SORT 34 (2) July-December 2010, 201-222

Bayes linear spaces

Karl Gerald van den Boogaart[1], Juan José Egozcue[2],
and Vera Pawlowsky-Glahn[3]

Abstract

Linear spaces consisting of σ-finite probability measures and infinite measures (improper priors and likelihood functions) are defined. The commutative group operation, called perturbation, is the updating given by Bayes theorem; the inverse operation is the Radon-Nikodym derivative. Bayes spaces of measures are sets of classes of proportional measures. In this framework, basic notions of mathematical statistics get a simple algebraic interpretation. For example, exponential families appear as affine subspaces with their sufficient statistics as a basis. Bayesian statistics, in particular some well-known properties of conjugated priors and likelihood functions, are revisited and slightly extended.

MSC: 60A10, 62E10

Keywords: Aitchison geometry, compositional data, exponential families, likelihood functions, probability measures, Radon-Nikodym derivative

1. Introduction

More than two decades ago, J. Aitchison (1986) noted that perturbation in the D-part simplex, the sample space of compositional data with a finite number of parts, *"is familiar in other areas of statistics ... as the operation of Bayes's formula to change a prior probability assessment into a posterior probability assessment through the perturbing influence of the likelihood function"* (Aitchison, 1986, p. 45). Recently, the linear space structure of the simplex has been recognised, with perturbation as the Abelian

[1] Institut für Stochastik, Fakultät für Mathematik und Informatik, TU Bergakademie Freiberg, Prüferstraße 9, D-09596 Freiberg, Germany, Tel.: 0049-(03731) 39-3225, Fax: 0049-(03731) 39-3598, boogaart@math.tu-freiberg.de

[2] Univ. Politécnica de Catalunya, Dep. Matemática Aplicada III, juan.jose.egozcue@upc.edu

[3] Univ. de Girona, Dep. Informática y Matemática Aplicada, vera.pawlowsky@udg.edu

Received: February 2010
Accepted: October 2010

group operation, and its Euclidean structure has been completed (Billheimer *et al.*, 2001; Pawlowsky-Glahn and Egozcue, 2001, 2002; Egozcue *et al.*, 2003) The extension of the underlying ideas to compositions of infinitely many parts is due to Egozcue *et al.* (2006). It leads to the study of probability densities with support on a finite interval, concluding with a Hilbert space structure based on the natural generalisation of the operations between compositions to operations between densities. The space contains both densities corresponding to finite measures, equivalent to probability measures, and densities corresponding to infinite measures, such as likelihood functions or improper (prior) densities. The extension to infinite support measures was suggested as an open problem and is now presented here.

Many different algebraic structures can be defined on sets of positive measures, and particulary on probability measures. For instance, certain classes of measures form a semi-group with respect to the ordinary sum or to the convolution (Bauer, 1992); Markov processes give rise to a semi-group of transition kernels (Markov-semigroups) (Bauer, 1992); $L^p(\lambda)$ can be seen as a space of densities of signed measures; random variables with variance constitute a Hilbert space (Witting, 1985; Small and Leish, 1994; Berlinet and Thomas-Agnan, 2004), which is relevant in statistical modelling; metric spaces are obtained defining distances such as Hellinger-Matusita (Hellinger, 1909; Matusita, 1955) or those based on Fisher-information. Finally, kernel reproducing Hilbert spaces (Whaba, 1990; Berlinet and Thomas-Agnan, 2004) are used for modelling stochastic processes, random measures and nonparametric functions, as well as linear observations of them, the inner product, reproducing kernel, and distance, being related to the variance of the process, and the elements of the space being realisations of stochastic processes (Whaba, 1990).

However, none of the above mentioned structures postulates Bayes updating as a group operation. Bayes theorem has two important characteristics that make it attractive as an operation between measures: (i) it has been considered as a paradigm of information acquisition, and (ii) it is a natural operation between densities (e.g. in probability, Bayesian updating; in system analysis, filtering in the frequency domain).

The primary goal of the present contribution is to provide a linear space structure for sets of classes of densities associated with positive measures of any support. The support of a density is treated as a measure itself, leading to a general and inclusive framework. In particular, linear spaces whose elements are classes of σ-additive positive measures – including probability measures, prior densities and likelihood functions – are introduced. Such spaces are suitable to review many issues of probabilistic modelling and statistics. We call them Bayes spaces because the Abelian group operation, or perturbation for short, corresponds to the operation implied in Bayes theorem. Section 2 defines Bayes linear spaces and Section 3 discusses their affine properties. Exponential families of distributions are identified as affine spaces in Section 4. In Section 5 a review of probabilistic models involved in Bayesian statistics is presented.

2. Bayes linear spaces

Standard tools of measure theory (Ash, 1972; Bauer, 1992, 2002; Shao, 1999) will be useful in the following development. Let λ be a σ-finite, positive measure on an arbitrary measurable space $(\Omega, \mathscr{B})$, where Ω is a non-empty set and $\mathscr{B}$ is a σ-field on Ω. The symbols λ and $\mathscr{B}$ have been chosen deliberately to associate them with the Lebesgue-measure and the Borelian σ-field, as they are a typical example for λ and $\mathscr{B}$. Measures with the same null-sets are called equivalent (Bauer, 1992). This is a very inclusive equivalence relation identifying e.g. the Lebesgue-measure – measuring the volume of a space portion – with any measure with positive density on the same measurable space. The class of measures equivalent to a reference measure, λ, is used to constitute the elements of the Bayes space:

Definition 1 (Equivalent measures) *Let λ and μ be σ-finite measures on $(\Omega, \mathscr{B})$. They are equivalent if, for all $R \in \mathscr{B}$, $\lambda(R) = 0$ if and only if $\mu(R) = 0$. The class of σ-finite measures on $(\Omega, \mathscr{B})$ equivalent to a given reference measure λ is denoted by $\mathscr{M}(\lambda)$ and its elements are called λ-equivalent measures.*

The Radon-Nikodym derivative theorem and the chain rule for densities are stated in the context of equivalent measures. The Radon-Nikodym-derivative is used to identify measures with functions:

Theorem 1 (Radon-Nikodym derivative) *Let λ be a σ-finite measure on $(\Omega, \mathscr{B})$, and μ a σ-finite λ-equivalent measure. Then, there exists a λ-almost-everywhere, $\lambda-$a.e., unique positive function $f : \Omega \to \mathbb{R}_+ = (0, \infty)$ such that, for any $R \in \mathscr{B}$, $\int_R d\mu = \int_R f \, d\lambda$. The function f is then called density, or Radon-Nikodym-derivative, of μ with respect to λ, and is denoted by*

$$\frac{d\mu}{d\lambda}(x) = f(x) \; .$$

Every measure in $\mathscr{M}(\lambda)$ can be represented by a unique density defined $\lambda-$a.e.. The chain rule is closely related to addition and difference in the Bayes linear space:

Theorem 2 (Chain rule for densities) *Let μ, ν be λ-equivalent measures. Then*

$$\frac{d\mu}{d\nu} = \frac{d\mu}{d\lambda} \, \frac{d\lambda}{d\nu} \; .$$

The aim of the following definitions is to build a linear space of classes of σ-finite measures represented either by probability measures or by infinite measures. The first step consists in identifying measures which differ only in a scale factor, leading to equivalence classes of proportional measures. As a consequence, finite measures can be represented by probability measures integrating to one. This idea has been previously used for densities on an interval in (Egozcue *et al.*, 2006) and goes back to a similar

idea which identifies equivalence classes of positive vectors with compositions (Barceló-Vidal *et al.*, 2001).

