Kent Academic Repository

Full text document (pdf)

Citation for published version

Leisen, Fabrizio and Rossini, Luca and Villa, Cristiano (2018) Objective Bayesian Analysis of the Yule-Simon Distribution with Applications. Computational Statistics, 33 . pp. 99-126. ISSN 0943-4062.

DOI

https://doi.org/10.1007/s00180-017-0735-1

Link to record in KAR

http://kar.kent.ac.uk/61711/

Document Version

Author's Accepted Manuscript

Objective Bayesian Analysis of the Yule–Simon Distribution with Applications

Fabrizio Leisen[a]   Luca Rossini[b,c]   Cristiano Villa[a]

[a] University of Kent, U.K.   [b] Free University of Bozen, Italy

[c] Ca' Foscari University of Venice, Italy

Abstract

The Yule–Simon distribution is usually employed in the analysis of frequency data. As the Bayesian literature, so far, has ignored this distribution, here we show the derivation of two objective priors for the parameter of the Yule–Simon distribution. In particular, we discuss the Jeffreys prior and a loss-based prior, which has recently appeared in the literature. We illustrate the performance of the derived priors through a simulation study and the analysis of real datasets.

Keywords: Kullback-Leibler divergence, Loss-based prior, Objective Bayes, Social Network daily returns, Text Analysis.

1 Introduction

In this work we aim to fill a gap in the Bayesian literature by proposing two objective priors for the parameter of the Yule–Simon distribution. The distribution was firstly discussed in Yule (1925) and then re-proposed in Simon (1955), and can be used in scenarios where the center of interest is some sort of frequency in the data. For example, Yule (1925) used it to model abundance of biological genera, while Simon (1955) exploited the distribution properties to model the addition of new words to a text. It goes without saying that other areas of applications can be considered where, for instance, frequencies represent the elementary unit of observation. For example, Gallardo et al. (2016) highlight that the heavy-tailed property of the Yule–Simon distribution allows for extreme values even for small sample sizes. In particular, they claim that the above property is suitable to model short survival times which, due to the nature of the problem, happen with relatively high frequency. In this paper we show the employment of the Yule–Simon distribution in modelling daily increments of social network stock options, surnames and 'superstar' success in the music industry.

Despite the wide range of applications, the literature on the Yule–Simon distribution appears to be limited. And, more surprisingly, to the best of our knowledge it seems that no attention has been given to the problem by the Bayesian community. Leisen et al. (2017) propose an explicit Gibbs sampling scheme when a Gamma prior is chosen for the shape parameter and illustrate this approach with an application to text

analysis and count data regression. Given the challenges that classical inference faces in estimating the parameter of the distribution (Garcia Garcia, 2011), the possibility of tackling the problem from a Bayesian perspective is, undoubtedly, appealing.

In addressing the estimation of the shape parameter of the Yule–Simon distribution by means of the Bayesian framework, we opted for an objective approach. We propose two priors: the first is the Jeffreys rule prior (Jeffreys, 1961), while the second is obtained by applying the loss-based approach discussed in Villa and Walker (2015). We would like to highlight that the purpose of this work is to present a Bayesian objective inferential procedure for the Yule-Simon distribution, without discussing whether the model is the best choice to represent some given set of observation. Although we formally introduce the Yule–Simon distribution and its derivation in the next Section, it is important to give an anticipation of the general idea here, so to fully appreciate the gain in adopting an objective approach. As nicely illustrated in Chung and Cox (1994), the shape parameter of the distribution is linked via a one-to-one transformation to the probability that the next observation will not take a value previously observed. For example, if we have observed n words in a text, we wish to make inference on the probability that the $(n + 1)$ observation is a word not yet encountered in the text, assuming this probability to be constant. It is then clear that the Yule–Simon distribution models extremely large events. As such, the information in the data about these events is limited and a "wrongly" elicited prior could end up dominating the data. On the other hand, a prior with minimal information content would allow the data "to speak", resulting in a more robust inferential procedure. We do not advocate that in every circumstance an objective approach is the only suitable. In fact, if reliable prior information is available, an elicited prior would represent, in general, the natural choice. Alas, in the presence of phenomena with extremely rare events, the above information is often insufficient or incomplete, and an objective choice would then represent the most sensible one.

The paper is organized as follows. In Section 2 we set the scene by introducing the Yule–Simon distribution and the notation that will be used throughout the paper. The proposed objective priors are derived and discussed in Section 3. Section 4 collects the analysis of the frequentist performances of the posterior distributions yielded by the proposed priors. Through a set of several simulation scenarios, we compare and analyse the inferential capacity of the objective priors here discussed. In Section 5 we illustrate the application of the priors to three real-data applications. Finally, Section 6 is reserved to concluding remarks and points of discussion.

2   Preliminaries

The most known functional form of the Yule–Simon distribution, possibly, is the following:

$$f(k; \rho) = \rho \, \mathrm{B}(k, \rho + 1), \qquad k = 1, 2, \ldots \text{ and } \rho > 0, \qquad (1)$$

where $B(\cdot, \cdot)$ is the beta function and ρ is the shape parameter. Note that the expectation of the Yule-Simon distribution exists for $\rho > 1$ and is equal to $\rho/(\rho - 1)$. The distribution in (1) was firstly proposed by Yule (1925) in the field of biology; in particular, to represent the distribution of species among genera in some higher taxon of biotic organisms. More recently, Simon (1955) noticed that the above distribution can be observed in other phenomena, which appear to have no connection among each others. These include, the distribution of word frequencies in texts, the distribution of authors by number of scientific articles published, the distribution of cities by population and the distribution of incomes by size. The derivation process followed by Yule (1925) was based on word frequencies, and it consisted of two assumptions:

(i) The probability that the $(n + 1)$-th word is a word observed k times in the first n words, is proportional to k; and

(ii) The probability that the $(n + 1)$-th word is new (i.e. not being observed in the first n words) is constant and equal to $\alpha \in (0, 1)$.

Yule (1925) shows that, under the condition of stationarity, the process defined by the above two assumptions yields (1) by setting $\rho = 1/(1 - \alpha)$, obtaining:

$$f(k; \alpha) = \frac{1}{1 - \alpha} B\left(k, \frac{1}{1 - \alpha} + 1\right). \tag{2}$$

An important consequence of the above assumption (ii) is that the shape parameter ρ of the distribution takes values in $(1, +\infty)$. In other words, should we use the model as in Yule (1925), which includes the possibility that $0 < \rho \leq 1$, we would loose the interpretation of the generating process described by the two assumptions above. In fact, for $\rho < 1$, the probability of observing a new word, expressed by α, would be negative; while for $\rho = 1$ the probability would be zero, rendering the process trivial (i.e. all the observed words will be equal to the first one observed). Furthermore, the expectation of the Yule–Simon distribution is defined only for values of the shape parameter larger than one, and this property is something one would expect in most applications. For all the above reasons, in this work we focus on the parametrization of the Yule-Simon given in (2), that is we will discuss prior distributions for α.

In addition to the parametrization of the Yule–Simon distribution as in (2), we will also consider the possibility of having the parameter α discrete. This is a common finding in literature, especially when implementations of the model are considered. See, for example, Simon (1955) and Garcia Garcia (2011). The discretization of α will be discussed in detail in Section 3.

3 Objective Priors for the Yule-Simon distribution

This section is devoted to the derivation of two objective priors for the Yule-Simon distribution: the Jeffreys prior and loss-based prior. The former assumes that

parameter space of α is continuous and it is based on the well-known invariance property proposed by Jeffreys (1961); the latter assumes the parameter space discrete and is based on Villa and Walker (2015). Based on the previous description of the Yule–Simon distribution, we consider the following Bayesian model,

$$k_1, \ldots, k_n | \alpha \overset{i.i.d}{\sim} f(k; \alpha)$$
$$\alpha \sim \pi(\alpha)$$

where $f(k; \alpha)$ is the Yule–Simon distribution described in 2 and $\pi(\alpha)$ is the objective prior distribution of the parameter of interest α, either the Jeffreys or the loss-based prior.

The likelihood function of the above model, conditionally to the parameter α, is the following:

$$L(\mathbf{k}|\alpha) = \prod_{i=1}^{n} f(k_i, \alpha) = \prod_{i=1}^{n} \frac{1}{1-\alpha} \mathrm{B}\left(k_i, \frac{1}{1-\alpha} + 1\right),$$

where $\mathbf{k} = (k_1, \ldots, k_n)$ is the vector of observations. In order to compute the Bayesian analysis of the model, we consider the posterior distribution for α:

$$\pi(\alpha|\mathbf{k}) \propto L(\mathbf{k}|\alpha)\pi(\alpha),$$

where $\pi(\alpha)$ is the objective prior as described in the following part of the section.

The Jeffreys Prior

The Jeffreys prior is defined in the following way (Jeffreys, 1961):

$$\pi(\alpha) \propto \sqrt{\mathcal{I}(\alpha)}$$

where $\mathcal{I}(\alpha) = \mathbb{E}_\alpha\left[-\frac{\partial^2 \log(f(k;\alpha))}{\partial \alpha^2}\right]$ is the Fisher Information. In the next Theorem (which proof is in the Appendix) an explicit expression of the Jeffreys prior for the Yule-Simon distribution is provided.

Theorem 3.1. *Let $f(k; \alpha)$ be the Yule-Simon distribution defined in equation (2), with $0 < \alpha < 1$. The Jeffreys prior for α is*

$$\pi(\alpha) \propto q(\alpha), \tag{3}$$

where

$$q(\alpha) = \frac{1}{1-\alpha}\sqrt{1 - \frac{1}{(2-\alpha)^2} {}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)}.$$

with ${}_3F_2$ being the hypergeometric distribution function.

4

The Jeffreys prior stated in Theorem 3.1 is a proper prior. In fact, let

$$\pi(\alpha) = \frac{q(\alpha)}{K},$$

where

$$K = \int_0^1 \frac{1}{1-\alpha}\sqrt{1 - \frac{1}{(2-\alpha)^2} \, {}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)} d\alpha$$

is the normalizing constant of $\pi(\alpha)$. It is not difficult to prove that

$$K < \infty.$$

Indeed,

$$K \leq \int_0^1 \sqrt{\frac{3-\alpha}{1-\alpha}} \frac{1}{2-\alpha} d\alpha = \frac{1}{3}\pi - \ln(2 - \sqrt{3}) < \infty.$$

The result above follows from the following inequality

$${}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right) \geq 1.$$

The properness of the prior in (3) ensures the properness of the yielded posterior distribution for α, as such suitable for inference.

