Approximating the Likelihood in Approximate Bayesian Computation

Christopher C Drovandi*, Clara Grazian†, Kerrie Mengersen*, Christian Robert‡

*School of Mathematical Sciences, Queensland University of Technology, Australia

†Nuffield Department of Medicine, University of Oxford, England

‡ Université Paris-Dauphine, PSL, France
Department of Statistics, University of Warwick, England

email: c.drovandi@qut.edu.au

March 20, 2018

ABSTRACT

The conceptual and methodological framework that underpins approximate Bayesian computation (ABC) is targetted primarily towards problems in which the likelihood is either challenging or missing. ABC uses a simulation-based non-parametric estimate of the likelihood of a summary statistic and assumes that the generation of data from the model is computationally cheap. This chapter reviews two alternative approaches for estimating the intractable likelihood, with the goal of reducing the necessary model simulations to produce an approximate posterior. The first of these is a Bayesian version of the synthetic likelihood (SL), initially developed by Wood (2010), which uses a multivariate normal approximation to the summary statistic likelihood. Using the parametric approximation as opposed to the non-parametric approximation of ABC, it is possible to reduce the number of model simulations required. The second likelihood approximation method we consider in this chapter is based on the empirical likelihood (EL), which is a non-parametric technique and involves maximising a likelihood constructed empirically under a set of moment constraints. Mengersen et al. (2013) adapt the EL framework so that it can be used to form an approximate posterior for problems where ABC can be applied, that is, for models with intractable likelihoods. However, unlike ABC and the Bayesian SL (BSL), the Bayesian EL (BC_{el}) approach can be used to completely avoid model simulations in some cases. The BSL and BCel methods are illustrated on models of varying complexity.

KEYWORDS:
Approximate Bayesian computation, empirical likelihood, importance sampling, BC_{el}, sequential Monte Carlo, synthetic likelihood

1 Introduction

Approximate Bayesian computation (ABC) is now a mature algorithm for likelihood-free estimation. It has been successfully applied to a wide range of real-world problems for which more standard analytic tools were unsuitable due to the absence or complexity of the associated likelihood. It has also paved the way for a range of algorithmic extensions that take advantage of appealing ideas embedded in other approaches. Despite the usefulness of ABC, the method does have a number of drawbacks. The approach is simulation intensive, requires tuning of the tolerance threshold, discrepancy function and weighting function, and suffers from a curse of dimensionality of the summary statistic. The latter issue stems from the fact that ABC uses a non-parametric estimate of the likelihood function of a summary statistic (Blum, 2010).

In this chapter we review two alternative methods of approximating the intractable likelihood function for the model of interest, both of which aim to improve computational efficiency relative to ABC. The first is the synthetic likelihood (SL, originally developed by Wood (2010)), which uses a multivariate normal approximation to the summary statistic likelihood. This auxiliary likelihood can be maximised directly or incorporated in a Bayesian framework, which we refer to as BSL. The BSL approach requires substantially less tuning than ABC. Further, BSL scales more efficiently with an increase in the dimension of the summary statistic due to the parametric approximation of the summary statistic likelihood. However, the BSL approach remains simulation intensive. In another chapter, Fasiolo et al. (2016) apply BSL to dynamic ecological models and compare it with an alternative Bayesian method for state space models. In this chapter, we provide a more thorough review of SL both in the classical and Bayesian frameworks.

The second approach we consider uses an empirical likelihood (EL) within a Bayesian framework (BC_{el}, see Mengersen et al. (2013)). This approach can in some cases avoid the need for model simulation completely and inherits the established theoretical and practical advantages of synthetic likelihood. This improvement in computational efficiency is at the expense of specification of constraints and making equivalence statements about parameters under the different models. Of note is that the latter enables, for the first time, model comparison using Bayes factors even if the priors are improper. In summary, in the Bayesian context, both of these approaches replace intractable likelihoods with alternative likelihoods that are more manageable computationally.

2 Synthetic Likelihood

The first approach to approximating the likelihood that is considered in this chapter is the use of a synthetic likelihood (SL), which was first introduced by Wood (2010). The key idea behind the SL is the assumption that the summary statistic conditional on a parameter value has a multivariate normal distribution with mean vector $\mu(\theta)$ and covariance matrix $\Sigma(\theta)$. That is, we assume that

$$p(s|\theta) = \mathcal{N}(s; \mu(\theta), \Sigma(\theta)),$$

where $\mathcal{N}$ denotes the density of the multivariate normal distribution. Of course, in general for models with intractable likelihoods, the distribution of the summary statistic is unknown and thus $\mu(\theta)$ and $\Sigma(\theta)$ are generally unavailable. However, it is possible to estimate these quantities empirically via simulation. Consider generating n independent and identically distributed (iid) summary statistic values, $s^{1:n} = (s^1, \ldots, s^n)$, from the model based on a particular value of θ, $s^{1:n} \overset{\text{iid}}{\sim} p(s|\theta)$. Then the mean and covariance matrix can be estimated via

$$\begin{aligned} \mu(\theta) \approx \mu_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} s^i, \\ \Sigma(\theta) \approx \Sigma_n(\theta) &= \frac{1}{n-1}\sum_{i=1}^{n}(s^i - \mu_n(\theta))(s^i - \mu_n(\theta))^\top, \end{aligned} \tag{1}$$

where the superscript $\top$ denotes transpose. The likelihood of the observed summary statistic, s_{obs}, is estimated via $p_n(s_{obs}|\theta) = \mathcal{N}(s_{obs}; \mu_n(\theta), \Sigma_n(\theta))$. We use the subscript n on $p_n(s_{obs}|\theta)$ to denote the fact that the approximate likelihood will depend on the choice of n. The larger the value of n the better the mean and covariance parameters of the multivariate normal distribution can be approximated. However, larger values of n need more computation to estimate the likelihood. It is likely that a suitable value of n will be problem dependent, in particular, it may depend on the actual distribution of the summary statistic and also the dimension of the summary statistic. The value of n must be large enough so that the empirical covariance matrix is positive definite.

Note that Wood (2010) described some extensions, such as using robust covariance matrix estimation to handle some non-normality in the summary statistics and robustifying the SL when the observed summary statistic falls in the tails of the summary statistic distribution (i.e. when a poor parameter value is proposed or when the model is mis-specified).

The SL may be incorporated into a classical or Bayesian framework, which are both described below. Then, attempts in the literature to accelerate the SL method are described. We finish the section with a real data example in cell biology.

2.1 Classical synthetic Likelihood

The approach adopted in Wood (2010) is to consider the following estimator

$$\hat{\theta}_n = \arg\max_{\theta} \mathcal{N}(s_{obs}; \mu_n(\theta), \Sigma_n(\theta)), \tag{2}$$

which is the maximum SL estimator. We use the subscript n to denote that the estimator will depend on the value of n, with higher accuracy likely to be obtained for larger values of n. We note that also because the likelihood is stochastic a different answer will be obtained for fixed n if a different random seed is applied. Since the optimisation in (2) is stochastic, Wood (2010) applied a Markov chain Monte Carlo (MCMC) to explore the space of θ and select the value of θ that produced the highest value of the SL. Some recent applications of the SL method have appeared in Hartig et al. (2014), who used the FORMIND model for explaining complicated biological processes that occur in natural forests, and Brown et al. (2014), who

considered models for the transmission dynamics of avian influenza viruses in different bird types.

