PROCEEDINGS OF SCIENCE

From χ^2 to Bayesian model comparison and Levy expansions of Bose-Einstein correlations in e^+e^- reactions

Michiel B. De Kock*, **Hans C. Eggers**
Department of Physics, University of Stellenbosch, ZA–7600 Stellenbosch, South Africa
E-mail: 14107112@sun.ac.za

Tamàs Csörgő

Wigner RCP, RMKI, H–1525 Budapest 114, P. O. Box 49, Hungary

The usual χ^2 method of fit quality assessment is a special case of the more general method of Bayesian model comparison which involves integrals of the likelihood and prior over all possible values of all parameters. We introduce new parametrisations based on systematic expansions around the stretched exponential or Fourier-transformed Lévy source distribution, and utilise the increased discriminating power of the Bayesian approach to evaluate the relative probability of these models to be true representations of a recently measured Bose-Einstein correlation data in e^+e^- annihilations at LEP.

The Seventh Workshop on Particle Correlations and Femtoscopy
September 20th to 24th, 2011
The University of Tokyo, Japan

*Speaker.

PoS(WPCF2011)033

1. Bayes factors

The Bayesian definition of probability differs radically from the conventional "frequentist" one, necessitating the overhaul of many concepts and techniques used in statistics and its applications. Since its introduction in 1900 [1], the χ^2 statistic has become the standard criterion for goodness of fit in physics and many other disciplines, while Laplace's Bayesian approach [2] remained largely forgotten until revived by Jeffreys [3]. Later refinements such as the Maximum Likelihood occupy a middle ground between the two approaches.

In this contribution, we demonstrate the use of one Bayesian technique in the simple context of fitting or, more generally, the quantitative assessment of evidence in favour of a hypothesis H_1 as a description of given data, compared to a rival hypothesis H_2. We do so by analysing the concrete example of binned data for the correlation function $C_2(Q)$ in the four-momentum difference $Q = \sqrt{-(p_1 - p_2)^2}$ as published recently by the L3 Collaboration [4].

Suppose we have data $\mathbf{D} = \{Q_1, \ldots, Q_n\}$ consisting of n measurements of particle four-momentum differences, assumed to be mutually independent as is customary in femtoscopy. Typically, the experimentalist will want to test how well various parametrisations fit the data. For the purposes of Bayesian analysis, a given parametrisation $y(Q\,|\,\boldsymbol{\theta}_m)$ with N_m free parameters $\boldsymbol{\theta}_m = \{\theta_{m1}, \theta_{m2}, \ldots, \theta_{mN_m}\}$ is considered a "model" or "hypothesis" H_m. The starting point is the *odds in favour of model H_m compared to a different model H_ℓ"*, defined as the ratio $p(H_m\,|\,\mathbf{D})/p(H_\ell\,|\,\mathbf{D})$, while the *evidence for H_m versus H_ℓ* is the logarithm[1] of the odds. Use of Bayes' Theorem for both hypotheses yields

$$\frac{p(H_m\,|\,\mathbf{D})}{p(H_\ell\,|\,\mathbf{D})} = \frac{p(\mathbf{D}\,|\,H_m)\,p(H_m)}{\sum_k p(\mathbf{D}\,|\,H_k)\,p(H_k)}\,\frac{\sum_k p(\mathbf{D}\,|\,H_k)\,p(H_k)}{p(\mathbf{D}\,|\,H_\ell)\,P(H_\ell)} = \frac{p(\mathbf{D}\,|\,H_m)}{p(\mathbf{D}\,|\,H_\ell)}\cdot\frac{p(H_m)}{p(H_\ell)}. \tag{1.1}$$

The evidence of H_m versus H_ℓ is therefore the same as the *Bayes factor* $B_{m\ell} = \lg[p(\mathbf{D}\,|\,H_m)/p(\mathbf{D}\,|\,H_\ell)]$ if there is no a priori reason to prefer H_m above H_ℓ and therefore $p(H_m) = p(H_\ell) = 1/2$. A large Bayes factor says that the evidence for H_m is stronger than the evidence for H_ℓ and vice versa. It can be written as a ratio of integrals over the respective parameter spaces of $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_\ell$,

$$B_{m\ell} = \lg\frac{p(\mathbf{D}\,|\,H_m)}{p(\mathbf{D}\,|\,H_\ell)} = \lg\frac{\int d\boldsymbol{\theta}_m\; p(\mathbf{D}|\boldsymbol{\theta}_m, H_m)\,p(\boldsymbol{\theta}_m|H_m)}{\int d\boldsymbol{\theta}_\ell\; p(\mathbf{D}|\boldsymbol{\theta}_\ell, H_\ell)\,p(\boldsymbol{\theta}_\ell|H_\ell)}. \tag{1.2}$$