The Loss-based Prior

Villa and Walker (2015) introduced a method for specifying and objective prior for discrete parameters. The idea is to assign a *worth* to each parameter value by objectively measuring what is lost if the value is removed, and it is the true one. The loss is evaluated by applying the well known result in Berk (1966) stating that, if a model is misspecified, the posterior distribution asymptotically accumulates on the model which is the nearest to the true one, in terms of the Kullback–Leibler divergence.

Given that the parameter $\alpha \in (0,1)$ of the Yule–Simon is in principle continuous, the above method can not be applied. However, the boundedness of the interval allows for an easy discretization, directly we can consider the set

$$\mathbb{D}_M = \left\{\alpha = \frac{i}{M} : i = 1, \ldots, M-1\right\}.$$

Therefore, the *worth* of the parameter value α is represented by the Kullback–Leibler divergence $D_{KL}(f(k|\alpha)\|f(k|\alpha'))$, where $\alpha' \neq \alpha$ is the parameter value that minimizes the divergence. To link the *worth* of a parameter value to the prior mass, Villa and Walker (2015) use the self-information loss function. This particular type of loss function measures the loss in information contained in a probability statement (Merhav and Feder, 1998). As we now have, for each value of α, the loss in information measured

in two different ways, we simply equate them obtaining the loss-based prior of Villa and Walker (2015):

$$\pi(\alpha) \propto \exp\left\{\min_{\alpha' \neq \alpha} D_{KL}(f(k|\alpha)\|f(k|\alpha'))\right\} - 1 \qquad \alpha, \alpha' \in \mathbb{D}_M, \tag{4}$$

where

$$D_{KL}(f(k|\alpha)\|f(k|\alpha')) = \log\left(\frac{1-\alpha'}{1-\alpha}\right) + \mathbb{E}_\alpha\left\{\log\left[B\left(k; \frac{1}{1-\alpha}+1\right)\right]\right\}$$
$$- \mathbb{E}_\alpha\left\{\log\left[B\left(k; \frac{1}{1-\alpha'}+1\right)\right]\right\}.$$

As the discretized parameter space is finite, no matter what value of M one chooses, the prior (4) is proper, hence, the yielded posterior will be proper as well.

An important aspect is that the value α' minimizing the Kullback–Leibler divergence can not be analytically determined, and the prior has to be computationally derived. However, even for large values of M, the computational cost is trifling compared to the whole Monte Carlo procedure necessary to simulate from the posterior distribution.

(a)   (b)
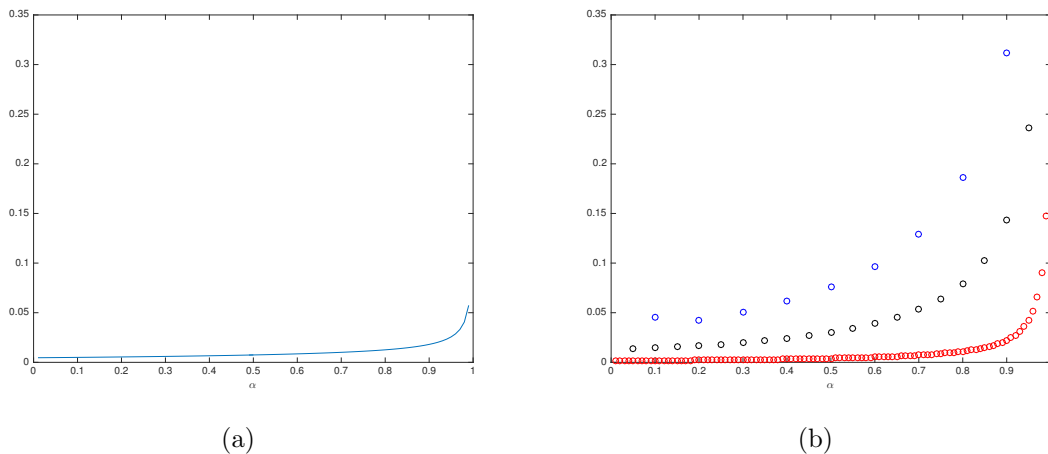
Figure 1: Prior distribution for α in panel obtained by applying, in panel (a), Jeffreys rule, while, in panel (b), the loss-based method with $M = 10$ (blue dots), with $M = 20$ (black dots) and with $M = 100$ (red dots).

To have a feeling of the prior distributions derived above, we have plotted them in Figure 1. The behaviour of the priors is similar, in the sense that they tend to increase as α increases.

4 Simulation Study

The objective priors defined in Section 3 are automatically derived by taking into consideration properties intrinsic to the Yule–Simon distribution. In other words, they do not depend on experts knowledge or previous observations. It is therefore necessary, in order to validate them, to assess the goodness of the priors by making inference on simulated data. This section is dedicated in performing a simulation study on the parameter α using observations obtained from fully known distributions.

We have considered different sample sizes, $n = 30$, $n = 100$ and $n = 500$, to analyse the behaviour of the prior distributions under different level of information coming from the data. Here we show the results for $n = 100$ only, as the sole differences in using $n = 30$ and $n = 500$ sample sizes are limited to the precision of the inferential results: relatively low for $n = 30$ and relatively high for $n = 500$, as one would expect. Besides that, the differences in the performance of the two priors noted for $n = 100$ remain for the other sample sizes. As the loss-based prior depends on the discretization of the parameter space, for illustration purposes, we have considered $M = 10$ and $M = 20$, that is $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$ and $\alpha \in \{0.05, 0.10, \ldots, 0.95\}$, respectively.

Both the Jeffreys prior and the loss-based prior yield posterior distributions for α which are not analytically tractable, hence, it is recommendable to use Monte Carlo methods. The reasons behind this choice are discussed in the following two remarks.

Remark 1. In the case of the Jeffreys' prior, one may use univariate numerical integration to compute quantities of interest, such as mean and credible intervals. However, in our experience, numerical integration may fail when the complexity on the integrand increases.

Remark 2. For the Loss based prior, one may be tempted to compute the posterior probabilities as

$$\pi(\alpha_j | \boldsymbol{k}) = \frac{L(\boldsymbol{k} | \alpha_j) \pi(\alpha_j)}{\sum_{\alpha_i \in \mathbb{D}_M} L(\boldsymbol{k} | \alpha_i) \pi(\alpha_i)}.$$

However, the Kullback-divergence in equation (4) may take values close to zero, leading to probabilities which can be interpreted as zero by the computational software employed. This is particularly true for moderately large values of M. Therefore, a Metropolis-Hastings in the logarithmic scale allows to overcome the above problem, where we choose a proposal transition kernel for the parameter α for both the priors. The proposal kernel is a Beta distribution, $\mathcal{B}e(a, b)$, with $a = b = 0.5$ for the Jeffreys' prior, while it is a discrete uniform distribution $\mathrm{DU}(1/M, M - 1/M)$ for the loss-based prior.

We have generated 100 samples from a Yule–Simon distribution with the parameter α set to every value in the parameter space, 9 for $M = 10$ and 19 for $M = 20$. For each sample we have simulated from the posterior distribution of α, under both priors, by running $10,000$ iterations, with a burn-in period of $2,000$ iterations.

To evaluate the priors we have considered two frequentist measures. The first is the frequentist coverage of the 95% credible interval. That is, for each posterior,

we compute the interval between the 0.025 and 0.975 quantiles and see if the true value of α is included in it. Over repeated samples, one would expect a proportion of about 95% of the posterior intervals to contain the true parameter value. The second frequentist measure gives an idea of the precision of the inferential process, and it is represented by the square root of the mean squared error (MSE) from the mean, relative to the parameter value: $\sqrt{\text{MSE}(\alpha)}/\alpha$. We have considered the MSE from the median as well but, due to the approximate symmetry of the posterior, the results are very similar to the MSE from the mean. Figure 2 details the results for the simulations

(a)   (b)
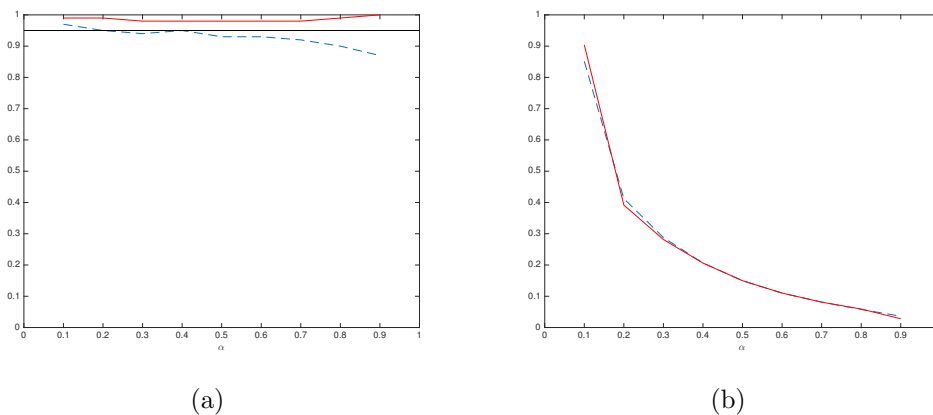
Figure 2: Frequentist properties of the Jeffreys prior (dashed line) and the loss-based prior (continuous line) for $n = 100$. The loss-prior is considered on the discretized parameter space with $M = 10$. The left plot shows the posterior frequentist coverage of the 95% credible interval, and the right plot represents the square root of the MSE from the mean of the posterior, relative to α.

with $n = 100$ and a parameter space for α discretized with increments of 0.1, that is $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. If we compare the coverage, we note that the loss-based prior tends to over-cover the credible interval, while the Jeffreys prior, although shows a better coverage for values of $\alpha < 0.5$, deteriorates in performance as the parameter tends to the upper bound of its space. Looking at the MSE, both priors appear to have very similar performance, and the (relative) error tends to decrease and α increases. In Figure 3 we have compared the frequentist performance of the Jeffreys prior with the loss-based prior defined over a more densely discretized parameter space, i.e. $\alpha = \{0.05, 0.10, \ldots, 0.95\}$. We note a smoother behaviour of the priors compared to Figure 2, which is obviously due to the denser characterization considered. The coverage still reveals a tendency of the loss-based prior to over-cover, although less pronounced than the previous case. Jeffreys prior does not present any significant difference from the previous case, as one would expect. For what it concerns the MSE, the differences between the two priors are negligible, and the only aspect we note, as mentioned above, is a smoother decrease of the error as the parameter increases.