The synthetic likelihood approach has a strong connection with indirect inference, which is a classical method for obtaining point estimates of parameters of models with intractable likelihoods. In the simulated quasi-maximum likelihood (SQML) approach of Smith (1993), an auxiliary model with a tractable likelihood function, $p_A(y|\phi)$, where ϕ is the parameter of that model, is used. Define the function $\phi(\theta)$ as the relationship between the parameter of the model of interest and the auxiliary model. This is often referred to as the binding function in the indirect inference literature. The SQML method aims to maximise the auxiliary likelihood rather than the intractable likelihood of the model of interest

$$\hat{\theta} = \max_{\theta} p_A(y_{obs}|\phi(\theta)).$$

Unfortunately the binding function is typically unavailable. However, it can be estimated by generating n iid datasets, $y_1, \ldots, y_n$, from the model of interest (the generative model) conditional on a value of θ. Define the auxiliary parameter estimate based on the ith simulated dataset as

$$\phi_{y_i} = \arg\max_{\phi} p_A(y_i|\phi).$$

Then we have

$$\phi(\theta) \approx \phi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi_{y_i}.$$

The binding function is defined as $\phi_n(\theta) \to \phi(\theta)$ as $n \to \infty$. The SQML estimator then becomes

$$\hat{\theta}_n = \max_{\theta} p_A(y_{obs}|\phi_n(\theta)).$$

The synthetic likelihood falls within the SQML framework but where y_{obs} has been reduced to s_{obs} and the density of the multivariate normal distribution is used for p_A.

2.2 Bayesian synthetic Likelihood

An intuitive approach to incorporating SL into a Bayesian framework involves combining the prior $\pi(\theta)$ with the synthetic likelihood, which induces the following approximate posterior

$$\pi_n(\theta|s_{obs}) \propto \mathcal{N}(s_{obs}; \mu_n(\theta), \Sigma_n(\theta))\pi(\theta),$$

where the subscript n denotes that the approximate posterior depends on the choice of n. Drovandi et al. (2015) consider a general framework called parametric Bayesian indirect likelihood (pBIL), where the likelihood of some auxiliary model with parameter ϕ, $p_A(y_{obs}|\phi(\theta))$, is used to replace the intractable likelihood of the actual or generative model, $p(y_{obs}|\theta)$. Since the binding function is generally not available in closed form, it can be estimated by simulation via drawing n iid datasets from the generative model and fitting the auxiliary model to this simulated data (as in the SQML method mentioned previously), producing $\phi_n(\theta)$.

Drovandi et al. (2015) demonstrate that the resulting approximate posterior depends on n, since in general $p_A(y_{obs}|\phi_n(\theta))$ is not an unbiased estimate of $p_A(y_{obs}|\phi(\theta))$ even when $\phi_n(\theta)$ is an unbiased estimate of $\phi(\theta)$. We note that when non-negative and unbiased likelihood estimates are used within Monte Carlo methods such as MCMC (Andrieu and Roberts, 2009) and sequential Monte Carlo (SMC, Chopin et al. (2013)) algorithms, the resulting target distribution is the posterior based on the originally intended likelihood function. Such approaches are referred to as pseudo-marginal or exact-approximate methods in the literature. BSL fits within the pBIL framework, but where the auxiliary model is applied at a summary statistic level rather than the full data level and that the auxiliary model is the multivariate normal distribution, so that the auxiliary parameter estimates have an analytic expression as shown in (1). Despite the fact that we use unbiased estimators for $\mu(\theta)$ and $\Sigma(\theta)$ (under the normality assumption) it is clear that $\mathcal{N}(s_{obs}; \mu_n(\theta), \Sigma_n(\theta))$ is not an unbiased estimate of $\mathcal{N}(s_{obs}; \mu(\theta), \Sigma(\theta))$. Therefore the BSL posterior is inherently dependent on n. However, under the assumption that the model is able to recover the observed statistic, Price et al. (2018) present extensive empirical evidence that the BSL posterior is remarkably insensitive to n. Further, some empirical evidence demonstrates that BSL shows some robustness to the lack of multivariate normality.

Price et al. (2018) developed a new BSL method that uses an exactly unbiased estimator of the normal likelihood, which is developed by Ghurye and Olkin (1969). Using the notation of Ghurye and Olkin (1969), let

$$c(k, v) = \frac{2^{-kv/2}\pi^{-k(k-1)/4}}{\prod_{i=1}^{k}\Gamma\left(\frac{1}{2}(v-i+1)\right)},$$

and for a square matrix A write $\psi(A) = |A|$ if $A > 0$ and $\psi(A) = 0$ otherwise, where $|A|$ is the determinant of A and $A > 0$ means that A is positive definite. The result of Ghurye and Olkin (1969) shows that an exactly unbiased estimator of $\mathcal{N}(s_{obs}; \mu(\theta), \Sigma(\theta))$ is (in the case where the summary statistics are normal and $n > d + 3$)

$$\hat{p}_A(s_{obs}|\phi(\theta)) = (2\pi)^{-d/2}\frac{c(d, n-2)}{c(d, n-1)(1-1/n)^{d/2}}|M_n(\theta)|^{-(n-d-2)/2} \\ \psi\left(M_n(\theta) - (s_y - \mu_n(\theta))(s_y - \mu_n(\theta))^\top/(1-1/n)\right)^{(n-d-3)/2},$$

where $M_n(\theta) = (n-1)\Sigma_n(\theta)$. It is interesting to note that this estimator is a mixture of a discrete and a continuous random variable (a realisation of the estimator can be identically 0 with positive probability). Thus, if this estimator is used within a Monte Carlo method, the target distribution is proportional to $\mathcal{N}(s_{obs}; \mu(\theta), \Sigma(\theta))\pi(\theta)$ regardless of the value of n (under the multivariate normality assumption). Price et al. (2018) referred to this method as uBSL, where 'u' denotes unbiased.

To sample from the BSL posteriors, an MCMC algorithm can be used, for example. We refer to this as MCMC BSL, which is shown in Algorithm 1. Given the insensitivity of the BSL posteriors to the value of n, it is of interest to maximise the computational efficiency of the MCMC method. For large n, the SL is estimated with high precision but the cost per iteration is high. Conversely, for small n, the cost per iteration is low but the SL is estimated less precisely, which reduces the MCMC acceptance rate. Price et al. (2018) found empirically

that the value of n that leads to an estimated log SL (at a θ with high BSL posterior support) with a standard deviation of roughly 2 produces efficient results. However, Price et al. (2018) also found that there a wide variety of n values that lead to similar efficiency. When the unbiased SL is used in place of the SL shown in Algorithm 1, the MCMC uBSL algorithm is obtained. In the examples of Price et al. (2018), MCMC BSL and MCMC uBSL have a similar efficiency. We also note that the MCMC BSL posteriors appear to exhibit very slow convergence when starting at a point with negligible posterior support. The reason for this is that the SL is estimated with a large variance when the observed statistic s_{obs} lies in the tail of the actual SL. Thus additional research is required on more sophisticated methods for sampling from the BSL posteriors.

which to use for binary discrete output of logistic regression

Algorithm 1 MCMC BSL algorithm. The inputs required are the summary statistic of the data, s_{obs}, the prior distribution, $p(\theta)$, the proposal distribution q, the number of iterations, T, and the initial value of the chain θ^0. The output is an MCMC sample $(\theta^0, \theta^1, \ldots, \theta^T)$ from the BSL posterior. Some samples can be discarded as burn-in if required.

