On Coarse Graining of Information and Its Application to Pattern Recognition

Ali Ghaderi

Telemark University College, Kjølnes Ring 56, NO-3918 Porsgrunn, Norway

Abstract. We propose a method based on finite mixture models for classifying a set of observations into number of different categories. In order to demonstrate the method, we show how the component densities for the mixture model can be derived by using the maximum entropy method in conjunction with conservation of Pythagorean means. Several examples of distributions belonging to the Pythagorean family are derived. A discussion on estimation of model parameters and the number of categories is also given.

Keywords: Mixture probability, Bayesian, Maximum entropy, Histogram, Clustering, Extensive property, Intensive property, Conservation laws, Pythagorean means, Pythagorean family of distributions.

PACS: 02.50.Cw,02.50.Tt

INTRODUCTION

One of the goals of any scientific study is to identify regularities in observations and classify them into possibly separate and simpler structures or categories. These categories can in turn be used to make inferences on the objects of interest. The major advantage of this approach is that one breaks down a complicated reality into a collection of simpler structures. In a similar way, in pattern recognition one is concern with discovery of regularities in data but through use of computer algorithms which can be used to classify the data into different categories [1]. Independent of ones point of view, any such analysis must start with definition of the categories. If one has sufficient information about the categories and their members, it is an easy task to establish a precise definition. However, for most real life situations this is not the case and the notion of category cannot be precisely defined. Under such conditions a fruitful approach is to consider a category as collection of objects which are likely to share the same properties. That is, in cases for which the information available is insufficient to reach certainty, we ought to quantify the degree to which we believe an object belongs to a given category. This degree of belief is described by probability distribution over the space of objects of interest, or sample space to be more precis.

The major bulk of the literature on the subject is dedicated to numerical aspect of the problem. While acknowledging that the numerical challenges can seriously compromise the applicability of a method, we believe that the fundamental problem of modelling categories under partial knowledge condition is just as important. In the following we will look at a class of problems in pattern recognition for which one is in possession of empirical distribution (histogram) over the objects of interests and a prior knowledge

on the number of categories involved. We propose an approach to modelling of the empirical distributions based on the finite mixture models which relies on identifying the relevant intensive properties of each category. In order to demonstrate this method, we will show how conservation of Pythagorean means, the most encountered class of intensive properties, in conjunction with maximum entropy method can be used to derive the functional form of the mixture model. We will also briefly discuss the extension to other conserved quantities and also give a short overview on numerical challenges related to the inference problem. In this article we restrict ourselves to positive univariate continuous quantities.

MIXTURE MODEL

In the situations where categories cannot be defined precisely, the probabilistic description might be the only possible option. In the probabilistic framework, we can only talk about the likelihood of an object belonging to a category. To this end, let us assume that by some experiment the observation X is made but it is not by itself sufficient to uniquely determine which category it belongs to. For example, the observations can be the height of people in certain region/country for which underlying categories are the age groups that each individual might belong to. In such cases one considers X as a random variable and tries to model its probability density function p. One approach to model p is based on the so called *finite mixture models* [2]. The underlying assumption in this approach is that p is a *convex combination* of k densities in which each density represents a single category. That is

$$p\left(x|\,\psi\right)=\sum_{j=1}^{k}\pi_j\left(\theta_j\right)f_j\left(x|\,\theta_j\right),\ x\in\mathscr{X} \tag{1}$$

where

$$\sum_{j=1}^{k}\pi_j\left(\theta_j\right)=1,\ \pi_j\geq 0 \tag{2}$$

and

$$\int_{\mathscr{X}}f_j\left(x|\,\theta_j\right)dx=1,\ f_j\left(x|\,\theta_j\right)\geq 0 \tag{3}$$

and

$$\psi=(\pi,\theta)=\left(\{\pi_1,\ldots,\pi_k\},\{\theta_1,\ldots,\theta_k\}\right). \tag{4}$$

In such cases, one says that X has a finite mixture distribution and that p is a finite mixture density function. The parameters π_j are called *mixing weights* and f_j the *component densities* of the mixture. In the context of pattern recognition, k is the number of categories and f_j is the density function describing the distribution of the members of the category j. It should be emphasis that the component densities do not necessarily belong to the same family of densities. Each component density represents our best guess about the structure of its respective category for which its existence is independent of the other categories.

