\boldsymbol{a}_i = \sqrt{\frac{i+1}{i}}\boldsymbol{e}_i$ where \boldsymbol{e}_i is the ith vector of the standard canonical basis of \mathbb{R}^{D-1} .

In other words, if under one hypothesis, the likelihood function is normally distributed on the Aitchison simplex¹³, it is also normally distributed for all the other hypotheses, with the same covariance matrix and the means are entirely determined by the covariance matrix. The proof of this result is given in Appendix B. This is a proof by induction where each density is recursively determined thanks to the recursive form of the bifurcation tree.

Since $\mu_1 = A^{-1}B \operatorname{vec}(\Sigma)$, the only parameter of the distributions is Σ which is a $(D-1) \times (D-1)$ symmetric positive definite matrix. Therefore, it corresponds to $\frac{D(D-1)}{2} = \binom{D}{2}$ 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 density. This relation between the mean vector and the covariance matrix, and how it is related to the divergences, extends what was presented in the two hypotheses case in Section 2.2 where $\mu = \frac{\sigma^2}{2}$, and is equal to the Kullback-Leibler divergence between the densities of the LLRs.

The covariance matrix of the ILRL distribution and the divergences. The covariance matrix Σ , i. e. the parameter of the Gaussian ILRL distributions, can be expressed in terms of the Kullback-Leibler divergences (D_{KL}) between each density. In a classification context, these divergences can be seen as the between class separabilities.

Let $\Delta = \{d_{i,j}\}_{1 \leq i,j \leq D} \in \mathcal{M}_{D \times D}(\mathbb{R}_+)$, where $d_{i,j} = D_{KL}\left(f_{\mathcal{L}_i} \| f_{\mathcal{L}_j}\right)$, be the matrix of Kullback-Leibler divergences between each density. Since the densities are multivariate Gaussian with the same covariance matrix, the divergences are symmetric and the Δ is therefore a symmetric matrix. Since $d_{i,j} = 0$ for i = j, the D diagonal elements are 0. Therefore, only $\frac{D(D-1)}{2} = \binom{D}{2}$ degrees of freedom remain for the matrix of divergences which is the same as for the covariance matrix Σ . The divergences can be expressed from the covariance matrix as follow:

$$\operatorname{vech}_{\neg \searrow}(\mathbf{\Delta}) = M \operatorname{vech}(\mathbf{\Sigma}), \tag{35}$$

^{13.} The multivariate normal distribution that appears with the ILR coordinate representation is also known as the additive logistic-normal distribution, the logistic-normal distribution (Aitchison and Shen, 1980), or simply the normal distribution on the simplex (Pawlowsky-Glahn et al., 2015).

where vech is the half-vectorization of a matrix, vech is the half-vectorization without the diagonal elements, and $M \in \mathcal{M}_{\frac{D(D-1)}{2} \times \frac{D(D-1)}{2}}(\mathbb{R})$ is a real square matrix. See Appendix C for more details. The divergences can therefore be computed from the parameter Σ^{14} .

Figure 5 shows a few examples of densities of the ILRL under each of the three hypotheses H_1 , H_2 , and H_3 . The blue rays mark out the maximum probability (likelihood) 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 in \mathbb{R}^2 with the ILR coordinate representation. The parameters of the densities are linked and constrained according to Theorem 4. Their parameters can be expressed in terms of the three divergences as explained above. When the separabilities between each density 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 the separability between all classes goes to 0, the densities collapse at $\mathbf{0}$ and tend to be a Dirac delta function.

In this Section, we extended the idempotence property and the definition of calibrated LLRs to their multiple hypotheses and multiclass counterpart: the isometric-log-ratio tranformed likelihood function (ILRL). Because likelihood functions and ILRLs are isomorph, those results can naturally be pulled back to probability-likelihood simplex.

As for the LLR, the idempotence leads to a constraint on the distribution of calibrated likelihood functions. We showed that if, under one hypothesis (or class), the likelihood function is normally distributed on the simplex, It is also normally distributed for the other hypotheses with the same covariance matrix and the means are entirely defined by this matrix. This give a reference distribution for the likelihood functions to be calibrated and to properly represent the statistical evidence.

These properties are in fact the generalization, to the multiple hypotheses case $(D \ge 2)$, of results that are well-known for the binary case (D = 2), and that can be widely found in the literature from different fields: for the weight-of-evidence since the 40s (Good, 1979, 1985), in the context of signal detectability since the 50s (Peterson et al., 1954; Birdsall, 1966), and in the context of LLRs calibration in forensic identification since the 2000s (van Leeuwen and Brümmer, 2013; Meester and Slooten, 2021).

In the next Section, we provide one application of these results by presenting how the discriminant space of a discriminant analysis can be designed to form the space of calibrated ILR transformed likelihood functions.

5 Application: Compositional discriminant analysis

In Noé et al. (2022), the authors presented how the idempotence property and its constraint on the distributions of the LLR can be used to design a non-linear discriminant analysis where the discriminant component forms a calibrated LLR¹⁵. Being based on results that were known only for the LLR, the approach was naturally limited to the two-classes case.

^{14.} Unfortunately, we did not prove that M is invertible. Assuming this is the case, the covariances can be expressed in terms of the divergences as: $\operatorname{vech}(\Sigma) = M^{-1} \operatorname{vech}_{\neg \sim}(\Delta)$.

^{15.} This has been presented in the context of privacy in speech technology. In the discriminant space, the LLR can be set to zero for hiding the evidence related to a binary attribute, in accordance with the concept of *perfect secrecy* (Shannon, 1949; Nautsch et al., 2020). See Noé (2023) for more details.

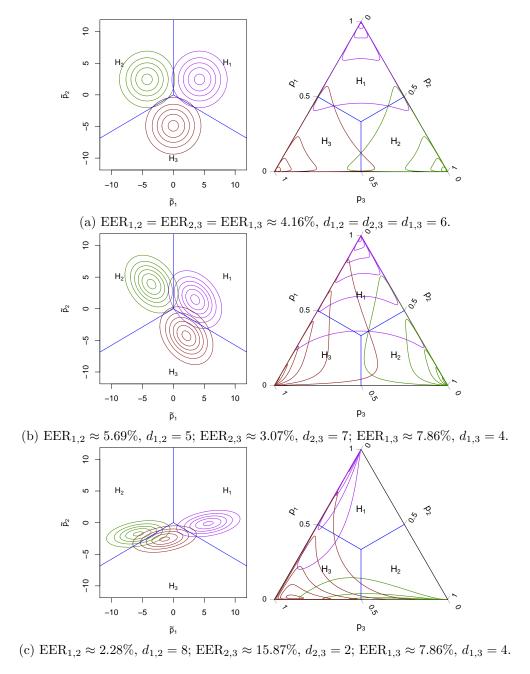


Figure 5: Few contours of Gaussian densities of the likelihood function in a three-hypotheses case. They are parameterized by a shared covariance matrix that can be expressed in terms of the three separabilities between each density. The densities on the ILR space (left) are with respect to the Lebesgue measure while the densities on the simplex (right) are with respect to the Aitchison measure (Mateu-Figueras et al., 2011).

However, with the results presented in Section 4, the discriminant analysis can now be defined for any number of classes. 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 Filzmoser et al. (2011) or in Section 8.4 of Pawlowsky-Glahn et al. (2015). The compositional nature of the CDA is on the treatment of the produced vector of likelihoods. In a nutshell, the idea is to design the discriminant space in accordance with Theorem 4 such that the discriminant components form a calibrated isometric-log-ratio transformed likelihood function (ILRL). This approach is presented in details in this section ¹⁶.

Let's go back to the linear discriminant analysis (LDA). For a two-classes case, as discussed in the example of Section 2.2, a LLR is computed (Equation 13). This is the same as projecting the feature vector onto the discriminant direction that minimizes the within-class variability and maximizes the between-class variability. However, since the observations are not necessarily normally distributed, violating therefore the Gaussian assumption, the computed LLRs may not be well-calibrated. Other discriminant analysis approaches relax the assumptions on the distribution of the data. The quadratic discriminant analysis (QDA) also assumes the features to be normally distributed for each class, but without the shared covariance assumption. The LLR computation becomes a quadratic form of the data, and the resulting discriminant function is non-linear. However, QDA is still based on Gaussian assumptions and the discriminant function is not necessarily invertible contrary to the LDA's mapping (Hastie and Zhu, 2001; Bishop, 2006)¹⁷. Some approaches make no explicit assumption on the distribution of the data. In Dorfer et al. (2016), 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 and statistical modeling nature of the LDA. In this sense, and since our work is set in the realm of generative classifiers, we do not consider this approach as a non-linear version of the LDA. Generalized (Stuhlsatz et al., 2012) and kernel-based (Mika et al., 1999) discriminant analysis are good candidates for generalizing the LDA in a non-linear manner. However, like the QDA, they do not have a trivial inverse mapping from the discriminant space back to the feature space (Weston et al., 2003). Works like Izmailov et al. (2020) and Ardizzone et al. (2020) propose generative classifiers by modeling the class-conditional distributions in the features space using normalizing flow (NF)¹⁸. In the base space, the class-conditional distributions are chosen to be multivariate normal. However, the choice of the normal distributions' parameters in the base space is arbitrary, and contrary to the CDA presented below, the resulting components in the base space can not be interpreted as a calibrated LLR or ILRL.

^{16.} This has been presented as a poster at CoDaWork2024, the 10th International Workshop on Compositional Data Analysis (Noé et al., 2024a).

^{17.} Having an invertible mapping is not required for applications that focus only on classification. However, some applications may require an invertible mapping, for doing data transformation by manipulating the discriminant space like in Noé et al. (2022) or like the interpolation example of Section 5.2.4; or for generation.

^{18.} Normalizing flow (NF) is a family of invertible neural networks that learn a diffeomorphism between the feature space and a *base* space. Some literature on NF uses the term *latent* rather than *base*. We were also doing so in Noé et al. (2022), however, agreeing with the argument proposed in Papamakarios et al. (2021), we have adopted the term *base*.

5.1 Proposed compositional discriminant analysis

Let's consider the set $C = \{C_1, C_2, \dots C_D\}$ of D classes and an observed vector x which belongs to one of these classes. The proposed discriminant analysis, named compositional discriminant analysis (CDA), is a generative classifier that models the distribution of the data under each class by learning an invertible and differentiable mapping between the feature space and a base space in which the class-conditional distributions are known. The class-conditional distributions in the base space are designed according to the idempotence property constraint of Theorem 4. In this way, the mapping transforms the observed vectors into a same-dimensional base space where the first D-1 dimensions form the isometric-logratio transformed likelihood function (ILRL), and the other dimensions form the residual meaning "everything in the data that is independent of the class variable".

