BIROn - Birkbeck Institutional Research Online

Levene, Mark and Kononovicius, A. (2019) Empirical survival Jensen-Shannon divergence as a goodness-of-fit measure for maximum likelihood estimation and curve fitting. Communications in Statistics - Simulation and Computation 50 (11), pp. 3751-3767. ISSN 0361-0918.

Empirical Survival Jensen-Shannon Divergence as a Goodness-of-Fit Measure for Maximum Likelihood Estimation and Curve Fitting

Mark Levene*

Department of Computer Science and Information Systems,
Birkbeck, University of London, London WC1E 7HX, U.K.

Aleksejus Kononovicius†

Institute of Theoretical Physics and Astronomy,
Vilnius University, Vilnius, Lithuania

Abstract

The coefficient of determination, known as R^2, is commonly used as a goodness-of-fit criterion for fitting linear models. R^2 is somewhat controversial when fitting nonlinear models, although it may be generalised on a case-by-case basis to deal with specific models such as the logistic model. Assume we are fitting a parametric distribution to a data set using, say, the maximum likelihood estimation method. A general approach to measure the goodness-of-fit of the fitted parameters, which is advocated herein, is to use a non-parametric measure for comparison between the empirical distribution, comprising the raw data, and the fitted model. In particular, for this purpose we put forward the *Survival Jensen-Shannon divergence* (SJS) and its empirical counterpart ($\mathcal{E}SJS$) as a metric which is bounded, and is a natural generalisation of the Jensen-Shannon divergence. We demonstrate, via a straightforward procedure making use of the $\mathcal{E}SJS$, that it can be used as part of maximum likelihood estimation or curve fitting as a measure of goodness-of-fit, including the construction of a confidence interval for the fitted parametric distribution. Furthermore, we show the validity of the proposed method with simulated data, and three empirical data sets.

Keywords: divergence measures, goodness-of-fit, maximum likelihood, curve fitting, survival Jensen-Shannon divergence

1 Introduction

We assume a general scenario, where we have some data from which we derive an empirical distribution that is fitted with maximum likelihood (Myung, 2003) or curve fitting (Fox, 2016) to some, possibly parametric distribution (Krishnamoorthy, 2015).

---

*email: mark@dcs.bbk.ac.uk

†corresponding author; email: aleksejus.kononovicius@tfai.vu.lt; website: `http://kononovicius.lt`; OR-CID: 0000-0001-6365-0252

The *coefficient of determination*, R^2 (Motulsky, 1995), is a well-known measure of goodness-of-fit for linear regression models. Despite its wide use, in its original form, it is not fully adequate for nonlinear models, (Anderson-Sprecher, 1994), where the author recommends to define R^2 as a comparison of a given model to the null model, claiming that this view allows for the generalisation of R^2. Further, in (Spiess & Neumeyer, 2010) the inappropriateness of R^2 for nonlinear models is clearly demonstrated via a series of Monte Carlo simulations. In (Cameron & Windmeijer, 1997), a novel R^2 measure based on the Kullback-Leibler divergence (Cover & Thomas, 2006) was proposed as a measure of goodness-of-fit for regression models in the exponential family. In addition, in Orelien & Edwards (2008) problems with using R^2 for assessing goodness-of-fit in linear mixed models with random effects were highlighted, and in Nakagawa & Schielzeth (2013) an improved extension was proposed in the context of both linear and generalised linear mixed models. Despite numerous proposals to address the issues with R^2 for nonlinear models, many of them are ad-hoc, and, as noted in Sapra (2014), should be applied with caution. In summary, there seems to be a lack of general purpose goodness-of-fit measures that could be applied to nonlinear models, which is the main issue we attempt to redress with the $\mathcal{ESJS}$.

Alternative nonparametric methods have also been proposed. In particular, the Akaike information criterion (AIC) and its counterpart the Bayesian information criterion (BIC) (Burnham & Anderson, 2002; Vrieze, 2012), are widely used estimators for model selection. Both AIC and BIC are asymptotically valid maximum likelihood estimators, with penalty terms to discourage overfitting. We stress that goodness-of-fit measures how well a single model fits the observed data, while model selection compares the predictive accuracy of two models relative to each other Burnham & Anderson (2002); Pitt & Myung (2002). The likelihood ratio test is also an established method for model selection between a null model and an alternative maximum likelihood model (Vuong, 1989; Lewis *et al.*, 2011). Despite the popularity of maximum likelihood methods, there is some controversy in their application as goodness-of-fit tests (Heinrich, 2003).

The *Jensen-Shannon divergence* (JSD) (Lin, 1991; Endres & Schindelin, 2003) is a symmetric form of the nonparametric Kullback-Leibeler divergence (Cover & Thomas, 2006), providing a measure of distance between two probability distributions. It has been employed in a wide range of applications such as detecting edges in digital images (Gómez-Lopera *et al.*, 2000), measuring the similarity of texts (Mehri *et al.*, 2015), training adversarial neural networks (Goodfellow *et al.*, 2014), comparison of genomes in bioinformatics (Sims *et al.*, 2009), distinguishing between quantum states in physics (Majtey *et al.*, 2005) and as a measure of distance between distributions in a social setting (Fenner *et al.*, 2018).

Here we generalise the JSD to the *survival Jensen-Shannon divergence* (SJS) and it empirical counterpart ($\mathcal{ESJS}$) by employing the survival functions of the constituent probability density functions. We apply the $\mathcal{ESJS}$ as an alternative measure of goodness-of-fit of a parametric distribution, acting as the model, to an empirical distribution, which comprises the raw data. The $\mathcal{ESJS}$ provides a direct measure of goodness-of-fit without the need of the maximum value of the likelihood function, as used in the AIC and BIC, or any linearity assumptions of the model being fitted, as often made when using R^2.

The rest of the paper is organised as follows. In Section 2, we introduce the *empirical survival Jensen-Shannon Divergence* ($\mathcal{ESJS}$) and some of its characteristics. In Section 3, we define the $\mathcal{ESJS}$ as a measure of goodness-of-fit within the context of distribution fitting and define the notion of the $\mathcal{ESJS}$ factor. In Section 4, we describe some experiments we carried out, with simulated data in Subsection 4.1 and empirical data in Subsection 4.2, to test the

viability of using the $\mathcal{E}SJS$ as a measure of goodness-of-fit. Finally, in Section 5, we give our concluding remarks.

