MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

On the Interpretation of $\mathcal{I}$-Divergence-Based Distribution-Fitting as a Maximum-Likelihood Estimation Problem

Jonathan LE ROUX, Hirokazu KAMEOKA,
Nobutaka ONO and Shigeki SAGAYAMA

(Communicated by Akimichi TAKEMURA)

METR 2008–11 March 2008

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

On the Interpretation of $\mathcal{I}$-Divergence-Based Distribution-Fitting as a Maximum-Likelihood Estimation Problem

Jonathan LE ROUX, Hirokazu KAMEOKA,
Nobutaka ONO and Shigeki SAGAYAMA

Department of Information Physics and Computing
Graduate School of Information Science and Technology
The University of Tokyo
leroux@hil.t.u-tokyo.ac.jp, kameoka@eye.brl.ntt.co.jp,
onono@hil.t.u-tokyo.ac.jp, sagayama@hil.t.u-tokyo.ac.jp

March 2008

Abstract

We investigate the mis-match in the classical interpretation of certain distribution-fitting problems as maximum-likelihood (ML) estimation problems in the particular case of the $\mathcal{I}$-divergence. The general relation between Bregman divergences and exponential families shown by Banerjee *et al.* [8] enables to consider distribution-fitting problems based on Bregman divergences as ML problems based on the corresponding exponential family. This interpretation is however only valid if the data is included in the support of the exponential family. This is the case for the $\mathcal{I}$-divergence, which is associated to the Poisson distribution, when applied to real-valued data. We explain more precisely the reason for this mis-match, and derive an asymptotically justifiable alternative to the usual workaround consisting in quantizing the data, by using the Gamma function as a normalization term.

1 Introduction

The possibility to interpret some distribution-fitting problems as maximum-likelihood (ML) problems is an important concept in signal processing and machine learning algorithms. It is well-known for example that the fitting on a domain D of a model $Q(x, \boldsymbol{\Theta})$ with parameters $\boldsymbol{\Theta}$ to observed data $W(x)$ based on the least-squared error can be interpreted as the ML estimation of the parameters $\boldsymbol{\Theta}$ of the model, assuming that the observation data points $W(x)$ are independently generated from a Gaussian distribution with mean

$Q(x, \boldsymbol{\Theta})$ and fixed variance. The parameters $\boldsymbol{\Theta}$ minimizing $\int_D ||Q(x, \boldsymbol{\Theta}) - W(x)||^2 dx$ and maximizing $\prod_{x \in D} \frac{1}{\sqrt{2\pi}c} \exp - \frac{||Q(x,\boldsymbol{\Theta}) - W(x)||^2}{2c^2}$ for any positive constant c (with an abuse of notation for the product sign) are indeed easily seen to be the same. This bridge between the two theories is particularly interesting as it justifies the use of penalty functions on the parameters as prior functions in a Maximum *a Posteriori* (MAP) framework, or more generally enables the use of Bayesian techniques in parameter estimation problems.

In a wide range of signal processing problems, such as audio signal processing [1, 2], linear inverse problems [3], deblurring [4] or problems relying on Non-negative Matrix Factorization (NMF) techniques [5], the distributions considered are intrinsically non-negative. Csiszár showed in [6] that in such situations, the only choice of discrepancy measure consistent with certain fundamental axioms such as locality, regularity and composition-consistency is the so-called $\mathcal{I}$-divergence [7]. The question of the possibility to interpret distribution-fitting problems based on the $\mathcal{I}$-divergence as ML estimation problems is thus very important.

Banerjee et al. [8] showed that there is a general one-to-one correspondence between a certain type of Bregman divergences, of which the $\mathcal{I}$-divergence is a particular case, and a certain type of exponential families, which are families of probability distributions including many common probability distribution families such as the Gamma, Gaussian, Binomial or Poisson distributions. This result justifies a general link between distribution-fitting using Bregman divergences and ML estimation using the corresponding exponential family. However, as noted in [8], such a relation is only useful for the instances which can be drawn from the exponential distribution, and the set of all these instances may be a strict subset of the domain of definition of the Bregman divergence.

In other words, for some divergence measures, the interpretation of the distribution-fitting problem as an ML estimation one may not be straightforward. This is actually the case for the $\mathcal{I}$-divergence, which is well-defined for real values of the distributions considered, but can be shown to correspond to ML estimation based on the Poisson distribution, which is only defined on the integers. Surprisingly, to our knowledge this has never been clearly stated in the literature, the usual workaround being to quantize and scale the data to get back to the discrete Poisson distribution [3, 4].