Let's introduce some notations. Let:

- $\mathcal{X} \subset \mathbb{R}^d$ be the d-dimensional feature space,
- $l(x) \in \mathcal{L} \subset \mathbb{R}^{D-1}$ be the ILRL of an observation $x \in \mathcal{X}$,
- $r(x) = [r_1(x), r_2(x), \dots r_{d-D+1}(x)]^T \in \mathcal{R} \subset \mathbb{R}^{d-D+1}$ be the residual of x.

We want to find a diffeomorphism that maps the data from the feature space to the base space $\mathcal{Z} = \mathcal{L} \oplus \mathcal{R}$ in which the first D-1 dimensions form a calibrated ILRL, representing therefore the statistical evidence about the classes, while the other dimensions form the residual.

5.1.1 Class-conditional distributions in the base space

In order to properly represent the statistical evidence, we want the first D-1 dimensions of the base space to form a calibrated ILRL. The class-conditional distributions in the base space are therefore chosen according to the idempotence property constraint of Theorem 4:

$$\forall i \in \{1, \dots D\}, \ \mathbf{z} \mid C_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{C}), \tag{36}$$

where:

- Σ is a $(D-1)\times(D-1)$ symmetric positive definite matrix, and is the only parameter of the distributions in the base space,
- the means $m_i \in \mathcal{Z} \subset \mathbb{R}^d$ are the concatenation of $\mu_i \in \mathcal{L}$ and the (d-D+1)-dimensional zero vector. μ_i is defined according to Theorem 4 and its expression is given by:

$$\forall i \in \{1, \dots D\}, \ \boldsymbol{\mu}_i = \boldsymbol{A}^{-1} \boldsymbol{B} \operatorname{vec}(\boldsymbol{\Sigma}) - \boldsymbol{\Sigma} \boldsymbol{a}_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1} \boldsymbol{\Sigma} \boldsymbol{a}_j,$$
 (37)

where $\mathbf{A} \in \mathcal{M}_{D-1,D-1}(\mathbb{R})$ and $\mathbf{B} \in \mathcal{M}_{D-1,(D-1)^2}(\mathbb{R})$ are fixed matrices, $\mathbf{a}_i = \sqrt{\frac{i+1}{i}}\mathbf{e}_i$ where \mathbf{e}_i is the *i*th vector of the standard canonical basis of \mathbb{R}^{D-1} and $\mathbf{a}_0 = \mathbf{0}$ is the (D-1)-dimensional zero vector (see Appendix B for more details),

 \bullet the covariance matrix C is the following block matrix:

$$C = \begin{bmatrix} \Sigma & \mathbf{0}_{D-1,d-D+1} \\ \mathbf{0}_{d-D+1,D-1} & \mathbf{I}_{d-D+1} \end{bmatrix}, \tag{38}$$

where 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 D-1 first dimensions are distributed according to Theorem 4 and the remaining dimensions are normally distributed with a zero mean and an identity covariance regardless of the class variable.

Lemma 5 With $z \mid C_i \sim \mathcal{N}(m_i, C) \ \forall i \in \{1, ... D\}$, as described above, the first D-1 dimensions of z form its isometric-log-ratio transformed likelihood function:

$$[z_1, \dots z_{D-1}] = \operatorname{ilr}\left([f_{\mathcal{Z}_1}(\boldsymbol{z}), \dots f_{\mathcal{Z}_D}(\boldsymbol{z})]\right). \tag{39}$$

See Appendix D for a proof.

5.1.2 DIFFEOMORPHISM BETWEEN THE FEATURE SPACE AND THE BASE SPACE

Let $q^{-1}: \mathcal{X} \mapsto \mathcal{Z}$ be a diffeomorphism that maps the data into the base space such that 19 :

$$\forall i \in \{1, \dots D\}, \ g^{-1}(\boldsymbol{x}) \mid C_i \sim \mathcal{N}(\boldsymbol{m}_i, \boldsymbol{C})$$
(40)

Theorem 6 With $g^{-1}(\mathbf{x}) \mid C_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{C}) \ \forall i \in \{1, \dots D\}$, the first D-1 dimensions of $\mathbf{z} = g^{-1}(\mathbf{x})$ form the isometric-log-ratio transformed likelihood function of \mathbf{x} :

$$[z_1, \dots z_{D-1}] = \operatorname{ilr}\left([f_{\mathcal{X}_1}(\boldsymbol{x}), \dots f_{\mathcal{X}_D}(\boldsymbol{x})]\right). \tag{41}$$

This means that the D-1 first dimension in the base space form the isometric-log-ratio transformed likelihood function of the data. Given Lemma 5, the proof is straightforward since the likelihood function of x and the likelihood function of z are the equivalent. Indeed, they are proportional, where the Jacobian determinant of the mapping is the scaling factor.

With the distributions in the base space defined by Equation 36 and thanks to Theorem 6, the first D-1 dimensions of z represent the statistical evidence in x about the classes in the form of a ILRL, while the other dimensions form the residual normally distributed with a zero mean vector and an identity covariance matrix regardless of the class.

The diffeomorphism g can be learned through Normalizing Flow (NF) (Papamakarios et al., 2021)²⁰. Let $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, c^{(1)}), \dots (\boldsymbol{x}^{(N)}, c^{(N)})\}$ be a training set of observed feature vectors with their corresponding class. The parameters θ_g of g, and Σ , are learned by maximizing the log-likelihood of the data:

$$\log f\left(\mathcal{D}; \theta_g, \mathbf{\Sigma}\right) = \sum_{i=1}^{D} \left(\sum_{(\boldsymbol{x}, c) \in \mathcal{D} | c = C_i} \log \left(f_{\mathcal{Z}_i}(\boldsymbol{z}; \mathbf{\Sigma}) \left| \det \left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}} \right) \right|^{-1} \right) \right), \tag{42}$$

^{19.} $g: \mathcal{Z} \mapsto \mathcal{X}$, where g stands for generator.

^{20.} In our experiments we used the RealNVP architecture (Dinh et al., 2017).

where the densities for \boldsymbol{x} are computed through the densities for \boldsymbol{z} and the change of variable formula. In our experiments, θ_g and $\boldsymbol{\Sigma}$ are learned through negative log-likelihood minimization with automatic differentiation and gradient descent. Regarding the initialization of $\boldsymbol{\Sigma}$, see Appendix E.

Remark: We have not made any explicit assumption on the distribution of the data. However, this does not mean there is no assumption at all. The use of the CDA implicitly assumes the existence of a diffeomorphism that would transform the distribution of the data into the target Gaussians and that the NF is flexible enough to reach this diffeomorphism.

5.1.3 Regarding the interpretability of the compositional discriminant analysis

The proposed discriminant analysis maps the data into a space where the first D-1 dimensions are discriminant and form the ILRL of the observation. With the standard LDA, the D-1 dimensions given by the non-zero-eigenvalue eigenvectors of the matrix $\Sigma_W^{-1}\Sigma_B$, where Σ_W is the shared within-class covariance matrix and Σ_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 CDA, 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 used bifurcation tree (we use here the one in Figure 3). The discriminations between the classes are given by the parameter Σ as discussed in Section 4.1.

Moreover, since the densities of the ILRL in the base space are designed to respect the idempotence constraint of Theorem 4, the approach tends to produce a set of ILRLs that is well-calibrated. The resulting classifier can therefore be used for uncertainty-aware predictions avoiding under or overconfident decisions. In addition, the first D-1 dimensions of the base space benefit from the Euclidean vector space structure of the Aitchison geometry, allowing distance measure, posterior probability distribution computation by simply shifting the likelihood by the prior, and straightforward and meaningful interpolation as we will give an example in Section 5.2.4.

5.2 Experiments

As a proof of concept, we report results of toy experiments. We first consider synthetic two-classes and two-dimensional datasets. Indeed, in this case, the base space is two-dimensional and can be fully visualized. We will then discuss a gaussian three-classes example. In a three-classes case, the discriminant subspace is two-dimensional and can therefore be fully visualized. Finally, we report discriminant and interpolation results on a ten-classes case with the hand-written digits dataset MNIST.

5.2.1 Concerning the log-likelihood-ratio cost and the so-called expected calibration error

Before going any further, we need to introduce the log-likelihood-ratio cost C_{llr} . Even though this metric is popular for the evaluation of forensic identification systems, it is still

little-known in the machine learning community. It has been introduced for the evaluation of systems that produce LLRs in the context of speaker verification (Brümmer and du Preez, 2006). The $C_{\rm llr}$ measures the goodness of a set of LLRs in terms of both discrimination and calibration. Among the several possible interpretations of the C_{llr} , one that will be familiar to the people working on calibration, is as an expected proper scoring rule (PSR) (Gneiting and Raftery, 2007; Bröcker, 2009; Silva Filho et al., 2023). To be more precise, the $C_{\rm llr}$ is the empirical expectation of the cross-entropy (with a $\log \, \mathrm{score}^{21})$ —also known as the binary cross-entropy loss in machine learning—where each probability is computed through the log-likelihood-ratio and a non-informative prior log-odds of 0. Like every expected PSR, it can be decomposed into two terms: a calibration term and a discrimination term. This decomposition is insured by the pool adjacent violator (PAV) algorithm which is known to minimize the expected PSR under a monotonicity constraint (Brümmer and Preez, 2013). The $C_{\rm llr}$ can therefore be written as $C_{\rm llr} = C_{\rm llr}^{\rm cal} + C_{\rm llr}^{\rm min}$, where $C_{\rm llr}^{\rm min}$ is obtained with the LLRs that have been calibrated through the PAV algorithm and informs on the discrimination quality of the LLRs. $C_{\rm llr}^{\rm cal} = C_{\rm llr} - C_{\rm llr}^{\rm min}$ informs on the calibration quality of the LLRs. However, the reader has to keep in mind that calibration only and discrimination only do not inform on the goodness of the LLRs. As an example, a set of LLRs can be perfectly calibrated but non-discriminant at all and therefore non-informative. The goodness of the LLRs should be assessed through the $C_{\rm llr}$ which incorporate both aspects and informs on the quality of the information provided by the LLRs (Brümmer and du Preez, 2006), like expected PSR should be used for evaluating the goodness of posterior probabilities (Ferrer and Ramos, 2025) regardless of their discrimination and calibration quality separately.