2   Survival Jensen-Shannon Divergence

First, in Subsection 2.1 we define the survival function with respect to the Lebesgue-Stieltjes measure, and then the empirical survival entropy, which generalises the entropy to continuous distributions. Building on these concepts, we then define, in Subsection 2.2 the survival Jensen-Shannon divergence, which generalises the standard Jensen-Shannon divergence (JSD), and derive a formula for computing its empirical counterpart.

2.1   Survival Entropy

We define the *survival function* $S = S_X$, for real-valued random variable X as

$$S_X(x) = P\left(X > x\right),$$

which represents the probability that X takes a value greater than x. It can be shown that there exists a unique probability measure P_S for S, with respect to the σ-field of Borel sets on the real line, defined as

$$P_S((a, b]) = S(a) - S(b). \tag{1}$$

The measure in (1) is known as the *Lebesgue-Stieltjes* measure; the detailed theory can be found in, for example, (Capiński & Kopp, 2004) and (Shorack, 2017) and an informative resource describing the various types of integral and induced measures can be found in (Burk, 2007).

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed random variables forming a random sample drawn from a population having survival function S. The *empirical survival function*, denoted by $\hat{S}_n$, is given by

$$\hat{S}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i > x\}}, \tag{2}$$

where I is the indicator function. It is well known, according to the Glivenko-Cantelli theorem, that $\hat{S}_n$ converges to S as n tends to infinity; for the details see, for example (Shorack, 2017).

The *empirical survival entropy*, (called the empirical cumulative entropy in (Rao *et al.*, 2004)), is given by

$$\mathcal{E}(\hat{S}_n) = -\int_0^\infty \hat{S}_n(x) \log \hat{S}_n(x) \; \mathrm{d}x, \tag{3}$$

using the convention that $0 \log 0 = 0$.

Following (Di Crescenzo & Longobardi, 2009) let $0 = X_{(0)} \leq X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ be the order statistics of the random sample, implying that

$$\mathcal{E}(\hat{S}_n) = -\sum_{i=1}^{n-1} \int_{X_{(i)}}^{X_{(i+1)}} \hat{S}_n(x) \log \hat{S}_n(x) \; \mathrm{d}x. \tag{4}$$

Moreover,

$$\hat{S}_n = 1 - \frac{i}{n}, \quad \text{when} \;\; X_{(i)} \leq x < X_{(i+1)}, \;\; \text{for} \;\; i = 1, 2, \ldots, n-1,$$

and thus on using (3) it follows that

$$\mathcal{E}(\hat{S}_n) = -\sum_{i=1}^{n-1} U_{i+1}\left(1 - \frac{i}{n}\right)\log\left(1 - \frac{i}{n}\right),\qquad(5)$$

where $U_i = X_{(i)} - X_{(i-1)}$ is the sample spacing of the random sample drawn from a population having survival function S; see also (Yari & Saghafi, 2012).

Again, by the Glivenko-Cantelli theorem, it can be shown that $\mathcal{E}(\hat{S}_n)$ would converge to $\mathcal{E}(S)$ as n tends to infinity, where $\mathcal{E}(S)$ is obtained by replacing $\hat{S}_n$ in (3) by S; see (Rao *et al.*, 2004; Di Crescenzo & Longobardi, 2009). As a result, here we will concentrate on the methodology and empirical analysis of large data sets, and leave the study of its application to small data sets as an open problem.

2.2 Survival Jensen-Shannon Divergence

Let P and Q be survival functions of the density function p and q, respectively, and let M be the survival function of the mixture $\frac{1}{2}(p + q)$. We define the *Survival Jensen-Shannon divergence* (SJS), as

$$SJS(P,Q) = \frac{1}{2}\int_0^\infty \left(P(x)\log\frac{P(x)}{M(x)} + Q(x)\log\frac{Q(x)}{M(x)}\right)\mathrm{d}x.\qquad(6)$$

It is easy to see that

$$SJS(P,Q) = \int_0^\infty \left(\frac{1}{2}P(x)\log P(x) + \frac{1}{2}Q(x)\log Q(x) - M(x)\log M(x)\right)\mathrm{d}x.\qquad(7)$$

The *empirical SJS* is thus given by

$$\mathcal{E}SJS(P,Q) = \frac{1}{2}\int_0^\infty \left(\hat{P}_n(x)\log\frac{\hat{P}_n(x)}{\hat{M}_n(x)} + \hat{Q}_n(x)\log\frac{\hat{Q}_n(x)}{\hat{M}_n(x)}\right)\mathrm{d}x,\qquad(8)$$

which by (7) can be stated as

$$\mathcal{E}SJS(P,Q) = -\int_0^\infty \left(\frac{1}{2}\mathcal{E}(\hat{P}_n) + \frac{1}{2}\mathcal{E}(\hat{Q}_n) - \mathcal{E}(\hat{M}_n)\right)\mathrm{d}x.\qquad(9)$$

Finally, on using (5) we have

$$\mathcal{E}SJS(P,Q) = \sum_{i=1}^{n-1}\left(\frac{1}{2}U[P]_{i+1} + \frac{1}{2}U[Q]_{i+1} - U[M]_{i+1}\right)\left(1 - \frac{i}{n}\right)\log\left(1 - \frac{i}{n}\right),\qquad(10)$$

where $U[P]_i$, $U[Q]_i$ and $U[M]_i$ are, respectively, the sample spacings drawn from populations having the survival functions P, Q and M.

An important fact to note is that the square root of the $\mathcal{E}SJS$ is a metric (Nguyen & Vreeken, 2015), generalising the result for the standard JSD (Endres & Schindelin, 2003). It is also bounded but with a different normalisation constant than that of the JSD. Moreover, we observe that it is often advantageous, as we do here, to use the survival function (or equivalently the cumulative distribution) instead of the probability density function as it may

be easier to interpret and manipulate, and it also acts to smooth the data. We further note that we are dealing with continuous distributions and therefore it makes more sense to use the $\mathcal{ESJS}$, based on the survival function, rather than on the empirical counterpart of the JSD based on the density function, which is only defined at discrete points. In this context we also note that our use of the empirical survival distribution is compatible with the empirical distribution function statistics as in Stephens (1986), which measure the absolute difference between the empirical distribution and an assumed distribution from the data it is sampled from.

3   Empirical Jensen-Shannon Divergence as a Goodness-of-Fit Measure

Making use of the $\mathcal{ESJS}$ as a measure of goodness-of-fit is quite straightforward. Assume that $P(D, \phi)$ is a sample from a parametric distribution D, with parameters ϕ, and that D is fitted with maximum likelihood (Myung, 2003) or curve fitting (Freund *et al.*, 2006) to an empirical distribution, $\hat{P}$.