We investigate here this mis-match between the distribution-fitting and the ML estimation problems in the particular case of the $\mathcal{I}$-divergence. We explain why the bridge between the two can not be perfectly filled, i.e., why it is sometimes impossible to interpret in full generality a distribution-fitting problem as an ML estimation one, and we derive a theoretical workaround which completes the practical quantization-based one. The goal of this report is simultaneously to clarify and attract the attention of the community on this easily over-looked but nonetheless important problem.

We first introduce the framework in Section 2, then give in Section 3 an insight on the reason for the non-existence of a normalization term giving the desired result. We finally show in Section 4 that the Gamma function can be used as a normalization term which asymptotically leads to the interpretation we are looking for.

2 Presentation of the framework

We consider a non-negative distribution W and a non-negative model $Q(\cdot, \boldsymbol{\Theta})$ parameterized by $\boldsymbol{\Theta}$, defined on a domain D. The $\mathcal{I}$-divergence [7] is a classical way to measure the "distance" between two such non-negative distributions:

$$\mathcal{I}(W|Q(\boldsymbol{\Theta})) \triangleq \int_D \left(W(x) \log \frac{W(x)}{Q(x;\boldsymbol{\Theta})} - \Big(W(x) - Q(x;\boldsymbol{\Theta}) \Big) \right) dx. \tag{1}$$

Distribution-fitting based on the $\mathcal{I}$-divergence amounts to looking for $\boldsymbol{\Theta}_{\text{opt}} = \text{argmin}_{\boldsymbol{\Theta}} \mathcal{I}(W|Q(\boldsymbol{\Theta}))$. Keeping only the terms depending on $\boldsymbol{\Theta}$ and reversing the sign of this expression, one defines the following function to maximize w.r.t. $\boldsymbol{\Theta}$:

$$\mathcal{J}(W, \boldsymbol{\Theta}) = \int_D \Big(W(x) \log Q(x;\boldsymbol{\Theta}) - Q(x;\boldsymbol{\Theta}) \Big) dx. \tag{2}$$

One would like to find a family of probability distributions with parameter $Q(x, \boldsymbol{\Theta})$, such that the corresponding likelihood for $\boldsymbol{\Theta}$, defined as the joint probability of all the variables $W(x)$ independently following the distribution of parameter $Q(x, \boldsymbol{\Theta})$, depends on $\boldsymbol{\Theta}$ only through $\mathcal{J}(W, \boldsymbol{\Theta})$. Remembering the case of least squares estimation and Gaussian distributions, we would like to define the log-likelihood of $\boldsymbol{\Theta}$ as $\mathcal{J}(W, \boldsymbol{\Theta})$ itself, up to a constant which only depends on the data:

$$\log P(W|\boldsymbol{\Theta}) \triangleq \mathcal{J}(W, \boldsymbol{\Theta}) + \int_D \log f(W(x)) dx. \tag{3}$$

Maximization of the log-likelihood of $\boldsymbol{\Theta}$ and maximization of $\mathcal{J}(W, \boldsymbol{\Theta})$ would then be equivalent. We thus need to look for a function f such that this indeed defines a probability measure with respect to W. The point here is that, for the equality (3) to be useful, the function f needs to be positive on the values taken by the data, as both terms of the equality would otherwise be equal to $-\infty$, thus hiding the contribution of the $\mathcal{I}$-divergence. The corresponding distribution density on $[0, +\infty)$, with parameter θ, is then

$$\mu_{f,\theta}(z) = e^{z \log \theta - \theta} f(z) = \theta^z e^{-\theta} f(z), \forall z \in [0, +\infty), \tag{4}$$

which needs to be a probability distribution density for any θ:

$$\forall\theta, \int_0^{+\infty} \theta^x e^{-\theta} f(x)dx = 1. \tag{5}$$

3 Non-existence of a continuous normalization

3.1 Relation with the Laplace transform

We show here that there is no solution to this problem with real-valued data, in the sense that there indeed exists a unique non-negative measure ν on $\mathbb{R}^+$ such that

$$\forall\theta, \int_0^{+\infty} \theta^x e^{-\theta} \nu(dx) = 1, \tag{6}$$

but it is supported by $\mathbb{N} = \{0, 1, 2, \dots\}$. This measure leads to none other than the discrete Poisson distribution. In the following, we thus look for a non-negative measure ν satisfying Eq. (6).

If we rewrite Eq. (6) with $\mu = \log\theta$, our problem amounts to looking for ν such that

$$\forall\mu, \int_0^{+\infty} e^{x\mu} d\nu(x) = e^{e^\mu}, \tag{7}$$

i.e., to looking for a measure ν whose Laplace transform is $\mu \mapsto e^{e^\mu}$.