In machine learning, a popular metric for evaluating the calibration of probabilistic predictions is the Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017). However, it is based on a suboptimal and unstable binning strategy, whereas the optimal one is given by the PAV algorithm (Brümmer and Preez, 2013; Dimitriadis et al., 2021). One possible reason for the popularity of the ECE is that it can be traced back to the decomposition of the expected Brier score in DeGroot and Fienberg (1983). However, in DeGroot and Fienberg (1983), the forecaster is allowed to choose a probability within a finite set of values: $\{0, 0.1, 0.2, ... 1\}$. In this case, the decomposition of the expected Brier score is natural and corresponds to the ECE's binning. In machine learning, predictions are not limited to this finite set of allowed values. ECE's binning strategy becomes therefore arbitrary and suboptimal. Consequently, we do not rely on ECE and instead report LLRs' quality in terms of the well-established $C_{\rm llr}$ decomposition²². Now, let's come back to our experiments.

5.2.2 Two-classes and two-dimensional examples

We provide a few two-classes and two-dimensional experiments to illustrate the CDA. We compare our approach—in terms of both discrimination and calibration—with the LDA and the QDA on three artificial datasets. The first dataset, called *Gaussians*, consists of two multivariate Gaussians with different means and covariance matrices. The second dataset,

^{21.} For the choice of the log score for assessing LLRs, the reader can refers to (Brümmer and du Preez, 2006).

^{22.} Basically, the $C_{\text{llr}}^{\text{cal}}$ can be seen as an ECE, but with an optimal binning, with the log score instead of the Brier score, and with a non-informative prior.

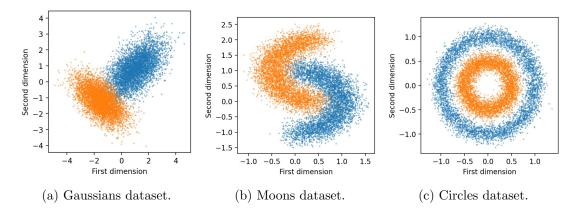


Figure 6: Training sets for the two-classes and two-dimensional examples. The color indicates to which class a sample belongs: blue for C_1 and orange for C_2 .

Table 1: C_{llr} measures of the discriminant analysis on the two-classes examples.

datasets	LD	A	QI)A	CDA	
	$C_{ m llr}^{ m min}$ [bit]	$C_{ m llr}$ [bit]	$C_{ m llr}^{ m min}$	$C_{ m llr}$	$C_{ m llr}^{ m min}$	$C_{ m llr}$
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

called *Moons*, consists of two interleaving noisy half-circles. The third one, called *Circles*, consists of a large noisy circle containing a smaller one²³. 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 $C_{\rm llr}$, $C_{\rm llr}^{\rm min}$, and scatter-plot visualizations.

Figure 6a shows the training set for the Gaussian example. Figure 7 shows the results of the maximum likelihood classification using LDA, QDA, and the proposed CDA. $C_{\rm llr}$ 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. The LDA has therefore a discrimination and a calibration that are not as good as the QDA. The $C_{\rm llr}^{\rm min}$ is 0.125 bit for the LDA while it is 0.115 bit for the QDA and the calibration cost $C_{\rm llr}^{\rm cal} = C_{\rm llr} - C_{\rm llr}^{\rm min}$ is 0.073 bit for the LDA and 0.011 for the QDA. Here, the QDA is also better than the CDA which has a $C_{\rm llr}^{\rm min}$ of 0.117 bit and a $C_{\rm llr}^{\rm cal}$ of 0.038 bit. The goodness of the QDA is here not surprising since the assumed data distribution and the actual distribution match.

Figure 6b shows the training set for the Moons example. Figure 8 shows the results of the maximum likelihood classification using LDA, QDA, and the CDA. $C_{\rm llr}$ measures are reported in Table 1. Both LDA and QDA hardly separate the two classes while the CDA does better in terms of both discrimination and calibration with $C_{\rm llr}=0.105$ bit and $C_{\rm llr}^{\rm cal}=C_{\rm llr}-C_{\rm llr}^{\rm min}=0.013$ bit.

^{23.} These datasets are generated with scikit-learn (Pedregosa et al., 2011).

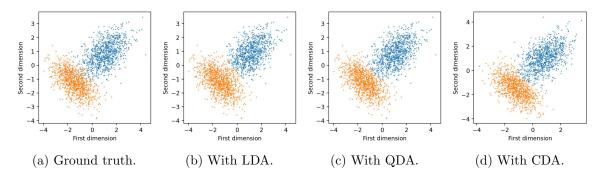


Figure 7: Maximum likelihood classification on the Gaussians dataset. In (a), the colors indicate the true label. For the other figures, the colors indicate the predicted class.

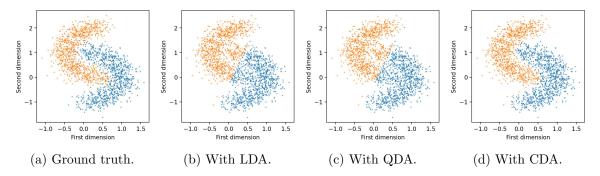


Figure 8: Maximum likelihood classification on the Moons dataset. In (a), the colors indicate the true label. For the other figures, the colors indicate the predicted class.

Figure 6c shows the training set for the Circles example. Figure 9 shows the results of the maximum likelihood classification. 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 $C_{
m llr}^{
m min}$ of 0.023, while the CDA has a slightly higher $C_{
m llr}^{
m min}$ of 0.040. We can indeed see in Figure 9d 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 since they are more points (see Figure 10). This is because 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 calibration of the CDA. The QDA has indeed a $C_{\rm llr}^{\rm cal}$ of 0.468 while the CDA has a $C_{\rm llr}^{\rm cal}$ of 0.037 on the testing set. In this example, the implicit assumption discussed in the remark of Section 5.1.2 is not fulfilled, the invertibility and differentiability constraints of the CDA limit the discrimination ability. However, having a low $C_{\rm llr}$, the CDA still produces better LLRs than the QDA. This is a typical example where a good discrimination does not necessarily implies a reliable extraction of the information: even if the QDA separates well the classes, the modeling of the data is bad, resulting in bad calibration. On the contrary, having a lower discrimination ability does not implies to

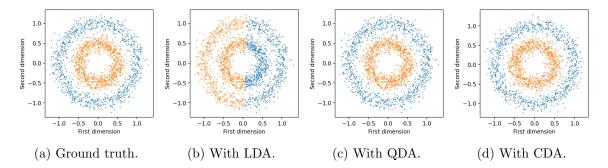


Figure 9: Maximum likelihood classification on the Circles dataset. In (a), the colors indicate the true label. For the other figures, the colors indicate the predicted class.

have a worst modeling quality.

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 11. The bottom part of Figure 11 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 zero. For visualizing the learned diffeomorphism, a set of points is generated homogeneously and regularly to form a grid 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 (top of Figure 11). In the feature space, the transformed regular grid approximates the manifold on which the data lives. The colors represent the true label of the samples. One can see, for the Circles example, the slice of miss-classification of the blue circle at the bottom left of the 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.

In the above examples, we discussed the two-classes case only²⁴. In this case, the discriminant subspace is one-dimensional and is the log-likelihood-ratio line with the properties that have been presented in Section 2. In the following, we present examples with more classes to fully appreciate our results presented in the multiclass setting in Section 4.

5.2.3 A Gaussian three-classes and four-dimensional example

Here we consider three-classes in a four-dimensional feature space. For each class, the data is normally distributed with different mean vectors and different covariance matrices. The discriminant subspace, i.e. the space of the isometric-log-ratio transformed likelihood functions (ILRL), is two-dimensional and the residual subspace too. For conciseness, we report only the visualization of the test data in the discriminant and the residual space. More details for this example can be found in Appendix F.

Figure 12 shows the base space on which the testing set is mapped using the learned transformation. Figure 12a shows the discriminant subspace of the CDA, i.e. the ILRL space. The first dimension discriminates the two first classes (blue and orange) while the

^{24.} For a two-classes example on real speech data, see Noé et al. (2022) and Noé et al. (2023).

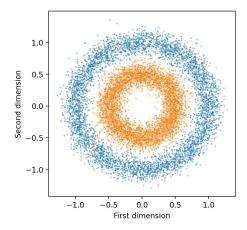


Figure 10: Compositional discriminant analysis on Circles training set (ground truth is given in Figure 6c). Since it is restricted to invertible and differentiable transformations, this discriminant analysis will never "perfectly" separate the two classes as the tiny slice of miss-classification illustrates.

second dimension discriminates the third class (green) from the two others without discriminating the first two. This is in accordance with the used Aitchison basis corresponding to the bifurcation tree of Figure 3. The distributions of the data in the discriminant subspace tend to respect the distributions of calibrated ILRL of Theorem 4 such that the discriminant component can be interpreted as a calibrated likelihood function. The residual components (in Figure 12b) are not discriminant and are normally distributed with zero mean and identity covariance matrix in accordance with the design of the CDA as presented in Section 5.1.1. The learned parameter Σ can be interpreted in terms of the following divergences, or separability between the classes, using Equation 35: $d_{1,2} = 35.5$, $d_{1,3} = 14.7$, and $d_{2,3} = 16.2$.

Figures 12c and 12d show the projection of the testing set using the LDA. The first two components shown in 12c are discriminant but are hardly interpretable by other means than the discriminative power given by the eigenvalues. The other dimensions seem less discriminant, but still contain class-related information, and are not identically distributed contrary to the CDA's residual space (Figure 12b)²⁵.

5.2.4 Hand-written digits recognition and interpolation

The MNIST database consists of grayscale images of size 28×28 . Each image is a handwritten digit between 0 and 9 (LeCun, 1998). The training set is made of 60000 samples and the testing set is made of 10000 samples. Figure 13 shows one randomly selected example for each class.

^{25.} 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 12. More results on this example, including the QDA, can be found in Appendix F.

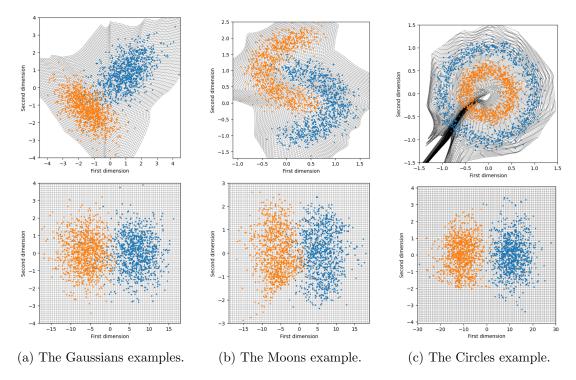


Figure 11: Grid vizualisation of the learned diffeomorphisms on the two-classes and two-dimensional examples. For each example, the samples from the testing sets are shown over a grid. The regular grids in the base space are on the bottom, and the transformed grids in the feature space are on the top.