In order to be able to adopt the mixture model to a specific problem, given that a priori one knows the number of categories, requires that one tackles two different problems.

The first problem is to determine how to achieve a quantitative description of state of partial knowledge, i.e. determining the functional form of the component densities. The second problem is to determine ψ based on the available evidence, i.e. the empirical density.

DETERMINATION OF COMPONENT DENSITIES

In general, objects in the same category are more similar to each other than to those in other categories. This similarity invokes the notion that there are properties at the coarser level which distinguishes the categories from each other. In fact, if we consider a category as a homogeneous[1] group in which the members are recognizably similar, then it is reasonable to assume that the properties that distinguish it from other categories should be intrinsic and independent of the *coarse graining* within the category itself. This coarse graining property is the key concept in finding the component distributions.

In general, coarse graining is achieved by first grouping the elements of the category into blocks, each having the same volume. Then following a predetermined rule, each block is replaced with a single element representing the elements of that block. This procedure is iterated *ad infinitum*. We call a property that is invariant under coarse graining as *intensive*. In this context, a category can be characterized and distinguished from others by its intensive properties. Identifying the relevant intensive properties are often challenging. Usually a less challenging approach is to first determine the so-called *extensive* properties of the category. An extensive property is a property that is additive under coarse graining. That is, under coarse graining, the elements that replace the blocks at each step, also inherit the sum of each of extensive properties of their respective block elements. Moreover, at each coarse graining step, due to similarity and homogeneity conditions which exist among the members of a category, the extensive properties scale independent of the choice of specific block. This, in general, results in greatly reducing the complexity of the analysis. However, it is conceivable that one might discover many extensive properties which might not be relevant to the classification problem at hand. In this respect, the choice of relevant properties are often problem dependent. Nevertheless, identifying and describing an extensive property means that one is able to find a function, up to a scaling factor, which captures the essential features of that property. It can be shown that the expectation of such a function is invariant with respect to coarse graining and hence it is intensive. For example, particle mass is an extensive property of a system consisting of a collection of particles. Whilst, the expected mass of a particle is intensive. In the following, we shall call the intensive properties that are expressed in the form of expectations as the *conservation laws*[2].

---

[1] Homogeneous in the sense that there is continuity between various members of the group.

[2] We adopt the view held by Steiner [3] that laws of conservation are simply not causal laws. They provide constraints on what is allowed to happen.

Conservation of Pythagorean means

Let $g(x)$ denote a function representing an extensive property of a category. Up to a scaling factor, some of the most encountered forms of g are

$$g(x) = x,\ g(x) = x^{-1},\ g(x) = \ln x. \tag{5}$$

The expected values of these functions constitute the so-called *Pythagorean means*. The Pythagorean means are the arithmetic, geometric and harmonic mean. More precisely, for a positive univariate continuous variable with density f, the Pythagorean means are defined as

$$\mu = \int_0^\infty x f(x)\, dx \tag{6a}$$

$$\ln \gamma = \int_0^\infty \ln x f(x)\, dx \tag{6b}$$

$$\eta^{-1} = \int_0^\infty x^{-1} f(x)\, dx \tag{6c}$$

where μ, γ and η are the arithmetic, geometric and harmonic means, respectively. In this regard, one can talk about two categories being similar with respect to some of the Pythagorean means.

Maximum entropy

Although the conserved quantities restrict the possible distribution of elements in a category, nonetheless, still there might be up to infinitely many distributions that satisfy the constraints. We are interested in the distribution that conserves the quantities of the interest while allowing maximum degree of freedom on the non-conserved quantities. It can be shown that among all the distributions that fulfill the constraints, the most uncommitted distribution is the one with largest relative entropy S

$$S[f,q] = -\int f(x) \ln \frac{f(x)}{q(x)} dx \tag{7}$$

where f is the unknown distribution and q, also known as the *prior*, defines what we mean by the uniform distribution in the sample space $\mathscr{X}$ [4, 5]. This method of finding a distribution is known as *maximum entropy* or in short *MaxEnt*.