Solving the high-dimensional integrals will often be an arduous task. Fortunately, the independence of the measurements implies that the likelihood $p(\mathbf{D}\,|\,\boldsymbol{\theta}_m, H_m)$ factorises into the product of likelihoods for individual data points, which by assumption have the same form,

$$p(\mathbf{D}\,|\,\boldsymbol{\theta}_m, H_m) = \prod_i p(Q_i\,|\,\boldsymbol{\theta}_m, H_m) \approx [p(Q\,|\,\boldsymbol{\theta}_m, H_m)]^n. \tag{1.3}$$

Due to the large exponent, even the slightest nonuniformity in $p(Q\,|\,\boldsymbol{\theta}_m, H_m)$ will lead to the development of a strong peak in parameter space for the overall likelihood, situated at the maximum likelihood point $\hat{\boldsymbol{\theta}}_m$. An asymmetric prior $p(\boldsymbol{\theta}_m\,|\,H_m)$ will shift the peak to a value $\boldsymbol{\theta}^*_m$, but it will not materially affect the width of the peak or its differentiability. Unless the shifted peak falls on a

[1] We use $\lg = \log_2$; other base units can be substituted as preferred.

boundary of the parameter space or happens to be nondifferentiable, it can therefore be expanded around $\boldsymbol{\theta}_m^*$ [5]:

$$p(\mathbf{D}\,|\,\boldsymbol{\theta}_m,H_m)\,p(\boldsymbol{\theta}_m\,|\,H_m)\simeq p(\mathbf{D}\,|\,\boldsymbol{\theta}^*,H_m)\,p(\boldsymbol{\theta}^*\,|\,H_m)\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_m-\boldsymbol{\theta}_m^*)\mathbf{A}^{-1}(\boldsymbol{\theta}_m-\boldsymbol{\theta}_m^*)\right] \quad (1.4)$$

where $\mathbf{A}^{-1}$ is the Hessian of the expansion

$$A_{ij}^{-1}=-\left.\frac{\partial^2\ln[p(\mathbf{D}\,|\,\boldsymbol{\theta}_m,H_m)\,p(\boldsymbol{\theta}_m\,|\,H_m)]}{\partial\theta_{mi}\,\partial\theta_{mj}}\right|_{\boldsymbol{\theta}_m^*} \quad (1.5)$$

and $\mathbf{A}$ is the parameter covariance matrix. As more data is accumulated, the peak narrows so that we can neglect the fact that parameters may have finite ranges. Integrating the above as if it were a Gaussian, one obtains Laplace's result [2]

$$\int_{-\infty}^{+\infty}d\boldsymbol{\theta}\;p(\mathbf{D}\,|\,\boldsymbol{\theta}_m,H_m)\,p(\boldsymbol{\theta}_m\,|\,H_m)\simeq p(\mathbf{D}\,|\,\boldsymbol{\theta}_m^*,H_m)\,p(\boldsymbol{\theta}_m^*\,|\,H_m)\sqrt{(2\pi)^{N_m}\det\mathbf{A}_m}\,, \quad (1.6)$$

which under the stated assumptions is a good approximation of the full-blown integral appearing in Eq. (1.2) if $n\gtrsim 20N_m$. The Bayes factor becomes simply the difference

$$B_{m\ell}\simeq h_\ell-h_m \quad (1.7)$$

$$h_k\equiv-\lg\left[p(\mathbf{D}\,|\,\boldsymbol{\theta}_k^*,H_k)\,p(\boldsymbol{\theta}_k^*\,|\,H_k)\sqrt{(2\pi)^{N_k}\det\mathbf{A}_k}\right]. \quad (1.8)$$