Definition 2 (*B*-equivalence) *Let μ and ν be measures in $\mathscr{M}(\lambda)$. They are B-equivalent, $\mu =_B \nu$, if and only if there exists a constant $c > 0$ such that, for any $R \in \mathscr{B}$, $\mu(R) = c \cdot \nu(R)$, using the convention $c \cdot (+\infty) = +\infty$. The set of $(=_B)$ equivalent classes is denoted as a quotient space $B(\lambda) = \mathscr{M}(\lambda)/(=_B)$.*

Theorem 3 *$(=_B)$ is an equivalence relation on $\mathscr{M}(\lambda)$.*

The elements of $B(\lambda) = \mathscr{M}(\lambda)/(=_B)$ are $(=_B)$-equivalence classes of measures in $\mathscr{M}(\lambda)$. From now on, no notational difference will be made between a measure and the equivalence class it represents. When a reference measure λ is fixed, a $(=_B)$-class of measures will be represented by a density (or Radon-Nikodym derivative with respect to λ) defined $\lambda-$a.e. and up to a positive constant. The equivalence symbol $(=_B)$ will be used for $\mu, \nu \in \mathscr{M}(\lambda)$ and for their respective densities, f_μ and f_ν. Thus, if $\mu =_B \nu$, then $f_\nu =_B f_\mu$, which means that there exists c such that $f_\nu(x) = c f_\mu(x)$ $\lambda-$a.e. Summarising, $(=_B)$ identifies a measure equivalence class with a density, and the measures are all seen as the same element of $B(\lambda)$. To build a linear space on $B(\lambda)$, the second step consists in introducing addition and multiplication by real scalars.

Definition 3 (Perturbation and powering) *Let μ and ν be measures in $B(\lambda)$. For every $R \in \mathscr{B}$, the perturbation of μ by ν is the measure in $B(\lambda)$ such that*

$$(\mu \oplus \nu)(R) = \int_R \frac{d\mu}{d\lambda} \frac{d\nu}{d\lambda} \, d\lambda \, . \tag{1}$$

For a scalar $\alpha \in \mathbb{R}$, the powering of μ is the measure in $B(\lambda)$ such that

$$(\alpha \odot \mu)(R) = \int_R \left(\frac{d\mu}{d\lambda} \right)^\alpha d\lambda \tag{2}$$

Theorem 4 *Perturbation and powering of σ-finite λ-equivalent measures are σ-finite.*

Proof: see appendix.

Perturbation and powering of σ-finite λ-equivalent measures are based on perturbation and powering in the simplex, introduced originally by J. Aitchison (Aitchison, 1986) and shown later to structure the simplex as a linear space (Martín-Fernández *et al.*, 1999; Billheimer *et al.*, 2001; Pawlowsky-Glahn and Egozcue, 2001; Aitchison *et al.*, 2002). The space is denoted $B(\lambda)$ (B for Bayes) recalling that perturbation, which plays the role of group operation, is essentially the operation in Bayes theorem.

The inverse operation of perturbation in $B(\lambda)$, i.e. substraction in $B(\lambda)$, is defined as $\ominus\mu =_B (-1) \odot \mu$. The use of densities representing the corresponding measures

generates alternative definitions of perturbation and powering. Let f_μ and f_ν be densities in $B(\lambda)$ and $\alpha \in \mathbb{R}$; then, perturbation, difference and powering are

$$(f_\nu \oplus f_\mu)(x) =_B f_\nu(x) f_\mu(x)\,, \tag{3}$$

$$(f_\nu \ominus f_\mu)(x) =_B \frac{f_\nu(x)}{f_\mu(x)}\,, \tag{4}$$

$$(\alpha \odot f_\nu)(x) =_B f_\nu(x)^\alpha\,. \tag{5}$$

Combining measures and densities we get equivalent expressions:

$$(f_\nu \oplus \mu) =_B \int_A f_\nu(x) d\mu(x)\,, \tag{6}$$

$$(\nu \ominus \mu)(x) =_B \frac{d\nu}{d\mu}\,. \tag{7}$$

A remarkable fact is that the difference (4), (7) is actually a Radon-Nikodym derivative due to the chain rule (Theorem 2).

When using densities representing measures, operations depend on the reference measure λ adopted. Therefore, whenever not clear from the context, a subscript will be used: $\oplus_\lambda$, $\ominus_\lambda$, $\odot_\lambda$, $=_{B(\lambda)}$.

Theorem 5 *With operations $\oplus$ and $\odot$, $B(\lambda)$ is a real vector space.*

Proof: see appendix.

Whatever the reference measure λ, the neutral element of $B(\lambda)$ with respect to perturbation is a constant density, or equivalently, the density with constant value 1. The perturbation-opposite of a density f_μ is B-equivalent to $1/f_\mu$.

Definition 4 (Bayes space) *The linear space $(B(\lambda), \oplus, \odot)$ is called Bayes space with reference measure λ.*

When the measurable space is $(\Omega, \mathscr{B}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with $\mathscr{B}(\mathbb{R})$ the Borel σ-field on $\mathbb{R}$, the most commonly used reference measure is the Lebesgue measure $\lambda_\mathbb{R}$. For constrained measurable spaces such as the positive real line, $\Omega = \mathbb{R}_+$, or the 3-part simplex, $\Omega = \mathscr{S}^3$, with the corresponding restricted Borelians, the Lebesgue measure restricted to them, λ_+, respectively $\lambda_{\mathscr{S}^3}$, may be readily used. These contexts are usual in probability theory and do not need further examples. Similarly, the measurable spaces of the integers or the non-negative integers, $(\mathbb{Z}, \mathbb{Z}_+)$, are normally used with the counting

measure as a reference. However, different but useful reference measures can be taken in $\mathbb{R}_+$ and in $\mathscr{S}^3$. As they are seldom used, they are given as examples.

Example 1 Consider $(\Omega, \mathscr{B}) = (\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+))$, being $\mathbb{R}_+$ the strictly positive real numbers. A natural reference is the relative measure, defined for any interval $[a,b] \subset \mathbb{R}_+$, as $\mu_+([a,b]) = \ln b - \ln a$, whose density with respect to λ_+ is

$$\frac{d\mu_+}{d\lambda_+} = \frac{d\ \ln(x)}{dx} = \frac{1}{x}\,.$$

The reference measure μ_+ corresponds to a constant density in the space $B(\mu_+)$. Moreover, in $B(\mu_+)$, the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \xi)^2}{2\sigma^2}\right), \tag{8}$$

represents a log-normal probability law with median $\exp(\xi)$ and logarithmic variance σ^2. It has been called the normal in $\mathbb{R}_+$ (Eaton, 1983; Mateu-Figueras *et al.*, 2002) and is accordingly denoted by $\mathscr{N}_+(\xi, \sigma^2)$. The positive real line, $\mathbb{R}_+$, can be structured as an Euclidean space taking into account that $\ln : \mathbb{R}_+ \to \mathbb{R}$ is a one-to-one mapping (Pawlowsky-Glahn and Egozcue, 2001). Then, μ_+ is induced by the Lebesgue measure in $\mathbb{R}$. Thus, the reference measure μ_+ corresponds to a relative scale in $\mathbb{R}_+$.