The goodness-of-fit of the distribution D, with parameters ϕ, to the empirical distribution $\hat{P}$ is now defined as

$$\mathcal{ESJS}(P(D, \phi), \hat{P}), \tag{11}$$

where $P(D, \phi)$ is a sample from the distribution D with parameters ϕ. We note that employing a sample from a parametric distribution in (11) does not restrict the $\mathcal{ESJS}$, and so it is also possible to measure one empirical distribution against any another.

The Bayes factor (Kass & Raftery, 1995) is a method for model comparison, taking the ratio of the models representing the likelihood of the data under the alternative hypothesis and likelihood of the data under the null hypothesis. In particular, the Bayes factor is advocated as an alternative method for null hypothesis significance testing, which depends only on the data and considers the models arising from both the null and alternative hypotheses (Jarosz & Wiley, 2014).

The *Empirical survival Jensen-Shannon divergence factor* is reformulation of the Bayes factor with the $\mathcal{ESJS}$, defined as

$$\frac{\mathcal{ESJS}(P(D_1, \phi_1), \hat{P})}{\mathcal{ESJS}(P(D_0, \phi_0), \hat{P})}, \tag{12}$$

which is the odds ratio of choosing the alternative hypothesis, $P(D_1, \phi_1)$, in preference to the null hypothesis, $P(D_0, \phi_0)$.

4   Experiments and Analysis

To assess the use of the $\mathcal{ESJS}$ as a goodness-of-fit measure, we provide experimental results with simulated and empirical data with respect to various parametric distributions including the Normal, Uniform, Log-normal, Gamma, Weibull, Beta, Exponential, Pareto (Krishnamoorthy, 2015) and q-Gaussian (Nadarajaha & Kotz, 2007; Tsallis, 2017) distributions. It is worth to note that q-Gaussian distribution is equivalent to Student's t-distribution (Krishnamoorthy, 2015). In the experiments we use the following parametrizations of these distributions:

- Normal distribution (two parameters are reported in the Tables: first parameter is the mean μ and the second parameter is standard deviation σ):

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{13}$$

- Uniform distribution (first parameter is the lower bound, x_l, and the second parameter is the upper bound, x_u):

$$p_{x_l,x_u}(x) = \begin{cases} \frac{1}{x_u - x_l} & \text{if } x_l \leq x \leq x_u \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

- Log-normal distribution (first parameter – μ, second parameter – σ):

$$p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}x} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \tag{15}$$

- Gamma distribution (first parameter – k, second parameter – τ):

$$p_{k,\tau}(x) = \frac{x^{k-1}}{\Gamma(k)\tau^k} \exp\left(-\frac{x}{\tau}\right). \tag{16}$$

- Weibull distribution (first parameter – k, second parameter – τ):

$$p_{k,\tau}(x) = \frac{kx^{k-1}}{\tau^k} \exp\left(-\frac{x^k}{\tau^k}\right). \tag{17}$$

- Beta distribution (first parameter – α, second parameter – β):

$$p_{\alpha,\beta}(x) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}. \tag{18}$$

- q-Gaussian distribution (first parameter – q, second parameter – λ):

$$p_{\lambda,x_0}(x) = \frac{\Gamma\left(\frac{\lambda}{2}\right)}{\sqrt{\pi x_0^2}\Gamma\left(\frac{\lambda-1}{2}\right)} \left(\frac{x_0^2}{x_0^2 + x^2}\right)^{\frac{\lambda}{2}}. \tag{19}$$

- Exponential distribution (only one parameter is reported – τ):

$$p_\tau(x) = \frac{1}{\tau} \exp\left(-\frac{x}{\tau}\right). \tag{20}$$

- Pareto distribution (only one parameter is reported – α):

$$p_\alpha(x) = \frac{\alpha}{x^{\alpha+1}}. \tag{21}$$

In the experiments we carry out, we will make use of the *bootstrap method* (Davison & Hinkley, 1997), which is a technique for computing a confidence interval that relies on random resampling with replacement from a given sample data set. The bootstrap method is usually nonparametric, making no distributional assumptions about the data set employed.

Our methodology for the experiments with simulated data (see Subsection 4.1) was as follows:

(i) First we generated a data set, say S_1, of size 10^6 from a *given distribution*, say D_1, with chosen parameters, say ϕ_1, which was then taken to be the empirical distribution.

(ii) We then considered S_1 to be distributed according to a *hypothesised distribution*, D_2, where D_2 may not be the same as D_1, and used the maximum likelihood method to obtain the parameters of S_1, say ϕ_2, assuming its distribution was D_2. (Obviously, if $D_2 = D_1$, then ϕ_2 is expected to be very close to ϕ_1.)

(iii) Next, assuming that S_1 was distributed according to D_2 with parameters ϕ_2, we generated a second data set, $S_2 = P(D_2, \phi_2)$, from distribution D_2 with parameters ϕ_2.

(iv) Finally, we evaluated $\mathcal{ESJS}(S_2, S_1)$ as a measure of the goodness-of-fit of D_2, with parameters ϕ_2, to S_1, and computed a 95% confidence interval for the $\mathcal{ESJS}$ from 1000 bootstrap resamples using the *basic bootstrap percentile method* (Davison & Hinkley, 1997, Section 5.3.1).

For the experiments with empirical data sets (see Subsection 4.2) we followed the same methodology, with the difference that the data set S_1 was an empirical data set rather than a generated one. In this case the survival function is estimated via a step function, known as the *Kaplan-Meier estimator* (Kaplan & Meier, 1958; Kleinbaum & Klein, 2012), making use of 10^6 bins.

For each set of experiments we followed the methodology described above for several possible alternative parametric distributions, D_2, and then computed the $\mathcal{ESJS}$ factor between the best and a lower performing distribution.

4.1 Experiments with Simulated Data

We now provide commentary on the results for the simulated data, shown in Tables 1, 2, 3, 4, 5, 6 and 7.