Evidence h_k can be determined for any single model H_k, but has no meaning on its own; only differences $h_\ell-h_m$ are meaningful in quantifying the probability for H_m to be true compared to H_ℓ,

$$\frac{p(H_m\,|\,\mathbf{D})}{p(H_\ell\,|\,\mathbf{D})}\simeq 2^{h_\ell-h_m}. \quad (1.9)$$

2. Relationship to χ^2 and the Maximum Likelihood

The Bayesian results obtained above differ from the traditional Maximum Likelihood Estimate (MLE), which ignores the priors $p(\boldsymbol{\theta}_m\,|\,H_m)$ and approximates the integral (1.2) to the maxima of the likelihoods,

$$B_{m\ell}=\lg\frac{\int d\boldsymbol{\theta}_m\;p(\mathbf{D}|\boldsymbol{\theta}_m,H_m)\,p(\boldsymbol{\theta}_m|H_m)}{\int d\boldsymbol{\theta}_\ell\;p(\mathbf{D}|\boldsymbol{\theta}_\ell,H_\ell)\,p(\boldsymbol{\theta}_\ell|H_\ell)}\simeq\lg\frac{p(\mathbf{D}|\hat{\boldsymbol{\theta}}_m,H_m)}{p(\mathbf{D}|\hat{\boldsymbol{\theta}}_\ell,H_\ell)}. \quad (2.1)$$

The traditional χ^2 goodness-of-fit is related to the above as follows. The measurements $\{Q_i\}$ are binned into bins $b=1,\ldots,B$ with bin midpoints Q_b, yielding the histogram version of the data, $\mathbf{D}=\{n_b\}_{b=1}^B$ with $\sum_b n_b=1$. The most general "parametrisation" of the histogram contents is then the multinomial with $\boldsymbol{\alpha}=\{\alpha_b\}_{b=1}^B$ the set of Bernoulli probabilities with $B-1$ degrees of freedom,

$$p(\boldsymbol{n}\,|\,\boldsymbol{\alpha},n)=n!\prod_{b=1}^{B}\frac{\alpha_b^{n_b}}{n_b!}\,, \quad (2.2)$$

which on use of the Stirling approximation becomes, up to a normalisation constant,

$$p(\boldsymbol{n}\,|\,\boldsymbol{\alpha},n) = c\cdot\exp\left[-\sum_b n_b \ln\frac{n_b}{n\alpha_b}\right]. \tag{2.3}$$

Expanding the free parameters $\boldsymbol{\alpha}$ around the measured data $\boldsymbol{n}$ and truncating

$$p(\boldsymbol{n}\,|\,\boldsymbol{\alpha},n) = c\cdot\exp\left[-\sum_b\left(\frac{(n\alpha_b-n_b)^2}{2n_b}-\frac{(n\alpha_b-n_b)^3}{3n_b^2}+\ldots\right)\right] \simeq c\cdot\exp\left[-\frac{1}{2}\sum_b\frac{(n\alpha_b-n_b)^2}{n_b}\right], \tag{2.4}$$

we can identify the multinomial quantities with the measured correlation functions at mid-bin points Q_b by setting[2] $n_b \to IC_2(Q_b)$, $C=\sum_b C_2(Q_b)$, and $n\to IC$. The n_b in the denominator is almost equal to the measured bin variances $n_b \simeq \sigma^2(n_b) = I^2\,\sigma^2(C_2(Q_b))$ so that the quadratic term is

$$\frac{(n\alpha_b-n_b)^2}{2\,n_b} \simeq \frac{[C_2(Q_b)-y(Q_b\,|\,\hat{\boldsymbol{\theta}}_m)]^2}{2\,\sigma(C_2(Q_b))^2}, \tag{2.5}$$

where $n\alpha_b/I \to y(Q_b\,|\,\hat{\boldsymbol{\theta}}_m)$, which includes all the constants, is the unnormalised parametrisation for $C_2(Q)$ in common use. Comparing this to the usual definition

$$\chi^2 = \sum_b \frac{[C_2(Q_b)-y(Q_b\,|\,\hat{\boldsymbol{\theta}}_m)]^2}{\sigma(C_2(Q_b))^2}, \tag{2.6}$$

we see that the maximum likelihood is approximately equal to

$$p(\mathbf{D}|\hat{\boldsymbol{\theta}}_m,H_m) \simeq e^{-\chi^2/2}, \tag{2.7}$$

so that χ^2 is seen to be an approximation of the Bayes formulation, using only a single point in the parameter space $\boldsymbol{\theta}^*_m \equiv \hat{\boldsymbol{\theta}}_m$ and thereby effectively assuming a uniform prior. Furthermore, χ^2 truncates the expansion of (2.4); this is probably the approximation most vulnerable to criticism.