Example 2 The unit 3-part simplex, $\mathscr{S}^3 \subset \mathbb{R}^3$, has elements which are vectors with 3 strictly positive components adding to 1. The simplex $\mathscr{S}^3$ has been shown to be a 2-dimensional Euclidean space using perturbation and powering (as operations of its elements) and the Aitchison metrics (Pawlowsky-Glahn and Egozcue, 2001; Billheimer *et al.*, 2001). Consequently, an orthonormal basis can be defined such that elements in the simplex can be represented by the corresponding coordinates. Once an orthonormal basis has been selected, the mapping assigning coordinates to each element of the simplex has been called isometric log-ratio transformation (ilr) (Egozcue *et al.*, 2003). A particular case of ilr can be used to define a new reference measure in $\mathscr{S}^3$ in the following way. Take $\Omega = \mathscr{S}^3$ and consider the one-to-one mapping $\mathrm{ilr} : \mathscr{S}^3 \to \mathbb{R}^2$ defined by

$$\mathrm{ilr}(\vec{x}) = \left(\frac{1}{\sqrt{2}}\ \ln\frac{x_1}{x_2}\,,\ \frac{1}{\sqrt{6}}\ \ln\frac{x_1 x_2}{x_3^2}\right),$$

where $\vec{x} = (x_1, x_2, x_3) \in \mathscr{S}^3$. Define the σ-field $\mathscr{B}(\mathscr{S}^3) = \mathrm{ilr}^{-1}(\mathscr{B}(\mathbb{R}^2))$ and a reference measure $\alpha_{\mathscr{S}^3}(\mathrm{ilr}^{-1}(R)) = \lambda_{\mathbb{R}^2}(R)$, for $R \in \mathscr{B}(\mathbb{R}^2)$. The measure $\alpha_{\mathscr{S}^3}$ is called Aitchison measure (Egozcue *et al.*, 2003; Mateu-Figueras *et al.*, 2003; Pawlowsky-Glahn, 2003). In this context, the additive logistic normal probability distribution (aln) (Aitchison,

1986) is represented by the density

$$f(\vec{x}) = \frac{1}{2\pi|\Sigma|}\exp\left(-\frac{1}{2}(\mathrm{ilr}(\vec{x})-\boldsymbol{\mu})^t\Sigma^{-1}(\mathrm{ilr}(\vec{x})-\boldsymbol{\mu})\right), \tag{9}$$

where vectors of three components in $\mathscr{S}^3$ are denoted using $(\vec{\ })$ and vectors in $\mathbb{R}^2$ are boldfaced. Σ is a $(2,2)$-covariance matrix, $|\Sigma|$ denotes its determinant, and $\boldsymbol{\mu}\in\mathbb{R}^2$ plays the role of a mean because $\mathrm{ilr}^{-1}(\boldsymbol{\mu})$ is actually the centre of the distribution. This probability measure corresponds to Aitchison's aln-probability law or logistic-normal distribution. However, the density (9) has been called normal in $\mathscr{S}^3$ (Mateu-Figueras *et al.*, 2003) because of the absence of the Jacobian of the ilr transformation, which is the density of the reference measure $\alpha_{\mathscr{S}^3}$ with respect to $\lambda_{\mathbb{R}^3}$.

3. Affine transformation and subsets of $B(\lambda)$

Changing the reference measure of $B(\lambda)$ to a B-equivalent one does not change the space. The transformation from $B(\lambda)$ to $B(\mu)$, being $\mu\in\mathscr{M}(\lambda)$, is an affine transformation and may be interpreted as a change of origin.

Theorem 6 *Let μ be a measure in $\mathscr{M}(\lambda)$. Then, $\mu =_B \lambda$ if and only if $B(\mu)$ and $B(\lambda)$ are equal as linear spaces.*

Proof: see appendix.

When changing the reference measure, or the origin, of the space $B(\lambda)$, the identification of density and measure is broken. Next theorem on change of origin is formulated in terms of measures, thus avoiding notation with densities.

Theorem 7 (Change of origin) *For all $\mu\in\mathscr{M}(\lambda)$ the spaces $B(\mu)$ and $B(\lambda)$ have the same elements and are equivalent as affine spaces. Consequently, changing the reference measure is a simple shift operation.*

Proof: see appendix.

In analytic geometry the elements of a linear space can be seen from two different points of view: points in the space and vectors or arrows. The first corresponds to affine geometry, the second to the vector space. In the present context, the elements of $B(\lambda)$ can be represented by measures, e.g. μ, ν. This representation by measures corresponds to *points*. Alternatively, the difference $\mu\ominus\nu =_B d\mu/d\nu$, which is actually a density, correspond to a *vector*, i.e. the difference between points is a *vector*. However, as in analytical geometry, there is no mathematical difference between *points* and *vectors* of any kind. The only practical difference arises when shifting the origin from λ to λ'. The vector representation $d\mu/d\lambda\in B(\lambda)$ of the *point* μ is then shifted by subtracting the new

origin represented as a *vector*: $(d\mu/d\lambda') = (d\mu/d\lambda)(d\lambda/d\lambda') =_B (d\mu/d\lambda) \ominus (d\lambda'/d\lambda)$. Therefore, the use of the density notation $f_\mu = d\mu/d\lambda$ makes sense only when the reference measure λ is clearly specified, because the density changes under change of origin.

The space $B(\lambda)$ contains $(=_B)$-classes of finite measures and other classes of infinite measures (σ-finite). A finite measure μ, can be represented by a probability measure $\mu/\mu(\Omega)$, being $\mu =_B \mu/\mu(\Omega)$. Infinite measures cannot be normalised in this way because the measure of the whole space Ω is then infinite. The latter $(=_B)$-classes contain measures like improper priors or improper likelihood functions appearing regularly in Bayesian statistics. In this context, $(=_B)$-equivalence achieves its full meaning as the likelihood principle that identifies proportional proper or improper densities (Birnbaum, 1962; Leonard and Hsu, 1999; Robert, 2001). This means that the space $B(\lambda)$ is decomposed into two well defined subsets: the set of classes of finite measures, $B_P(\lambda)$ containing proper probability measures; and $B_I(\lambda)$ containing classes of infinite measures. By definition $B_P(\lambda)$ and $B_I(\lambda)$ constitute a partition of $B(\lambda)$. The different role that proper and improper densities play in statistics motivates the following properties concerning $B_P(\lambda)$ and $B_I(\lambda)$. Some properties are related to other two important subsets of $B(\lambda)$, namely the set of measures whose density is upper bounded $\lambda-$a.e., $B_u(\lambda)$, and the set of measures whose densities are double bounded, i.e. such that if f a density in $B(\lambda)$, then there exist a positive constant, b, such that $0 < 1/b < f < b < +\infty$ ($\lambda-$a.e.); this subset is denoted by $B_b(\lambda)$.

Theorem 8

1. $B_P(\lambda)$, $B_I(\lambda)$ *is a partition of* $B(\lambda)$.
2. $B_P(\lambda)$ *is convex.*
3. $B_b(\lambda)$ *is a subspace of* $B(\lambda)$.
4. $B_u(\lambda)$ *is a convex cone.*
5. $B_P(\lambda) \oplus B_u(\lambda) = B_P(\lambda)$.
6. $B_I(\lambda) \ominus B_u(\lambda) = B_I(\lambda)$.
7. $\mu \in B(\lambda)$ *if and only if* $B_P(\mu) = B_P(\lambda)$ *as sets of measures.*
8. $\mu \in B_b(\lambda)$ *if and only if* $B_b(\mu) = B_b(\lambda)$ *as sets of measures.*
9. $\mu \in B_P(\mu)$ *if and only if* $B_b(\mu) \subset B_P(\mu)$.
10. $\mu \in B_I(\mu)$ *if and only if* $B_b(\mu) \subset B_I(\mu)$.

Proof: see appendix.