8

(a)   (b)
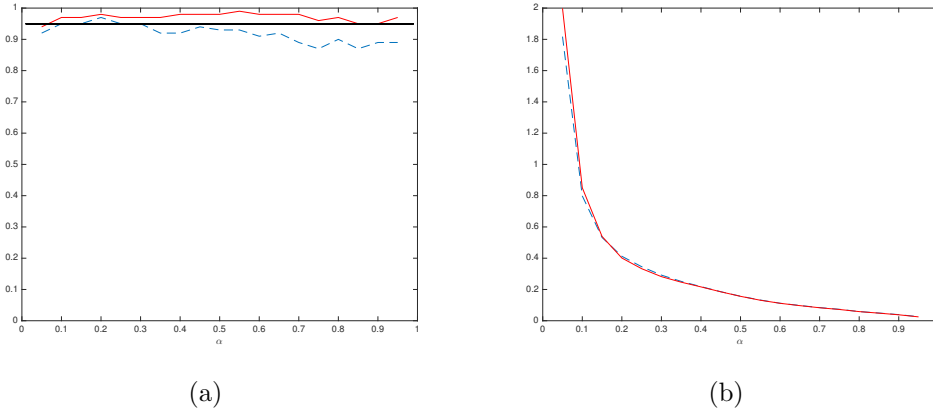
Figure 3: Frequentist properties of the Jeffreys prior (dashed line) and the loss-based prior (continuous line) for $n = 100$. The loss-prior is considered on the discretized parameter space with $M = 20$. The left plot shows the posterior frequentist coverage of the 95% credible interval, and the right plot represents the square root of the MSE from the mean of the posterior, relative to α.

We look more into the details of the objective approach by analysing two i.i.d. samples. In particular, we consider a random sample of size $n = 100$ from a Yule–Simon distribution with $\alpha = 0.40$ and a sample, of the same size, from a Yule–Simon with $\alpha = 0.68$.

In both cases, we have sampled from the posterior distribution via Monte Carlo methods with $10,000$ iterations and a burn-in period of $2,000$ iterations. Figure 4 shows the posterior samples and posterior histograms derived by applying the Jeffreys prior and the loss-based prior with two different discretizations, that is $M = 10$ and $M = 20$. The summary statistics of the three posteriors are reported in Table 1, where we have the mean, the median, and the 95% credible interval. By comparing

Prior	Mean	Median	95% C.I.
Jeffreys	0.40	0.41	(0.23,0.53)
Loss-based ($M = 10$)	0.40	0.4	(0.2,0.5)
Loss-based ($M = 20$)	0.40	0.41	(0.22,0.56)

Table 1: Summary statistics of the posterior distributions for the parameter α of the simulated data from a Yule-Simon distribution with $\alpha = 0.40$.

the mean of the posterior distributions, we see that they are all centered around the true parameter value. The credible interval yielded by the loss-based priors with the most dense discretization ($M = 20$) is larger than the other two intervals. However, the difference is very small and we can conclude that the three prior distributions result in posteriors which carry the same uncertainty. In other words, the three objective

9
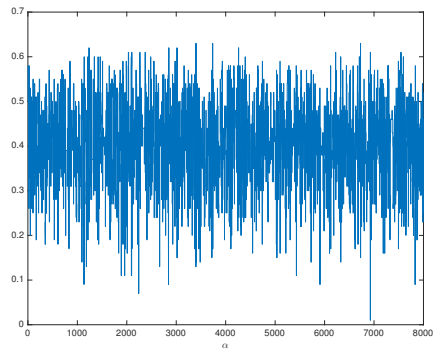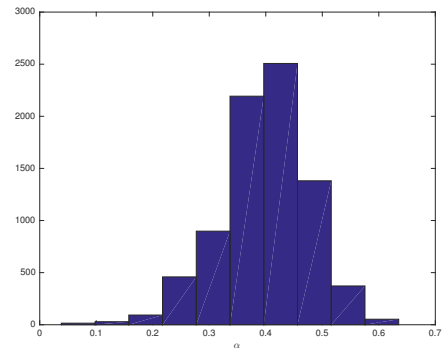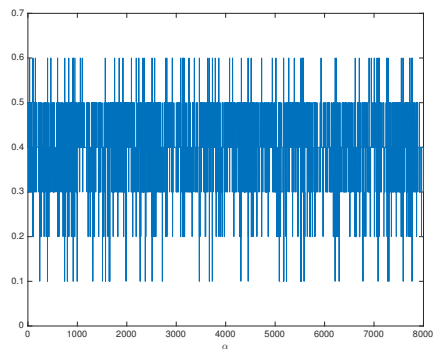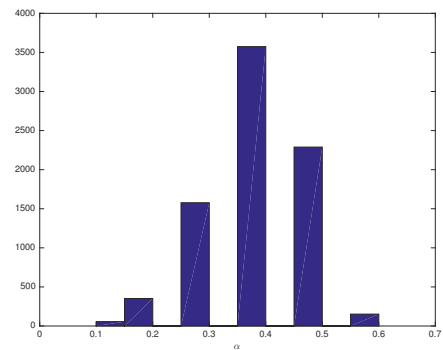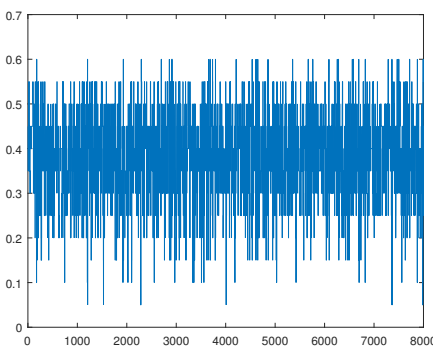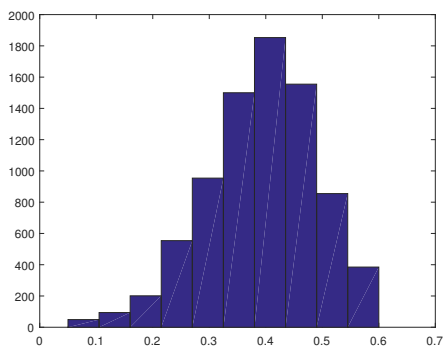
Figure 4: Posterior samples (left) and histograms (right) of the analysis of an i.i.d. sample of size $n = 100$ from a Yule–Simon distribution with $\alpha = 0.40$. From top to bottom, we have Jeffreys prior, loss-based prior with $M = 10$ and loss-based prior with $M = 20$.
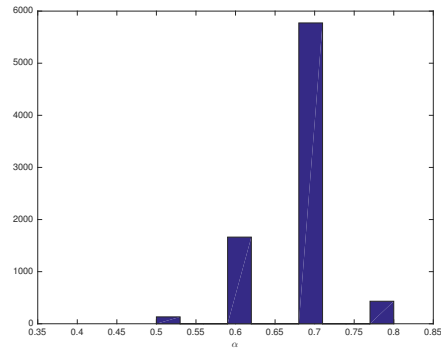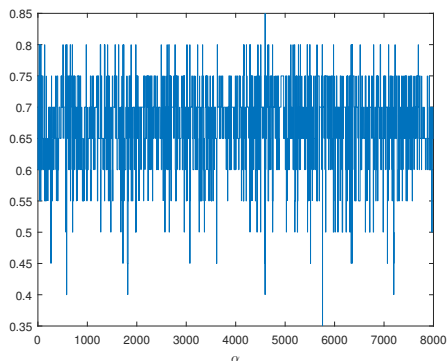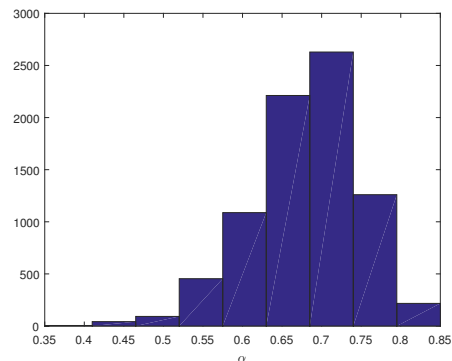
priors perform in the same way.
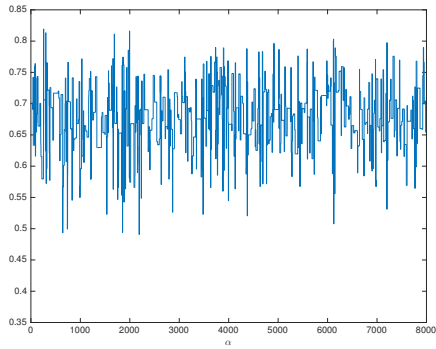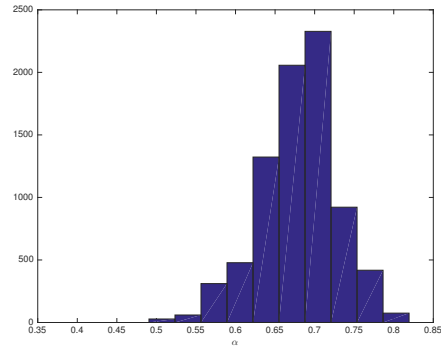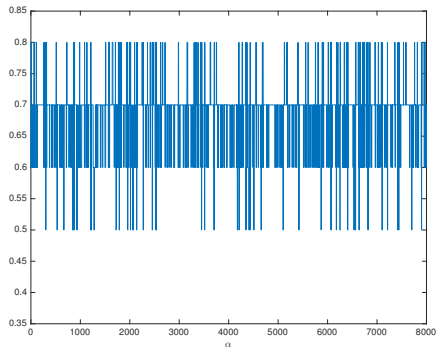
Figure 5: Posterior samples (left) and histograms (right) of the analysis of an i.i.d. sample of size $n = 100$ from a Yule–Simon distribution with $\alpha = 0.68$. From top to bottom, we have Jeffreys prior, loss-based prior with $M = 10$ and loss-based prior with $M = 20$.

Prior	Mean	Median	95% C. I.
Jeffreys	0.68	0.68	(0.57,0.77)
Loss-based ($M = 10$)	0.68	0.7	(0.6,0.8)
Loss-based ($M = 20$)	0.68	0.68	(0.55,0.79)

Table 2: Summary statistics of the posterior distributions for the parameter α of the simulated data from a Yule-Simon distribution with $\alpha = 0.68$.

Similar considerations can be made for the case where we have sampled $n = 100$ observations from a Yule–Simon distribution with $\alpha = 0.68$. By inspecting Figure 5 and Table 2, we note a very similar behaviour of the three priors, in the sense that the posterior distributions are still centered around the true value of α and that the credible intervals do not present important differences. Note that the choice of a true parameter value which would have not been included in any of the two discretized sample spaces, upon which the loss-prior is based, allows to show that the inferential process appears to be not affected by the discretization, hence motivating it.

To conclude, the simulation study shows no tangible differences in the performance of the prior distributions, in the spirit of objective Bayesian analysis.

5 Real Data Application

To illustrate the proposed priors, both the Jeffreys and the loss-based prior for the Yule-Simon distribution, we analyze three datasets. The first dataset concerns daily increments of four popular social networks stock indexes in the US market, the second contains the frequencies of surnames observed in the 1990 US Census, and the last dataset consists of 'number one' hits in the US music industry. Note that the convergence diagnostics for the last two datasets are illustrated in Appendix .