Simulate $s_{1:n} \overset{\text{iid}}{\sim} p(\cdot|\theta^0)$
Compute $\phi^0 = (\mu_n(\theta^0), \Sigma_n(\theta^0))$
for $i = 1$ to T **do**
  Draw $\theta^* \sim q(\cdot|\theta^{i-1})$
  Simulate $s^*_{1:n} \overset{\text{iid}}{\sim} p(\cdot|\theta^*)$
  Compute $\phi^* = (\mu_n(\theta^*), \Sigma_n(\theta^*))$
  Compute $r = \min\left(1, \frac{\mathcal{N}(s_{obs};\mu_n(\theta^*),\Sigma_n(\theta^*))p(\theta^*)q(\theta^{i-1}|\theta^*)}{\mathcal{N}(s_{obs};\mu_n(\theta^{i-1}),\Sigma_n(\theta^{i-1}))p(\theta^{i-1})q(\theta^*|\theta^{i-1})}\right)$
  if $\mathcal{U}(0,1) < r$ **then**
    Set $\theta^i = \theta^*$ and $\phi^i = \phi^*$
  else
    Set $\theta^i = \theta^{i-1}$ and $\phi^i = \phi^{i-1}$
  end if
end for

The BSL method has been applied in the literature. Fasiolo et al. (2016) used BSL for posterior inference for state space models in ecology and epidemiology based on data reduction and compared it with particle Markov chain Monte Carlo (Andrieu et al., 2010). Hartig et al. (2014) implemented BSL for a forest simulation model.

BSL could be seen as a direct competitor with ABC as they are both simulation-based methods and differ only in the way the intractable summary statistic likelihood is approximated. Importantly, BSL does not require the user to select a discrepancy function, as one is naturally induced via the multivariate normal approximation. The simulated summary statistics in BSL are automatically scaled, whereas an appropriate weighting matrix to compare summary statistics in ABC must be done manually. As noted in Blum (2010) and Drovandi et al. (2015), ABC uses a non-parametric approximation of the summary statistic likelihood based on similar procedures used in kernel density estimation. From this point of view, the ABC approach may be more accurate when the summary statistic s_{obs} is low dimensional, however the accuracy/efficiency trade-off is less clear when the summary statistic s_{obs} is high dimensional. Price et al. (2018) demonstrated on a toy example that BSL becomes increasingly more computationally efficient than ABC as the dimension of the summary statistic grows

beyond 2. Furthermore, Price et al. (2018) demonstrated that BSL outperformed ABC in a cell biology application with 145 summary statistics.

2.3 Accelerating synthetic likelihood

As with ABC, the SL method is very simulation intensive. There have been several attempts in the literature to accelerate the SL method by reducing the number of model simulations required. Meeds and Welling (2014) assumed that the summary statistics are independent and during their MCMC BSL algorithm fit a Gaussian process (GP) to each summary statistic output as a function of the model parameter θ. The Gaussian process (GP) is then used to predict the model output at proposed values of θ, provided that the prediction is accurate enough. If the GP prediction cannot be performed with sufficient accuracy, more model simulations are taken at that proposed θ and the GP is re-fit for each summary statistic. The independence assumption of the summary statistics is questionable, and may overstate the information contained in s_{obs}.

In contrast, Wilkinson (2014) used a GP to model the SL as a function of θ directly and use the GP to predict the SL at new values of θ. The GP is fit using a history matching approach (Craig et al., 1997). Once the final GP fit is obtained, an MCMC algorithm is used with the GP prediction used in place of the SL.

Moores et al. (2015) considered accelerating Bayesian inference for the Potts model, which is a complex single parameter spatial model. Simulations are performed across a pre-defined grid with the mean and standard deviation of the summary statistic (which turns out to be sufficient in the case of the Potts model, as it belongs to the exponential family) estimated from these simulations. Non-parametric regressions are then fitted individually to the mean and standard deviation estimates in order to produce an estimate of the mappings $\mu(\theta)$ and $\sigma(\theta)$ across the space of θ, where σ is the standard deviation of the summary statistic. The regressions are then able to predict the mean and standard deviation of the summary statistic at θ values not contained in the grid. Further, the regression also smooths out the mappings, which are estimated using a finite value of n. The estimated mapping is then used in a sequential Monte Carlo Bayesian algorithm.

2.4 Example

Cell motility, cell proliferation and cell-to-cell adhesion play an important role in collective cell spreading, which is critical to many key biological processes, including skin cancer growth and wound healing (e.g. Cai et al. (2007); Treloar et al. (2013)). The main function of many medical treatments is to influence the biology underpinning these processes (Decaestecker et al., 2007). In order to measure the efficacy of such treatments, it is important that estimates of the parameters governing these cell spreading processes can be obtained along with some characterisation of their uncertainty. Agent-based computational models are frequently used to interpret these cell biology processes since they can produce discrete image-based and movie-based information which is ideally suited to collaborative investigations involving applied mathematicians and experimental cell biologists. Unfortunately, the likelihood functions for these models are computationally intractable, so standard statistical inferential methods

for these models are not applicable.

To deal with the intractable likelihood, several papers have adopted an ABC approach to estimate the parameters (Johnston et al., 2014; Vo et al., 2015a,b). One difficulty with these cell biology applications is that the observed data are typically available as sequences of images and therefore it is not trivial to reduce the dimension of the summary statistic to a suitable level for ABC while simultaneously retaining relevant information contained in the images. For example, Johnston et al. (2014) considered data collected every 5 minutes for 12 hours but only analyse images at 3 time points. Vo et al. (2015a) reduced images initially down to a 15 dimensional summary statistic, but perform a further dimension reduction based on the approach of Fearnhead and Prangle (2012) to ensure there is one summary statistic per parameter.

Here we will re-analyse the data considered in Treloar et al. (2013) and Vo et al. (2015a). The data consist of images of spatially expanding human malignant melanoma cell populations. To initiate each experiment, either 20,000 or 30,000 cells are approximately evenly distributed within a circular barrier, located at the centre of the well. Subsequently, the barriers are lifted and population-scale images are recorded at either 24 hours or 48 hours, independently. Furthermore, there are two types of experiments conducted. The first uses a treatment in order to inhibit cells giving birth (cell proliferation) while the second does not use the treatment. Each combination of initial cell density, experimental elapsed time and treatment is repeated 3 times, for a total of 24 images. The reader is referred to Treloar et al. (2013) for more details on the experimental protocol. For simplicity, we consider here the 3 images related to using 20,000 initial cells, 24 hours elapsed experimental time and no cell proliferation inhibitor.