4. Exponential families as affine spaces

Many commonly used distribution families, including multinomial, normal, beta, gamma and Poisson, are exponential families. A common general definition can be given as follows (Witting, 1985):

Definition 5 (Exponential family) *For λ a measure on a measurable space $(\Omega,\mathscr{B})$, consider a strictly positive measurable function $g:(\Omega,\mathscr{B})\to(\mathbb{R}^+,\mathscr{B}(\mathbb{R})|_{\mathbb{R}^+})$; a vector of measurable functions $\vec{T}=(T_1,T_2,\ldots,T_k)$ with $T_i:(\Omega,\mathscr{B})\to(\mathbb{R},\mathscr{B}(\mathbb{R}))$, $i=1,\ldots,k$; and a function $\vec{\boldsymbol{\theta}}=(\theta_1,\theta_2,\ldots,\theta_k)$, where $\theta_i:A\to\mathbb{R}$ and A is a parameter space. A k-parametric exponential family of distributions, $P_{\vec{\boldsymbol{\alpha}}}$, $\vec{\boldsymbol{\alpha}}\in A$, on $(\Omega,\mathscr{B})$ is given by*

$$\frac{dP_{\vec{\boldsymbol{\alpha}}}}{d\lambda}(x)=f_{\vec{\boldsymbol{\alpha}}}(x)=C(\vec{\boldsymbol{\alpha}})\cdot g(x)\cdot\exp\left[\sum_{j=1}^{k}\theta_j(\vec{\boldsymbol{\alpha}})T_j(x)\right],$$

with a normalisation constant

$$C(\vec{\boldsymbol{\alpha}})=\left(\int\exp\left[\sum_{j=1}^{k}\theta_j(\vec{\boldsymbol{\alpha}})T_j(x)\right]\,g(x)\,d\lambda(x)\right)^{-1}. \tag{10}$$

The exponential family is denoted $\mathrm{Exp}(\lambda,g,\vec{T},\vec{\boldsymbol{\theta}})$. *If k is minimal, the family is called strictly k-parametric.*

The function $\kappa(\vec{\boldsymbol{\alpha}})=-\ln C(\vec{\boldsymbol{\alpha}})$ is called the cumulant function of the family. Classically, the parameter space A is restricted to values of $\vec{\boldsymbol{\alpha}}$ for which $C(\vec{\boldsymbol{\alpha}})$ exists. Frequently, λ is called reference measure, and is typically the Lebesgue measure on $\mathbb{R}$ when the support of the random variable is $\mathbb{R}$, or a counting measure when the support is discrete; $\vec{T}(x)$ defines a set of statistics; and $\vec{\boldsymbol{\theta}}(\vec{\boldsymbol{\alpha}})$ is a mapping of the used parameters $\vec{\boldsymbol{\alpha}}\in A$ into the so-called natural parameters, $\theta_i(\vec{\boldsymbol{\alpha}})$, of the family. The normal family of distributions is a typical case: $g(x)$ is constant; $\vec{T}(x)=(x,x^2)$; $\vec{\boldsymbol{\alpha}}=(m,\sigma^2)$, where m is the mean and σ^2 is the variance; and $\vec{\boldsymbol{\theta}}(\vec{\boldsymbol{\alpha}})=(\theta_1(\vec{\boldsymbol{\alpha}}),\theta_2(\vec{\boldsymbol{\alpha}}))=(m/\sigma^2,-1/(2\sigma^2))$.

As mentioned, classical exponential families are defined only for those $\vec{\boldsymbol{\alpha}}$ for which $\kappa(\vec{\boldsymbol{\alpha}})$ or $C(\vec{\boldsymbol{\alpha}})$ in (10) exists. However, the idea of Bayes spaces permits to drop this condition and infinite measures can be considered natural members of exponential families. A definition of such extended exponential families is the following.

Definition 6 (Extended exponential family) *Using the notation in definition 5, an extended exponential family, denoted* $\mathrm{Exp}_B(\lambda, g, \vec{T}, \vec{\boldsymbol{\theta}})$, *contains the densities*

$$\frac{dP_{\vec{\boldsymbol{\alpha}}}}{d\lambda}(x)=_B f_{\vec{\boldsymbol{\alpha}}}(x)=_B g(x)\cdot\exp\left[\sum_{j=1}^{k}\theta_j(\vec{\boldsymbol{\alpha}})T_j(x)\right].$$

If k is minimal, the family is called strictly k-parametric.

Densities in the extended family may or may not correspond to probability measures. Particularly, the elements with finite integral form the exponential family in the ordinary sense. Next theorems account for the properties of the extended exponential families.

Theorem 9 *An extended exponential family* $\mathrm{Exp}_B(\lambda, g, \vec{T}, \vec{\boldsymbol{\theta}})$ *is a finite dimensional affine subspace of the Bayes space* $B(\lambda)$.

Proof: see appendix.

Theorem 10 *Any k-dimensional affine subspace S of* $B(\lambda)$ *is a strictly k-parametric extended exponential family.*

Proof: see appendix.

When an extended exponential family is viewed as an affine space, g can be identified as the origin of the affine space. Also, the change of origin of $B(\lambda)$ from λ to $\mu =_B \lambda \oplus g$, where g is taken as a density of a σ-finite measure, transforms the exponential family into a subspace of $B(\mu)$ because the constant density or neutral element for $\oplus$ is now an element of the family. Another important aspect is that the natural parameters $\theta_j(\vec{\boldsymbol{\alpha}})$ are the coordinates of $\mu_{\vec{\boldsymbol{\alpha}}}$ expressed in the basis elements $V_j(x)$. The restriction of the parameter space of exponential families, due to the integrability condition for the existence of the normalisation constant, is not any more needed in this context. Non integrable elements correspond to densities of infinite measures in $B_I(\lambda)$. When exponential families must be used as families of probability distributions, improper distributions can be just ignored and restrictions to the parameters apply.

Example 3 For $\Omega = \mathbb{R}_+$, and using the notation of Example 1, the log-normal exponential family is

$$f_{\xi,v}(x) = \frac{dP_{\xi,v}}{d\lambda_+}(x) = \frac{1}{\sqrt{2\pi v}} \cdot \frac{1}{x} \cdot \exp\left(-\frac{(\ln x - \xi)^2}{2v}\right),$$

where v is the logarithmic variance and $C(\xi, v) = \exp(-\frac{1}{2v}\xi^2)/\sqrt{(2\pi v)}$, $g(x) = 1/x$, $\vec{\boldsymbol{\theta}} = (\xi/v, -1/(2v))$ and $\vec{T} = (\ln x, (\ln x)^2)$. However, for real values of ξ and positive values of v, $\theta_2 = -1/(2v) < 0$; this means that the family only spans half of the affine space, an affine cone, in $B(\lambda_+)$. The whole affine space is spanned accepting values $v < 0$; for these values, $f_{\xi,v}(x)$ is no longer a probability density but it belongs to $B_I(\lambda_+) \subset B(\lambda_+)$. Additionally, changing the origin from λ_+ to $\mu_+ =_B 1/x$ the family adopts the form

$$\frac{dP_{\xi,v}}{d\mu_+} =_{B(\mu_+)} \exp\left(-\frac{(\ln x - \xi)^2}{2v}\right),$$

which is again the normal in $\mathbb{R}_+$ (8) given in Example 1. The family can be expressed as a subspace of $B(\mu_+)$,

$$\frac{dP_{\xi,v}}{d\mu_+} =_{B(\mu_+)} \left(\frac{\xi}{v} \odot e^x\right) \oplus \left(\frac{1}{v} \odot \frac{dP_{0,1}}{d\mu_+}\right),$$

whereas the family span is an affine subspace of $B(\lambda_+)$,

$$\frac{dP_{\xi,\nu}}{d\lambda_+} =_{B(\lambda_+)} \frac{1}{x} \oplus \left(\frac{\xi}{\nu} \odot e^x\right) \oplus \left(\frac{1}{\nu} \odot \frac{dP_{0,1}}{d\lambda_+}\right) .$$