In the first experiment the given distribution was Normal with mean $\mu = 0$ and standard deviation $\sigma = 1$, and the hypothesised distributions were Normal and Uniform. The $\mathcal{ESJS}$ factor between the $\mathcal{ESJS}$s of the Normal and Uniform distributions is 803.9692, which can be derived from Table 1.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	0.0003	0.9995	**0.0002**	0.0002	0.0005
Uniform	-4.7467	4.8122	*0.1947*	0.1911	0.1950

Table 1: Experiment 1: Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

In the second experiment the given distribution was Log-normal with mean $\mu = 0$ and standard deviation $\sigma = 1$, and the hypothesised distributions were Normal, Uniform, Log-normal,

Gamma and Weibull. The $\mathcal{ESJS}$ factor between the $\mathcal{ESJS}$s of the Log-normal distribution, which is the smallest, and the Gamma distribution, whose $\mathcal{ESJS}$ is the closest to it, is 207.5271, which can be derived from Table 2.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	1.6492	2.1497	0.1520	0.1513	0.1527
Uniform	0.0054	104.6298	0.9001	0.8816	0.9005
Log-normal	0.0001	1.0006	**0.0002**	0.0002	0.0005
Gamma	1.1373	1.4501	*0.0481*	0.0477	0.0484
Weibull	1.0002	1.6494	0.0543	0.0540	0.0547

Table 2: Experiment 2: Log-normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

In the third experiment the given distribution was Gamma with shape $\alpha = 2$ and scale $\theta = 2$, and the hypothesised distributions were Normal, Uniform, Log-normal, Weibull and Gamma. The $\mathcal{ESJS}$ factor between the $\mathcal{ESJS}$s of the Gamma distribution, which is the smallest, and the Weibull distribution whose $\mathcal{ESJS}$ is the closest to it, is 85.8914, which can be derived from Table 3.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	3.9999	2.8282	0.0743	0.0741	0.0746
Uniform	0.0020	34.4431	0.5539	0.5137	0.5543
Log-normal	1.1164	0.8020	0.0308	0.0305	0.0311
Gamma	2.0031	1.9969	**0.0002**	0.0002	0.0005
Weibull	1.4831	4.4386	*0.0150*	0.0147	0.0153

Table 3: Experiment 3: Gamma distribution with shape $\alpha = 2$ and scale $\theta = 2$.

In the fourth experiment the given distribution was Gamma with shape $\alpha = 50$ and scale $\theta = 2$, and the hypothesised distributions were Normal, Uniform, Log-normal, Weibull and Gamma. The $\mathcal{ESJS}$ factor between the $\mathcal{ESJS}$s of the Gamma distribution, which is the smallest, and the Log-normal distribution whose $\mathcal{ESJS}$ is the closest to it, is 17.9510, which can be derived from Table 4.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	100.0013	14.1418	0.0132	0.0130	0.0135
Uniform	44.9060	186.2970	0.2073	0.2019	0.2103
Log-normal	4.5952	0.1421	*0.0059*	0.0057	0.0062
Gamma	50.0161	1.9994	**0.0003**	0.0002	0.0006
Weibull	7.2817	106.2044	0.04636	0.0460	0.04667

Table 4: Experiment 4: Gamma distribution with shape $\alpha = 50$ and scale $\theta = 2$.

In the fifth experiment the given distribution was Beta with parameters $\alpha = 2$ and $\beta = 2$, and the hypothesised distributions were Normal, Log-normal, Gamma, Weibull and Beta. The

$\mathcal{ESJS}$ factor between the Beta distribution which is the smallest, and the Normal distribution, whose $\mathcal{ESJS}$ is the closest to it, is 99.0703, which can be derived from Table 5.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	0.4999	0.2237	*0.0246*	0.0244	0.0248
Log-normal	-0.8340	0.6020	0.0583	0.0581	0.0586
Gamma	3.7158	0.1345	0.0399	0.0396	0.0401
Beta	1.9973	1.9988	**0.0002**	0.0002	0.0005
Weibull	2.3816	0.5628	0.0255	0.0253	0.0258

Table 5:   Experiment 5: Beta distribution with $\alpha = 2$ and $\beta = 2$.

In the sixth experiment the given distribution was Beta with parameters $\alpha = 50$ and $\beta = 50$, and the hypothesised distributions were Normal, Log-normal, Gamma, Weibull and Beta. The $\mathcal{ESJS}$ factor between the Beta distribution, which is the smallest, and the Normal distribution, whose $\mathcal{ESJS}$ is the closest to it, is 6.4432, which can be derived from Table 6.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	0.5000	0.0497	*0.0010*	0.0008	0.0013
Log-normal	-0.6982	0.1007	0.0128	0.0125	0.0131
Gamma	99.7537	0.0050	0.0086	0.0083	0.0088
Beta	50.0429	50.0496	**0.0002**	0.0002	0.0004
Weibull	10.6949	0.5225	0.0369	0.0365	0.0372

Table 6:   Experiment 6: Beta distribution with $\alpha = 50$ and $\beta = 50$.

In the seventh and final experiment the given distribution was Beta with parameters $\alpha = 60$ and $\beta = 30$, and the hypothesised distributions were Normal, Log-normal, Gamma, Weibull and Beta. The $\mathcal{ESJS}$ factor between the Beta distribution, which is the smallest, and the Normal distribution, whose $\mathcal{ESJS}$ is the closest to is, is 30.8739, which can be derived from Table 7.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Normal	0.6666	0.0494	*0.0064*	0.0062	0.0067
Log-normal	-0.4084	0.0750	0.0158	0.0156	0.0161
Gamma	179.5350	0.0037	0.0127	0.0125	0.0130
Beta	60.0885	30.0540	**0.0002**	0.0002	0.0005
Weibull	14.6275	0.6893	0.0326	0.0323	0.0329

Table 7:   Experiment 7: Beta distribution with $\alpha = 60$ and $\beta = 30$.

We note that all the computations described in this subsection were carried out using the Matlab software package. Table 8 summarises, for all experiments, the $\mathcal{ESJS}$ factors between the best (highlighted in bold) and second best (highlighted in italics) performing distributions. In all cases, apart from experiment 6 for the Beta distribution with $\alpha = 50$ and $\beta = 50$, shown

in Table 6, the $\mathcal{ESJS}$ factor overwhelmingly supports the given distribution, as we would expect. The reason for the relatively low $\mathcal{ESJS}$ factor in this particular case is the known fact that the Beta distribution can be approximated by the Normal distribution when α and β are large (Peizer & Pratt, 1968).

Experiment	Distribution	$\mathcal{ESJS}$ factor
1	Normal	803.9692
2	Log-normal	270.5271
3	Gamma	85.8914
4	Gamma	17.9510
5	Beta	99.0703
6	Beta	6.4432
7	Beta	30.8739

Table 8: The $\mathcal{ESJS}$ factors between the best and second best performing distributions for the simulated data.