In order to summarise an image, Vo et al. (2015a) considered 6 sub-regions along a transect of each image. The position of the cells in these regions is mapped to a square lattice. The number of cells in each sub-region is counted, together with the number of isolated cells. A cell is identified as isolated if all of its nearest neighbours (north, south, east, west) are unoccupied. For each region, these summary statistics are then averaged over the three independent replicates. We refer to these 12 summary statistics as $\{c_i\}_{i=1}^6$ and $\{p_i\}_{i=1}^6$, where c_i and p_i are the number of cells and the percentage of isolated cells (averaged over the 3 images) for region i, respectively. Vo et al. (2015a) also estimated the radius of the entire cell colony using image analysis. Thus Vo et al. (2015a) included three additional summary statistics, $(r_{(1)}, r_{(2)}, r_{(3)})$, which are the estimated and ordered radii for the three images. For more details on how these summary statistics are obtained, the reader is referred to Vo et al. (2015a). This creates a total of 15 summary statistics, which is computationally challenging to deal with in ABC. As mentioned earlier, Vo et al. (2015a) found it beneficial to apply the technique of Fearnhead and Prangle (2012), which uses a regression to estimate the posterior means of the model parameters from the initial summaries, which are then used as summary statistics in a subsequent ABC analysis. Here we attempt to see whether or not BSL is able to accommodate the 15 summary statistics, and compare the results with the ABC approach of Vo et al. (2015a).

Treloar et al. (2013) and Vo et al. (2015a) considered a discretised time and space (two-dimensional lattice) stochastic model to explain the cell spreading process of melanoma cells. For more details on this model, the reader is referred to Treloar et al. (2013) and Vo et al. (2015a). The model contains three parameters: P_m (probability that an isolated agent can

move to a neighbouring lattice site in one time step), P_p (probability that an agent will attempt to proliferate and deposit a daughter at a neighbouring lattice site within a time step) and q (the strength of cell-to-cell adhesion, that is, cells sticking together). These model parameters can then be related to biologically relevant parameters such as cell diffusivity and the cell proliferation rate. Here we will report inferences in terms of the parameter $\theta = (P_m, P_p, q)$.

Here we consider a simulated dataset with $P_m = 0.1$, $P_p = 0.0012$ and $q = 0.2$ (same simulated data as analysed in Vo et al. (2015a)) and the real data. We ran BSL using Algorithm 1 with independent $\mathcal{U}(0, 1)$ prior distributions on each parameter. We used a starting value and proposal distribution for the MCMC based on the results provided in Vo et al. (2015a), so we do not apply any burn-in. We also applied the uBSL algorithm. The BSL approaches were run with $n = 32$, 48, 80 and 112 (the independent simulations were farmed out across 16 processors of a computer node). To compare the efficiency of the different choices of n we considered the effective sample size (ESS) for each parameter divided by the total number of model simulations performed multiplied by a large constant scalar to increase the magnitude of the numbers (we refer to this as the normalised ESS).

Marginal posterior distributions for the parameters obtained from BSL and uBSL for different values of n are shown in Figures 1 and 2, respectively. It is evident that the posteriors are largely insensitive to n, which is consistent with the empirical results obtained in Price et al. (2018). The normalised ESS values and MCMC acceptance rates for the BSL approaches are shown in Table 1 for different values of n. The efficiency of BSL and uBSL appears to be similar. The optimal value of n out of the trialled values appears to be 32 or 48. However, even $n = 80$ is relatively efficient. For $n = 112$ the increase in acceptance rate is relatively small given the extra amount of computation required per iteration.

We also applied the BSL approaches with similar settings to the real data. The posterior results are presented in Figures 3 and 4. Again we found the results are relatively insensitive to n. Table 2 suggests that $n = 48$ or $n = 80$ are the most efficient choices for n out of those trialled. However, it is again apparent that there are a wide variety of n values that lead to similar efficiency.

Table 1: Sensitivity of BSL/uBSL to n for the simulated data of the cell biology example with regards to MCMC acceptance rate, normalised ESS for each parameter.

n	acc. rate (%)	ESS P_m	ESS P_p	ESS q
32	17/17	96/114	86/113	115/126
48	27/32	95/103	93/92	110/115
80	35/37	82/76	74/67	106/89
112	38/40	61/65	61/58	68/70

The results, in comparison to those obtained in Vo et al. (2015a), are shown in Figure 5 for the simulated data and Figure 6 for the real data. From Figure 5 it can be seen that BSL approaches produced results similar to that of ABC for the simulated data. It appears that BSL is able to accommodate the 15 summary statistics directly without further dimension reduction. However, it is clear that the dimension reduction procedure of Vo et al. (2015a) performs well. From Figure 6 (real data) it is evident that ABC and the BSL approaches produce similar posterior distributions for P_p and q. For P_m, there is a difference of roughly

Table 2: Sensitivity of BSL/uBSL to n for the real data of the cell biology example with regards to MCMC acceptance rate, normalised ESS for each parameter.

n	acc. rate (%)	ESS P_m	ESS P_p	ESS q
32	8/9	46/51	38/45	41/43
48	17/18	76/71	56/63	70/54
80	27/28	66/64	66/60	68/58
112	32/33	58/60	51/54	54/43

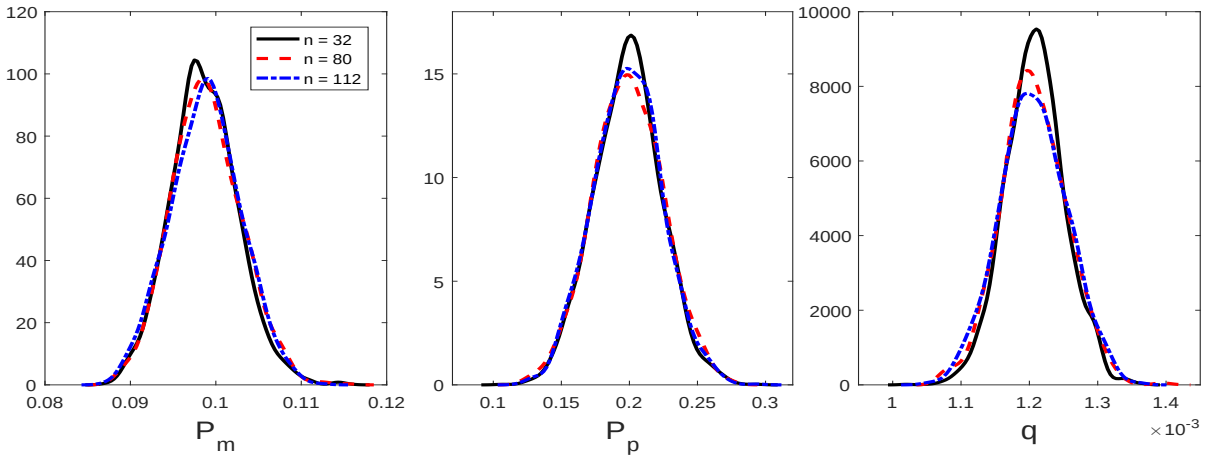


Figure 1: Posterior estimates for P_m, P_p and q based on the simulated data for the melanoma cell biology application using MCMC BSL for different values of n.