5. Bayes theorem is summing information

The following context is inspired by Bayesian statistics, however it is also relevant in likelihood function based statistics. For the observations consider a measurable space $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$, with $\mathscr{B}(\mathscr{X})$ a σ-field on $\mathscr{X}$, and a reference measure on it denoted by λ. Let $\vec{x} = (x_1, x_2, \ldots, x_n) \in \mathscr{X}^n$ be the vector of observations modelled by independent random variables X_i with values in $\mathscr{X}$ and probability law given by the measure $P_\theta \in B_P(\lambda)$, distribution for short, depending on a set of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ with values in a measurable space $(\Theta, \mathscr{B}(\Theta))$ of parameters. Denote by P_{prior} a prior distribution on $(\Theta, \mathscr{B}(\Theta))$, by P_{post} the posterior, by

$$L_{x_i}(\theta) = \frac{dP_\theta}{d\lambda}(x_i) ,$$

the individual likelihood functions, and by $L_{\vec{x}}(\theta) = \prod_i L_{x_i}(\theta)$ the joint likelihood function. According to the likelihood principle (Leonard and Hsu, 1999), a likelihood L_{x_i} and its scaled version αL_{x_i} should give the same result in the analysis. Thus $=_B$ for functions of θ is a natural equivalence relation for likelihood functions. Consider a reference measure $\tau \in \mathscr{M}(P_{prior})$ on $(\Theta, \mathscr{B}(\Theta))$. Now, two different Bayes spaces are relevant in this situation:

- The Bayes space $B(\lambda)$ containing the family $\{P_\theta : \theta \in \Theta\}$ of distributions for the observations, being $P_\theta \in \mathscr{M}(\lambda)$ on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$.
- The Bayes space $B(\tau)$ containing the distributions of the parameters P_{prior}, P_{post}, for the reference measure τ on $(\Theta, \mathscr{B}(\Theta))$.

Theorem 11 *If the distributions of the family $\{P_\theta : \theta \in \Theta\}$ are in $B(\lambda)$, then $L_{x_i} \in B(\tau)$, $P_\theta - a.e.$*

Proof: see appendix.

In this context, the Bayes formula can be written

$$\frac{dP_{post}}{d\tau}(\theta) = \frac{\frac{dP_{prior}}{d\tau}(\theta) \prod_{i=1}^n L_{x_i}(\theta)}{\int_\Theta \frac{dP_{prior}(\theta)}{d\tau(\theta)} \prod_{i=1}^n L_{x_i}(\theta) d\tau(\theta)} .$$

The denominator is a constant not depending on θ, accordingly,

$$\frac{dP_{post}}{d\tau}(\theta) =_B \frac{dP_{prior}}{d\tau}(\theta) \prod_{i=1}^{n} L_{x_i}(\theta),$$

which, using Bayes space operations, simplifies to the following theorem:

Theorem 12 (Bayes theorem in terms of Bayes spaces) *If $P_\theta \in B(\lambda)$ and the prior $P_{prior} \in B(\tau)$ then, $\bigotimes_{i=1}^{n} P_\theta(x_i)$-a.e.,*

$$P_{post} =_B P_{prior} \oplus \bigoplus_{i=1}^{n} L_{x_i} \tag{11}$$

Bayes theorem has several well-known and interesting direct implications. Here, Theorem 12 is an elegant form of Bayes formula: it is a sum in a vector space and, consequently, Bayesian updating is associative, commutative, invertible and has a neutral element (the non-informative experiment here represented by the measure τ). Also, the addition of the prior is invertible, as the prior can be subtracted and another prior can be added. Thus, adding information in terms of Bayes statistics is nothing but summing vectors in a space of information, here represented by $B(\tau)$. This means that the three densities P_{prior}, $L_{\vec{x}}$ and P_{post} represent information: before the experiment, provided by the experiment, and updated from the experiment respectively. Furthermore, Bayes formula as expressed in Theorem 12, admits both proper or improper priors and improper intermediate posteriors. Also the likelihood function of a repeated independent observation takes the form of a sum:

Corollary 1 *In the conditions of Theorem 12,*

$$L_{\vec{x}} =_B \bigoplus_{i=1}^{n} L_{x_i}$$

6. Bayes theorem and exponential families

In order to simplify the notation, the natural parameters of an exponential family will be used instead of the dependence on general parameters $\vec{\boldsymbol{\theta}}(\vec{\boldsymbol{\alpha}})$; then, arguments of functions of the parameters will be expressed simply as $\vec{\boldsymbol{\theta}}$. The components of the boldfaced vectors of parameters, statistics and observations, are denoted with the same text letters subscripted to indicate component.

Theorem 13 *Let x_i, $i = 1,\ldots,n$, be repeated independent observations from a strictly k parametric exponential family* $\mathrm{Exp}_{B(\lambda)}(\lambda, g, \vec{\boldsymbol{\theta}}, \vec{T})$,

$$P_{\vec{\boldsymbol{\theta}}}(x) = C(\vec{\boldsymbol{\theta}}) \cdot g(x) \cdot \exp\left(\sum_{j=1}^{k} \theta_j T_j(x)\right),$$

then, the joint likelihood $L_{\vec{x}}(\vec{\boldsymbol{\theta}})$, as a function of $\vec{\boldsymbol{\theta}}$, is a $k+1$-parametric family $\mathrm{Exp}_{B(\tau)}(\tau, g^*, \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$, *with $g^*(\vec{\theta}) = 1$, $\vec{\boldsymbol{\theta}}^* = (\ln C(\vec{\boldsymbol{\theta}}), \vec{\boldsymbol{\theta}})$, and $\vec{T}^*(\vec{x}) = (n, \sum_{i=1}^{n} \vec{T}(x_i))$. The family is strictly k_1-parametric with $k \le k_1 \le k+1$.*

Proof: see appendix.

A remarkable fact is that the initial statistic $\vec{T}$ plays the role of the vector of natural parameters, $\vec{T}^*$, in the resulting exponential family. Also, note that the first element in $\vec{\boldsymbol{\theta}}^*$ is the negative cumulant function $\kappa(\vec{\boldsymbol{\theta}}) = -\ln C(\vec{\boldsymbol{\theta}})$. Theorem 13 allows the identification of conjugated families of priors and densities of observations.

Theorem 14 *In the conditions of Theorem 13, a prior density $P_{prior}(\vec{\boldsymbol{\theta}})$ in* $\mathrm{Exp}_{B(\tau)}(\tau, g^*, \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$ *generates a posterior density through the Bayes theorem*

$$P_{post} =_{B(\tau)} L_{\vec{x}} \oplus P_{prior},$$

which is also in $\mathrm{Exp}_{B(\tau)}(\tau, g^*, \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$, *i.e.* $\mathrm{Exp}_{B(\tau)}(\tau, g^*, \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$ *and* $\mathrm{Exp}_{B(\lambda)}(\lambda, g, \vec{\boldsymbol{\theta}}, \vec{T})$ *are conjugated families.*

Proof: see appendix.

It is well known that, for exponential families of densities of observations, an exponential family of conjugated priors exists such that it also contains the posteriors (Leonard and Hsu, 1999). Next theorem goes a little bit further, stating that, regardless of the prior, the possible posterior densities are in an extended exponential family.

Theorem 15 *If the likelihood function of a multiple observation $L_{\vec{x}}$ satisfies the conditions of Theorem 13, for any prior $P_{prior} \in B(\tau)$, the posterior, $P_{post} =_{B(\tau)} L_{\vec{x}} \oplus P_{prior}$, is in* $\mathrm{Exp}_{B(\tau)}(\tau, P_{prior}(\vec{\boldsymbol{\theta}}), \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$.

Proof: see appendix.

Next theorem is also a new result for exponential families of posteriors stating the converse of Theorem 15.