It is evident that the larger the size of the data set the more accurate the $\mathcal{ESJS}$ will be. In Table 9 we demonstrate how the accuracy of the $\mathcal{ESJS}$ increases while the data set size increases, when both the given and hypothesised distribution are Normal with mean $\mu = 0$ and standard deviation $\sigma = 1$; it is also noticeable that the maximum likelihood estimation of parameters converges to the correct one as the size of the data increases. In Figure 1 we show that the decrease of the $\mathcal{ESJS}$, empirically, follows a power-law distribution with an exponent of approximately 0.5, which is $\sqrt{x}$, where x represents the data size.

Parameter 1	Parameter 2	$\mathcal{ESJS}$	Data set size
-0.3125	0.9571	0.0444	32
-0.0171	1.0037	0.0350	64
0.0098	1.0424	0.0233	128
-0.0153	1.0361	0.0197	256
-0.0119	1.0407	0.0112	512
-0.0022	1.0049	0.0079	1024
-0.0140	0.9894	0.0071	2048
-0.0084	0.9915	0.0050	4096
0.0178	0.9819	0.0042	8192
0.0091	0.9895	0.0033	16384
0.0090	0.9924	0.0018	32768
0.0076	0.9950	0.0017	65536
0.0053	0.9936	0.0011	131072
0.0061	0.9962	0.0007	262144
0.0028	0.9976	0.0004	524288
0.0012	0.9992	0.0003	1048576

Table 9: The $\mathcal{ESJS}$ resulting from increasing the data set size, where the given distribution is Normal with mean $\mu = 0$ and standard deviation $\sigma = 1$.
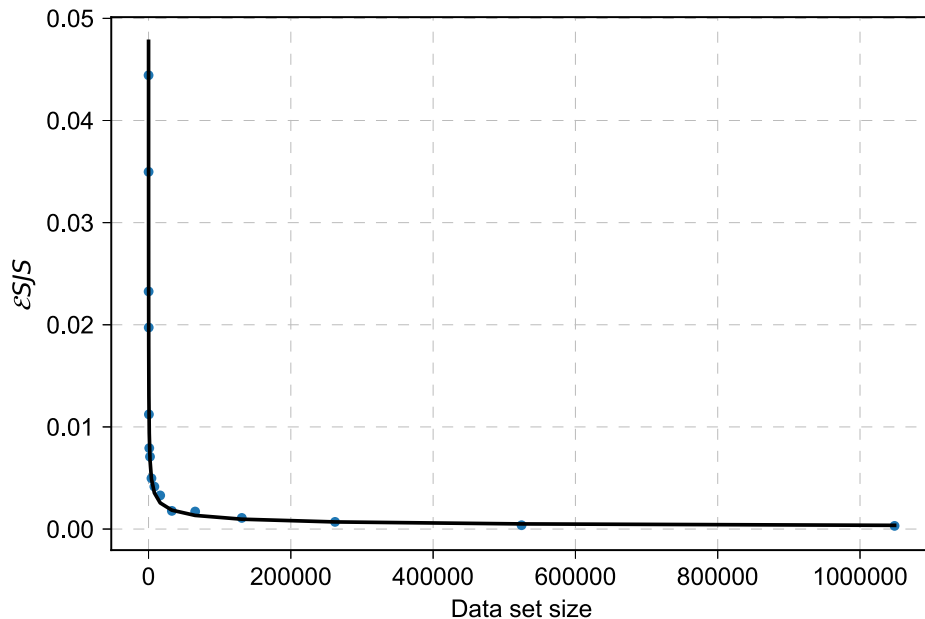
Figure 1: Power-law fit, $0.2426x^{-0.4691}$, for data in Table 9.

4.2 Analysis of Empirical Data

We now provide commentary on the results for the empirical data sets, shown in Tables 10, 11 and 12. We note that all the computations described in this subsection were carried out using Python. Table 13 summarises, for all three data sets, the $\mathcal{ESJS}$ factors between the best (highlighted in bold) and a lower performing distribution (highlighted in italics).

The first empirical data set we consider, contains detailed voting results of party vote shares in different polling stations, during the Lithuanian parliamentary election of 1992 (the data was obtained from (Kononovicius, 2017b)); for each party under consideration we have examined 2061 data points. For this data set we use the interval $(0,1)$ for quantifying the vote shares, since these are naturally bounded to this value range.

Note that we consider only the top three parties and have renormalised the original data so that the total vote share of the top three parties would sum to one in each polling station. This data set was first considered in (Kononovicius, 2017a), where an agent-based model generating the Beta distribution, and reasonably well reproducing detailed election results, was proposed. In (Kononovicius, 2018) a statistical comparison between the four commonly used distributions in sociophysics (Sen & Chakrabarti, 2014), the Normal, Log-normal, Beta and Weibull, was carried out using the Watanabe-Akaike information criterion ($WAIC$) (Watanabe, 2013), which is a generalisation of the AIC. The comparison concluded that the Beta and Weibull distributions provide the best fits for the empirical data. However, their respective WAIC scores were within each other's confidence intervals, and therefore no final conclusion was made. Here we also obtain a similar result, the Beta and Weibull distributions clearly have the overall best scores, however, as before, their confidence intervals overlap (see Table 10). As

11

was noted in (Kononovicius, 2018), the Beta and Weibull distributions are similar when the observed mean is close to 0.5 and the observed variance is reasonably small. In the empirical analysis this similarity is further increased when the sample size is small. In addition, for the estimated parameter values, the Gamma and Weibull distributions behave similarly when $x \in (0, 1)$. Therefore, we report the $\mathcal{ESJS}$ factor between the best performing distribution (highlighted in bold) and the next best distribution which is neither a Beta, Gamma nor a Weibull distribution (see Table 13).

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
SK – Sąjūdžio koalicija					
Normal	0.2412	0.1132	0.0265	0.0229	0.0307
Log-normal	-1.5556	0.5721	*0.0243*	0.0203	0.0290
Gamma	3.9083	0.0617	0.0126	0.0097	0.0161
Beta	3.0771	9.7045	0.0091	0.0067	0.0125
Weibull	2.2492	0.2722	**0.0089**	0.0068	0.0127
LKDP – Lietuvos krikščionių demokratų partija					
Normal	0.1575	0.0921	0.0308	0.0276	0.0339
Log-normal	-2.0435	0.6884	*0.0168*	0.0138	0.0202
Gamma	2.7196	0.0579	0.0060	0.0044	0.0084
Beta	2.3282	12.4475	**0.0057**	0.0044	0.0082
Weibull	1.7869	0.1773	0.0081	0.0060	0.0110
LDDP – Lietuvos demokratinė darbo partija					
Normal	0.6013	0.1516	*0.0169*	0.0135	0.0236
Log-normal	-0.5459	0.2892	0.0525	0.0468	0.0582
Gamma	13.5242	0.0445	0.0423	0.0361	0.0485
Beta	5.4454	3.6014	0.0106	0.0093	0.0181
Weibull	4.5344	0.6588	**0.0089**	0.0081	0.0172

Table 10: Empirical data set 1: Lithuanian parliamentary election 1992.