In the following experiment, each image is flattened and normalized into a 784-dimensional feature vector²⁶. One can see from Figure 13 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, the $C_{\rm llr}$ measures can still be reported for each of the three pairs of classes (Appendix F). In the current example, there are 10 classes corresponding to 45 pairs of classes. Thus, we instead report the empirical expected log scoring rule (or cross-entropy loss) with a uniform prior. It is a multiclass extension of the $C_{\rm llr}$ and is defined as:

$$C_{mc} = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{X}_i|} \sum_{\boldsymbol{x} \in \mathcal{X}_i} s_i(\boldsymbol{x}), \tag{43}$$

where $\mathcal{X}_i = \{x \in \mathcal{X} \mid c = C_i\}$ and $s_i(\boldsymbol{x})$ is the log-likelihood prediction, for input \boldsymbol{x} and class C_i , given by the classifier: $s_i(\boldsymbol{x}) = \left(\operatorname{ilr}^{-1}\left(g^{-1}(\boldsymbol{x})_{1:D-1}\right)\right)_i = \left(\operatorname{ilr}^{-1}\left(\boldsymbol{z}_{1:D-1}\right)\right)_i$. Note

^{26.} State-of-the-art discriminative approaches for classification on MNIST are based on CNN. Even if coupling layers Dinh et al. (2017) can be made of convolutions, we do not consider this way of processing the images.

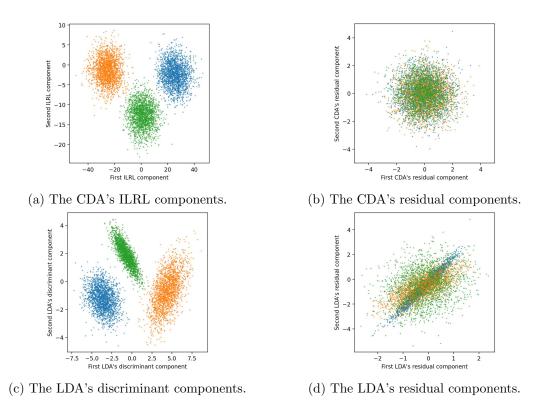


Figure 12: Testing set in the LDA and CDA base spaces for the three-classes non-shared covariance Gaussian example.

0123456789

Figure 13: Samples from the MNIST database.

that in the two classes case, the PAV algorithm calibrates a set of LLRs and minimizes the empirical expected scoring rule, without changing the discrimination quality. This allows the decomposition into a calibration term and a discrimination term. However, when more than two classes are involved, there is no available method to obtain perfectly calibrated probabilities or likelihoods as reference for the decomposition. Thus, C_{mc} can not be decomposed into a calibration term and a discrimination term, but can still be used to summarise the amount of useful information given by the recognizer as done for instance in the context of language recognition (Rodríguez-Fuentes et al., 2013).

In our experiments, we report the empirical expected scoring rule C_{mc} . In addition, since we do not have a minimum expected scoring rule as a discrimination measure, we report the accuracy of the system for a maximum likelihood decision rule²⁷.

^{27.} Be aware that the cross-entropy-based measures and the accuracy differ by nature. The accuracy measures the goodness of hard decisions while the cross-entropy 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 expected scoring rule.

Table 2: Cross-entropy (C_{mc}) 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	2.2310-1	94.43

Table 2 gives the C_{mc} in nat^{28} and the accuracy on the testing set for the LDA, the QDA, and CDA. All the systems result in a cross-entropy loss 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 loss. This confirms that having a good accuracy does not mean that a system models well the data and is good for making rational decisions in general i. e. expected cost minimizing decisions. The CDA results in the lowest cross-entropy which shows that it extracts the most useful information from the images.

Being respectively 9-dimensional and 31-dimensional, the discriminant subspace and the residual subspace can not be visualized in 2-dimensional plots. We therefore report dimension-reduction based visualization using the uniform manifold approximation and projection method (UMAP) (McInnes et al., 2018). Figure 14a shows an UMAP visualization of the nine first dimensions i.e. of the estimated ILRL. One can see the ten clusters. Note that the components given by the UMAP can not be interpreted, the figures are just for illustration and cluster visualization purpore. Figure 14b shows an UMAP visualization of the residual where no clusters appear. This suggests, as expected, that the digit-related information is concentrated in the ILRL components. Table 3 shows the estimated Kullback-Leibler divergences between the digit's class-conditional distributions in the base space. These divergences are computed from the estimated Σ using Equation 35. This informs us about the separability between the classes.

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 directly used for generating images. The interpretability of the base space allows intuitive manipulation or generation of images.

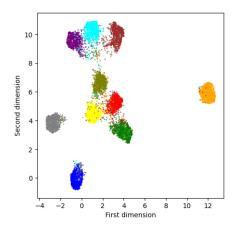
Interpolation²⁹: The Euclidean vector space structure of the base space allows easy interpolation between digits. This can be done with linear interpolation between the digits' centroid. The interpolation between the digit i and the digit j in the base space is given by:

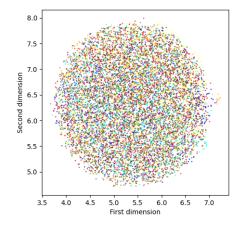
$$\boldsymbol{z}_{i,j}(\alpha) = \alpha \boldsymbol{m}_i + (1 - \alpha) \boldsymbol{m}_j, \tag{44}$$

where $\alpha \in [0, 1]$ and m_i and m_j are the learned digits' centroid as defined in Section 5.1.1. The image can then be constructed by mapping $z_{i,j}(\alpha)$ into the feature space using the

^{28.} 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.

^{29.} We trained the CDA without pre-dimensionality reduction (i. e. on the 784-dimensional features vectors) resulting in a C_{mc} of 3.8110^{-1} nat.





- (a) UMAP visualization of the estimated ILRLs.
- (b) UMAP visualization of the residual.

Figure 14: 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.

Table 3: Estimated Kullback-Leibler divergences between the digit's conditional distributions 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

learned diffeomorphism g: $\mathbf{x}_{i,j}(\alpha) = g(\mathbf{z}_{i,j}(\alpha))$. The feature vector is then unflattened to produce the image. Figure 15 shows two examples³⁰ of digit interpolation for $\alpha = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. Because the discriminant space is a linear space of calibrated likelihood functions, the interpolated images have a neat interpretation: as an example, for the interpolation from 0 to 7, with $\alpha = 0.5$, the interpolated image can be interpreted as equally likely to come from a 0 or a 7; alternativelely, with $\alpha = 0.2 = 1/5$, it is 5 times more likely to come from a 0 than a 7.

In this section, we saw how the discriminant space of a discriminant analysis can be constrained according to Theorem 4 to form the space of calibrated likelihood functions over the set of classes. We illustrated the relevance of this approach with simple experiments.

^{30.} More examples can be found in Noé (2023).



Figure 15: Examples of digit interpolation in the linear space of likelihood functions. From 0 to 7 (top), from 6 to 8 (bottom).

6 Conclusion and perspectives

This paper introduced the concept of calibrated likelihood functions. While this has been known for the binary case, i.e. when the likelihood function can be written in the form of a log-likelihood-ratio, we extended the definition of calibration to likelihood functions for any number of countable hypotheses. We also showed in Theorem 4, that if, under one hypothesis, calibrated likelihood functions are normally distributed on the simplex, they are also normally distributed for the other hypotheses with some additional constraints on their parameter. This extends a result that has been known for decades in the binary case, i.e. for the weight-of-evidence and for calibrated log-likelihood-ratios. In order to do so, we have used the Aitchison geometry of the simplex, which has its origins in the field of compositional data analysis. It provides, to the probability simplex, an Euclidean vector space structure and recovers the additive form of the Bayes' rule. This allowed us to extend the concept of LLR, in a vector form, to any number of hypotheses; and therefore to extend the definition of the calibration, and the constraint on the distribution of calibrated likelihood functions.

The core of the paper is mainly conceptual and theoretical. However, an application of these results to machine learning has been presented. We introduced the Compositional Discriminant Analysis as a non-linear discriminant analysis where the discriminant subspace is designed to form a calibrated likelihood function over the classes. The distribution of the data in the discriminant space is constrained according to Theorem 4 and the discriminant mapping is learned through normalizing flow. This results in an easy-to-interpret and reliable discriminant analysis.

However, our contributions are more general and not limited to discriminant analysis. Theorem 4 gives a reference distribution for the likelihood functions to be calibrated. We therefore expect, in the future, several applications of our results to the calibration and the evaluation of probabilistic predictions in a multiple hypotheses and multiclass setting. Indeed, the results are not restricted to the space of likelihood functions since, given a prior, the posterior probability distribution and the likelihood function are isomorphic under a scale-invariant equivalence relation. Our results can therefore be pullbacked to the set of probabilistic prediction (by simply translating/perturbating the set of likelihood functions by the prior), which is more familiar to the calibration and statistical machine learning community.

Independently of the calibration, we expect that the use of the Aitchison geometry of the simplex will find many applications in machine learning, especially in multiclass settings, and as an alternative to the suboptimal one-vs-rest or one-vs-one approaches. As

an example, the Aitchison geometry of the simplex has been used to extend the concept of Shapley value for explaining a probabilistic prediction in a multiclass context (Noé et al., 2024b).

Appendix A. The distribution of calibrated LLRs

In this section we provide a detailed proof for Theorem 2. The proof is similar as the one in van Leeuwen and Brümmer (2013). Alternative proofs can be found for the weight-of-evidence in Good (1985), Peterson et al. (1954), and Meester and Slooten (2021).

Proof.

Starting from the idempotence property, we have:

$$l = \log \frac{f_{\mathcal{L}_1}(l)}{f_{\mathcal{L}_2}(l)} \iff f_{\mathcal{L}_2}(l) = e^{-l} f_{\mathcal{L}_1}(l).$$

$$(45)$$

Let the density for the LLR under the first hypothesis be a Gaussian:

$$l \mid H_1 \sim \mathcal{N}(\mu, \sigma^2), \text{ where } \mu \ge 0$$

$$f_{\mathcal{L}_1}(l) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right). \tag{46}$$

Thanks to Expression 45, we have:

$$f_{\mathcal{L}_{2}}(l) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l-\mu)^{2}}{2\sigma^{2}}\right) \exp\left(-l\right),$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(l-\left(\mu-\sigma^{2}\right)\right)^{2}}{2\sigma^{2}}\right) \exp\left(\frac{\sigma^{2}}{2}-\mu\right).$$
(47)

Since $f_{\mathcal{L}_2}(\cdot)$ is a probability density function, its integral is one:

$$\int_{-\infty}^{+\infty} f_{\mathcal{L}_2}(l) \, dl = 1 \iff \exp\left(\frac{\sigma^2}{2} - \mu\right) = 1,$$

$$\iff \sigma^2 = 2\mu.$$
(48)

Therefore,

$$f_{\mathcal{L}_2}(l) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(l - (-\mu))^2}{2\sigma^2}\right),$$

$$l \mid H_2 \sim \mathcal{N}(-\mu, \sigma^2),$$
(49)

and
$$\sigma^2 = 2\mu$$
.