5.1 Social network stock indexes

We analyze different data in the social media marketing, in particular we focus on Facebook, Twitter, Linkedin and Google. These four major companies are the most powerful social networks in the world and are listed in the Wall Street exchange market (http://finance.yahoo.com). We analyze the daily increments for the stocks and, in particular, we consider the adjusted closing price from the 1st of October 2014 to the 11th of March 2016, for a total of $n = 365$ observations. The daily increments are obtained by applying $z_t = |r_t/r_{t-1} - 1| \cdot 100$, for $t = 2, \ldots, 365$, where r_t is the adjusted closing price for the index at day t, and we built our frequency on it. These are shown in Figure 6, while Figure 7 shows the histogram of the frequencies of the discretized data. The discretization has been done by counting the number of times a daily return took a value truncated at the second decimal digit. For example, if two observed daily returns are 1.2494 and 1.2573, they were both considered as two occurrences of the same value. By inspecting the histograms in Figure 7 it seems that the (transformed)
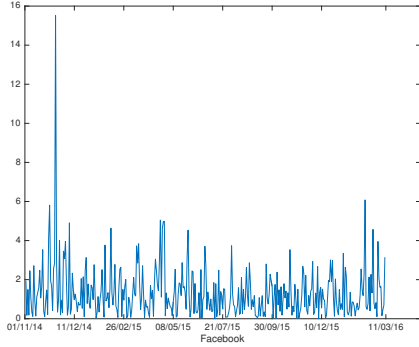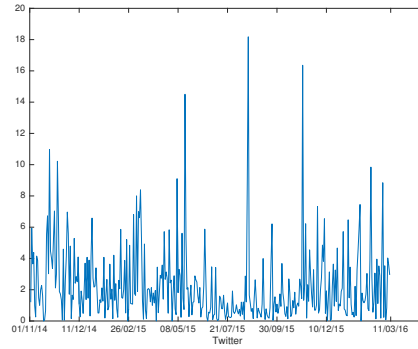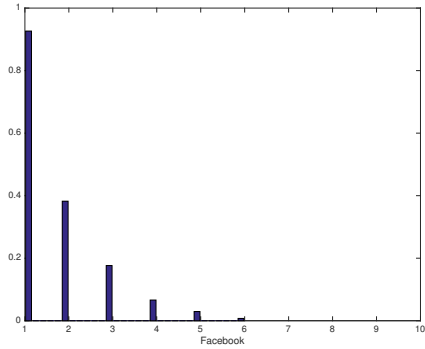
(a) Facebook
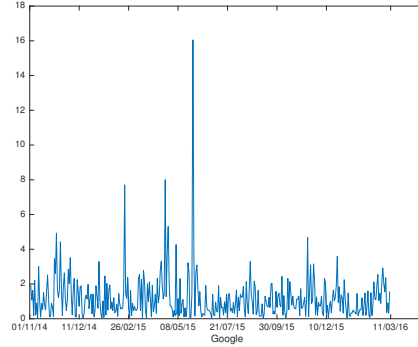
(b) Google
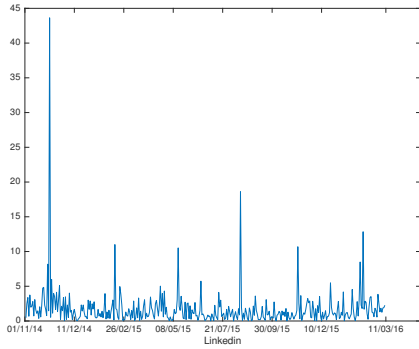
(c) Linkedin

(d) Twitter

Figure 6: Daily increments for Facebook, Google, Linkedin and Twitter from the 1st of October 2014 to the 11th of March 2016.

Yule–Simon distribution might be a suitable statistical model to represent the data. Although the daily increments are time dependent, the autocorrelation function and partial autocorrelation function plots (Figure 8) show that the data can be treated as if it were an i.i.d. sample. In fact, the plots highlight the absence of significant serial correlation with the lag series. We apply the Bayesian framework and obtain the posterior distribution for the parameter of interest as

$$\pi(\alpha|\mathbf{k}) \propto L(\mathbf{k}|\alpha)\pi(\alpha),$$

where $\mathbf{k} = (k_1, \ldots, k_n)$ represents the set of observations, i.e. the frequencies of the discretized daily returns, $L(\mathbf{k}|\alpha)$ the likelihood function and $\pi(\alpha)$ the prior distribution which, in turn, has the form of the Jeffreys prior in (3) or the loss-based prior (4). We have obtained the posterior distributions for the parameter α of the transformed Yule-Simon distribution by Monte Carlo methods. We run 25,000 iterations with a burn-in period of 5,000 iterations. We have reported the chain and the histogram

13
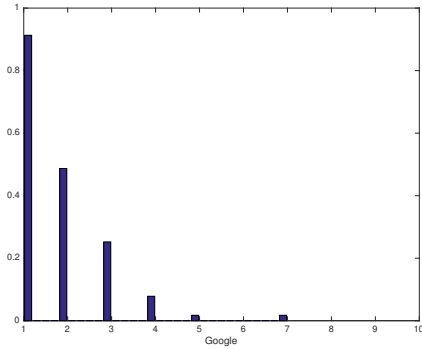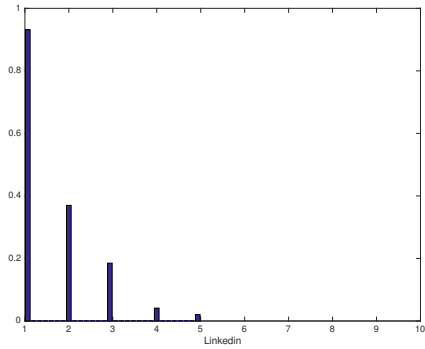
(a) Facebook
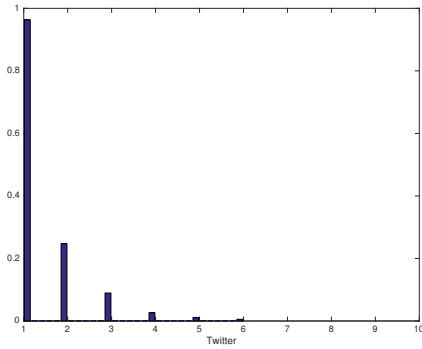
(b) Google

(c) Linkedin

(d) Twitter

Figure 7: Histograms of the discretized daily returns for Facebook, Google, Linkedin and Twitter.

of the posterior distributions in Figure 9 and in Figure 10, with the corresponding summary statistics in Table 3. Note that, with the purpose of limiting the amount of space used, we have included the plots of the Facebook and Google daily returns only.

For all the four assets we notice that the results for α are very similar, as can be inferred by the minimal (or absence of) difference between the means and the medians. The credible intervals, as well, are very similar, with a slight larger size for the case where the loss-based prior with $(M = 20)$ is applied. One way of interpreting the results is as follows. The parameter α can be seen as the probability that the next observation is different from the ones observed so far, and therefore we note that Twitter has the highest chance to take a daily increment not yet observed, while Google has the smallest.

14
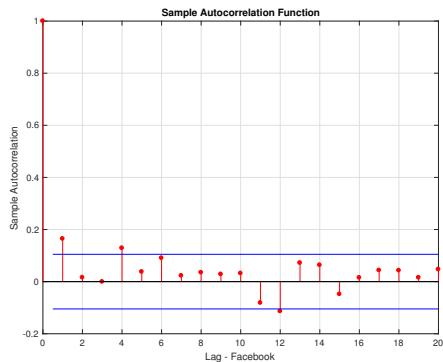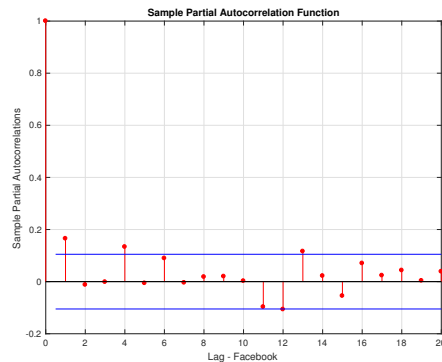
(a) Facebook

(b) Facebook
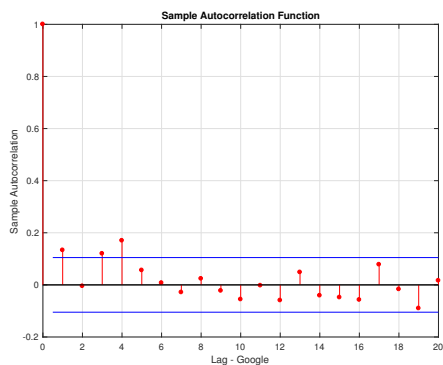
(c) Google
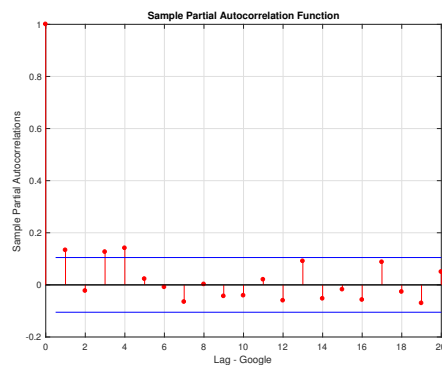
(d) Google

Figure 8: Sample autocorrelation function (left panel) and sample partial autocorrelation function (right panel) for Facebook and Google.