Theorem 1 *Let all the three Pythagorean means be conserved. Then the MaxEnt distribution is*

$$f(x;\lambda_1,\lambda_2,\lambda_3) = \frac{q(x)}{Z_q(\lambda_1,\lambda_2,\lambda_3)} x^{\lambda_3-1} \exp\left(-\lambda_1 x - \lambda_2 x^{-1}\right) \tag{8}$$

where

$$Z_q(\lambda_1,\lambda_2,\lambda_3) = e^{\lambda_0} = \int_{\mathscr{X}} q(x) x^{\lambda_3-1} \exp\left(-\lambda_1 x - \lambda_2 x^{-1}\right) dx \tag{9}$$

is the partition function*, which acts as normalization factor.*

Proof. *This is equivalent to finding the maximum of the Lagrangian L with respect to* f

$$\begin{aligned} L[f] = &-\int_{\mathscr{X}} f(x)\ln\frac{f(x)}{q(x)}dx-(\lambda_0-1)\left(\int_{\mathscr{X}} f(x)\,dx-1\right) \\ &-\lambda_1\left(\int_{\mathscr{X}} xf(x)\,dx-\mu\right)-\lambda_2\left(\int_{\mathscr{X}} x^{-1}f(x)\,dx-\eta^{-1}\right) \\ &-(1-\lambda_3)\left(\int_{\mathscr{X}}\ln xf(x)\,dx-\ln\gamma\right) \end{aligned} \tag{10}$$

where $\lambda_0\ldots\lambda_3$ *are the four Lagrange multipliers corresponding to the four constraints*[3]*. It can be shown that maximizing the functional L is equivalent to solving the corresponding* Euler-Lagrange *equation of the calculus of variations [6] which results in statement of the theorem.* ■

We shall say a distribution that share the same functional form as (8) belongs to *Pythagorean family of distributions*. Note that q can be even improper with non-compact support as long as the distribution in (8) is normalizable. If we know, up to a normalization constant, the functional form of the prior q and the values of the Pythagorean means then f can be uniquely determined. Moreover, note that if q is very narrow then $f\approx q$, whilst if q is very broad then its influence is negligible and can be considered to be the uniform distribution. The following corollary is a direct consequence of Eq. (8).

Corollary 2 *Let q be the improper uniform distribution on the positive real line. Then for all* $x\in\mathbb{R}^+$

$$f(x;\alpha,\beta,\lambda)=\frac{1}{2\alpha K_\lambda(\beta)}\left(\frac{x}{\alpha}\right)^{\lambda-1}\exp\left\{-\frac{\beta}{2}\left(\frac{x}{\alpha}+\frac{\alpha}{x}\right)\right\},\lambda\in\mathbb{R},\alpha>0,\beta>0 \tag{11}$$

where K_λ *is the modified Bessel function of the second kind and*

$$\lambda=\lambda_3,\alpha=\sqrt{\frac{\lambda_2}{\lambda_1}},\beta=2\sqrt{\lambda_1\lambda_2}. \tag{12}$$

In literature the distribution (11) is known as *generalized inverse Gaussian (GIG)* distribution [7]. Some of its well-known sub-classes are the *inverse Gaussian (IG)* ($\lambda=-1/2$), the *reciprocal inverse Gaussian (RIG)* ($\lambda=-1/2$) and the *hyperbolic (H)* ($\lambda=0$) distributions (see Fig.1). Other familiar distributions arise when only some of the Pythagorean means are conserved. For example, if one drops the constraint on arithmetic mean, that is $\lambda_1=0$ in (10), the distribution is known as *inverse gamma* distribution. If the constraint on the harmonic mean is dropped, that is $\lambda_2=0$ in (10), the distribution is the *gamma* distribution. The list is longer than this but the above examples demonstrate the abundance of different variety of distributions belonging to Pythagorean family.