Appendix B. The distribution of calibrated ILRLs

In this section, we provide a proof for Theorem 4. Note that the use of the specific basis obtained with the Gram-Schmidt procedure in no way excludes the general aspect of the

following results since Aitchison orthonormal bases are related through unitary transformations Egozcue et al. (2003).

Let's first recall Theorem 4. Let $A \in \mathcal{M}_{D-1,D-1}(\mathbb{R})$ be the following real square matrix:

$$\mathbf{A} = \{\alpha_{ij}\}_{1 \le i, j \le D-1}$$

$$\alpha_{ij} = \begin{cases} 2\sqrt{\frac{i+1}{i}}, & \text{if } i = j\\ \frac{2}{\sqrt{j(j+1)}}, & \text{if } j < i\\ 0, & \text{otherwise} \end{cases}$$

$$(50)$$

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

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}^{(1)} & \boldsymbol{B}^{(2)} & \dots & \boldsymbol{B}^{(D-1)} \end{bmatrix} \tag{51}$$

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

$$B^{(b)} = \{\beta_{ij}^{(b)}\}_{1 \le i, j \le D-1}$$

$$\beta_{ij}^{(b)} = \begin{cases} \frac{b+1}{b}, & \text{if } i = j = b\\ 2\sqrt{\frac{i+1}{ib(b+1)}}, & \text{if } (i = j) \land (b < i)\\ \frac{1}{jb\sqrt{(j+1)(b+1)}}, & \text{if } (b < i) \land (j < i)\\ 0, & \text{otherwise.} \end{cases}$$
(52)

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

If $l \mid H_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, then $\forall i \in \{2, \dots D\}$, $l \mid H_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where

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

and $\boldsymbol{\mu}_1 = \boldsymbol{A}^{-1}\boldsymbol{B}\operatorname{vec}(\boldsymbol{\Sigma})$, where the $(D-1)^2$ -dimensional vector $\operatorname{vec}(\boldsymbol{\Sigma})$ is the vectorization of the covariance matrix $\boldsymbol{\Sigma}$ and $\forall i \in \{1, \dots D-1\}$, $\boldsymbol{a}_i = \sqrt{\frac{i+1}{i}}\boldsymbol{e}_i$ where \boldsymbol{e}_i is the ith vector of the standard canonical basis of \mathbb{R}^{D-1} .

Proof. Let's recall some notations. Let $\mathbf{w}_{\theta}(x) = \left[f_{\theta_{\mathcal{X}_i}}(x)\right]_{1 \leq i \leq D}$ be the likelihood vector and $\tilde{\mathbf{w}}_{\theta}(x) = \operatorname{ilr}(\mathbf{w}_{\theta}(x)) = \mathbf{l}_{\theta}$ its ILR transformation. With Equation 27, the *i*th ILR component of a likelihood vector can be written as:

$$l_{\theta i} = \tilde{w}_{\theta}(x)_{i} = \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod_{j=1}^{i} f_{\theta_{\mathcal{X}_{j}}}(x)}{(f_{\theta_{\mathcal{X}_{i+1}}}(x))^{i}} \right).$$
 (53)

Using the idempotence property we can replace every $f_{\theta_{\mathcal{X}_i}}(x)$ by $f_{\mathcal{L}_i}(l_{\theta})$:

$$l_{\theta i} = \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod_{j=1}^{i} f_{\mathcal{L}_{j}}(\boldsymbol{l}_{\theta})}{\left(f_{\mathcal{L}_{i+1}}(\boldsymbol{l}_{\theta})\right)^{i}} \right)$$
(54)

After rewriting this expression and setting $\mathbf{a}_i = \sqrt{\frac{i+1}{i}}\mathbf{e}_i$ where \mathbf{e}_i is the *i*th vector of the standard canonical basis for \mathbb{R}^{D-1} i.e. with zero everywhere except with 1 at the *i*th position, we get:

$$f_{\mathcal{L}_{i+1}}(\boldsymbol{l}_{\theta}) = \exp\left(-\boldsymbol{a}_{i}^{T}\boldsymbol{l}\right) \sqrt[i]{\prod_{j=1}^{i} f_{\mathcal{L}_{j}}(\boldsymbol{l}_{\theta})}.$$
 (55)

We thus have a recursive way to get any $f_{\mathcal{L}_i}$ from $f_{\mathcal{L}_1}$. With $\mathbf{l} \sim \mathcal{N}(\boldsymbol{\mu}_1 \mid \boldsymbol{\Sigma})$ and:

$$f_{\mathcal{L}_1}(\boldsymbol{l}) = \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{l} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{l} - \boldsymbol{\mu}_1)\right),$$
(56)

we can use Equation 55 to recursively compute the densities $f_{\mathcal{L}_i}$ for $i \in \{2, \dots D\}$.

The main idea of the proof is to show—by induction and using the recursive relation of Expression 55—that:

 $\forall i \in \{1, \dots D - 1\},$ $f_{\mathcal{L}_{i+1}}(\boldsymbol{l}) = \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{l} - \boldsymbol{\mu}_{i+1})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{l} - \boldsymbol{\mu}_{i+1})\right),$ (57)

$$egin{align} oldsymbol{\mu}_{i+1} &= rac{1}{i} \sum_{j=1}^i oldsymbol{\mu}_j - oldsymbol{\Sigma} oldsymbol{a}_i, \ oldsymbol{\mu}_{i+1}^T oldsymbol{\Sigma}^{-1} oldsymbol{\mu}_{i+1} &= oldsymbol{\mu}_1^T oldsymbol{\Sigma}^{-1} oldsymbol{\mu}_1. \end{align}$$

for all integer $D \geq 2$ we have:

The base case:

From Equation 55 and Equation 56 we get:

$$f_{\mathcal{L}_{2}}(\boldsymbol{l}) = \exp\left(-\boldsymbol{a}_{1}^{T}\boldsymbol{l}\right) f_{\mathcal{L}_{1}}(\boldsymbol{l}),$$

$$= \frac{1}{(2\pi)^{\frac{D-1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{l} - \boldsymbol{\mu}_{1})^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{l} - \boldsymbol{\mu}_{1}) - \boldsymbol{a}_{1}^{T}\boldsymbol{l}\right),$$

$$= \frac{1}{(2\pi)^{\frac{D-1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{l}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{l} + \boldsymbol{l}^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{1} - \boldsymbol{\Sigma}\boldsymbol{a}_{1}) - \frac{1}{2}\boldsymbol{\mu}_{1}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{1}\right),$$

$$= \frac{1}{(2\pi)^{\frac{D-1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{l} - \boldsymbol{\mu}_{2})^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{l} - \boldsymbol{\mu}_{2})\right) \exp\left(\frac{1}{2}\boldsymbol{\mu}_{2}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{2} - \frac{1}{2}\boldsymbol{\mu}_{1}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{1}\right),$$
(58)

where $\mu_2 = \mu_1 - \Sigma a_1$. Since $f_{\mathcal{L}_2}$ is a probability density function, its integral is one:

$$\int_{\boldsymbol{l} \in \mathbb{R}^{D-1}} f_{\mathcal{L}_{2}}(\boldsymbol{l}) d\boldsymbol{l} = 1$$

$$\iff \exp\left(\frac{1}{2}\boldsymbol{\mu}_{2}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{2} - \frac{1}{2}\boldsymbol{\mu}_{1}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{1}\right) = 1,$$

$$\iff \boldsymbol{\mu}_{2}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{2} = \boldsymbol{\mu}_{1}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{1}.$$
(59)

We just showed that $\boldsymbol{l} \mid H_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma} \boldsymbol{a}_1$ and $\boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1$.

The induction step:

Let's assume that for an integer K we have:

$$\forall i \in \{1, \dots K - 1\},\$$

$$f_{\mathcal{L}_{i+1}}(\mathbf{l}) = \frac{1}{(2\pi)^{\frac{D-1}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{l} - \boldsymbol{\mu}_{i+1})^T \mathbf{\Sigma}^{-1} (\mathbf{l} - \boldsymbol{\mu}_{i+1})\right)$$

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

$$\boldsymbol{\mu}_{i+1}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_1^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_1.$$
(60)

Let's show that this still holds for K + 1. Using Equation 55 we can write:

$$f_{\mathcal{L}_{K+1}}(\boldsymbol{l}) = \exp\left(-\boldsymbol{a}_{K}^{T}\boldsymbol{l}\right) \sqrt{\prod_{j=1}^{K} f_{\mathcal{L}_{j}}(\boldsymbol{l})},$$

$$= \exp\left(-\boldsymbol{a}_{K}^{T}\boldsymbol{l}\right) \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \sqrt{\prod_{j=1}^{K} \exp\left(-\frac{1}{2}(\boldsymbol{l} - \boldsymbol{\mu}_{j})^{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{l} - \boldsymbol{\mu}_{j})\right)},$$

$$= \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\boldsymbol{a}_{K}^{T}\boldsymbol{l} - \frac{1}{2K} \sum_{j=1}^{K} (\boldsymbol{l} - \boldsymbol{\mu}_{j})^{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{l} - \boldsymbol{\mu}_{j})\right),$$

$$= \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\boldsymbol{a}_{K}^{T}\boldsymbol{l} - \frac{1}{2K} \sum_{j=1}^{K} (\boldsymbol{l}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{l} + \boldsymbol{\mu}_{j}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{j} - 2\boldsymbol{\mu}_{j}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{l}\right)\right).$$

$$(61)$$

Since for all $j \in \{1, ... K\}$, $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1$, we have:

$$f_{\mathcal{L}_{K+1}}(\boldsymbol{l}) = \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\boldsymbol{a}_{K}^{T} \boldsymbol{l} - \frac{1}{2} \boldsymbol{l}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{l} - \frac{1}{2} \boldsymbol{\mu}_{1}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{1} + \frac{1}{K} \left(\sum_{j=1}^{K} \boldsymbol{\mu}_{j}\right)^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{l}\right),$$

$$= \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{l}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{l} - \frac{1}{2} \boldsymbol{\mu}_{1}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{1} + \boldsymbol{l}^{T} \boldsymbol{\Sigma}^{-1} \left(\frac{1}{K} \left(\sum_{j=1}^{K} \boldsymbol{\mu}_{j}\right) - \boldsymbol{\Sigma} \boldsymbol{a}_{K}\right)\right).$$
(62)

Setting $\mu_{K+1} = \frac{1}{K} \sum_{j=1}^{K} \mu_j - \Sigma a_K$ we get:

$$f_{\mathcal{L}_{K+1}}(\boldsymbol{l}) = \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{l} - \boldsymbol{\mu}_{K+1})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{l} - \boldsymbol{\mu}_{K+1})\right) \times \exp\left(\frac{1}{2} \boldsymbol{\mu}_{K+1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{K+1} - \frac{1}{2} \boldsymbol{\mu}_{1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{1}\right),$$

$$(63)$$

Since $f_{\mathcal{L}_{K+1}}$ is a probability density function, its integral is one, which leads to:

$$\boldsymbol{\mu}_{K+1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{K+1} = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1. \tag{64}$$

Therefore, 60 holds also for K+1. We hence have proved by induction the expressions 57.