Theorem 16 *Assume that the posterior density is obtained from the Bayes formula $P_{post}(\vec{\boldsymbol{\theta}}) =_{B(\tau)} L_{\vec{x}}(\vec{\boldsymbol{\theta}}) \oplus_{\tau} P_{prior}(\vec{\boldsymbol{\theta}})$, where $P_{prior}(\vec{\boldsymbol{\theta}})$ is the prior and the likelihood function is $L_{\vec{x}}(\vec{\boldsymbol{\theta}}) = \prod_{i=1}^{n} L_{x_i}(\vec{\boldsymbol{\theta}})$. If $P_{post}(\vec{\boldsymbol{\theta}}|\vec{x}) \in \mathrm{Exp}_{B(\tau)}(\tau, h, \vec{\boldsymbol{\theta}}, \vec{S})$, then $L_x(\vec{\boldsymbol{\theta}})$, as a func-*

tion of x, is in $\mathrm{Exp}_{B(\lambda)}(\lambda,1,\vec{T},\vec{\boldsymbol{\theta}})$*, for some statistic* $\vec{T}(x)$*. If* $\mathrm{Exp}_{B(\tau)}(\tau,h,\vec{\boldsymbol{\theta}},\vec{S})$ *is k-dimensional, then* $\mathrm{Exp}_{B(\lambda)}(\lambda,1,\vec{T},\vec{\boldsymbol{\theta}})$ *is* k_1*-dimensional with* $k_1 \leq k$*.*

Proof: see appendix.

Corollary 2 *A family of* λ*-equivalent distributions is in an exponential family if and only if, for any prior, the family of its posteriors (perturbation of prior and a member of the family) is an extended exponential family.*

Example 4 Consider $\mathbb{Z}_+$, the non-negative integers, as space of observations, and the counting measure ν as a reference measure on it, i.e. $\nu(\{x\}) = 1$ for any single point $\{x\}$ in $\mathbb{Z}_+$. Define the two-parametric exponential family

$$\mathrm{Exp}(\nu, g(x), (\theta_1,\theta_2), (T_1(x), T_2(x)))\,,$$

with $g(x) = (x!)^{-1}$, $\theta_1 = \ln\phi$, $T_1 = x$, $T_2 = \delta(x)$, with $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. A density of this exponential family has the expression

$$f(x|\phi,\theta_2) = C(\phi,\theta_2)\cdot\frac{1}{x!}\cdot\exp\left(x\ln\phi + \delta(x)\theta_2\right),\ \phi > 0\,, \tag{12}$$

being the normalising constant

$$C(\phi,\theta_2) = \frac{1}{\exp(\theta_2)+\exp(\phi)-1}\,.$$

The density (12) is the Bayes-perturbation in $B(\nu)$ of a Poisson density of parameter ϕ by a step-density $\exp(\theta_2\delta(x))$, the latter in $B_I(\nu)$. However, the whole family is in $B_P(\nu)$ according to Theorem 8, number 5. Note that, for $\theta_2 = 0$, the family reduces to the standard Poisson exponential family. The exponential family (12) may be called *zero-inflated Poisson* family (Lambert, 1992) because it can be written

$$f(x|\phi,\theta_2) = (1-p)\cdot\delta(x) + p\cdot\frac{\phi^x e^{-\phi}}{x!}\,,\ \theta_2 = \ln\left[(1-p)e^{\phi}+p\right]\,,$$

as a mixture of a Dirac and a Poisson distributions, although from the latter expression it is difficult to deduce its exponential character. This zero-inflated Poisson family can also be expressed as an affine subspace of $B(\nu)$

$$f(x|\phi,\theta_2) \;=_{B(\nu)}\; \frac{1}{x!}\oplus(\ln\phi\odot e^x)\oplus\left(\theta_2\odot e^{\delta(x)}\right)\,,$$

or, alternatively, taking $\mu = \nu\ominus(1/x!)$ as reference measure, the family is a subspace of $B(\mu)$

$$f(x|\phi,\theta_2) =_{B(\mu)} (\ln\phi \odot e^x) \oplus \left(\theta_2 \odot e^{\delta(x)}\right) .$$

In both cases, with $\theta_2 = 0$, the extended Poisson family is obtained.

A natural question is which is the conjugated family of prior densities. Theorem 13 implies that this family is 3-parametric and the densities are

$$P_{prior}(\theta_1,\theta_2) =_B \exp\left(t_0 \ln C(e^{\theta_1},\theta_2) + t_1\theta_1 + t_2\theta_2\right) ,$$

where the parameters t_0, t_1 and t_2 have the following meaning: t_0 corresponds to the sample size; t_1 stands for the total sum of the observations, $\sum x_i$, and t_2 is the number of null observations, $\sum \delta(x_i)$. This family of priors contains both proper and improper priors because the t_i are arbitrary real numbers. The family of prior densities, as functions of the natural parameters of the family (12), i.e. (θ_1,θ_2), is in $B(\lambda_{\mathbb{R}^2})$. Finally, note that (12) may be expressed using the measure whose density is $(x!)^{-1}$ as a reference. In this case, the expression (12) remains the same but removing the factorial.

7. Conclusion

Classes of proportional σ-finite measures, including probability measures, have been structured as Bayes linear spaces. These classes can be represented by densities, including probability densities, likelihood functions and improper priors. The group operation, perturbation, is Bayes updating, thus defining a meaningful and interpretable structure. The affine subspaces are identified with extended exponential families, which include standard probability densities (or measures) and, additionally, infinite measures. Standard theorems of Bayesian statistics are revisited and slightly extended using this new algebraic-geometric point of view. The idea that Bayes theorem is the paradigm of information acquisition is now interpreted as an addition in the formal sense, being this possible because (proper and improper) probability densities and likelihood functions share the same Bayes space.

The presented framework permits a new interpretation of the standard probability theory, justifies the use of improper probability densities and opens up the study of some subspaces which may have richer structures with a metric or even a Hilbert space structure. The examples presented refer to quite usual probabilistic models, like normal and log-normal distributions; other distributions, although well-known and useful in practice (logistic normal, zero-inflated Poisson) need a more detailed mathematical development. The presented methodology, when applied to these examples, illustrates the new perspective introduced, namely how to deal with probability models in the framework of Bayes spaces. In particular, the idea that exponential families constitute an advanced mathematical tool in mathematical statistics, is here reduced to a very simple model, i.e. in the new framework they are linear affine subspaces.

Acknowledgements

This research has been supported by the Spanish Ministry of Education and Science under projects Ref.: MTM2009-13272 and Ref.: *Ingenio Mathematica (i-MATH)* No. CSD2006-00032 *(Consolider - Ingenio 2010)*, and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under project Ref: 2009SGR424.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 pp.

Aitchison, J., Barceló-Vidal, C., Egozcue, J. J. and Pawlowsky-Glahn, V. (2002). A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. See Bayer *et al.*, pp. 387-392.

Ash, R. B. (1972). *Real Analysis and Probability*. Academic Press, New York, NY (USA). 476 pp.

Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 - The 6th annual conference of the Int. Ass. for Mathematical Geology*, pp. 20. CD-ROM.

Bauer, H. (1992). *Maß- und Integrationstheorie, 2 überarb. Auflage*. de Gruyter, Berlin (DE). 260 pp.

Bauer, H. (2002). *Wahrscheinlichkeitstheorie, 5 Auflage*. de Gruyter, Berlin (DE). 520 pp.

Bayer, U., Burger, H. and Skala, W. (Eds.) (2002). *Proceedings of IAMG'02 - The 8th annual conference of the Int. Ass. for Math. Geol.*, Volume I and II. Alfred-Wegener-Stiftung, Berlin (DE), ISSN 0946-8978, 1106 pp.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Pub. 378 pp.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456), 1205-1214.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57, 269-326.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons.

Egozcue, J. J., Díaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22(4), 1175-1182.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279-300.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 36, 210-271.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1-14.

Leonard, T. and Hsu, J. S. J. (1999). *Bayesian Methods: an analysis for statisticians and interdisciplinary researchers*. Cambridge Series in Statistical and Probabilistical Mathematics. New York: Cambridge U. Press. 333 pp.

Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1999). A measure of difference for compositional data based on measures of divergence. In S. J. Lippard, A. Næss, and R. Sinding-Larsen (Eds.), *Proceedings of IAMG'99 - The 5th annual conference of the Int. Ass. for Math. Geol.*, Volume I and II, pp. 211-216. Tapir, Trondheim (N), 784 pp.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Barceló-Vidal, C. (2003). Distributions on the simplex. See Thió-Henestrosa and Martín-Fernández (2003).