---

[3] Note that λ_0-1 and $1-\lambda_3$ are used instead of λ_0 and λ_3 as a matter of convenience.

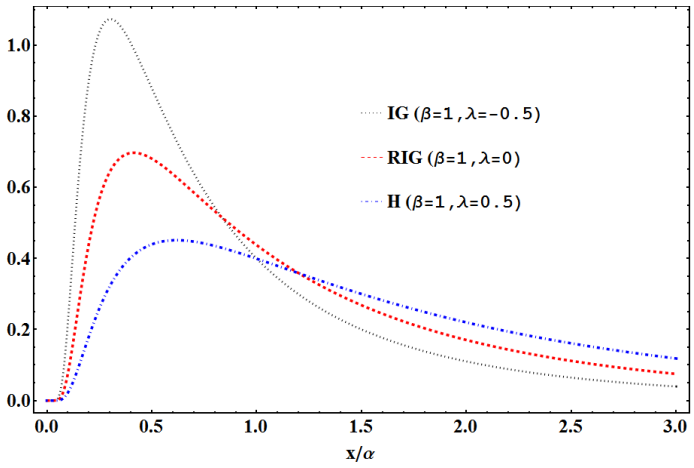

FIGURE 1. Three of the known sub-classes of the generlized inverse Gaussian distribution (GIG).

Other conservation laws

For the sake of clarity we narrowed the discussions to the conservation of Pythagorean means. But other conservation laws are possible and are even at use. The MaxEnt method can handle other conserved quantities as well. Nonetheless, it is recommended that one should always conduct an assessment on conservation of Pythagorean means at the start of the analysis. The outcome can be used as prior q in (7) along with other conserved quantities to derive the functional form of the component densities.

BAYESIAN INFERENCE

We have not touched the numerical aspect of this problem. It is often case dependent and difficult to discuss without getting into the specifics. However, the statement of the most important problems using the rules of probability is quite simple.

Determination of model parameters ψ

Let I summarize the information about the functional form of the component densities and their number. Technically, once I is known, determining ψ in (1) becomes a standard problem in statistical inference. To this end, assume that the observations are randomly generated from $p(x|\psi,I)$. Then the normalized histogram of the data, say $h(x)$, can be considered as the empirical estimate for $p(x|\psi,I)$. Consequently, the unknown ψ can be estimated from h by using the Bayesian methods. Indeed, it follows from *Bayes rule* that

$$p(\psi|h,I) \propto p(\psi|I)\, p(h|\psi,I) \tag{13}$$

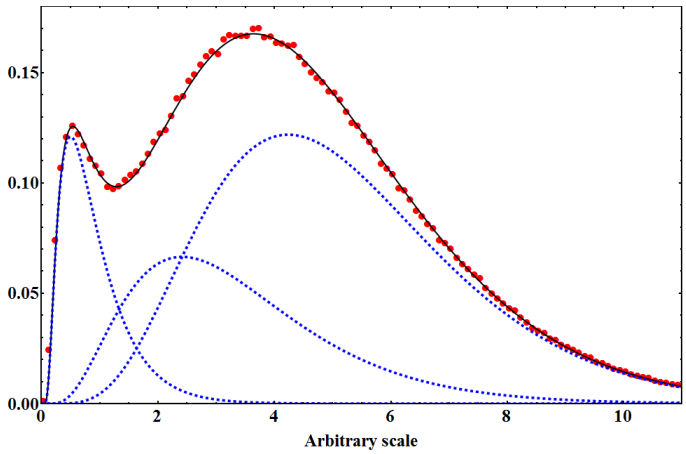

FIGURE 2. The dots represent the histogram of 50000 numbers, simulated from the mixture of three GIG-variates. The dashed curves are the component densities of the variates times their respective mixing weights. The sum of the three dashed curves is the mixture model in red.