A general formula for the means:

We will here proof by induction the following expression for the derivation of the mean vectors:

for all integer $D \geq 2$:

$$\forall i \in \{2, \dots D\},$$

$$\boldsymbol{\mu}_{i} = \boldsymbol{\mu}_{1} - \boldsymbol{\Sigma} \boldsymbol{a}_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1} \boldsymbol{\Sigma} \boldsymbol{a}_{j}, \text{ where } \boldsymbol{a}_{0} = \boldsymbol{0} \text{ the zero vector.}$$
(65)

The base case is straightforward so we provide only the induction step. Let's assume that the expression is true for an integer K:

$$\forall i \in \{2, \dots K\},$$

$$\boldsymbol{\mu}_{i} = \boldsymbol{\mu}_{1} - \boldsymbol{\Sigma} \boldsymbol{a}_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1} \boldsymbol{\Sigma} \boldsymbol{a}_{j}.$$
(66)

Let's show that this holds also for K + 1. From Expression 57 that we proofed above, we know that:

$$\boldsymbol{\mu}_{K+1} = -\boldsymbol{\Sigma}\boldsymbol{a}_K + \frac{1}{K} \sum_{j=1}^K \boldsymbol{\mu}_j, \tag{67}$$

we replace μ_j according to Expression 66:

$$\mu_{K+1} = -\Sigma a_{K} + \frac{1}{K} \sum_{j=1}^{K} \left(\mu_{1} - \Sigma a_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \Sigma a_{k} \right),$$

$$= \mu_{1} - \Sigma a_{K} - \frac{1}{K} \sum_{j=2}^{K} \Sigma a_{j-1} - \frac{1}{K} \sum_{j=3}^{K} \sum_{k=1}^{j-2} \frac{1}{k+1} \Sigma a_{k},$$

$$= \mu_{1} - \Sigma a_{K} - \frac{1}{K} \sum_{j=1}^{K-1} \Sigma a_{j} - \frac{1}{K} \sum_{j=1}^{K-2} \frac{K-1-j}{j+1} \Sigma a_{j},$$

$$= \mu_{1} - \Sigma a_{K} - \frac{1}{K} \Sigma a_{K-1} - \frac{1}{K} \sum_{j=1}^{K-2} \left(\Sigma a_{j} + \frac{K-1-j}{j+1} \Sigma a_{j} \right),$$

$$= \mu_{1} - \Sigma a_{K} - \frac{1}{K} \Sigma a_{K-1} - \sum_{j=1}^{K-2} \frac{1}{j+1} \Sigma a_{j},$$

$$= \mu_{1} - \Sigma a_{K} - \sum_{j=1}^{K-1} \frac{1}{j+1} \Sigma a_{j},$$

$$(69)$$

Hence,

$$\forall i \in \{2, \dots K+1\},$$

$$\mu_i = \mu_1 - \Sigma a_{i-1} - \sum_{j=1}^{i-2} \frac{1}{j+1} \Sigma a_j,$$
(70)

The general expression 65 has therefore been proved by induction.

About matrices A and B

In Proposition 4, the mean vector μ_1 is expressed in terms of the covariance matrix Σ and two constant matrices A and B as follow:

$$\boldsymbol{\mu}_1 = \boldsymbol{A}^{-1} \boldsymbol{B} \operatorname{vec}(\boldsymbol{\Sigma}), \tag{71}$$

where $A \in \mathcal{M}_{D-1,D-1}(\mathbb{R})$ and is defined as:

$$\mathbf{A} = \{\alpha_{ij}\}_{1 \le i, j \le D-1}$$

$$\alpha_{ij} = \begin{cases} 2\sqrt{\frac{i+1}{i}}, & \text{if } i = j\\ \frac{2}{\sqrt{j(j+1)}}, & \text{if } j < i\\ 0, & \text{otherwise} \end{cases}$$

$$(72)$$

and where $\boldsymbol{B} \in \mathcal{M}_{D-1,(D-1)^2}(\mathbb{R})$ is a block matrix:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}^{(1)} & \boldsymbol{B}^{(2)} & \dots & \boldsymbol{B}^{(D-1)} \end{bmatrix} \tag{73}$$

where $\boldsymbol{B}^{(b)} \in \mathcal{M}_{D-1,D-1}(\mathbb{R})$ is the *b*th block and is defined as:

$$\mathbf{B}^{(b)} = \{\beta_{ij}^{(b)}\}_{1 \le i, j \le D-1}
\beta_{ij}^{(b)} = \begin{cases} \frac{b+1}{b}, & \text{if } i = j = b \\ 2\sqrt{\frac{i+1}{ib(b+1)}}, & \text{if } (i = j) \land (b < i) \\ \frac{1}{jb\sqrt{(j+1)(b+1)}}, & \text{if } (b < i) \land (j < i) \\ 0, & \text{otherwise} \end{cases}$$
(74)

In this paragraph, we show how these matrices are derived. The following system of equations, that comes from the expressions 57, can be written in a matrix form:

$$\forall i \in \{1, \dots D - 1\},
\boldsymbol{\mu}_{i+1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1.$$
(75)

Using the general expression 65 for the means, the system of equations becomes:

$$\forall i \in \{1, \dots D-1\},\$$

$$2\boldsymbol{a}_{i}^{T}\boldsymbol{\mu}_{1} + 2\boldsymbol{\mu}_{1}^{T}\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{i-1}\frac{1}{j+1}\boldsymbol{\Sigma}\boldsymbol{a}_{j} = \boldsymbol{a}_{i}^{T}\boldsymbol{\Sigma}\boldsymbol{a}_{i} + 2\boldsymbol{a}_{i}^{T}\sum_{j=1}^{i-1}\frac{1}{j+1}\boldsymbol{\Sigma}\boldsymbol{a}_{j} + \sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\frac{1}{(j+1)(k+1)}\boldsymbol{a}_{j}^{T}\boldsymbol{\Sigma}\boldsymbol{a}_{k},$$
(76)

Since $x^T \Sigma y = \text{vec}(\Sigma)^T \text{vec}(xy^T)$ and by setting $\theta_{\Sigma} = \text{vec}(\Sigma)$ we get:

$$\forall i \in \{1, \dots D-1\},\,$$

$$\left(2a_{i} + 2\sum_{j=1}^{i-1} \frac{1}{j+1}a_{j}\right)^{T} \mu_{1}$$

$$= \left(\operatorname{vec}(a_{i}a_{i}^{T}) + 2\sum_{j=1}^{i-1} \frac{1}{j+1}\operatorname{vec}(a_{i}a_{j}^{T}) + \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} \frac{1}{(j+1)(k+1)}\operatorname{vec}(a_{j}a_{k}^{T})\right)^{T} \theta_{\Sigma}.$$
(77)

 $a_i a_j^T$ is a $(D-1) \times (D-1)$ matrix with zero everywhere except the element at the *i*th row and *j*th column which is $\sqrt{\frac{(i+1)(j+1)}{ij}}$. Its vectorization is therefore the $(D-1)^2$ -dimensional vector with zero everywhere except the ((j-1)(D-1)+i)th element which is $\sqrt{\frac{(i+1)(j+1)}{ij}}$. Let's now rewrite this system in a matrix form:

$$A\mu_1 = B\theta_{\Sigma},\tag{78}$$

where $\mathbf{A} \in M_{D-1,D-1}(\mathbb{R})$, $\mathbf{B} \in M_{D-1,(D-1)^2}(\mathbb{R})$. In 77, the vector on the left side of $\boldsymbol{\mu}$ is the *i*th row of the matrix \mathbf{A} and the vector on the left side of $\boldsymbol{\theta_{\Sigma}}$ is the *i*th row of \mathbf{B} . This is straightforward that \mathbf{A} is triangular with diagonal elements $2\sqrt{\frac{i+1}{i}}$ for $i \in \{1, \ldots D-1\}$. Consequently, its determinant is non zero, and therefore \mathbf{A} is invertible. The mean vector $\boldsymbol{\mu}_1$ can therefore be written in terms of the variances and covariances as follow:

$$\mu_1 = \mathbf{A}^{-1} \mathbf{B} \boldsymbol{\theta}_{\Sigma} = \mathbf{A}^{-1} \mathbf{B} \operatorname{vec}(\boldsymbol{\Sigma}). \tag{79}$$

Appendix C. The covariance matrix of the ILRL distribution and the divergences

In Section 4.1, we have seen that the covariance matrix Σ —which is the only parameter of the densities of normally distributed ILRLs—can be expressed in terms of the Kullback-Leibler divergences between each density. In this section, we provide details of the computation. The divergence between the density for the hypothesis i and the density for the hypothesis j can be written as:

$$d_{i,j} = \frac{1}{2} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$$= \frac{1}{2} (\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - 2\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i),$$
(80)

and since $\forall i \in \{1, \dots D-1\}$, $\boldsymbol{\mu}_{i+1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1$ (see the Appendix B),

$$d_{i,j} = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j.$$
 (81)

Replacing μ_i and μ_j with the expression given in Theorem 4 we get:

$$d_{i,j} = \boldsymbol{\zeta}_{i,j}^T \boldsymbol{\mu}_1 - \boldsymbol{\eta}_{i,j}^T \operatorname{vec}(\boldsymbol{\Sigma})$$
(82)

where

$$\zeta_{i,j} = \left(\mathbf{a}_{i-1} + \mathbf{a}_{j-1} + \sum_{k=1}^{i-2} \frac{1}{k+1} \mathbf{a}_k + \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{a}_k \right),$$

$$\eta_{i,j} = \left(\operatorname{vec}(\mathbf{a}_{i-1} \mathbf{a}_{j-1}^T) + \sum_{k=1}^{j-2} \frac{1}{k+1} \operatorname{vec}(\mathbf{a}_{i-1} \mathbf{a}_k^T) + \sum_{k=1}^{i-2} \frac{1}{k+1} \operatorname{vec}(\mathbf{a}_{j-1} \mathbf{a}_k^T) + \sum_{k=1}^{i-2} \frac{1}{k+1} \operatorname{vec}(\mathbf{a}_{j-1} \mathbf{a}_k^T) \right)$$

$$+ \sum_{k=1}^{i-2} \sum_{k=1}^{j-2} \frac{1}{(k+1)(l+1)} \operatorname{vec}(\mathbf{a}_k \mathbf{a}_l^T) \right). \tag{83}$$