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Martín-Fernández, J. A. (2002). Normal in $\Re^+$ vs lognormal in $\Re$. See Bayer *et al.* (2002), pp. 305-310.

Matusita, K. (1955). Decision rules based on the distance for problems of fit, two samples and estimation. *The Annals of Mathematical Statistics*, 26, 631-640.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. See Thió-Henestrosa and Martín-Fernández (2003).

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5), 384-398.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2002). BLU estimators and compositional data. *Mathematical Geology*, 34(3), 259-274.

Robert, C. P. (2001). *The Bayesian Choice. A Decision Theoretic Motivation*. New York, NY (USA): Springer V. 436 pp.

Shao, J. (1999). *Mathematical Statistics*. Springer, New York (USA). 529 pp.

Small, C. G. and Leish, D. L. M. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. New York, NY (USA): Wiley-Interscience. 270 pp.

Thió-Henestrosa, S. and Martín-Fernández, J. A. (Eds.) (2003). *Compositional Data Analysis Workshop - CoDaWork'03, Proceedings*. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Whaba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Soc. for Industrial & Applied Math. 165 pp.

Witting, H. (1985). *Mathematische Statistik I. Parametrische Verfahren bei festem Stichprobenumpfang*. Stuttgart (DE): B. G. Teubner. 538 pp.

Appendix A. Proofs of theorems

Theorem 4

Proof. Perturbation: Since λ is σ-finite, there exists a family A_i, $i = 1,\ldots,n$, of sets increasing to Ω such that $\lambda(A_i) < +\infty$. Since μ and ν are in $\mathscr{M}(\lambda)$, they have $\lambda-$a.e. finite λ-equivalent densities f_μ and f_ν. Choose a version of these densities being everywhere finite and define families of sets $B_i := \{\omega \in \Omega : f_\mu(\omega) < i\}$, $C_i := \{\omega \in \Omega : f_\nu(\omega) < i\}$ increasing to Ω. Furthermore, consider the family of sets $D_i = A_i \cap B_i \cap C_i$; it is also increasing to Ω and

$$(\mu \oplus \nu)(D_i) =_B \int_{D_i} f_\mu f_\nu d\lambda \leq i^2 \lambda(A_i) < +\infty.$$

Thus, $\mu \oplus \nu$ is σ-finite.

Powering: Analogously, consider again the increasing family A_i, as well as the families $B_i := \{\omega \in \Omega : i^{-1} < f_\mu(\omega) < i\}$ and $C_i = A_i \cap B_i$. Then,

$$(\alpha \odot \mu)(A_i \cap B_i) = \int f_\mu^\alpha d\lambda \leq i^{|\alpha|} \lambda(A_i) < +\infty.$$

Thus, $\alpha \odot \mu$ is σ-finite. □

Theorem 5

Proof. According to the definition of Radon-Nikodym derivatives, expressions of $\oplus$ and $\odot$ using measures (1), (2), and using the respective densities, (6), (5), are equivalent. The operations are well defined on the equivalence classes since, for real constants k_1, k_2 and α,

$$(k_1 f_1 \oplus k_2 f_2)(x) = k_1 k_2 (f_1(x) f_2(x)) =_B (f_1 \oplus f_2)(x)\,,$$
$$(\alpha \odot k_1 f)(x) = k_1^\alpha f(x)^\alpha =_B (\alpha \odot f)(x)\,.$$

Linear space axioms follow from straightforward calculations:

- The neutral element is given by $\lambda =_B d\lambda/d\lambda =_B 1$.
- The opposite (negative) element is given by $(\ominus f_\mu) =_B (1/f_\mu) =_B d\lambda/d\mu$. □

Theorem 6

Proof. For measures, the equivalence relation $(=_B)$ does not depend on the reference measure λ; therefore, the quotient set $\mathscr{M}(\lambda)/(=_B)$ is equal to both $B(\mu)$ and $B(\lambda)$. In fact, any measure $\nu \in \mathscr{M}(\lambda)$ is represented in $B(\lambda)$ and $B(\mu)$ by B-equivalent densities; i.e. $\mu = k\lambda$, implies $d\lambda/d\mu = k$, $\lambda-$a.e., and then

$$\frac{d\nu}{d\mu} = \frac{d\nu}{d\lambda}\frac{d\lambda}{d\mu} = k\frac{d\nu}{d\lambda} \quad (\lambda\text{−a.e.})\,,$$

where $\lambda-$a.e. is equivalent to $\mu-$a.e. due to $\mu \in \mathscr{M}(\lambda)$. Therefore, operations $\oplus$ and $\odot$, expressed using densities, give proportional results when expressed in $B(\mu)$ or $B(\lambda)$. □

Theorem 7

Proof. Since $\mu \in \mathscr{M}(\lambda)$, $\mathscr{M}(\mu) = \mathscr{M}(\lambda)$. Furthermore, $(=_B)$-equivalence classes are the same in $\mathscr{M}(\mu)$ and in $\mathscr{M}(\lambda)$, and affine equivalence holds since there exists an affine mapping $g : B(\mu) \to B(\lambda)$, given by $g(\nu) :=_{B(\lambda)} \nu \ominus_\lambda \mu$, which is linear. Using the fact that $\ominus_\lambda \mu = d\lambda/d\mu$, and that any $\nu \in B(\mu)$ has the representation $(d\nu/d\mu)(d\mu/d\lambda)$ in $B(\lambda)$, linearity is given by:

$$g((\alpha \odot_\mu \nu_1) \oplus_\mu \nu_2) =_B g\left(\left(\frac{d\nu_1}{d\mu}\right)^\alpha \frac{d\nu_2}{d\mu}\right) =_B \underbrace{\left(\frac{d\nu_1}{d\mu}\frac{d\mu}{d\lambda}\frac{d\lambda}{d\mu}\right)^\alpha}_{g(\nu_1)} \underbrace{\frac{d\nu_2}{d\mu}\frac{d\mu}{d\lambda}\frac{d\lambda}{d\mu}}_{g(\nu_2)}$$
$$=_B (\alpha \odot_\lambda g(\nu_1)) \oplus_\lambda g(\nu_2)\,,$$

where the subscripts of $\oplus$ and $\odot$ indicate the reference measure of the space where the operation is carried out. □

Theorem 8

Proof.

1. $\mu =_B \nu$ is equivalent to $\mu(\Omega) = k\nu(\Omega)$; therefore, μ, ν are either finite or infinite and then B_P and B_I are well defined and they constitute the whole space.
2. For any densities f, g, in $B_P(\lambda)$ and for any value $0 \leq \alpha \leq 1$, the statement is equivalent to
$$(\alpha \odot f) \oplus ((1-\alpha) \odot g) = \int f^\alpha g^{1-\alpha} d\lambda \leq \int f d\lambda + \int g d\lambda < +\infty \,.$$
3. Boundedness is preserved by arbitrary powering and perturbation with bounded values.
4. The same holds for upper boundedness as long as the exponents are positive.
5. It follows from the inequality $fg < bf$ ($\lambda-$a.e.).
6. It follows from the inequality $f/g < f/b$ ($\lambda-$a.e.).
7. ($\Rightarrow$): $\nu(\Omega)$ does not depend on λ.
($\Leftarrow$): λ and μ are λ-equivalent and then $\mu \in B(\lambda)$.
8. If $\mu \in B_b(\lambda)$, then $b_1^{-1} \leq d\mu/d\lambda < b_1$, and if $\nu \in B_b(\mu)$, then $b_2^{-1} \leq d\nu/d\mu < b_2$; combining both expressions, $(b_1 b_2)^{-1} \leq d\nu/d\lambda = (d\nu/d\mu)(d\mu/d\lambda) \leq b_1 b_2$ and then $\nu \in B(\lambda)$.
9. ($\Rightarrow$): If $\nu \in B_b(\mu)$ with density f, $0 < b^{-1} \leq f \leq b$ and $\int f \, d\mu \leq b\mu(\Omega) < +\infty$.
($\Leftarrow$): $\nu \in B_b(\mu) \subset B(\mu)$ implies $+\infty > \int f d\mu \geq b^{-1}\mu(\Omega)$, then $\mu(\Omega) < +\infty$.
10. Similar to the previous statement. □