5.2   Census Data - Surname analysis

The second example we examine concerns with the frequency of surnames in the US (http://www.census.gov/en.html). From the population censuses (Maruka et al., 2010), we focus on the US Census completed in 1990 and consider the first 500 most common surnames. Refer to Table 4 for a list of the first 10 most frequent surnames. Briefly, the process followed by Maruka et al. (2010) to obtain the data converts the surname with Senior (SR), Junior (JR) or a number in the last name field (f.e. Moore Sr or Moore Jr or Moore III are converted to Moore) and, in addition, the authors examined each name entry for the possibility of an inversion (e.g. a first name appearing in the last name fields or vice-versa). However, as there is the possibility of having many surnames that also inverted can sound absolutely right, the authors considered also the surname of the spouse, obtaining additional information to invert the name field of the entire family.
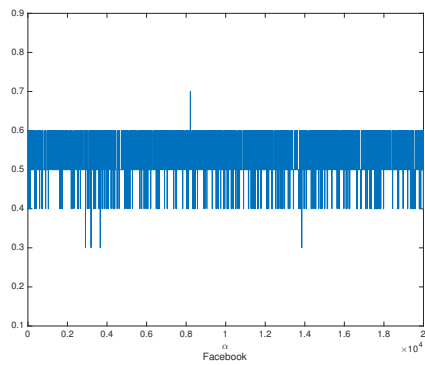
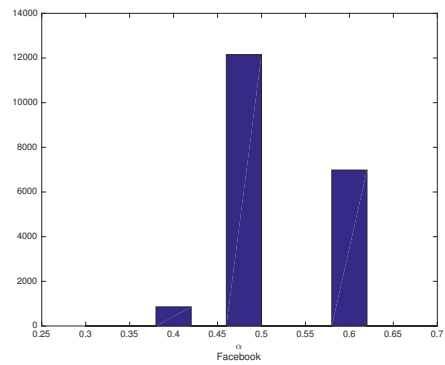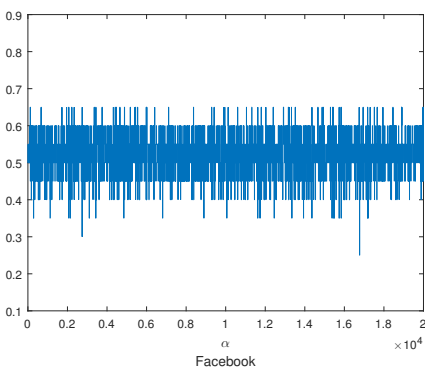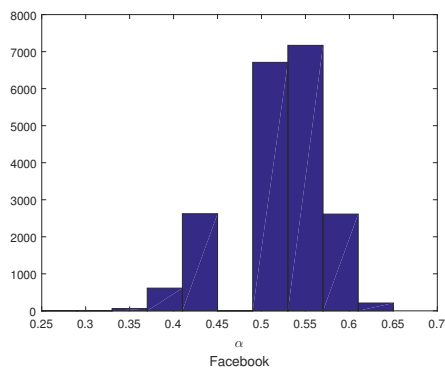The analysis has been performed by running both the Markov Chain Monte Carlo

15
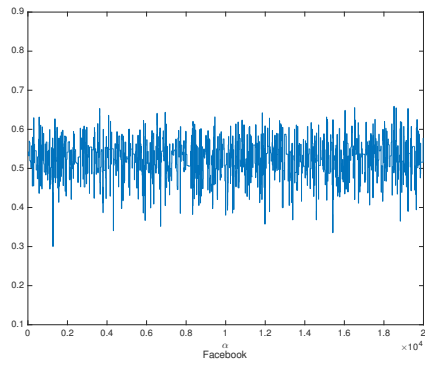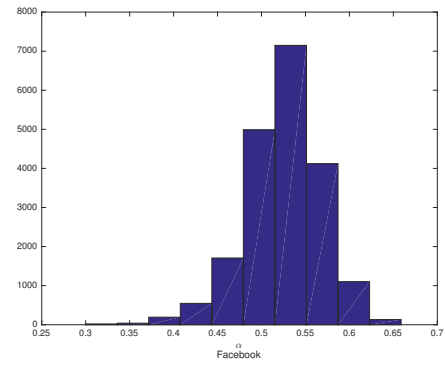
Figure 9: Posterior samples (left) and posterior histograms (right) for the Facebook daily returns obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).
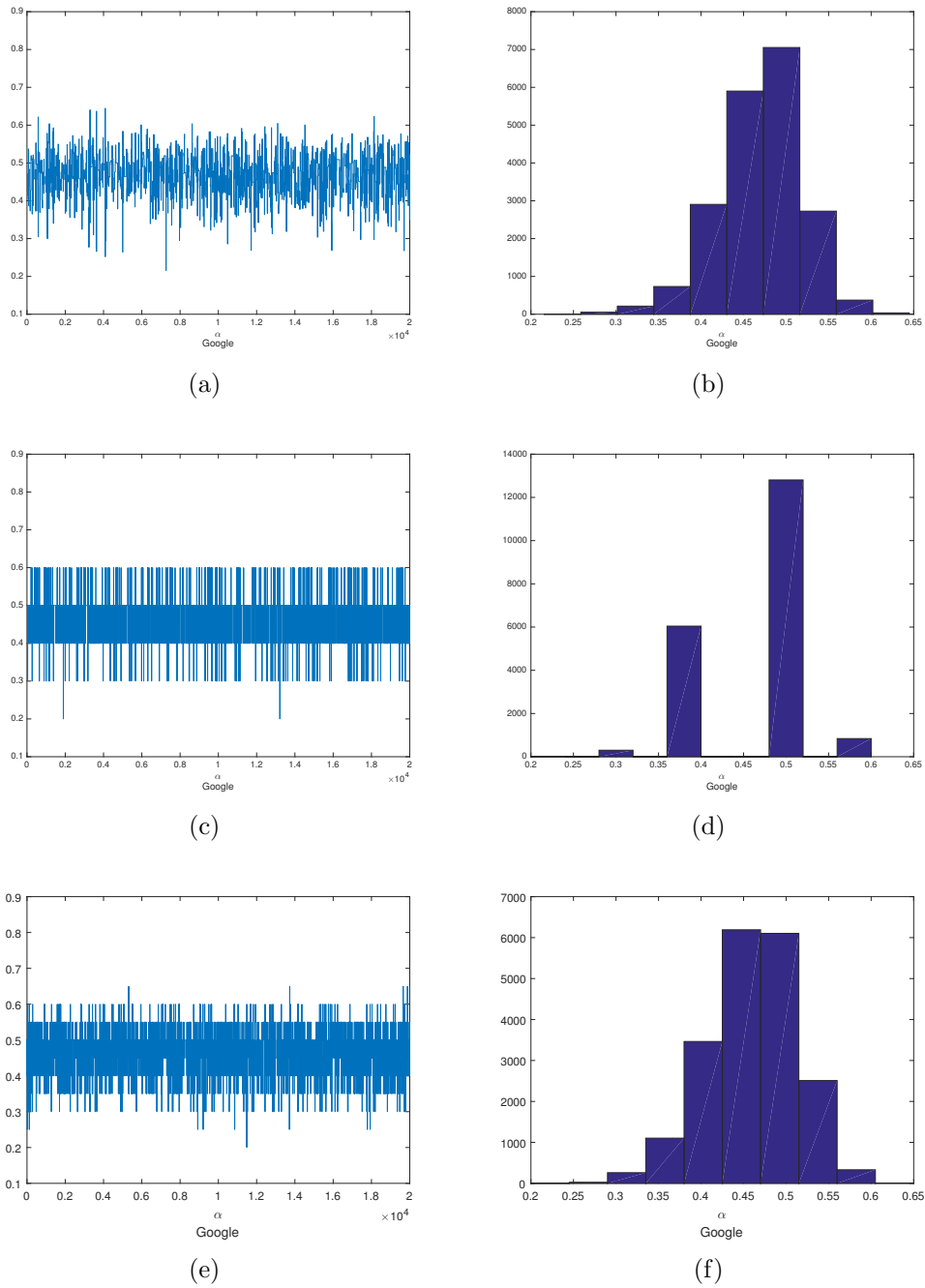
16

(a)

(b)

(c)

(d)

(e)

(f)

Figure 10: Posterior samples (left) and posterior histograms (right) for the Google daily returns obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

17

Company	Prior	Mean	Median	95% C.I.
Facebook	Jeffreys	0.53	0.53	(0.43, 0.61)
Facebook	Loss-based ($M = 10$)	0.53	0.5	(0.4, 0.6)
Facebook	Loss-based ($M = 20$)	0.52	0.55	(0.40, 0.60)
Google	Jeffreys	0.47	0.47	(0.37, 0.55)
Google	Loss-based ($M = 10$)	0.47	0.5	(0.4, 0.6)
Google	Loss-based ($M = 20$)	0.47	0.46	(0.35, 0.55)
Linkedin	Jeffreys	0.56	0.57	(0.47, 0.64)
Linkedin	Loss-based ($M = 10$)	0.57	0.6	(0.5, 0.6)
Linkedin	Loss-based ($M = 20$)	0.56	0.55	(0.45, 0.65)
Twitter	Jeffreys	0.68	0.68	(0.62, 0.73)
Twitter	Loss-based ($M = 10$)	0.69	0.7	(0.6, 0.7)
Twitter	Loss-based ($M = 20$)	0.68	0.70	(0.60, 0.75)

Table 3: Summary statistics of the posterior distribution for the parameter α of the social network stock index data.

#	Surname	Frequency	#	Surname	Frequency
1	Smith	2502021	6	Davis	1193807
2	Johnson	2014550	7	Miller	1054530
3	Williams	1738482	8	Wilson	843126
4	Jones	1544488	9	Moore	775975
5	Brown	1544488	10	Taylor	773488

Table 4: Ten most common Surname in United States from the Census 1990 analysis.

for 25,000 iterations, with a burn-in of 5,000 iterations.

The posterior samples and the posterior histograms are shown in Figure 11, with the corresponding summary statistics of the posterior distributions reported in Table 5. We again notice similarities to the simulation study and the analysis of daily increments, in the sense that means and medians are very similar for each prior, and the 95% credible interval obtained by applying the loss-based prior with $M = 20$ is slightly larger than the one obtained by using either the Jeffreys prior or the loss-based prior with $M = 10$.