A direct computation gives the Laplace transform of the Poisson distribution $p(\cdot, \theta) = \sum_{k\in\mathbb{N}} \frac{\theta^k e^{-\theta}}{k!}\delta_k(\cdot)$ of parameter θ with $k \in \mathbb{N}$,

$$\int_0^{+\infty} e^{x\mu} p(x,\theta)dx = e^{-\theta} \sum_{k\in\mathbb{N}} \frac{e^{kt}\theta^k}{k!} = e^{-\theta} e^{\theta e^t}. \tag{8}$$

Up to the constant factor e^{-1}, the Poisson distribution with mean parameter 1 is thus a solution to (6), and conversely any solution to (6) must have (up to a constant factor) the same Laplace transform as the Poisson distribution of mean 1. But this Laplace transform is holomorphic in a neighborhood of 0 (it is in fact an entire function, holomorphic on the whole complex plane), and thus, as shown in [9] (Chap. 30), determines the measure it is associated with. In other words, a measure with such a Laplace transform is unique, which shows that the unique probability distribution satisfying Eq. (6) is $\nu = e\, p(\cdot, 1)$, leading for $\mu_{\nu,\theta}$ to none other than the classical Poisson distribution family $p(\cdot, \theta)$, which is supported by $\mathbb{N}$.

3.2 Consequences on the interpretation of $\mathcal{I}$-divergence-based fitting as an ML estimation problem

As there is no function taking positive values on $[0; +\infty)$ which verifies (5), we cannot directly interpret $\mathcal{I}$-divergence-based distribution-fitting problems with real-valued data as ML estimation problem. If nothing is done, this means that all non-integer data points will have zero likelihood, even if the model fits them perfectly. The usual workaround [3,4] is to quantize the data and to perform a suitable scaling of the data and the model so as to act as if the data were integer, justified by the fact that computers quantize the data anyway. Quantization is a practical justification, but it may seem rather disappointing and inelegant, as it looses the continuous quality of the problem. We derive in the following section a theoretical justification which retains more of the continuity of the problem on real-valued data.

4 Asymptotically satisfactory normalization using the Gamma function

By analogy with the discrete case, for which for any $\theta \in \mathbb{R}^+$, $\frac{\theta^n e^{-\theta}}{\Gamma(1+n)}$ is a probability density distribution on $n \in \mathbb{N}$ called the Poisson distribution, we consider the distribution of the variable $x \in [0, +\infty)$ with parameter θ,

$$f_\theta(x) = \frac{\theta^x e^{-\theta}}{\Gamma(1+x)}, \tag{9}$$

where Γ is the Gamma function. Note that if we reverse the roles of x and θ, it is nothing else than the Gamma distribution.

This distribution is unfortunately not a probability distribution (it could not be, as shown in Section 3), and needs a normalizing constant. Let us consider the evolution of this normalizing constant with respect to the parameter θ. We denote by

$$g(\theta) = \int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \tag{10}$$

the mass of the distribution f_θ and by

$$h(\theta) = \int_0^{+\infty} \frac{x\theta^x e^{-\theta}}{\Gamma(1+x)} dx \tag{11}$$

its first-order moment.

4.1 Limit of g at the origin

We notice that $\forall \eta > 0$,

$$\int_0^\eta \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \leq \int_0^\eta \frac{1}{\Gamma(1+x)} dx \underset{\eta \to 0}{\longrightarrow} 0,$$

and $\forall \theta < 1, \forall \eta > 0$,

$$\int_\eta^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \leq \theta^\eta \int_0^{+\infty} \frac{e^{-\theta}}{\Gamma(1+x)} dx \leq C\theta^\eta \underset{\theta \to 0}{\longrightarrow} 0.$$

Thus, for any $\epsilon > 0$, by choosing η_0 such that

$$\forall \eta < \eta_0, \int_0^{\eta} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \leq \epsilon$$

and then θ_0 such that

$$\forall \theta < \theta_0, \int_{\eta_0}^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \leq \epsilon,$$

we show that

$$g(\theta) \underset{\theta \to 0}{\longrightarrow} 0. \tag{12}$$

4.2 Rewriting $g(\theta)$

As we can write

$$\forall \zeta > 0, \forall \theta > 0,\ g(\theta) = g(\zeta) + \int_{\zeta}^{\theta} g'(t) dt, \tag{13}$$

we look more closely at the derivative of g w.r.t. θ:

$$\begin{aligned} g'(\theta) &= \int_0^{+\infty} \frac{x\theta^{x-1}e^{-\theta} - \theta^x e^{-\theta}}{\Gamma(1+x)} dx \\ &= \int_0^{+\infty} \frac{x\theta^{x-1}e^{-\theta}}{\Gamma(1+x)} dx - \int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \\ &= \int_{-1}^{+\infty} \frac{\theta^u e^{-\theta}}{\Gamma(1+u)} du - g(\theta) \\ &= \int_{-1}^{0} \frac{\theta^u e^{-\theta}}{\Gamma(1+u)} du. \end{aligned} \tag{14}$$