The second data set we consider contains the log-returns of two different exchange rates. We consider the BTC/JPY exchange rate during the time period between July 4, 2017 and July 4, 2018 on the bitFlyer exchange (the data was obtained from (Bitcoincharts.com, 2018)), as well as the EUR/USD exchange rate during the time period between June 1, 2000 and September 1, 2010 (the data was obtained from (HistData.com, 2012)). We have taken into consideration the daily and one minute log-returns. For this data set we use the interval $[-100, 100]$ to quantify the returns for the unbounded distributions (Normal and q-Gaussian distributions), and the interval $[0.01, 100]$ to quantify the returns for the bounded distributions (Log-normal, Gamma and Weibull distributions). We have selected this value range, since the maximum observed absolute return value in this data set is approximately 85. The number of return data points we examine is as follows: 364 for the daily BTC/JPY exchange rate, 526541 for the 1 min BTC/JPY exchange rate, 2544 for the daily EUR/USD exchange rate and 3665755 for the 1 min EUR/USD exchange rate. For this data set we employ the moving block bootstrap (Kreiss & Lahiri, 2012) with a block size of one day.

In the econophysics literature (Slanina, 2014) it is commonly accepted that the log-returns are power-law distributed (Cont, 2001). One of the commonly used fits for the log-returns is

so-called q-Gaussian distribution (Tsallis, 2017), which we add to our analysis for this empirical data set.

However, as can be seen in Table 11, we find that the Gamma and Weibull distributions noticeably outperform the q-Gaussian distribution. Performance of the Gamma and Weibull distributions is similar, due to the fact that for the estimated parameter values both of these distributions behave reasonably similarly; for these parameter values they are reasonably close to the Exponential distribution. Therefore, we report the $\mathcal{ESJS}$ factor between the best performing distribution (highlighted in bold) and the next best distribution, which is neither a Gamma nor a Weibull distribution (see Table 13).

For our fourth sample in the second empirical data set, i.e. the EUR/USD one minute log-returns, unexpectedly the Log-normal and q-Gaussian distributions had the best performance. Though they are far from being similar for the considered observable value range and parameter values, they, most likely attained similar scores due to the shape of the empirical distribution. The Log-normal distribution seems to represent smaller log-returns well, while the q-Gaussian is better at describing the tail events.

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
Daily – BTC/JPY (bitFlyer)					
Normal	0.0415	1.0000	0.0099	0.0055	0.0144
Log-normal	-0.9039	1.1431	*0.0030*	0.0017	0.0050
Gamma	1.1015	0.6174	**0.0020**	0.0010	0.0028
Weibull	1.0276	0.6882	0.0024	0.0010	0.0030
q-Gaussian	3.9028	1.0413	0.0061	0.0048	0.0116
One minute – BTC/JPY (bitFlyer)					
Normal	0.0011	1.0000	0.0140	0.0121	0.0158
Log-normal	-1.2561	1.5771	*0.0075*	0.0070	0.0077
Gamma	0.7717	0.7993	**0.0033**	0.0026	0.0029
Weibull	0.8468	0.5654	0.0034	0.0028	0.0031
q-Gaussian	2.8964	0.5871	0.0082	0.0069	0.0097
Daily – EUR/USD (Forex)					
Normal	0.0179	1.0000	0.0038	0.0026	0.0053
Log-normal	-0.7227	1.1086	0.0049	0.0042	0.0052
Gamma	0.6102	1.2489	0.0010	0.0007	0.0014
Weibull	1.1579	0.8021	**0.0006**	0.0004	0.0011
q-Gaussian	7.8686	2.2146	*0.0022*	0.0018	0.0039
One minute – EUR/USD (Forex)					
Normal	0.0004	1.0000	0.0138	0.0136	0.0142
Log-normal	-0.3236	0.6165	**0.0086**	0.0085	0.0087
Gamma	2.3423	0.3881	0.0116	0.0113	0.0116
Weibull	1.3548	1.0057	0.0145	0.0142	0.0144
q-Gaussian	2.9820	0.6415	*0.0092*	0.0087	0.0093

Table 11: Empirical Data Set 2: Daily and one minute log-returns of EUR/USD and BTC/JPY exchange rates.

Similar conclusions follow from the WAIC scores. Using the same methodology as in (Kononovicius, 2018) we have also calculated WAIC scores for the second empirical data set. For the first three samples in the data set, we observe that the Gamma and Weibull distributions have the best WAIC scores and, in addition, have overlapping confidence intervals. Moreover, in all three cases, the Log-normal distribution has the next best WAIC score. For the fourth sample the Log-normal distribution has the lowest score, while the q-Gaussian distribution is the next best.

The third data set we consider is the European soccer data set (Mathien, 2016), which contains 25 thousand matches played in European national championships throughout 2008–2016. From this data set we have extracted five random teams and computed inter–goal times for each team. We have used the interval $[1, 505]$ to quantify the inter–gaol times, since the longest inter–goal time observed in the considered series was 501 minutes. The number of inter–goal times under consideration for each team is as follows: 471 for TOT, 397 for GLA, 564 for MUN, 475 for VAL and 64 for ELC.

We introduce two new distributions for the third data set analysis. First of all we include exponential distribution, which should arise if goal scoring rate is independent of previous goal scoring history. Yet, if goal scoring rate is increasing as more goals are scored, we could expect to observed power–law distribution. Power–law inter–event time distributions are observed in many social systems, e.g., human communication time series exhibit such pattern Barabási (2005).

We have treated goals scored during extra time as scored on the 45th minute (if scored during the first half) and the 90th minute (if scored during the second half). In this analysis we have added the Exponential and Pareto distributions. For the estimated parameter values the Gamma and Weibull distributions behave similarly to the Exponential distribution. Note that the shape parameter values of the Gamma and Weibull distributions are very close to 1 and the respective scale parameter values are similar. In this case it is known that Gamma and Weibull distributions are equivalent to the Exponential distribution with the appropriate scale parameter value. We therefore report the $\mathcal{ESJS}$ factor between the best performing distribution (highlighted in bold) and the next best distribution, which is neither an Exponential, Gamma nor a Weibull distribution (see Table 13). We observe that for the ELC sample, the obtained $\mathcal{ESJS}$ factor is the lowest and the $\mathcal{ESJS}$ score is the largest. This is most likely due to this team having played opponents with a larger variety of skill. In particular, it played in the top and the second tiers of the national championship during the considered time period, resulting in a goal scoring rate with a higher variation.