When $2 \le i, j \le D$ 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 $\boldsymbol{\mu}_1 = \mathbf{A}^{-1}\mathbf{B}\operatorname{vec}(\boldsymbol{\Sigma})$, the divergences can be written as follow:

$$\forall i \in \{1, \dots D-1\}, \ \forall j \in \{i+1, \dots D\}, \quad d_{i,j} = \left(\boldsymbol{\zeta}_{i,j}^T \boldsymbol{A}^{-1} \boldsymbol{B} - \boldsymbol{\eta}_{i,j}^T\right) \operatorname{vec}(\boldsymbol{\Sigma}).$$
 (84)

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

$$\operatorname{vech}_{\neg \searrow}(\boldsymbol{\Delta}) = \underbrace{\begin{bmatrix} \boldsymbol{\zeta}_{1,2}^{T} \boldsymbol{A}^{-1} \boldsymbol{B} - \boldsymbol{\eta}_{1,2}^{T} \\ \boldsymbol{\zeta}_{1,3}^{T} \boldsymbol{A}^{-1} \boldsymbol{B} - \boldsymbol{\eta}_{1,3}^{T} \\ \vdots \\ \boldsymbol{\zeta}_{N-1,N}^{T} \boldsymbol{A}^{-1} \boldsymbol{B} - \boldsymbol{\eta}_{D-1,D}^{T} \end{bmatrix} \boldsymbol{D}_{D-1} \operatorname{vech}(\boldsymbol{\Sigma}),$$

$$\operatorname{vech}_{\neg \searrow}(\boldsymbol{\Delta}) = \boldsymbol{M} \operatorname{vech}(\boldsymbol{\Sigma}),$$

$$(85)$$

where D_{N-1} is the duplication matrix (Magnus and Neudecker, 1999) such that $\text{vec}(\Sigma) = D_{N-1} \text{ vech}(\Sigma)$ and $M \in \mathcal{M}_{\frac{D(D-1)}{2} \times \frac{D(D-1)}{2}}(\mathbb{R})$ is a real square matrix.

Appendix D. Proof that the base space's first dimensions form the ILRL

This section gives a proof for Lemma 5. It shows that with the class-conditional distributions as defined in Equation 36, the first D-1 dimensions of $z \in \mathbb{Z}$ form the ILRL.

Proof. The *i*th component of the ILRL vector of
$$z$$
 is given by:

$$\forall i \in \{1, \dots D-1\}, \quad l_i(\boldsymbol{z}) = \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod\limits_{j=1}^{i} f_{\mathcal{Z}_j}(\boldsymbol{z})}{f_{\mathcal{Z}_{i+1}}(\boldsymbol{z})^i} \right)$$

$$= \frac{1}{\sqrt{i(i+1)}} \log \left(\frac{\prod\limits_{j=1}^{i} \exp\left(-\frac{1}{2} (\boldsymbol{z} - \boldsymbol{m}_j)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - \boldsymbol{m}_j)\right)}{\exp\left(-\frac{i}{2} (\boldsymbol{z} - \boldsymbol{m}_{i+1})^T \boldsymbol{C}^{-1} (\boldsymbol{z} - \boldsymbol{m}_{i+1})\right)} \right), \tag{86}$$

where m_i and C are respectively the mean vector and the covariance matrix as defined in the following of Equation 36,

$$l_{i}(z) = \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(-\frac{1}{2} (z - m_{j})^{T} C^{-1} (z - m_{j}) \right) + \frac{i}{2} (z - m_{i+1})^{T} C^{-1} (z - m_{i+1}) \right),$$

$$= \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(m_{j}^{T} C^{-1} z - \frac{1}{2} m_{j}^{T} C^{-1} m_{j} \right) + \frac{i}{2} m_{i+1}^{T} C^{-1} m_{i+1} - i m_{i+1}^{T} C^{-1} z \right),$$

$$= \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(\mu_{j}^{T} \Sigma^{-1} z_{1:D-1} - \frac{1}{2} \mu_{j}^{T} \Sigma^{-1} \mu_{j} \right) + \frac{i}{2} \mu_{i+1}^{T} \Sigma^{-1} \mu_{i+1} - i \mu_{i+1}^{T} \Sigma^{-1} z_{1:D-1} \right),$$

$$(87)$$

where $\mathbf{z}_{1:D-1} = [z_1, z_2, \dots z_{D-1}]^T$ is the vector of the first D-1 components of \mathbf{z} . Since $\forall i \in \{1, \dots D\}, \, \boldsymbol{\mu}_i^T \boldsymbol{\Sigma} \boldsymbol{\mu}_i = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma} \boldsymbol{\mu}_1 \text{ (see Appendix B), we have:}$

$$l_i(\boldsymbol{z}) = \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}_{1:D-1} \right) - i \boldsymbol{\mu}_{i+1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}_{1:D-1} \right),$$
(88)

using Equation 37, we get:

$$l_{i}(z) = \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(\mathbf{A}^{-1} \mathbf{B} \operatorname{vec}(\mathbf{\Sigma}) - \mathbf{\Sigma} \mathbf{a}_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{\Sigma} \mathbf{a}_{k} \right)^{T} \mathbf{\Sigma}^{-1} \mathbf{z}_{1:D-1} - i \left(\mathbf{A}^{-1} \mathbf{B} \operatorname{vec}(\mathbf{\Sigma}) - \mathbf{\Sigma} \mathbf{a}_{i} - \sum_{k=1}^{i-1} \frac{1}{k+1} \mathbf{\Sigma} \mathbf{a}_{k} \right)^{T} \mathbf{\Sigma}^{-1} \mathbf{z}_{1:D-1} \right),$$

$$= \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \left(-\mathbf{a}_{j-1}^{T} \mathbf{z}_{1:D-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{a}_{k}^{T} \mathbf{z}_{1:D-1} \right) + i \mathbf{a}_{i}^{T} \mathbf{z}_{1:D-1} \right) + i \mathbf{a}_{i}^{T} \mathbf{z}_{1:D-1} \right).$$

$$(89)$$

In the next paragraph, we will see that:

$$\sum_{j=1}^{i} \left(-\boldsymbol{a}_{j-1}^{T} \boldsymbol{z}_{1:D-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \boldsymbol{a}_{k}^{T} \boldsymbol{z}_{1:D-1} \right) + i \sum_{k=1}^{i-1} \left(\frac{1}{k+1} \boldsymbol{a}_{k}^{T} \boldsymbol{z}_{1:D-1} \right) = 0.$$
 (90)

We therefore have:

$$l_{i}(z) = \frac{i}{\sqrt{i(i+1)}} \boldsymbol{a}_{i}^{T} \boldsymbol{z}_{1:D-1} = \frac{i}{\sqrt{i(i+1)}} \sqrt{\frac{i+1}{i}} \boldsymbol{e}_{i}^{T} \boldsymbol{z}_{1:D-1} = \boldsymbol{e}_{i}^{T} \boldsymbol{z}_{1:D-1},$$

$$= z_{i},$$
(91)

the *i*th component of the ILRL is therefore the *i*th component of z for all $i \in \{1, ..., D-1\}$.

Proof of Expression 90:

In the following, we will show by induction that Expression 90 is true for all $i \in \{1, ..., D-1\}$ which is equivalent to show that:

$$\forall i \in \{1, \dots D - 1\}, \qquad \sum_{j=1}^{i} \left(-\mathbf{a}_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{a}_k \right) + i \sum_{k=1}^{i-1} \left(\frac{1}{k+1} \mathbf{a}_k \right) = \mathbf{0}.$$
 (92)

The base case of the proof by induction is straightforward, we therefore focus only on the induction step. We assume that the expression is true for a i = n where $n \in \mathbb{N}$:

$$\sum_{j=1}^{n} \left(-\boldsymbol{a}_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \boldsymbol{a}_k \right) + n \sum_{k=1}^{n-1} \left(\frac{1}{k+1} \boldsymbol{a}_k \right) = \mathbf{0}.$$
 (93)

Let's show this is still true for i = n + 1:

$$\sum_{j=1}^{n+1} \left(-\mathbf{a}_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{a}_k \right) + (n+1) \sum_{k=1}^{n} \left(\frac{1}{k+1} \mathbf{a}_k \right)$$

$$= \sum_{j=1}^{n} \left(-\mathbf{a}_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{a}_k \right) + n \sum_{k=1}^{n-1} \left(\frac{1}{k+1} \mathbf{a}_k \right)$$

$$- \mathbf{a}_n - \sum_{k=1}^{n-1} \left(\frac{1}{k+1} \mathbf{a}_k \right) + \sum_{k=1}^{n-1} \left(\frac{1}{k+1} \mathbf{a}_k \right) + \frac{n+1}{n+1} \mathbf{a}_n$$

$$= \sum_{j=1}^{n} \left(-\mathbf{a}_{j-1} - \sum_{k=1}^{j-2} \frac{1}{k+1} \mathbf{a}_k \right) + n \sum_{k=1}^{n-1} \left(\frac{1}{k+1} \mathbf{a}_k \right)$$

$$= \mathbf{0} \text{ according to Equation 93.}$$
(94)

Appendix E. Regarding the initialisation and estimation of the covariance matrix

In our experiments, the training of the CDA turned out to be very sensitive to the initialization of Σ . Here, we present an initialization strategy for starting the optimization with a Σ that we expect to be not too eccentric.

Section 4.1, we saw how the covariance matrix Σ can be expressed by the Kullback-Leibler divergences (D_{KL}) within each pair of classes³¹. We propose here to initialize the mapping g as the identity function and to initialize Σ with the D_{KL} measured in the feature space assuming that each class-conditional distributions 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 Σ and g, these initializations appeared

^{31.} Keep in mind that since the densities in the base space are Gaussian with the same covariance matrix, the Kullback-Leibler divergences are symmetric.

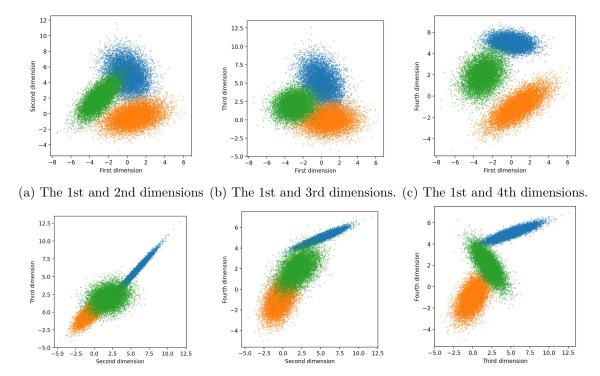
to be effective in our experiments. The intuition is that we initialize Σ with a kind of approximated divergence matrix not too eccentric and not too far from the "true" one³².