3. Parametrisations and Lévy-based polynomial expansions

We now apply the above general ideas to the specific case of the various parametrisations shown in Table 1 for the correlation function data for two-jet events published by the L3 Collaboration [4]. Hypotheses H_1 to H_3 are taken from the L3 paper. Realising that it is important to quantify the degree of deviation of Bose-Einstein correlation data from the Gaussian or the exponential shape, the L3 Collaboration also studied a "Laguerre expansion" as well as the symmetric Lévy source distribution, characterized by the stretched-exponential correlation function of hypothesis H_2. In H_4 and H_5, we propose a new expansion technique that measures deviations from H_2 in terms of a series of "Lévy polynomials" that are orthogonal to the characteristic function of symmetric Lévy distributions, generalising the results presented in Ref. [6].

$$L_1(x\,|\,\alpha) = \det\begin{pmatrix}\mu_{0,\alpha} & \mu_{1,\alpha}\\ 1 & x\end{pmatrix} \qquad L_2(x\,|\,\alpha) = \det\begin{pmatrix}\mu_{0,\alpha} & \mu_{1,\alpha} & \mu_{2,\alpha}\\ \mu_{1,\alpha} & \mu_{2,\alpha} & \mu_{3,\alpha}\\ 1 & x & x^2\end{pmatrix} \qquad \text{etc.} \tag{3.1}$$

[2] I is an arbitrary large integer to ensure that $IC_2(Q_b)$ is an integer. As it eventually cancels out, its size is immaterial.

where $\mu_{r,\alpha} = \int_0^\infty dx\, x^r f(x|\alpha) = \frac{1}{\alpha}\Gamma(\frac{r+1}{\alpha})$. These reduce, up to a normalisation constant, to the Laguerre polynomials for $\alpha = 1$. Figure 1 displays two examples for various values of α. Polynomials cannot be both orthogonal and derivatives for transcendental weight functions [9], and therefore in H_6 and H_7 we also investigated nonorthogonal derivative functions of the stretched exponential[3].

	Hypothesis	Functional form	N_m
H_1	Gauss	$\gamma[1+\varepsilon Q]\left[1+\lambda e^{-R^2Q^2}\right]$	4
H_2	Stretched Exponential	$\gamma[1+\varepsilon Q]\left[1+\lambda e^{-R^\alpha Q^\alpha}\right]$	5
H_3	Simplified τ-model	$\gamma[1+\varepsilon Q]\left[1+\lambda e^{-R^{2\alpha}Q^{2\alpha}}\cos[\tan(\alpha\pi/2)R^{2\alpha}Q^{2\alpha}]\right]$	5
H_4	1st-order Lévy polynomial	$\gamma\left[1+\lambda e^{-R^\alpha Q^\alpha}[1+c_1L_1(Q\|\alpha,R)]\right]$	5
H_5	3rd-order Lévy polynomial	$\gamma\left[1+\lambda e^{-R^\alpha Q^\alpha}[1+c_1L_1(Q\|\alpha,R)+c_3L_3(Q\|\alpha,R)]\right]$	6
H_6	1st-order derivative	$\gamma\left[1+\lambda e^{-R^\alpha Q^\alpha}+c_1\frac{d}{dQ}e^{-R^\alpha Q^\alpha}\right]$	5
H_7	3rd-order derivative	$\gamma\left[1+\lambda e^{-R^\alpha Q^\alpha}+c_1\frac{d}{dQ}e^{-R^\alpha Q^\alpha}+c_3\frac{d^3}{dQ^3}e^{-R^\alpha Q^\alpha}\right]$	6