The estimated value of α, on the basis of the 500 most common surnames in the US (and if we consider the mean) is, roughly, 1/2. In other words, there are about 50% chances that the next observed surname is not in the list of the 500. Obviously, a larger sample size would yield a smaller posterior mean, as the number of surnames is finite and the more we observe, the harder is to find a "new" one.
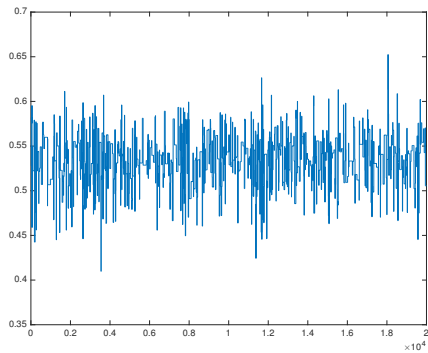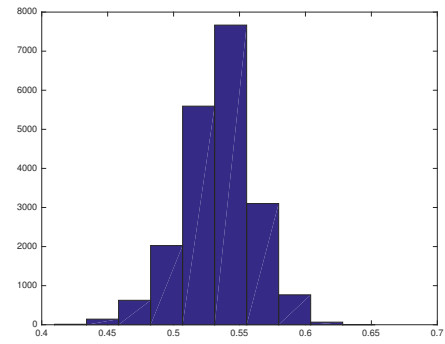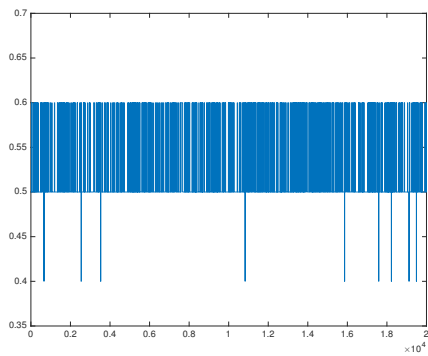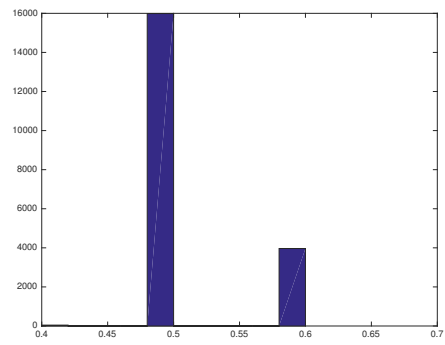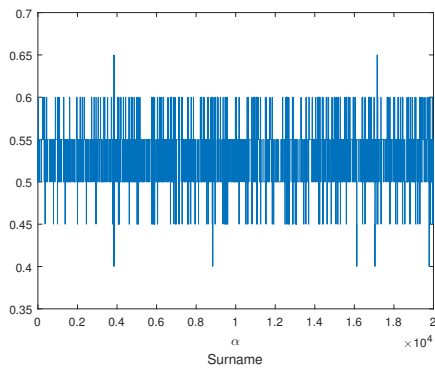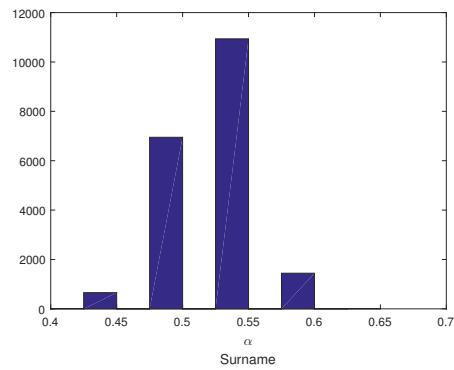
Figure 11: Posterior sample (left) and posterior histogram (right) for the surname data set obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

Prior	Mean	Median	95% C. I.
Jeffreys	0.53	0.54	(0.47, 0.58)
Loss-based ($M = 10$)	0.52	0.5	(0.5, 0.6)
Loss-based ($M = 20$)	0.53	0.55	(0.45, 0.60)

Table 5: Summary statistics of the posterior distributions for the parameter α of the Census surname analysis.

5.3 'Superstardom' analysis

The last example consists in modelling the number of 'number one' hits a music artist had in the period 1955–2003 on the Billboard Hot 100 chart. The data, which is displayed in Table 6, has been used by Chung and Cox (1994) and Spierdijk and Voorneveld (2009) to show an apparent absence of correlation between talent and success in the music industry.

| Hits	Observations	Hits	Observations
1	119	9	4
2	57	10	2
3	30	11	1
4	13	12	2
5	10	13	1
6	4	14	1
7	1	15	1
8	1	16	1

Table 6: Number of 'number one' hits per artist from 1955 to 2003.

We have run the Monte Carlo simulation for 25,000 iterations, with a burn in period of 5,000, for each of the considered priors. The posterior samples and histograms are shown in Figure 12, with the correspondinf statistic summaries in Table 7.

Prior	Mean	Median	95% C.I.
Jeffreys	0.08	0.07	(0.004, 0.24)
Loss-based ($M = 10$)	0.13	0.1	(0.1, 0.3)
Loss-based ($M = 20$)	0.11	0.10	(0.05, 0.25)
Loss-based ($M = 100$)	0.10	0.08	(0.01, 0.29)

Table 7: Summary statistics of the posterior distribution for the parameter α of the analysis of the music 'number one' hits.
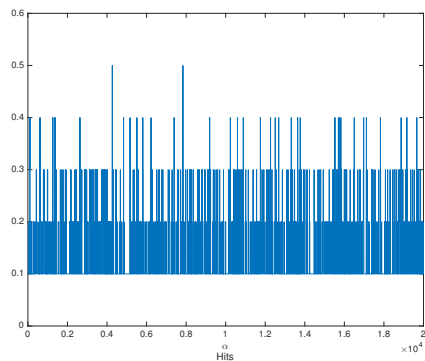
This example of the music hits allows for some interesting points of dicussion. First, we note that the posterior distributions of for α are skewed; therefore, the posterior median represents a better centrality index than the posterior mean. Second, it is
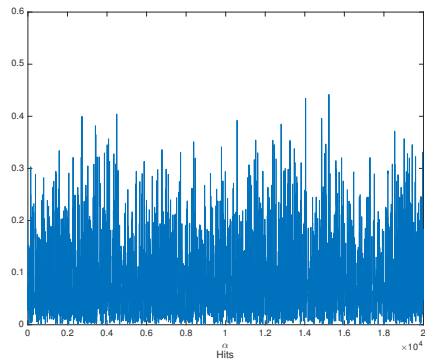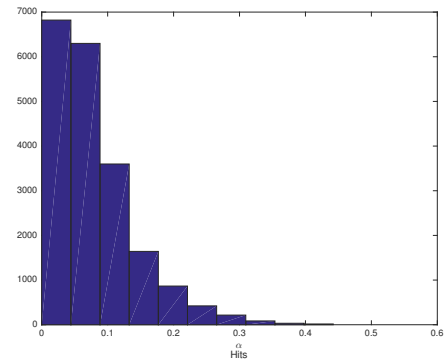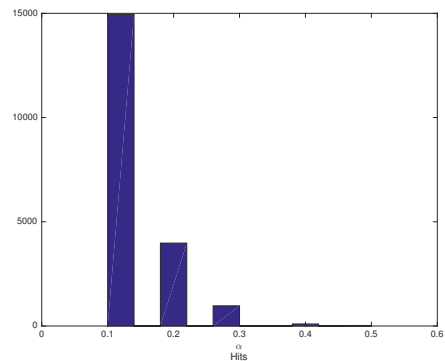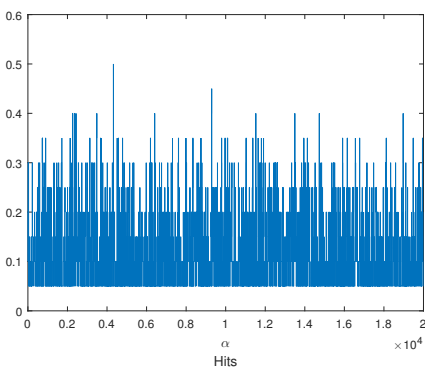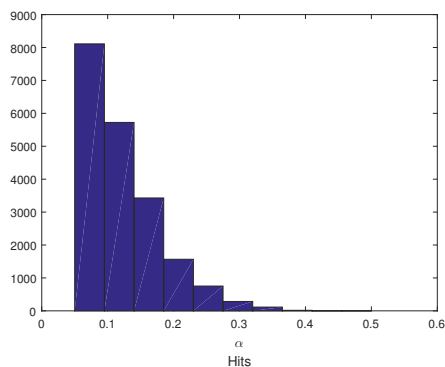
Figure 12: Posterior sample (left) and posterior histogram (right) for the music 'number one' hits data set obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

clear that the "true" value of α may be close to zero. As such, in order to explore better the parameter space when the loss-based prior is used, a denser discretization is more appropriate. We have then considered $M = 100$, resulting the posterior summary statistics in Table 7. We note now that the posterior median is similar to the one obtained using the Jeffreys prior. It is therefore recomendable that, when the inference on α indicates values near the parameter space boudaries, the level of discretization to be considered should be relatively dense.

6 Discussions

It is surprising how, from time to time, the Bayesian literature presents gaps even for problems which appear to be straightforward. The Yule–Simon distribution has undoubtedly many possibilities of application, as the discussed examples and the refereed papers show, and therefore demanded for a satisfactory discussion within the Bayesian framework.

Given the importance that objective Bayesian analysis can have in applications, and not only (Berger, 2006), we have presented two priors which are suitable in scenarios with minimal prior information. The first prior is the Jeffreys prior which, as it is well known, has the appealing property of being invariant under monotone differentiable transformations of the parameter of interest. The second prior is derived considering the loss in information one would incur if the 'wrong' model was selected. Although the latter requires a discretization of the parameter space, we have shown through simulation studies that the performance of the yielded posterior are very similar, both between the Jeffreys and the loss-based prior, and between different structures of the discretized parameter space. This is not surprising as both priors, i.e. the Jeffreys and the loss-based, have a similar behaviour, in the sense that they increase as the parameter α increases. Given that the performance of the prior is virtually the same, one may choose a preferred one on other grounds. For example, the Jeffreys' prior presents the property of being invariant under reparametrisation, while the loss-based prior is computationally more efficient.

We have limited our analysis to the case where the shape parameter of the Yule–Simon distribution, ρ, is strictly larger than one. Doing so, we allow for a more convenient parametrization of the distribution where the new parameter $\alpha = (\rho - 1)/\rho$ has the interpretation of being the probability that the next observation takes a value not observed before.

Besides through a simulation study, we have compared the objective priors by applying them on three data sets: the first related to financial data, the second to surnames in the US and the third one on the number of hits in the music industry. All comparisons allowed to show that the two proposed objective priors lead to similar results, in terms of posterior distributions. For obvious reasons, we have not considered if the choice of the Yule–Simon is the *best* model to represent the data, but limited our analysis to make inference for the unknown parameter α.

Acknowledgements

References

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis* **1,** 385–402.

Berk, R. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. of Math. Statist.* **37,** 51–58.

Chung, K. and Cox, R. (1994). A stochastic model of superstardom: an application of the yule distribution. *Review of Economics and Statistics* **76,** 771–775.

Gallardo, D. I., Gomex, H. W., and Bolfarine, H. (2016). A new cure rate model based on the Yule-Simon distribution with application to a melanoma data set. *Journal of Applied Statistics* **10.1080/02664763.2016.1194385,**.

Garcia Garcia, J. M. (2011). A fixed-point algorithm to estimate the yule-simon distribution parameter. *Applied Mathematics and Computation* **217,** 8560–8566.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, J. F. M., editors, *Bayesian Statistics 4*, pages 169–193. Oxford University Press, Oxford.

Gradshteyn, I. and Ryzhik, I. (2007). *Table of Integrals, Series and Products.* Academic Press.

Jeffreys, H. (1961). *Theory of Probability.* London: Oxford University Press.

Leisen, F., Rossini, L., and Villa, C. (2017). A note on the inference for the Yule–Simon distribution. *Journal of Statistical Computation and Simulation* **Forthcoming,**.

Maruka, Y. E., Shnerb, N. M., and Kessler, D. A. (2010). Universal features of surname distribution in a subsample of a growing population. *Journal of Theoretical Biology* **262,** 245–256.

Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Trans. Inf. Theory* **44,** 2124–2147.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6,** 7–11.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika* **42,** 425–440.

Spierdijk, L. and Voorneveld, M. (2009). Superstars without talent? the yule distribution controversy. *The Review of Economics and Statistics* **91,** 648–652.

Villa, C. and Walker, S. (2015). An Objective approach to prior mass function for discrete parameter spaces. *Journal of the American Statistical Association* **110,** 1072–1082.

Yule, G. U. (1925). A Mathematical theory of evolution, based on the conclusion of Dr. J.C. Willis. *Philosophical Transactions of the Royal Society of London, Series B* **213,** 21–87.

A   Appendix

Proof of Theorem 3.1. First of all, we note that

$$\frac{\partial^2 \log(f(k;\alpha))}{\partial \alpha^2} = \frac{1}{(1-\alpha)^2} + \frac{2}{(1-\alpha)^3}\left[\psi^{(0)}\left(\frac{1}{1-\alpha}+1\right)\right.$$
$$\left. - \psi^{(0)}\left(\frac{1}{1-\alpha}+k+1\right)\right] + \frac{1}{(1-\alpha)^4}\left[\psi^{(1)}\left(\frac{1}{1-\alpha}+1\right)\right.$$
$$\left. - \psi^{(1)}\left(\frac{1}{1-\alpha}+k+1\right)\right]$$

where $\psi^{(i)}$ is the polygamma function:

$$\psi^{(i)}(x) = \frac{\partial^{i+1}}{\partial x^{i+1}}\log(\Gamma(x)) = (-1)^{i+1}i!\sum_{k=0}^{\infty}\frac{1}{(x+k)^{i+1}} \quad i = 1, 2, \ldots.$$

It's easy to see that

$$\psi^{(0)}\left(\frac{1}{1-\alpha}+1\right) - \psi^{(0)}\left(\frac{1}{1-\alpha}+k+1\right) = \sum_{j=1}^{k}\frac{1}{\frac{1}{1-\alpha}+j}$$

and

$$\psi^{(1)}\left(\frac{1}{1-\alpha}+1\right) - \psi^{(1)}\left(\frac{1}{1-\alpha}+k+1\right) = \sum_{j=0}^{k-1}\frac{1}{(\frac{1}{1-\alpha}+1+j)^2}$$

Therefore, we have that the Fisher information is:

$$\mathcal{I}(\alpha) = -\frac{1}{(1-\alpha)^2} + \frac{2}{(1-\alpha)^3}\mathbb{E}_\alpha\left[\sum_{j=1}^{k}\frac{1}{\frac{1}{1-\alpha}+j}\right]$$
$$- \frac{1}{(1-\alpha)^4}\mathbb{E}_\alpha\left[\sum_{j=0}^{k-1}\frac{1}{(\frac{1}{1-\alpha}+1+j)^2}\right] \tag{A.1}$$

In order to compute the Jeffreys prior, we need compute the two expected value of equation A.1 separately.

$$\mathbb{E}_\alpha\left[\sum_{j=1}^{k}\frac{1}{(\frac{1}{1-\alpha}+j)}\right] = \sum_{k=1}^{\infty}\sum_{j=1}^{k}\frac{1}{(\frac{1}{1-\alpha}+j)}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right)$$
$$= \sum_{j=1}^{\infty}\frac{1}{(\frac{1}{1-\alpha}+j)}\sum_{k=j}^{\infty}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right) \tag{A.2}$$

The second summation with respect to k in equation A.2 can be rewritten as:

$$\sum_{k=j}^{\infty}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right) = \sum_{k=j}^{\infty}\frac{1}{1-\alpha}\int_{0}^{1}x^{k-1}(1-x)^{\frac{1}{1-\alpha}}dx$$

$$= \sum_{l=0}^{\infty} \frac{1}{1-\alpha} \int_0^1 x^l x^{j-1} (1-x)^{\frac{1}{1-\alpha}} dx$$

$$= \frac{1}{1-\alpha} \int_0^1 x^{j-1} (1-x)^{\frac{1}{1-\alpha}-1} dx$$

$$= \frac{\Gamma(j)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+j)} \tag{A.3}$$

Finally, we have that :

$$\mathbb{E}_\alpha\left[\sum_{j=1}^{k} \frac{1}{(\frac{1}{1-\alpha}+j)}\right] = \Gamma(\frac{1}{1-\alpha}+1)\sum_{j=1}^{\infty} \frac{\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j)(\frac{1}{1-\alpha}+j)}$$

$$= \Gamma\left(\frac{1}{1-\alpha}+1\right)\sum_{j=1}^{\infty} \frac{\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j+1)}$$

$$= \sum_{j=1}^{\infty} \left(\frac{1-\alpha}{1-\alpha}\right)\frac{\Gamma(\frac{1}{1-\alpha}+1)\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j+1)}$$

$$= (1-\alpha)\sum_{j=1}^{\infty} \frac{1}{1-\alpha}\frac{\Gamma(\frac{1}{1-\alpha}+1)\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j+1)} =$$

$$= (1-\alpha) \tag{A.4}$$

where the summation $\sum_{j=1}^{\infty}(\frac{1}{1-\alpha})B(j,\frac{1}{1-\alpha}+1) = 1$, since we are summing over all the possible values of the probability function of the Yule-Simon distribution.

As we have done with the first expected value of A.1, now we compute the second expected value of equation A.1:

$$\mathbb{E}_\alpha\left[\sum_{j=0}^{k-1} \frac{1}{(\frac{1}{1-\alpha}+1+j)^2}\right] = \sum_{k=1}^{\infty}\sum_{j=0}^{k-1} \frac{1}{(1+\frac{1}{1-\alpha}+j)^2}\frac{1}{1-\alpha}\frac{\Gamma(k)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+1+k)}$$

$$= \sum_{k=1}^{\infty}\sum_{j=1}^{k} \frac{1}{(\frac{1}{1-\alpha}+j)^2}\left(\frac{1}{1-\alpha}\right)\frac{\Gamma(k)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+1+k)}$$

$$= \sum_{j=1}^{\infty} \frac{1}{(\frac{1}{1-\alpha}+j)^2}\sum_{k=j}^{\infty} \frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right)$$

$$= \sum_{j=1}^{\infty} \frac{1}{(\frac{1}{1-\alpha}+j)^2}\frac{\Gamma(j)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+j)} \tag{A.5}$$

where the last equality follows from (A.3).Finally we obtain the following form:

$$\mathbb{E}_\alpha\left[\sum_{j=0}^{k-1} \frac{1}{(\frac{1}{1-\alpha}+1+j)^2}\right] = \sum_{j=1}^{\infty} \frac{B(j,\frac{1}{1-\alpha}+1)}{\frac{1}{1-\alpha}+j} \tag{A.6}$$

The equation A.6 can be written in a more simple way as a function of an Hypergeometric function.

$$\sum_{j=1}^{\infty} \frac{B(j, \frac{1}{1-\alpha}+1)}{\frac{1}{1-\alpha}+j} = \sum_{j=1}^{\infty} \frac{1}{\frac{1}{1-\alpha}+j} \int_0^1 x^{j-1}(1-x)^{\frac{1}{1-\alpha}} dx$$

$$= \int_0^1 (1-x)^{\frac{1}{1-\alpha}} \sum_{j=1}^{\infty} \frac{1}{\frac{1}{1-\alpha}+j} x^{j-1} dx \qquad (A.7)$$

Looking at the summation we have:

$$\sum_{j=1}^{\infty} \frac{1}{\frac{1}{1-\alpha}+j} x^{j-1} dx = \sum_{l=0}^{\infty} \frac{x^l}{\frac{1}{1-\alpha}+l+1} \qquad (A.8)$$

But the denominator can be written as a ratio of Pochhammer representations:

$$\frac{1}{\frac{1}{1-\alpha}+l+1} = \frac{\left(\frac{1}{1-\alpha}+1\right)_l}{\left(\frac{1}{1-\alpha}+2\right)_l} \frac{1}{\frac{1}{1-\alpha}+1}$$

Hence the equation A.8 is written as:

$$\sum_{l=0}^{\infty} \frac{x^l}{\frac{1}{1-\alpha}+l+1} = \frac{1}{\frac{1}{1-\alpha}+1} \sum_{l=0}^{\infty} \frac{\left(\frac{1}{1-\alpha}+1\right)_l}{\left(\frac{1}{1-\alpha}+2\right)_l} x^l \frac{(1)_l}{l!} =$$

$$= \frac{1}{\frac{1}{1-\alpha}+1} {}_2F_1\left(1, \frac{1}{1-\alpha}+1, \frac{1}{1-\alpha}+2, x\right)$$

where ${}_2F_1(\alpha, \beta, \gamma, x)$ is the hypergeometric function. So we have that equation A.7 can be rewritten as:

$$\int_0^1 \frac{(1-x)^{\frac{1}{1-\alpha}}}{\frac{1}{1-\alpha}+1} {}_2F_1\left(1, \frac{1}{1-\alpha}+1, \frac{1}{1-\alpha}+2, x\right) =$$

$$= \frac{(1-\alpha)^2}{(2-\alpha)^2} {}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)$$

where the last equality follows from 7.512.5 of Gradshteyn and Ryzhik (2007). Summing up,

$$\mathcal{I}(\alpha) = -\frac{1}{(1-\alpha)^2} + \frac{2}{(1-\alpha)^3}(1-\alpha) +$$

$$-\frac{1}{(1-\alpha)^4} \frac{(1-\alpha)^2}{(2-\alpha)^2} {}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)$$

and this concludes the proof. $\qquad\square$

B Convergence Analysis

In this section of the appendix, we describe the converge analysis we have performed for the Census surname dataset and the 'Superstardom' hits data set.

The convergence analysis for the Census and 'Superstardom' datasets has been done using the R Coda Package (Plummer et al., 2006). In particular, we have computed the Geweke convergence test, the Gelman–Rubin's plot and test statistics, the autocorrelation and partial autocorrelation function. Tables B.1 and B.2 report the results of the Geweke's convergence test (Geweke, 1992) for the two datasets. The test statistics show no convergence issues. To perform the Gelman–Rubin test of convergence, we have run multiple chains with sparse starting points. Figure B.1 plots the srinking factor for the Census data set, while Figure B.3 shows the factor for the 'Superstardom' data set. In both cases we do not see any indication of failed convergence under each objective prior. As a final check, we have plotted the autocorrelation function and the partial autocorrelation function for both datasets under each prior. These plots are shown in Figure B.2 and B.4. We again see no indication of convergence problems.

Prior	Test
Jeffreys	0.3823
Loss-based ($M = 10$)	-0.6470
Loss-based ($M = 20$)	0.7761

Table B.1: Geweke's test statistics for the posterior chain of the parameter α for the Census surname dataset.