Similar conclusions follow from the WAIC scores. Using the same methodology as in (Kononovicius, 2018) we have also calculated WAIC scores for the third empirical data set. For all samples in the third the data set, we observe that the Gamma, Weibull and Exponential distributions have the best WAIC scores and, in addition, have overlapping confidence intervals. We note that, for the ELC sample, using WAIC, we have not observed any significant differences between the distributions; this is likely, since the sample size for the ELC is small (i.e. 64 goals).

5 Concluding remarks

We have proposed the empirical survival Jensen-Shannon divergence ($\mathcal{ESJS}$) as a goodness-of-fit measure for data fitted with maximum likelihood estimation or curve fitting. Our ex-

Distribution	Parameter 1	Parameter 2	$\mathcal{ESJS}$	lb	ub
TOT – Tottenham Hotspur (English Premier League)					
Normal	57.7346	58.2559	0.0381	0.0326	0.0423
Gamma	1.0565	54.6477	0.0071	0.0055	0.0109
Weibull	1.0183	58.1866	0.0069	0.0054	0.0102
Exponential	57.7346	-	**0.0066**	0.0057	0.0115
Pareto	1.6809	-	*0.0249*	0.0200	0.0310
GLA – Borussia Monchengladbach (German Bundesliga)					
Normal	61.5088	61.3039	0.0390	0.0334	0.0424
Gamma	1.0186	60.3877	0.0075	0.0063	0.0111
Weibull	1.0012	61.5400	0.0072	0.0062	0.0106
Exponential	61.5088	-	**0.0071**	0.0062	0.0127
Pareto	1.6850	-	*0.0235*	0.0187	0.0303
MUN – Manchester United (English Premier League)					
Normal	48.3174	49.8385	0.0366	0.0315	0.0394
Gamma	1.0497	46.0306	0.0070	0.0051	0.0101
Weibull	1.0084	48.4960	0.0068	0.0050	0.0094
Exponential	48.3174	-	**0.0067**	0.0050	0.0103
Pareto	1.6886	-	*0.0242*	0.0203	0.0301
VAL – Valencia CF (Spanish La Liga)					
Normal	57.2779	52.8406	0.0334	0.0284	0.0365
Gamma	1.1212	51.0879	**0.0053**	0.0048	0.0090
Weibull	1.0725	58.8645	0.0054	0.0047	0.0093
Exponential	57.2779	-	0.0066	0.0051	0.0122
Pareto	1.6547	-	*0.0287*	0.0238	0.0347
ELC – Elche CF (Spanish La Liga/La Liga 2)					
Normal	102.1875	87.1523	0.0402	0.0263	0.0658
Gamma	1.3220	77.2953	0.0173	0.0161	0.0349
Weibull	1.1929	108.4610	**0.0158**	0.0157	0.0332
Exponential	102.1875	-	0.0281	0.0192	0.0581
Pareto	1.6154	-	*0.0400*	0.0257	0.0632

Table 12: Empirical data set 3: Inter–goal times recorded in the European soccer data set.

Data set	Sample	Distribution	$\mathcal{ESJS}$ factor
1	SK	Weibull	2.7157
1	LKDP	Beta	2.9374
1	LDDP	Weibull	1.8943
2	Daily BTC/JPY	Gamma	1.5027
2	1 min BTC/JPY	Gamma	2.2476
2	Daily EUR/USD	Weibull	3.3437
2	1 min EUR/USD	Log-normal	1.0564
3	TOT	Exponential	3.7628
3	GLA	Exponential	3.2758
3	MUN	Exponential	3.6089
3	VAL	Weibull	5.2841
3	ELC	Weibull	2.5242

Table 13: The $\mathcal{ESJS}$ Factors between the best and a lower performing distribution for the empirical data sets.

periments with simulated and empirical data in Section 4, for a variety of parametric distributions commonly employed in sociophysics and econophysics, show that for simulated data the method is unequivocal in its preference for the true distribution (see Subsection 4.1), and for empirical data the method is effective in selecting the more likely distributions from a selection of hypothesised distributions (see Subsection 4.2).

As we have shown in Section 2 the $\mathcal{ESJS}$ can be formally defined, building on the survival entropy as a generalisation of the standard Jensen-Shannon divergence, and the $\mathcal{ESJS}$ factor has an intuitive meaning in terms of an odds ratio, in analogy to the Bayes factor. Moreover, the implementation of the $\mathcal{ESJS}$ as a measure of goodness-of-fit or for model comparison is relatively straightforward; see (Kononovicius & Levene, 2018) for a Python implementation of the $\mathcal{ESJS}$.

Ultimately, for a definitive assessment of how the $\mathcal{ESJS}$ performs in practice, an extensive comparison with existing goodness-of-fit measures should be carried out, and more experience with empirical data sets is needed.

References

Anderson-Sprecher, R. 1994. Model comparisons and R^2. *The american statistician*, **48**, 113–117.

Barabási, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. *Nature*, **435**, 207–211.

Bitcoincharts.com. 2018. `http://api.bitcoincharts.com/v1/csv/`.

Burk, F. 2007. *A garden of integrals*. Providence, RI: The Mathematical Association of America.

Burnham, K.P., & Anderson, D.R. 2002. *Model selection and multimodel inference: A practical information-theoretic approach.* second edn. New York: Springer-Verlag.

Cameron, A.C., & Windmeijer, F.A.G. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, **77**, 329–342.

Capiński, M., & Kopp, E. 2004. *Measure, integral and probability.* second edn. Springer Undergraduate Mathematics Series. Berlin: Springer-Verlag.

Cont, R. 2001. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative finance*, **1**, 1–14.

Cover, T.M., & Thomas, J.A. 2006. *Elements of information theory.* second edn. Wiley Series in Telecommunications. Hoboken, New Jersey: John Wiley & Sons.

Davison, A.C., & Hinkley, D.V. 1997. *Bootstrap methods and their applications.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press.