Note that this does not mean that an additional assumption on how the feature vectors are distributed is made. This is only for the initialization of the optimization. The parameters of g and Σ are then free to take any value under the constraints of differentiability and invertibility for g and symmetric positive definiteness for Σ .

The symmetric positive definiteness of Σ is insured by optimizing instead the lower triangular matrix \boldsymbol{L} from the Cholesky decomposition $\Sigma = \boldsymbol{L}\boldsymbol{L}^T$, and since the diagonal elements of \boldsymbol{L} must be positive, the log-Cholesky parametrization is used (Pinheiro and Bates, 1996). In our experiments, the estimation of \boldsymbol{L} and the parameters of g is done with automatic differentiation and gradient descent.

Appendix F. A Gaussian three-classes and four-dimensional example

In this Appendix, we provide complete the results of the Gaussian three-classes and four-dimensional example of Section 5.2.3. In this example, each class is generated by a multi-variate normal distribution with its own mean and covariance matrix.



(d) The 2nd and 3rd dimensions. (e) The 2nd and 4th dimensions (f) The 3rd and 4th dimensions.

Figure 16: 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 .

^{32.} Note that the initialization is deterministic.

Table 4: C_{llr} measures for the non-shared covariance example. Samples from the non-concerned class are discarded.

compared classes	LDA		QDA		CDA	
	$C_{ m llr}$ [bit]	$C_{ m llr}^{ m min}$ [bit]	$C_{ m llr}$	$C_{ m llr}^{ m min}$	$C_{ m llr}$	$C_{ m llr}^{ m min}$
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	2.0010 ⁻¹	$1.76_{10^{-2}}$	$6.37_{10^{-4}}$	0.0	$8.46_{10^{-3}}$	$5.31_{10^{-3}}$

We already discussed Figure 12 in Section 5.2.3. QDA's results were not given because QDA does not have an information-preserving mapping of the data into a same-dimensional space. However, it can still be used to compute LLRs. The LLRs of class i against class j is given by:

$$\log \frac{f(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{f(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \frac{1}{2} \boldsymbol{x}^T \left(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1} \right) \boldsymbol{x} + \boldsymbol{x}^T \left(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j \right) + \frac{1}{2} \left(\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \right) + \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|}$$

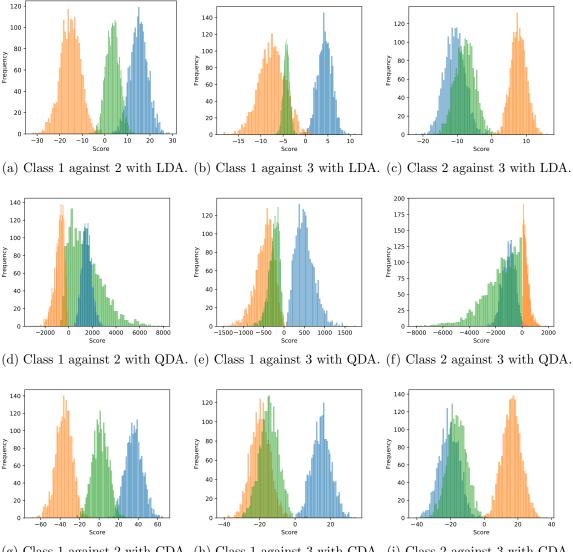
$$(95)$$

Figure 17 shows the histograms of the LLRs obtained with LDA (Figures 17a,17b and 17c), QDA (Figures 17d,17e and 17f), and CDA (Figures 17g, 17h and 17i). 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 between the two classes³³. For the LDA and the CDA, the class-conditional distributions of the LLRs look Gaussian as expected. However, for the LDA, they are not symmetric as required by the idempotence property. We therefore expect the LDA's LLRs to have a lower calibration quality. For the QDA the histograms are not symmetric but this does not suggest that the scores are not calibrated. Indeed, the idempotence constraint of Theorem 2 and Theorem 4 are for normally distributed LLRs, while they are here not Gaussian for the QDA³⁴.

To better assess the discrimination and calibration quality of the LLRs, Table 4 provides $C_{\rm llr}$ measures. The LDA has the worst discrimination and calibration which is not surprising since it is based on the shared covariance assumption. QDA models the best the data and the resulting LLRs have the best discrimination and calibration which is again not surprising since the data is actually distributed as described by the model. However, as mentioned above, the QDA does not provide an information-preserving transformation necessary for data generation or conversion. On the contrary, the CDA does, and still has good discrimination and calibration performance.

^{33.} 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 27: its first component is $\frac{1}{\sqrt{2}}$ times the log-ratio.

^{34.} To be more precise, since the data is here normally distributed and the mapping is quadratic, the LLRs are distributed according to a generalised chi-squared distribution.



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

Figure 17: Histograms of the LLRs 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.

References

John Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), 44(2):139–177, 1982.

John Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. Biometrika, 67(2):261–272, 1980.

- Colin Aitken and Franco Taroni. Statistics and the evaluation of evidence for forensic scientists. John Wiley & Sons, 2004.
- Colin Aitken, Franco Taroni, and Silvia Bozza. The role of the bayes factor in the evaluation of evidence. *Annual Review of Statistics and Its Application*, 11(1):203–226, 2024.
- Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. Training normalizing flows with the information bottleneck for competitive generative classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7828–7840. Curran Associates, Inc., 2020.
- James O Berger and Robert L Wolpert. The likelihood principle. IMS, 1988.
- Theodore Gerald Birdsall. The theory of signal detectability: ROC curves and their character. University of Michigan, 1966.
- Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg, 2006.
- Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 135(643):1512–1519, 2009.
- Niko Brümmer. Measuring, refining and calibrating speaker and language information extracted from speech. PhD thesis, Stellenbosch, University of Stellenbosch, 2010.
- Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. Computer Speech & Language, 20(2):230–275, 2006.
- Niko Brümmer and J.A. Preez. The PAV algorithm optimizes binary proper scoring rules. 2013.
- A. Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- Bruno de Finetti. Theory of Probability: A Critical Introductory Treatment. John Wiley & Sons, 1975.
- Morris H. DeGroot. Optimal Statistical Decisions. John Wiley & Sons, Inc., 1970.
- Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. Journal of the Royal Statistical Society. Series D (The Statistician), 32(1/2):12–22, 1983.
- Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8): e2016191118, 2021. doi: 10.1073/pnas.2016191118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016191118.

- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In Proc. ICLR International Conference on Learning Representations, 2017.
- Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. In *Proc. ICLR International Conference on Learning Representations*, 2016.
- Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3):279–300, 2003.
- Juan José Egozcue and Vera Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005.
- Juan José Egozcue and Pawlowsky-Glahn Vera. Evidence functions: a compositional approach to information. SORT-Statistics and Operations Research Transactions, 1(2): 101–124, Dec. 2018.
- Luciana Ferrer and Daniel Ramos. Evaluating posterior probabilities: Decision theory, proper scoring rules, and calibration. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=qbrE0LR7fF.
- Peter Filzmoser, Karel Hron, and Matthias Templ. Discriminant analysis for compositional data and robust parameter estimation. volume 27, page 17, 12 2011.
- Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- I.J. Good. Studies in the history of probability and statistics. XXXVII A.M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396, 1979.
- I.J. Good. Weight of evidence: A brief survey. Bayesian statistics, 2:249–270, 1985.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proc. ICML International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.
- Trevor Hastie and M Zhu. Dimension reduction and visualization in discriminant analysis discussion. Australian & New Zealand Journal of Statistics, 43:179–185, 06 2001.
- Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semisupervised learning with normalizing flows. In *Proc. ICML - International Conference* on *Machine Learning*, pages 4615–4630. PMLR, 2020.
- E. T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, Cambridge, 2003.
- Yann LeCun. The mnist database of handwritten digits. 1998.
- Dennis V. Lindley. *Understanding Uncertainty*. Wiley-Interscience, 2006.

- Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, second edition, 1999.
- Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.
- Glòria Mateu-Figueras, Vera Pawlowsky-Glahn, and Juan José Egozcue. The principle of working on coordinates. *Compositional data analysis: Theory and applications*, pages 29–42, 2011.
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. 2018.
- Ronald Meester and Klaas Slooten. Probability and Forensic Evidence: Theory, Philosophy, and Applications. Cambridge University Press, 2021.
- 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*, pages 41–48, 1999.
- Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.
- 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*, pages 1698–1702, 2020.
- Paul-Gauthier Noé. Representing evidence for attribute privacy: bayesian updating, compositional evidence and calibration. PhD thesis, Université d'Avignon, 2023.
- 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.
- Paul-Gauthier Noé, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, et al. Compositional discriminant analysis through calibrated evidence functions. In *The 10th International Workshop on Compositional Data Analysis (CoDaWork2024)*, 2024a.
- Paul-Gauthier Noé, Miquel Perelló-Nieto, Jean-François Bonastre, and Peter Flach. Explaining a probabilistic prediction on the simplex with shapley compositions. In *ECAI* 2024, pages 1124–1131. IOS Press, 2024b.
- 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*, pages 3094–3098, 2022.

- George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- W. Peterson, T. Birdsall, and W. Fox. The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, 4(4):171–212, 1954.
- José C. Pinheiro and Douglas M. Bates. Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296, 1996.
- Daniel Ramos. Forensic evaluation of the evidence using automatic speaker recognition systems. PhD thesis, Madrid, Universidad Politécnica de Madrid, 2007.
- Luis Javier Rodríguez-Fuentes, Niko Brümmer, Mikel Penagarikano, Amparo Varona, Germán Bordel, and Mireia Diez. The albayzin 2012 language recognition evaluation. pages 1497–1501, 2013.
- C. E. Shannon. Communication theory of secrecy systems. The Bell System Technical Journal, 28(4):656–715, 1949.
- Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.
- André Stuhlsatz, Jens Lippel, and Thomas Zielke. Feature extraction with deep neural networks by a generalized discriminant analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):596–608, 2012.
- John R Thornbury, Dennis G Fryback, and Ward Edwards. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. *Radiology*, 114(3):561–565, 1975.
- David van Leeuwen and Niko Brümmer. The distribution of calibrated likelihood-ratios in speaker recognition. In *Proc. ISCA Interspeech*, pages 1619–1623, 2013.
- Jason Weston, Bernhard Schölkopf, and Gökhan Bakir. Learning to find pre-images. In S. Thrun, L. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems, volume 16. MIT Press, 2003.
- Robert L. Winkler and Allan H. Murphy. "good" probability assessors. *Journal of Applied Meteorology*, 7(5):751–758, 1968.

Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2002.