Prior	Test
Jeffreys	0.5647
Loss-based ($M = 10$)	-0.9505
Loss-based ($M = 20$)	0.5647

Table B.2: Geweke's test statistics for the posterior chain of the parameter α for the 'Superstardom' hits dataset.
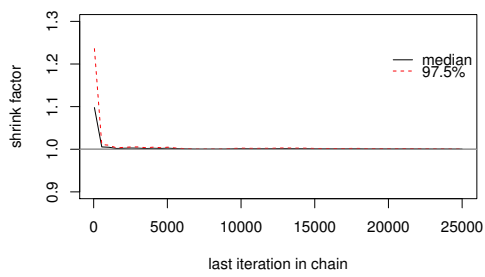
(a) Jeffreys
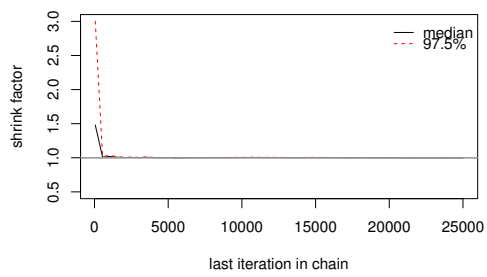
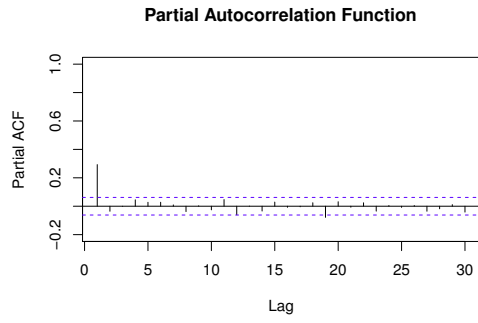(b) loss-prior with $M = 10$
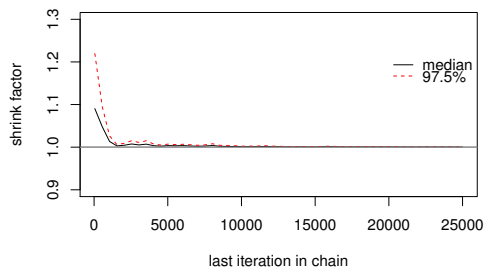
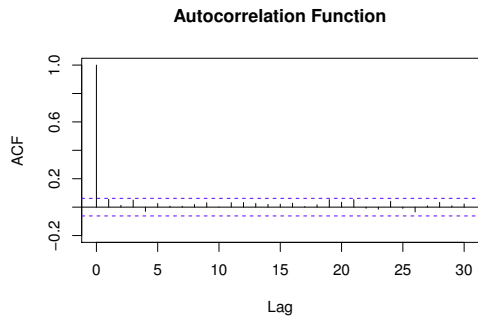(c) loss-prior with $M = 20$

Figure B.1: Gelman–Rubin plot for the Census surname dataset.
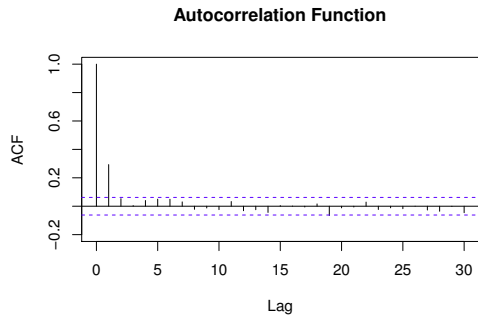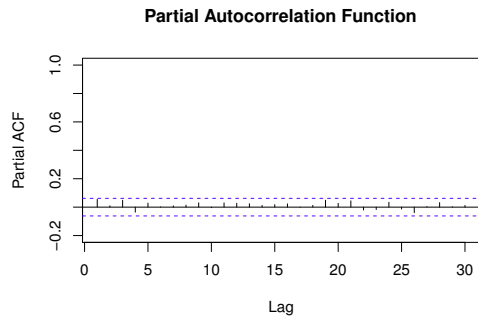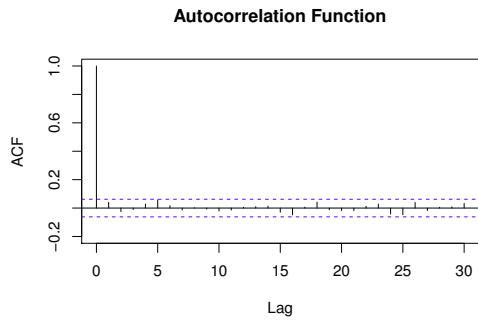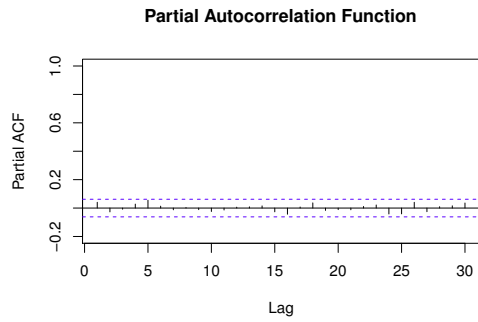
(a) Jeffreys

(b) Jeffreys

(c) loss-prior with $M = 10$

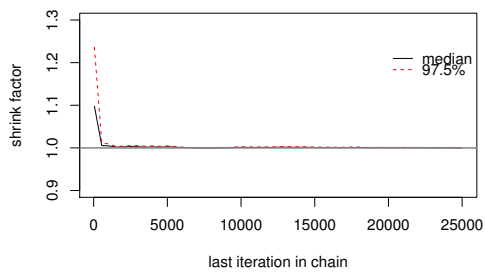(d) loss-prior with $M = 10$

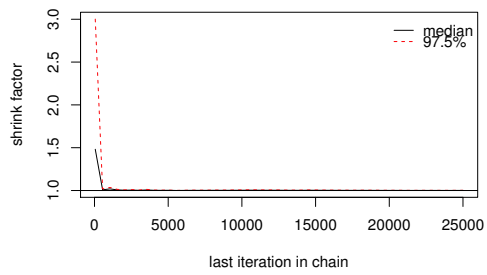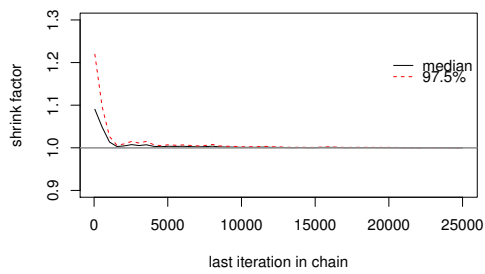(e) loss-prior with $M = 20$

(f) loss-prior with $M = 20$

Figure B.2: Autocorrelation and partial autocorrelation functions of the chains for the Census surname dataset analysis under each objective prior.
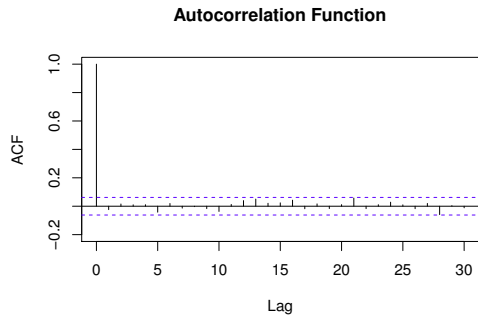
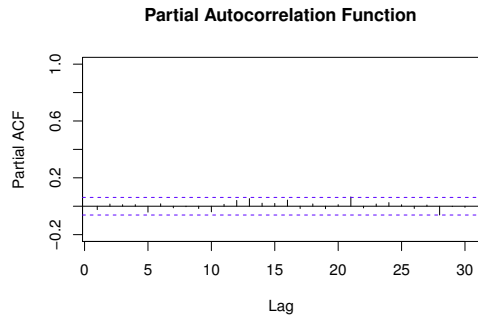(a) Jeffreys

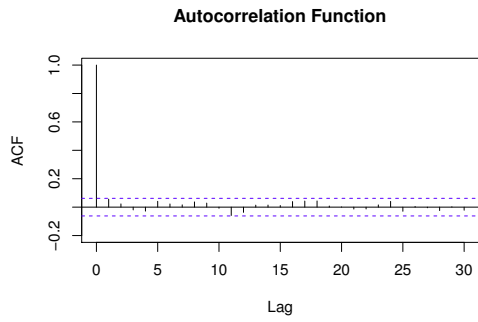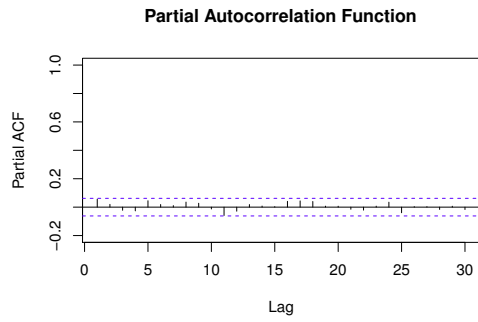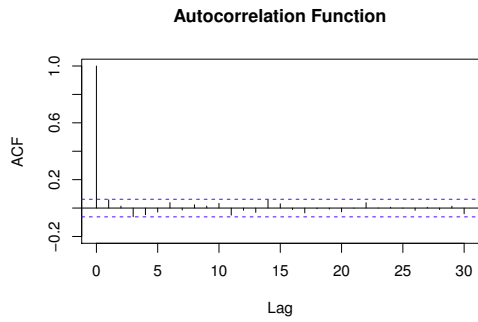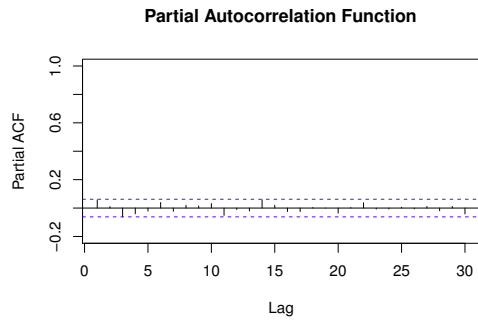(b) loss-prior with $M = 10$

(c) loss-prior with $M = 20$

Figure B.3: Gelman–Rubin plot for the'Superstardom' hits dataset.

31

Autocorrelation Function

Partial Autocorrelation Function

(a) Jeffreys

(b) Jeffreys

Autocorrelation Function

Partial Autocorrelation Function

(c) loss-prior with $M = 10$

(d) loss-prior with $M = 10$

Autocorrelation Function

Partial Autocorrelation Function

(e) loss-prior with $M = 20$

(f) loss-prior with $M = 20$

Figure B.4: Autocorrelation and partial autocorrelation functions of the chains for the 'Superstardom' hits dataset analysis under each objective prior.