By using the definition of the Gamma function

$$\Gamma(z) = \int_0^{+\infty} e^{-t} t^{z-1} dt$$

and Euler's reflection formula

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin \pi z},$$

we can perform the following derivation:

$$\begin{aligned} g'(t) &= \int_0^{1} \frac{t^{-u}e^{-t}}{\Gamma(1-u)} du \\ &= \int_0^{1} \frac{\sin \pi u}{\pi} t^{-u} e^{-t} \Gamma(u) du \\ &= \int_0^{1} \frac{\sin \pi u}{\pi} t^{-u} e^{-t} \int_0^{+\infty} e^{-s} s^{u-1} ds\, du \\ &= \int_0^{+\infty} \frac{e^{-s}e^{-t}}{\pi s} \left(\int_0^{1} \left(\frac{s}{t}\right)^u \sin(\pi u) du \right) ds. \end{aligned} \tag{15}$$

The inside integral can be analytically computed. If we note $\alpha = \log \frac{s}{t}$, then

$$\begin{aligned}
\int_0^1 \left(\frac{s}{t}\right)^u \sin(\pi u) du &= \int_0^1 \frac{e^{(\alpha+i\pi)v} - e^{(\alpha-i\pi)v}}{2i} du \\
&= -\frac{1}{2i}\left(\frac{1}{\alpha+i\pi} - \frac{1}{\alpha-i\pi}\right)(1+e^{\alpha}) \\
&= \pi \frac{1+e^{\alpha}}{\pi^2+\alpha^2} \\
&= \pi \frac{1+\frac{s}{t}}{\pi^2+(\log\frac{s}{t})^2}.
\end{aligned} \tag{16}$$

We get

$$g'(t) = \int_0^{+\infty} e^{-s} e^{-t} \frac{\frac{1}{s}+\frac{1}{t}}{\pi^2+(\log t - \log s)^2} ds. \tag{17}$$

Altogether, for $\theta > 0$ and $\zeta > 0$,

$$g(\theta) = g(\zeta) + \int_\zeta^\theta \int_0^{+\infty} e^{-t} e^{-s} \frac{\frac{1}{t}+\frac{1}{s}}{\pi^2+(\log t - \log s)^2} ds\, dt. \tag{18}$$

By letting ζ go to 0 in this last expression, we deduce from Eq. (12) another expression for $g(\theta)$:

$$g(\theta) = \int_0^\theta \int_0^{+\infty} e^{-t} e^{-s} \frac{\frac{1}{t}+\frac{1}{s}}{\pi^2+(\log t - \log s)^2} ds\, dt. \tag{19}$$

We can further simplify Eq. (19). We perform a change of variables $u = \log s/t$,

$$\begin{aligned}
g(\theta) &= \int_0^\theta \int_{-\infty}^{+\infty} e^{-te^u} e^{-t} \frac{1+e^u}{\pi^2+u^2} du\, dt \\
&= \int_{-\infty}^{+\infty} \frac{1}{\pi^2+u^2} \int_0^\theta (1+e^u) e^{-(1+e^u)t} dt\, du \\
&= \int_{-\infty}^{+\infty} \frac{1-e^{-(1+e^u)\theta}}{\pi^2+u^2} du \\
&= \int_{-\infty}^{+\infty} \frac{1}{\pi^2+u^2} du - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2+u^2} du \\
&= \frac{1}{\pi}\left[\arctan(\frac{x}{\pi}) - \arctan(\frac{x}{\pi})\right]_{-\infty}^{+\infty} - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2+u^2} du \\
&= 1 - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2+u^2} du.
\end{aligned} \tag{20}$$

and we conclude that

$$\int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx = 1 - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2+u^2} du. \tag{21}$$

We note that this result can be also interpreted as an alternative expression for the Laplace transform of the function $x \mapsto \frac{1}{\Gamma(1+x)}$. Indeed, if we write $s = \log\theta$, we have

$$\int_0^{+\infty} e^{xs} \frac{1}{\Gamma(1+x)} dx = e^{e^s} - \int_{-\infty}^{+\infty} \frac{e^{-e^{s+u}}}{\pi^2 + u^2} du. \tag{22}$$

4.3 Asymptotic behavior

We can easily see that $g(\theta)$ is concave and increasing by looking at its derivatives. We see from Eq. (21) that g is bounded by 1. It thus converges to a finite value.