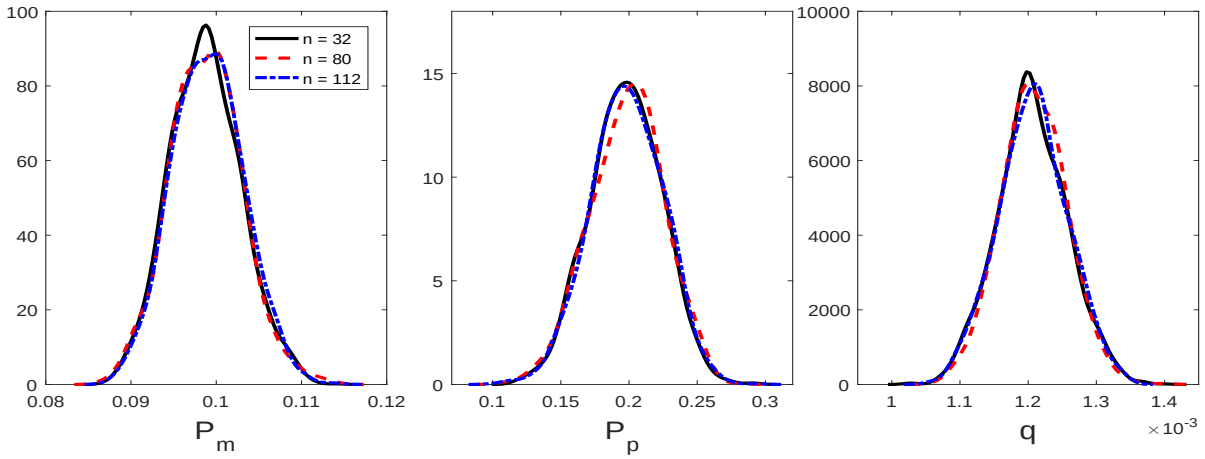


Figure 2: Posterior estimates for P_m, P_p and q based on the simulated data for the melanoma cell biology application using MCMC uBSL for different values of n.

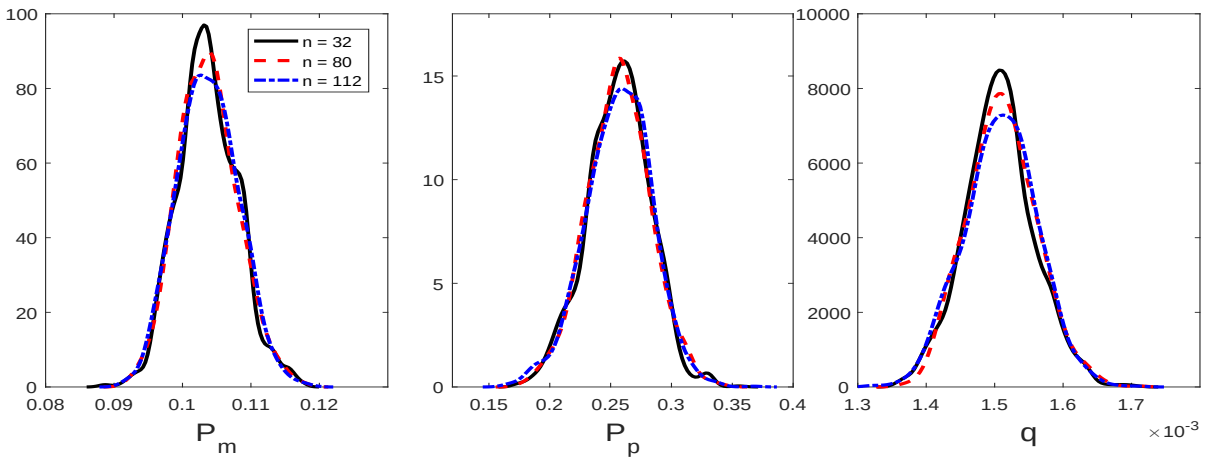


Figure 3: Posterior estimates for P_m, P_p and q based on the real data for the melanoma cell biology application using MCMC BSL for different values of n.

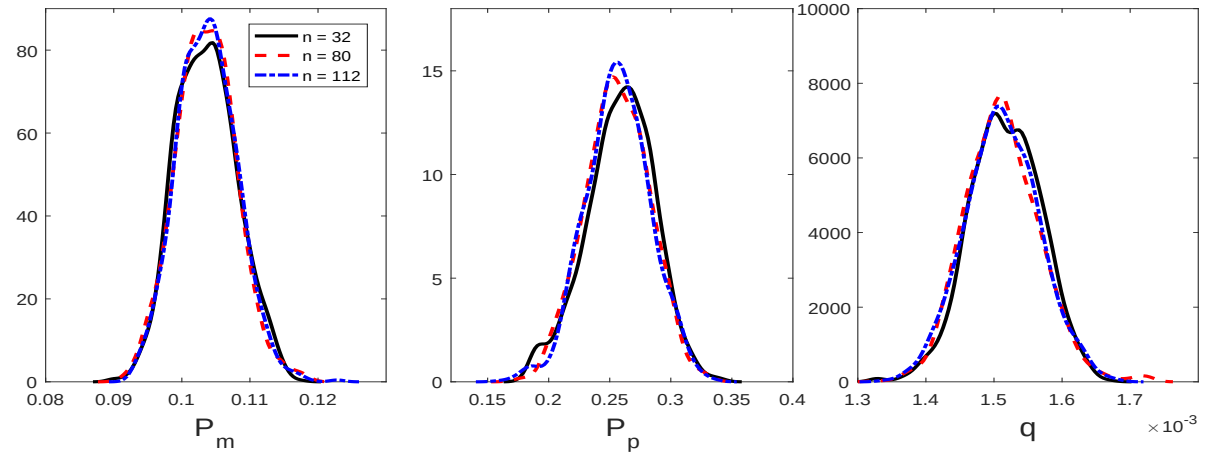


Figure 4: Posterior estimates for P_m, P_p and q based on the simulated data for the melanoma cell biology application using MCMC uBSL for different values of n.