where $p(\psi|I)$ is the prior for ψ. Usually we just have some rough knowledge about the domain of ψ and therefore it is common to assume that $p(\psi|I)$ is uniformly distributed over that domain. The likelihood function $p(h|\psi,I)$ depends on our assessment of the sources that contribute to deviation between the model and data and, in general, is problem specific [8, 9]. The most likely estimate for ψ is the one which coincides with the global maximum of $p(\psi|h,I)$ which is called *maximum a posteriori probability (MAP) estimate*. Usually, due to intractability of analytical form of the posterior distribution the methods for estimating MAP are *Monte Carlo* based [10]. For illustration purpose, in Fig. 2, we have plotted an example of a mixture model and its three component densities versus their joint simulated histogram.

Determination of number of categories k

In the above discussions, we assumed that the number of categories are known. However, often we do not know this number and we need to estimate it. In the Bayesian framework this is known as *model selection problem*. Indeed, in order to estimate the number of categories we need to evaluate the posterior distribution for k conditional on h. By the Bayes rule we have

$$p\left(k|h,I'\right) \propto p\left(k|I'\right) p\left(h|k,I'\right) \tag{14}$$

where I' summarize the information about the functional form of the component densities. Note that $I = (k,I')$. Now, by *marginalization* and *product rule* we have

$$p\left(h|k,I'\right) = \int_{\Psi} p\left(h,\psi|k,I'\right) d\psi = \int_{\Psi} p(\psi|I)\, p(h|\psi,I)\, d\psi \tag{15}$$

and hence

$$p\left(k|h,I'\right) \propto p\left(k|I'\right) \int_{\Psi} p(\psi|I)\, p(h|\psi,I)\, d\psi. \tag{16}$$

The integral on the right hand side of (16) is known as *evidence* and is equal to normalization factor on the right hand side of (13). If one assumes $p(k|I')$ to be uniform then the most probable value of k is the one which corresponds to the model with largest evidence. It is often quite challenging to get a good estimate of evidence. Most methods are Monte Carlo based and have their own pros and cons. Therefore, the choice of the method is very much application dependent. It is not uncommon that one uses several different methods in order to find a good estimate. For an overview over the most used methods the reader is referred to [11].

CONCLUSION

In situations where we have partial knowledge about the categories, the probabilistic description based on the finite mixture model is a possible approach. In order to determine the component densities of the model one can start with finding the relevant extensive properties of each category under coarse graining. Taking the expectation of these extensive properties will lead to the right conservation laws and in conjunction with MaxEnt to component densities. Then the model parameters can be estimated from the empirical density data using the standard Bayesian methods.

ACKNOWLEDGMENTS

I would like to thank Nicholas Armstrong for insightful advice on calculation of evidence.

REFERENCES

1. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, 2006.
2. D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons Ltd, 1985.
3. M. Steiner, *Noûs* **12**, 17–28 (1978).
4. A. Caticha, Entropic inference and the foundations of physics, The Brazilian chapter of the international society for Bayesian analysis (ISBrA), Sao Paulo, Brazil (2012), URL `http://www.albany.edu/physics/ACaticha-EIFP-book.pdf`, invited monograph.
5. D. S. Sivia, *Data Analysis: A Bayesian Tutorial*, Clarendon Press, Oxford, 1996.
6. G. B. Arfken, and H. J. Weber, *Mathematical Methods for Physicists*, Harcourt/Academic Press, 2001, fifth edn.
7. B. Jørgensen, *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Lecture notes in statistics, Springer-Verlag, 1982.
8. P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach With Mathematica® Support*, Cambridge University Press, Cambridge, 2005, first edn.
9. W. von der Linden, V. Dose, and U. von Toussaint, *Bayesian probability theory, applications in physical sciences*, Cambridge, 2014, first edn.
10. C. P. Robert, and G. Casella, *Monte Carlo Statistical Methods*, Springer Science+BusineSpringer, LLC, 2004, second edn.
11. N. Friel, and J. Wyse, *Statistica Neerlandica* **66**, 288–308 (2012).