Di Crescenzo, A., & Longobardi, M. 2009. On cumulative entropies and lifetime estimations. *Pages 132–141 of: Proceedings of international work-conference on the interplay between natural and artificial computation (iwinac).*

Endres, D., & Schindelin, J. 2003. A new metric for probability distributions. *Ieee transactions on information theory*, **49**, 1858–1860.

Fenner, T., Levene, M., & Loizou, G. 2018. A multiplicative process for generating the rank-order distribution of UK election results. *Quality & quantity*, **52**, 1069âĂŞ–1079.

Fox, J. 2016. *Applied regression analysis and generalized linear models.* 3rd edn. Thousand Oaks, Ca.: Sage Publications.

Freund, R.J., Wilson, W.J., & Sa, P. 2006. *Regression analysis: Statistical modeling of a response variable.* Second edn. San Diego, CA.: Academic Press.

Gómez-Lopera, J.F., Martínez-Aroza, J., Robles-Pérez, A.M., & Román-Roldán, R. 2000. An analysis of edge detection by using the Jensen-Shannon divergence. *Journal of mathematical imaging and vision*, **13**, 35–56.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Ozair, D. Warde-Farley S., Courville, A., & Bengio, Y. 2014. Generative adversarial nets. *Pages 2672–2680 of: Proceedings of advances in neural information processing systems (nips) 27.*

Heinrich, J. 2003. Pitfalls of goodness-of-fit from likelihood. *Pages 52–55 of: Proceedings of the conference on statisical problems in particle physics, astrophysics and cosmology (phystat2003).*

HistData.com. 2012. `http://www.histdata.com/download-free-forex-data/`.

Jarosz, A.F., & Wiley, J. 2014. What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of problem solving*, **7**, Article 2, 8pp.

Kaplan, E.L., & Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the american statistical association*, **53**, 457–481.

Kass, R.E., & Raftery, A.E. 1995. Bayes factors. *Journal of the american statistical association*, **90**, 773–795.

Kleinbaum, D.G., & Klein, M. 2012. *Survival analysis, a self-learning text.* third edn. New York, NY: Springer Science+Business Media, LLC.

Kononovicius, A. 2017a. Empirical analysis and agent-based modeling of the Lithuanian parliamentary elections. *Complexity*, **2017**, 7354642.

Kononovicius, A. 2017b. *Lithuanian parliamentary election data.* `http://github.com/akononovicius/lithuanian-parliamentary-election-data`.

Kononovicius, A. 2018. Modeling of the parties' vote share distributions. *Acta physica polonica a*, **133**, 1450–1458.

Kononovicius, A., & Levene, M. 2018. *PyJSD: Python implementation of the Jensen-Shannon divergence.* `http://github.com/akononovicius/pyjsd`.

Kreiss, J.-P., & Lahiri, S.N. 2012. Bootstrap methods for time series. *Pages 3–26 of:* Rao, T.S., Rao, S.S., & Rao, C.R. (eds), *Handbook of statistics volume 30, time series analysis: Methods and applications.* Oxford: North-Holland.

Krishnamoorthy, K. 2015. *Handbook of statistical distributions with applications.* Second edn. Boca Raton, FL: CRC Press.

Lewis, F., Butler, A., & Gilbert, L. 2011. A unified approach to model selection using the likelihood ratio test. *Methods in ecology and evolution*, **2**, 155–162.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *Ieee transactions on information theory*, **37**, 145–151.

Majtey, A.P., Lamberti, P.W., & Prato, D.P. 2005. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. *Physical review a*, **72**, 052310–1–052310–6.

Mathien, H. 2016. *European soccer data set.* `http://www.kaggle.com/hugomathien/soccer`.

Mehri, A., Jamaati, M., & Mehri, H. 2015. Word ranking in a single document by Jensen–Shannon divergence. *Physics letters a*, **379**, 1627–1632.

Motulsky, H. 1995. *Intuitive biostatistics.* Oxford: Oxford University Press.

Myung, I.J. 2003. Tutorial on maximum likelihood estimation. *Journal of mathematical psychology*, **47**, 90–100.

Nadarajaha, S., & Kotz, S. 2007. On the q-type distributions. *Physica a*, **377**, 465–468.

Nakagawa, S., & Schielzeth, H. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, **4**, 133–142.

Nguyen, H.-V., & Vreeken, J. 2015. Non-parametric Jensen-Shannon divergence. *Pages 173–189 of: Proceedings of european conference on machine learning and principles and practice of knowledge discovery in databases (ecml pkdd)*.

Orelien, J.G., & Edwards, L.J. 2008. Fixed-effect variable selection in linear mixed models using R^2 statistics. *Computational statistics & data analysis*, **52**, 1896–1907.

Peizer, D.B., & Pratt, J.W. 1968. A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. *Journal of the american statistical association*, **63**, 1416–1456.

Pitt, M.A., & Myung, I.J. 2002. When a good fit can be bad. *Trends in cognitive sciences*, **6**, 421–425.

Rao, M., Chen, Y., & Vemuri, B.C. 2004. Cumulative residual entropy: A new measure of information. *Ieee transactions on information theory*, **50**, 1220–1228.

Sapra, R.L. 2014. Using R^2 with caution. *Current medicine research & practice*, **4**, 130–134.

Sen, P., & Chakrabarti, B.K. 2014. *Sociophysics: An introduction*. Oxford: Oxford University Press.

Shorack, G.R. 2017. *Probability for statisticians*. second edn. Springer Texts in Statistics. Basel: Springer-Verlag.

Sims, G.E., Jun, S.-R., Wu, G.A., & Kim, S.-H. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the national academy of sciences of the united states of america*, **106**, 2677âĂŞ–2682.

Slanina, F. 2014. *Essentials of econophysics modelling*. Oxford: Oxford University Press.

Spiess, A.-N., & Neumeyer, N. 2010. An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *Bmc pharmacology*, **10**. 11 pages.

Stephens, M.A. 1986. Test based on EDF statistics. *Chap. 4, pages 97–193 of:* D'Agostino, R.B., & Stephens, M.A. (eds), *Goodnes-of-fit techniques*. New York, NY: Marcel Dekker.

Tsallis, C. 2017. Economics and finance: q-Statistical stylized features galore. *Entropy*, **19**(9), 457.

Vrieze, S.I. 2012. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, **17**, 228–243.

Vuong, Q.H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.

Watanabe, S. 2013. A widely applicable Bayesian information criterion. *Journal of machine learning research*, **14**, 867–897.

Yari, G., & Saghafi, A. 2012. Unbiased Weibull modulus estimation using differential cumulative entropy. *Communications in statistics – simulation and computation*, **41**, 1372–1378.