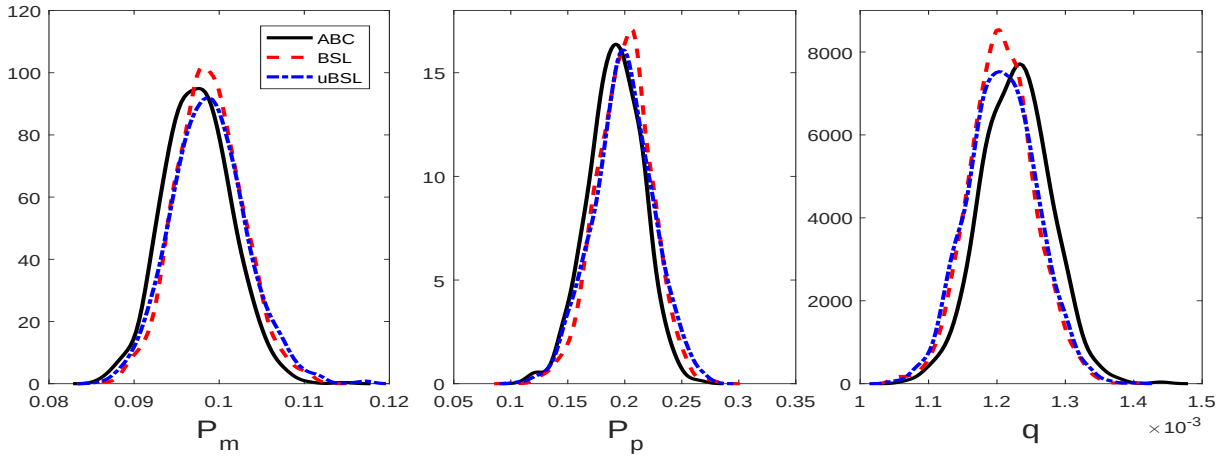


Figure 5: Posterior estimates for P_m, P_p and q for the melanoma cell biology application using the ABC approach of Vo et al. (2015a) (solid), BSL (dash) and uBSL (dot-dash) based on simulated data with $P_m = 0.1$, $P_p = 0.0012$ and $q = 0.2$. The BSL results are based on $n = 48$.

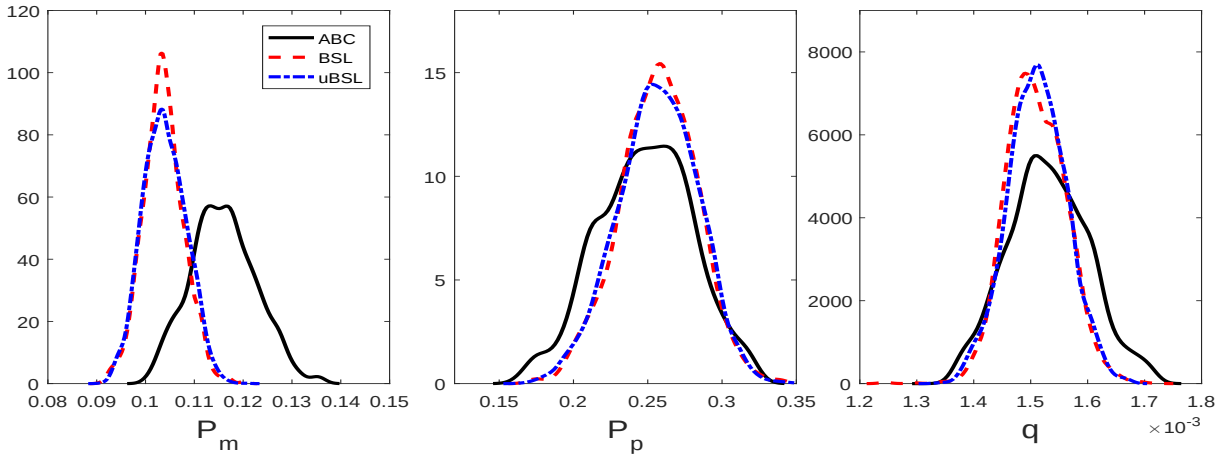


Figure 6: Posterior estimates for P_m, P_p and q for the melanoma cell biology application using the ABC approach of Vo et al. (2015a) (solid), BSL (dash) and uBSL (dot-dash) based on real data. The BSL results are based on $n = 48$.

0.01 between the posterior means of the BSL and ABC approaches and an increase in precision for BSL. This discrepancy for the real data not apparent in the simulated data requires further investigation. One potential source of error for BSL is the multivariate normal assumption. The estimated marginal distributions of the summary statistics (using $n = 200$) when the parameter is $\theta = (0.1, 0.0015, 0.25)$ is shown in Figure 7. All distributions seem quite stable but there is an indication of non-normality for some of the summary statistics. Given the results in Figure 5 and 6, it appears that BSL is showing at least some robustness to this lack of normality.

3 Further Reading

Ong et al. (2018) developed a stochastic optimisation algorithm to obtain a variational approximation of the BSL posterior. The authors utilise an unbiased estimator of the log of the multivariate normal density due to Ripley (1996, pp. 56). Ong et al. (2018) demonstrated that significant computational savings can be achieved relative to MCMC BSL, at the expense of resorting to a parametric approximation of the posterior. This work has been extended by Ong et al. (2017) to higher dimensional summary statistic and parameter spaces.

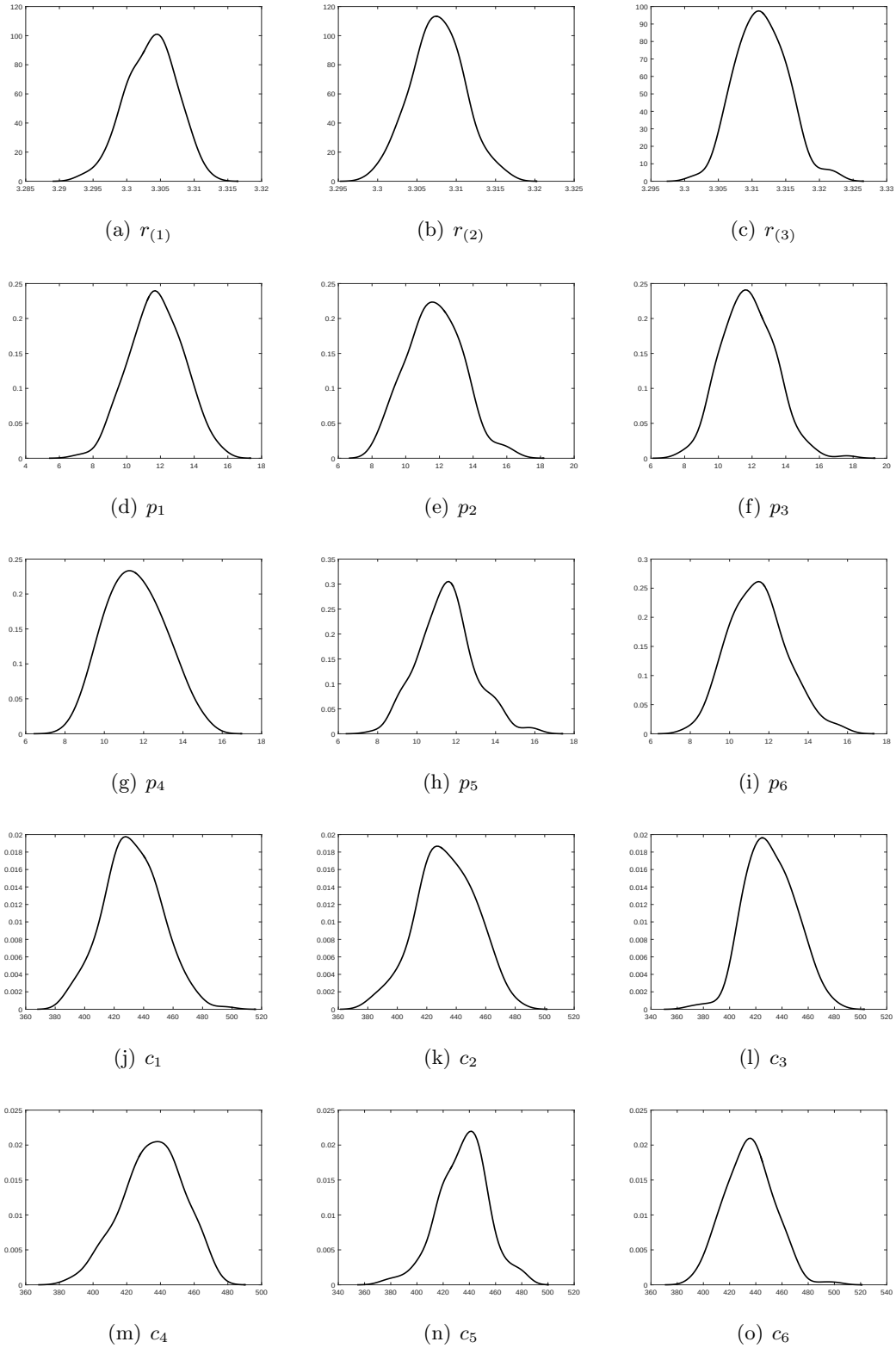
An et al. (2016) and Ong et al. (2017) considered shrinkage estimators of the covariance matrix of the model summary statistic in order to reduce the number of simulations required to estimate the synthetic likelihood.