Table 1: Summary of parametrisations tested

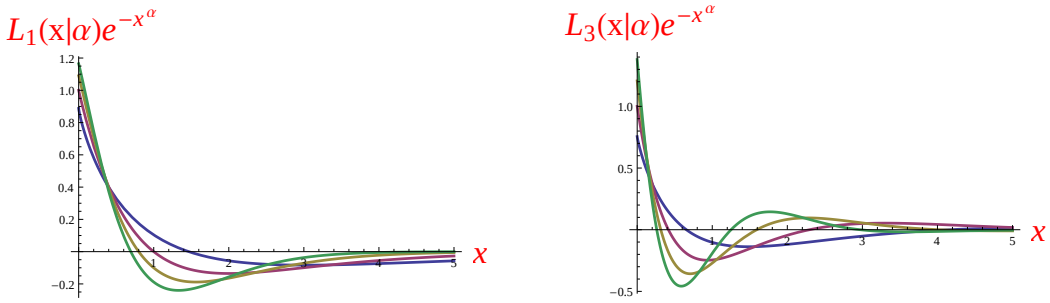

Figure 1: Lévy polynomials of first and third order times the weight function e^{-x^α} for $\alpha = 0.8, 1.0, 1.2, 1.4$.

4. Application to L3 binned data

In Table 2, we show the results of applying the Laplace approximation (1.6) to the L3 two-jet data, which is provided in terms of 100 binned values for the correlation function $C(Q_b)$ together with standard errors $\sigma(C(Q_b))$ in the range $0 < Q < 4$ GeV. Throughout, we used a Gaussian prior $p(\boldsymbol{\theta}^*_m|H_m)$ with a width which was determined by numerical integration over one of the L3 data points. To illustrate the contributions of the likelihood, prior and determinant factors entering h_m

[3]Note the absence of the $[1+\varepsilon Q]$ long-range correction term. L3 demonstrated that this term vanishes if the dip, the non-positive definiteness of $C_2(Q)-1$, is taken into account by the parametrisation elsewhere, e.g. by the cosine in H_3 and by the first-order polynomials in H_4 and H_5, resulting in ε values consistent with zero.

in (1.8), we have listed their logarithmic contributions separately in the three columns headed L, P and F. These quantities are therefore the building blocks for calculating the odds between any two competing hypotheses. Thus one can, for example, deduce that the odds for H_7 compared to H_6 are $2^{100.6-97.0} \simeq 12{:}1$. Also included in Table 2 are the traditional χ^2 measure (C) and its associated confidence level (CL).

Hypothesis	N_m	L	P	F	h_m	C	CL
H_1 Gauss	4	177.8	-3.6	32.2	206.5	2.57	3.4×10^{-13}%
H_2 Stretched Exponential	5	138.5	-0.5	34.0	172.0	2.02	1.5×10^{-6}%
H_3 Simplified τ-model	5	68.2	-3.4	37.0	101.8	1.00	49.1%
H_4 1st-order Lévy polynomial	5	66.2	2.2	30.3	98.8	0.97	57.3%
H_5 3rd-order Lévy polynomial	6	65.9	3.8	41.6	111.3	0.97	55.7%
H_6 1st-order derivative	5	67.3	4.2	29.1	100.6	0.98	53.0%
H_7 3rd-order derivative	6	60.4	4.9	31.7	97.0	0.89	77.0%

Table 2: Results of fitting parametrisations listed in Table 1.
Legend: $\mathrm{L} \equiv -\lg P(\mathbf{D}\,|\,\boldsymbol{\theta}^*_m, H_m) \equiv \chi^2/(2\ln 2)$ $\quad h_m \equiv L+P+F$
$\mathrm{P} \equiv -\lg P(\boldsymbol{\theta}^*_m\,|\,H_m)$ $\quad \mathrm{C} \equiv \chi^2/(B-N_m)$
$\mathrm{F} \equiv -\lg\sqrt{(2\pi)^{N_m}\det\mathbf{A}}$ $\quad \mathrm{CL} \equiv$ confidence level

It is inappropriate to generalise conclusions based on one specific dataset with its specific circumstances. The fact that in the two-jet L3 data the correlation function $C_2(Q)$ drops well below 1.0 for $0.5 < Q < 2$ GeV, for example, is probably the dominant influence on the goodness of fit. Under this caveat, we make the following observations regarding the results shown in Table 2:

1. At first sight, the Bayes factor and the χ^2 methodologies deliver judgements which are rather similar: H_7 is consistently ranked best, while H_1 and H_2 are ranked worst (least likely). The two methodologies yield vastly different numbers when one hypothesis is bad. As shown below, there are surprising variations even among the better ones.
2. The determinant plays an important role. For example, factor F$=41.6$ for H_5 is significantly larger than that of similar models H_4 and H_6 even though the three log likelihoods are similar. This can be traced to the fact that the uncertainty in the parameters for H_5 is larger, as expressed in the width of its Gaussian (1.4). While χ^2, based only on the likelihood, can hardly distinguish between H_4 and H_5, the contribution of the large H_5 determinant ensures that the Bayesian odds for H_4 versus H_5 are 5800:1. In other words, by taking into account not only the best parameter values $\boldsymbol{\theta}^*_5$ but also their uncertainties, the Bayes factor could distinguish what χ^2 could not.
3. Our Bayes factor calculation takes the experimental standard errors $\sigma(C(Q_b))$ into account by using (2.5) in the exponent of the likelihood; in other words, we assume that they are Gaussian. We can improve on this approximation by doing a more complete Bayesian analysis using not the binned data but the pair momenta $\{Q_i\}$ themselves.
4. As Fig. 1 shows, the Lévy polynomials introduced here are well suited to describe one-sided strongly-peaked data. It may be helpful to use them, as we have done here, merely as part of parametrisations of data to which they show some resemblance. More systematic use in Gram-Charlier or other expansions will be faced with issues inherent in all asymptotic series [7, 8].

5. Conclusions

1. In hypotheses H_4 to H_7, we have presented new techniques to study deviations from a stretched exponential or Fourier-transformed Lévy shape. Details will be published elsewhere.
2. The standard measures of fit quality like χ^2 or CL are useful in rejecting models which are inconsistent with a given dataset. Where two or more models are consistent with the data, however, they are unable to select the more probable. The Bayes factor (1.9) permits quantification of the evidence (relative probability) for the validity of models.
3. Besides the likelihood, the prior and determinant also play a role, sometimes decisively so.
4. The Laplace approximation (1.4) is usually fairly accurate, but the assumption of Gaussian errors for count data (2.4), which is made by truncation of the Taylor expansion in the data, is of dubious quality.
5. By integrating over parameter space, Bayesian evidence takes into account *all* possible values of the parameters, while χ^2 and Maximum Likelihood do not.
6. Bayes factors depend linearly on the two priors. This is good in that they are made explicit, but bad in the sense that results can and do change depending on the choice of priors.
7. The omission of priors in χ^2 is to its disadvantage as it discards important information.
8. It may appear that χ^2 does not need any alternative hypothesis to be of use. This is not so, however: the alternative implicit in χ^2 is the "Bernoulli class" of multinomials [10].

Acknowledgements: We thank the L3 collaboration for making its results available electronically [4] and the organizers of WPCF 2011 for support and an excellent atmosphere. This work was supported in part by the South African National Research Foundation and by the Hungarian OTKA grant NK–101438.

References

[1] K. Pearson, *On a criterion that a given system of deviations ... is such that it can be reasonably supposed to have arisen in random sampling*, Phil. Mag. **(5) 50** (1900) 157.

[2] P.S. Laplace, *Mémoires de Mathématique et de Physique, Tome Sixième* (1774).

[3] R. Jeffreys, *Theory of Probability*, Oxford University Press (1961).

[4] L3 Collaboration, P. Achard et al., *Test of the τ-Model of Bose-Einstein Correlations and Reconstruction of the Source Function in Hadronic Z-boson Decay at LEP*, Eur. Phys. J. C**71** (2011) 1648 [arXiv:1105.4788], see also http://l3.web.cern.ch/l3/

[5] R.E. Kass and A.E. Raftery, *Bayes Factors*, J. American Statistical Association **90** (1995) 773.

[6] T. Csörgő and S. Hegyi, *Model independent shape analysis of correlations in 1, 2 or 3 dimensions*, Phys. Lett. B **489** (2000) 15.

[7] M.B. de Kock, *Gaussian and non-Gaussian-based Gram-Charlier and Edgeworth expansions for correlations of identical particles in HBT interferometry*, M.Sc., University of Stellenbosch (2009).

[8] H.C. Eggers, M.B. de Kock and J. Schmiegel, *Determining source cumulants in femtoscopy with Gram-Charlier and Edgeworth series*, Mod. Phys. Lett. A **26** (2011) 1771 [arXiv:1011.3950].

[9] A. Erdélyi et al., *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York (1953).

[10] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press (2003).