Theorem 9

Proof. Let $\mu_{\vec{\boldsymbol{\alpha}}} \in \mathrm{Exp}_B(\lambda, g, \vec{T}, \vec{\boldsymbol{\theta}})$ be a measure. By definition $\mu_{\vec{\boldsymbol{\alpha}}}$ is λ-equivalent and $\mu_{\vec{\boldsymbol{\alpha}}} \in B(\lambda)$. Then, it can be expressed as

$$\mu_{\vec{\boldsymbol{\alpha}}} =_B g \oplus \bigoplus_{j=1}^{k} (\theta_j(\vec{\boldsymbol{\alpha}}) \odot V_j(x)) \,,$$

with $V_j =_B \exp(T_j)$. Therefore, the exponential family corresponds to the affine subspace of $B(\lambda)$

$$g \oplus \mathrm{span}\{V_j, j = 1, \ldots, k\} \,,$$

where the natural parameters $\theta_j(\vec{\boldsymbol{\alpha}})$ are the coordinates of $\mu_{\vec{\boldsymbol{\alpha}}}$ with respect to the basis elements V_j. □

Theorem 10

Proof. Let $g \in S$ be a density and V_j, $j = 1, 2, \ldots, k$, be a basis of the subspace $S \ominus g$. Any element $\mu \in S$ is expressed as $\mu =_B g \oplus \bigoplus_{j=1}^{k} (\alpha_j \odot V_j)$, thus spanning exactly S. Then, $\mu \in \mathrm{Exp}_B(\lambda, g, \ln\vec{V}, \vec{Id})$, with $\ln\vec{V} = (\ln V_1, \ldots, \ln V_k)$ and $\vec{Id}$ the identity mapping. The parametrisation is strict, since the coordinates with respect to a basis are unique. □

Theorem 11

Proof. The statement is proven if, L_{x_i} is a τ-equivalent density of a σ-finite and τ-equivalent measure $P_\theta(x_i)$-a.e. For $\theta \in \Theta$, $L_{x_i} > 0$ since $P_\theta \in B(\lambda)$. Thus, it is τ-equivalent. It is in $B(\tau)$ if it corresponds to a σ-finite measure. To prove that L_{x_i} is a density of a σ-finite measure, consider any finite measure $\tau' \in B_P(\tau)$. If $P(x_i, \theta)$ is the joint probability distribution of X_i and θ constructed from τ' as marginal distribution, then

$$L_{x_i}(\theta) = \frac{dP(x_i, \theta)}{d\tau'(\theta) d\lambda(x_i)},$$

because P_θ is the conditional distribution and τ' plays the role of a marginal distribution for θ. Fubini theorem implies $\int L_{x_i} d\tau < +\infty$ (λ-a.e.), or, equivalently, P_θ-a.e. Then, $L_{x_i} \in B(\tau')$ and represents a finite measure μ_{x_i} (P_θ-a.e.). According to Theorem 7 on shift of origin, from $B(\tau)$ to $B(\tau')$, we get $L_{x_i} =_{B(\tau)} \mu_{x_i} \ominus_\tau \tau'$ and thus $L_{x_i} \in B_P(\tau)$. □

Theorem 13

Proof. The likelihood function can be written

$$\begin{aligned} L_{\vec{x}}(\vec{\boldsymbol{\theta}}) &= C^n(\vec{\boldsymbol{\theta}}) \cdot \prod_{i=1}^{n} g(x_i) \cdot \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{k} \theta_j T_j(x_i) \right) \\ &=_{B(\tau)} C^n(\vec{\boldsymbol{\theta}}) \cdot \exp\left(\sum_{j=1}^{k} \theta_j \left[\sum_{i=1}^{n} T_j(x_i) \right] \right). \end{aligned} \tag{13}$$

If $C(\vec{\boldsymbol{\theta}})$ is in the span of $\exp(\vec{\boldsymbol{\theta}})$, $L_{\vec{x}}(\vec{\boldsymbol{\theta}})$ corresponds to a k-dimensional subspace of $B(\tau)$ with $g^*(\vec{\boldsymbol{\theta}}) =_{B(\tau)} 1$, $\vec{\boldsymbol{\theta}}^* = \vec{\boldsymbol{\theta}}$ and $\vec{T}^* = \vec{T}$. Otherwise, taking $g^*(\vec{\boldsymbol{\theta}}) = 1$, $\vec{\boldsymbol{\theta}}^* = (\ln C(\vec{\boldsymbol{\theta}}), \vec{\boldsymbol{\theta}})$, and $\vec{T}^*(\vec{x}) = (n, \sum_{i=1}^{n} \vec{T}(x_i))$, Eq. $L_{\vec{x}}(\vec{\boldsymbol{\theta}})$ corresponds to a $(k+1)$-dimensional subspace of $B(\tau)$. In both cases, Theorem 9 implies the statement. □

Theorem 14

Proof. The family $\mathrm{Exp}_{B(\tau)}(\tau, g^*, \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$ is a subspace of $B(\tau)$ because $g^* =_B 1$. Since subspaces are invariant under perturbation of elements of the subspace, the posterior $P_{post}(\vec{\boldsymbol{\theta}})$ is in the subspace. □

Theorem 15

Proof. The likelihood $L_{\vec{x}}(\vec{\boldsymbol{\theta}})$, as a function of $\vec{\boldsymbol{\theta}}$, is in the extended exponential family $\mathrm{Exp}_{B(\tau)}(\tau, g^*, \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$ that has been identified as a subspace of $B(\tau)$. Application of Bayes theorem is a perturbation, i.e. a shifting, and the result is the affine space $\mathrm{Exp}_{B(\tau)}(\tau, P_{prior}(\vec{\boldsymbol{\theta}}), \vec{T}^*, \vec{\boldsymbol{\theta}}^*)$, where the origin coincides with P_{prior} because $g^* =_{B(\tau)} 1$. □

Theorem 16

Proof. The posterior density in the extended exponential family is expressed as

$$P_{post}(\vec{\boldsymbol{\theta}}) =_{B(\tau)} h \oplus \exp\left(\sum_{j=1}^{k} S_j(\vec{x})\theta_j\right) .$$

Combining this expression with the Bayes formula, the likelihood function is

$$L_{\vec{x}}(\vec{\boldsymbol{\theta}}) =_{B(\tau)} (h \ominus P_{prior}(\vec{\boldsymbol{\theta}})) \oplus \exp\left(\sum_{j=1}^{k} S_j(\vec{x})\theta_j\right) .$$

In $B(\lambda)$ it can be rewritten as

$$L_{\vec{x}}(\vec{\boldsymbol{\theta}}) =_{B(\lambda)} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{k} T_j(x_i)\theta_j\right) ,$$

where $S_j(\vec{x}) = \sum_{i=1}^{n} T_j(x_i)$. The existence of the statistics T_j comes from the multiplicative form of the likelihood function and the fact that the expression should be valid for any arbitrary n. Therefore,

$$L_x(\vec{\boldsymbol{\theta}}) =_{B(\lambda)} 1 \cdot \exp\left(\sum_{j=1}^{k} T_j(x)\theta_j\right) ,$$

where the perturbation of k terms may collapse in $k_1 \leq k$ terms for equal T_j's. □