Pham et al. (2014) replaced the ratio of intractable summary statistic likelihoods of the Metropolis-Hastings ratio in an MCMC algorithm with the outcome of a classification algorithm. Datasets are simulated under the current parameter and proposed parameter with the former observations labelled as class 1 and the latter labelled as class 2. A classifier, Pham et al. (2014) used random forests, is then applied. From the fitted classifier, the odds for the value of the observed summary statistic s_{obs} is computed and used as a replacement to the ratio of intractable likelihoods. Pham et al. (2014) noted that BSL is a special case of this approach when classical quadratic discriminant analysis is adopted as the classifier.

Everitt et al. (2017) suggested that the SL can be used to perform Bayesian model selection in doubly intractable models, which contain an intractable normalising constant that is a function of θ. Such models can occur in complex exponential family models such as exponential random graph models for social networks and the Potts model for image analysis. Everitt et al. (2017) developed computational algorithms in order to produce an SL approximation to the evidence $p(s_{obs}) = \int_\theta p(s_{obs}|\theta)\pi(\theta)d\theta$ for each model.

4 Bayesian empirical likelihood

ABC is a popular computational method of choice not only when there is no likelihood, but also when the likelihood is available but difficult or impossible to evaluate. Another popular idea is to replace the likelihood itself with an empirical alternative. This so-called empirical likelihood (EL) can be embedded within an ABC algorithm or provide an alternative to ABC. The approach is appealing even for less complex models if there is a concern that the model is poorly specified. For instance, if the likelihood is a mixture but is misspecified as



(a) $r_{(1)}$ (b) $r_{(2)}$ (c) $r_{(3)}$

(d) p_1 (e) p_2 (f) p_3

(g) p_4 (h) p_5 (i) p_6

(j) c_1 (k) c_2 (l) c_3

(m) c_4 (n) c_5 (o) c_6

Figure 7: Estimated marginal distributions of the 15 summary statistics using $n = 200$ for the melanoma cell biology applications when $P_m = 0.1$, $P_p = 0.0015$ and $q = 0.25$.

a single normal distribution, the corresponding parameter estimates, intervals and inferences may exhibit unacceptably poor behaviour (Chen and Quin, 2003). In this case, normal approximation confidence intervals perform poorly in the area of interest, i.e. the lower tail, but intervals based on an EL are shown to perform as well as intervals based on a correctly specified mixture model.

EL has been shown to have good small sample performance compared with methods that rely on asymptotic normality. Moreover, it enables distribution-free tests without simulation, and provides confidence intervals and regions that have appealing theoretical and computational properties (Owen, 1988, 2001). Some of these features are discussed in more detail below.

Close parallels with the EL approach have been drawn with estimating equations (Qin and Lawless, 2001; Grendar and Judge, 2007), kernel smoothing in regression (Chen and Van Keilegom, 2009a,b; Haardle, 1990; Fan and Gijbels, 1996), maximum entropy (Rochet, 2012) and functional data analysis (Lian, 2012). We do not elaborate on these associations in this chapter, but refer the interested reader to the cited references.

EL approaches have been developed in both frequentist and Bayesian contexts. This section provides a brief overview of the method under both paradigms. We then focus on a particular algorithm, BC_{el}, proposed by Mengersen et al. (2013), which was first conceived as part of an ABC algorithm but was then developed independently of the ABC architecture.

4.1 Empirical Likelihood

Empirical likelihood (EL) has been a topic of active research and investigation for over a quarter of a century. Although similar ideas were established earlier (see, for example, the proposal of a "scale-load" method for survey sampling by Hartley and Rao (1968)), EL was formally introduced by Owen (1988) as a form of robust likelihood ratio test.

Following Owen (1988) and Owen (2001), assume that we have i.i.d. data $Y_i, i = 1, ..., n$ from a distribution F. An EL denoted $L(F)$ is given by

$$L(F) = \prod_{i=1}^{n} F(\{y_i\}).$$

The likelihood ratio and corresponding confidence region are given by, respectively,

$$R(F) = L(F)/L(\hat{F}) \quad \text{and}$$

$$\{T(F)|R(F) \geq r\}$$

where $\hat{F}$ is the empirical distribution function and for some appropriate value of r.

Given parameters of interest θ and an appropriate sufficient statistic $T(F)$ for it, a profile likelihood and corresponding confidence region become, respectively,

$$\mathcal{R}(\theta) = \sup\{R(F)|T(F) = \theta\} \quad \text{and}$$

$$\{\theta|\mathcal{R}(F) \geq r\}.$$

If there are no ties, we let $p_i = F(\{y_i\})$, $p_i > 0$, $\sum_{i=1}^n p_i = 1$, and find that

$$L(F) = \prod_{i=1}^n p_i \quad ; \quad L(\hat{F}) = \prod_{i=1}^n 1/n$$

$$R(F) = \prod_{i=1}^n np_i \quad ; \quad \mathcal{R}(\theta) = \sup\left\{\prod_{i=1}^n np_i | T(F) = \theta\right\} .$$

Obvious adjustments are made to $L(F)$ and $L(\hat{F})$ if there are ties.

A fundamental result obtained by Owen (1988) is that if the mean θ_0 of the distribution F is finite and its covariance matrix is finite with rank $q > 0$, then as $n \to \infty$,

$$-2\log R(\theta_0) \to \chi^2_q.$$

This is the same as that obtained by Wilks' Theorem for the parametric setup. Thus for a $100(1-\alpha)\%$ confidence region, $r = r_0 = \exp(-\chi^2_{q,\alpha/2})$.