By noticing that

$$\forall \theta > 0, \forall u \in \mathbb{R},\ 0 \leq \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2 + u^2} du \leq 1,$$

we can conclude from Eq. (21) that

$$\lim_{\theta \to +\infty} \int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx = 1. \tag{23}$$

The normalization factor of f_θ thus converges to 1.

We also notice that $h(\theta) = \theta(g(\theta) + g'(\theta))$, from which we deduce in the same way that

$$\lim_{\theta \to +\infty} h(\theta) - \theta = 0. \tag{24}$$

So, asymptotically, f_θ behaves in the same way as the Poisson distribution, i.e., its mass converges to 1 and its first moment is asymptotically equal to its parameter.

4.4 Justifying again the cross-interpretation

As evoked in Section 3.2, several authors present their work with the $\mathcal{I}$-divergence directly on discretized data, enabling them to fall back to the discrete Poisson distribution after proper scaling of the data and the model. The above results on the normalization factor and the mean justify in a different way the bridge between $\mathcal{I}$-divergence-based fitting and ML estimation for real-valued data without quantization.

For sufficiently large values of the model, the distribution is indeed almost a probability distribution which behaves like the discrete Poisson distribution. If one can ensure that the values taken by the model will be bounded from below by a positive value, then by rescaling both the model and the data by multiplying them by a large constant, the continuous distribution $f_{Q(x,\mathbf{\Theta})}$ parameterized by the model $Q(x, \mathbf{\Theta})$ can be made as close to a probability distribution as desired for all the values taken by the model.

Meanwhile, the optimal parameters $\mathbf{\Theta}$ are not changed by the scaling operation, as the log-likelihoods before and after scaling are equal up to scaling and addition of a constant which does not depend on the parameters:

$$\int (\alpha W \log \alpha Q(\mathbf{\Theta}) - \alpha Q(\mathbf{\Theta})) = \alpha \int (W \log Q(\mathbf{\Theta}) - Q(\mathbf{\Theta})) + C \qquad (25)$$

where $\alpha > 0$ is the scaling parameter.

One can ensure that the model is bounded from below by a positive value for example if the data are themselves bounded from below by such a value and if the model is well-designed, such that it should not take infinitely small values if the data to fit is large enough. One can also add to both the model and the data a small value $\epsilon > 0$. The optimal parameters for this shifted problem can be made as close as desired to the optimal parameters for the original problem by choosing ϵ small enough, while, for ϵ fixed, the shifted problem can be made as close to a righteous ML problem as desired through scaling.

The interpretation of $\mathcal{I}$-divergence-based fitting as a ML problem is thus justified. In particular, it justifies the use of prior distributions in a MAP framework such as performed in [1, 2].

5 Conclusion

We presented the inherent mis-match occurring in the interpretation of some distribution-fitting problems as ML estimation ones, focusing on the example of the $\mathcal{I}$-divergence. We gave insights on the reason why distribution-fitting based on the $\mathcal{I}$-divergence on real-valued data cannot be seen directly as an ML estimation problem, and derived a theoretical workaround to this issue using the Gamma function, thus justifying MAP estimation as performed in [1, 2]. We plan to use this result in a forthcoming paper on missing-data problems.

Acknowledgments

The authors would like to thank Prof. Philippe Flajolet for valuable comments on Section 3.1, Prof. Akimichi Takemura for helpful comments and Dr. Vincent Plagnol for fruitful discussions.

References

[1] H. Kameoka, T. Nishimoto, and S. Sagayama, “A multipitch analyzer based on harmonic temporal structured clustering,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[2] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1135–1145, May 2007.

[3] K. Choi, "Minimum *I*-divergence methods for inverse problems," Ph.D. dissertation, Georgia Institute of Technology, 2005.

[4] D. L. Snyder, T. Schulz, and J. O'Sullivan, "Deblurring subject to nonnegativity constraints," *IEEE Trans. Signal Processing*, vol. 40, pp. 1143–1150, 1992.

[5] D. Lee and H. Seung, "Learning of the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[6] I. Csiszár, "Why least squares and maximum entropy? - an axiomatix approach to inverse problems," *Ann. Stat.*, vol. 19, pp. 2033–2066, 1991.

[7] ——, "*I*-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[8] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[9] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.