As a concrete example of EL, suppose that interest is in estimation of the mean, i.e., $\theta = E[Y]$. Then $T(\hat{F}) = n^{-1}\sum_{i=1}^n y_i$, with confidence region and profile likelihood given by, respectively,

$$\left\{\sum_{i=1}^n p_i y_i | p_i \geq 0, \sum_{i=1}^n p_i = 1, \prod_{i=1}^n np_i > r\right\} \quad \text{and}$$

$$R(\theta) = \sup\left\{\prod_{i=1}^n np_i | p_i > 0, \sum_{i=1}^n p_i = 1, \prod_{i=1}^n np_i = \theta\right\} .$$

Thus under certain conditions, a $(1-\alpha)$-level EL confidence interval for $\theta_0 = \bar{Y}$ is given by

$$\{\theta | r(\theta) \leq \chi^2_1(\alpha))$$

where $r(\theta) = -2\sum \log(n\hat{p}_i)$ is the log EL function and $\chi^2_1(\alpha)$ is the upper α quantile of the χ^2 distribution with one degree of freedom.

The above set-up can also be seen as an estimating equation problem, where the true value θ_0 satisfies the estimating equation

$$E_F[m(Y;\theta)] = 0$$

with $m(Y;\theta)$ denoting a vector-valued (estimating) function. Hence we can take $m(Y;\theta) = Y - \theta$ to indicate a vector mean, $m(Y;\theta) = I_{Y\in A} - \theta$ for $Pr(Y \in A)$, $m(Y;\theta) = I_{Y<\theta} - \alpha$ for the αth quantile of Y if Y is continuous, $m(Y;\theta) = I_{Y\leq\theta} - 0.5$ for the median, and so on.

More generally, we have one or more constraints of the form $E_F[h(Y,\theta)] = 0$, where the dimension of h sets the number of constraints in unequivocally defining the parameters of interest θ. Then the EL is defined as

$$L_{el}(\theta|y) = \max_p \prod_{i=1}^n p_i$$

for $p \in [0,1]^n$, with constraints

$$\sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i h(y_i, \theta) = 0 \ .$$

Perhaps surprisingly, there are relatively few Bayesian formulations of EL in the published literature. An earlier Bayesian ABC approach using an approximation of the EL based on the pairwise score equation was proposed by Pauli and Adimara (2010). The authors focused on establishing the validity of the procedure, arguing that its asymptotic properties were preferred over the approximations employed by Pauli et al. (2011). See also Ruli et al. (2016). Owen (2001) (Ch. 9) noted some parallels between EL and the Bayesian bootstrap (Rubin, 1981), and Rochet (2012) has suggested a Bayesian approach to generalised empirical likelihood, and generalised method of moments, via a form of maximum entropy. Chaudhuri and Ghosh (2011) describe Bayesian EL approaches in a spatial modelling context, as discussed in more detail below.

More direct research into Bayesian EL comprise a Monte Carlo study (Lazar, 2003) and two probabilistic studies (Schennach, 2005; Grendar and Judge, 2007). In contrast to the reported Bayesian bootstrap-type approaches of Owen (2001), Schennach (2005) and Rasuga (2006), Grendar and Judge (2007, 2009) proposed a Bayesian large deviations (law of large numbers) probabilistic interpretation and justification of EL. They showed that, in a parametric estimating equations setting, the EL method is an asymptotic instance of the Bayesian non-parametric maximum a-posteriori approach.

4.2 Features of EL

Since Owen's paper in 1988, the properties of EL have been comprehensively investigated and reviewed (Hall and La Scala, 1990; Owen, 2001).

EL methods have been favourably contrasted with common alternatives for estimation of complex models. For example, a natural competitor is calibration, which proceeds by choosing, by some method, parameter values that match selected features of the observed data. This can be difficult for richly parametrised models with strong correlation structure. EL can be perceived as a more statistically formal method of calibration in that it uses moments for matching. Another common competitor, maximum likelihood, requires the definition, estimation and maximisation of a likelihood and can be both analytically and computationally demanding for complex models. In contrast, EL requires only summary (moment) statistics and can perform inference on an approximate likelihood, but inherits the properties of standard likelihood (Owen, 2001). These properties of standard likelihood are principally obtained by appeal to Wilks' Theorem (Qin and Lawless, 2001).

As described above, likelihood ratio confidence regions can be constructed by EL that often do not require estimation of the variance (Chen and Van Keilegom, 2009a,b) and have the same order of magnitude error as their parametric counterparts. This also applies for more general regression contexts (Chen, 1993, 1994; Chen and Cui, 2006; Chen and Gao, 2007). The confidence regions constructed in this manner respect the boundaries of the support of the target parameter and are more natural in shape and orientation of the data since they

contour a log-likelihood ratio. In particular, they are often superior to confidence regions based on asymptotic normality when the sample size is small. The confidence regions can be further improved by applying Bartlett's correction, $(1_a/n)\chi^2_{q,\alpha}$, where a involves higher order moments of Y (DiCiccio et al., 1991).

A key assumption underlying standard EL is that the random variables are independent with a common distribution. An analogue, the weighted EL or exponentially tilted distribution accommodates data that are independent but not necessarily identically distributed. This approach was introduced by Schennach (2005) and has been taken up by a large number of authors (Owen, 2001; Kitamura, 2006; Glenn and Zhao, 2007). Chaudhuri and Ghosh (2011) contrast the two approaches as follows. They frame the EL as

$$l(\theta) = \prod_{i=1}^{m} \hat{w}_i(\theta)$$

where

$$\hat{w}(\theta) = \arg\max_{w \in \mathcal{W}_0} \sum_{i=1}^{m} f\{w_i(\theta)\}$$

for some specified function f. They then consider standard EL as a form of constrained maximum of a nonparametric likelihood since for a given θ, $l(\theta)$ equals the EL when $f(w_i) = \log(w_i)$, and the exponentially tilted likelihood as a form of maximum entropy such that $f(w_i) = -w_i \log(w_i)$. As these authors discussed, the exponentially tilted likelihood can also be seen as a profile likelihood for θ.

Moreover, Schennach (2005) shows that this reformulation of the maximisation problem of the EL allows for a probabilistic interpretation which justifies its use in a Bayesian setting. More precisely, the posterior distribution for a parameter of interest θ may be seen as

$$\pi(\theta|y) = \pi(\theta) \int_{\Psi} L(\theta, \psi|y)\pi(\psi|\theta)d\psi,$$

where ψ represents a (potentially infinite-dimensional) nuisance parameter which absorbs all those aspects of the model not described by the parameter of interest θ. The information contained in the nuisance parameter may be discretised by a vector of parameter $\xi = (\xi_1, \ldots, \xi_N)$ with $N \to \infty$. The nuisance parmater ξ may then be given a prior which shares the Dirichlet prior's property of providing posteriors which assign probability one to distributions supported by the sample. Schennach (2005) shows then this reformulation has a computationally convenient representation, for which the posterior of the parameter of interest θ may be obtained through

$$\pi(\theta|y) = \pi(\theta) \prod_{i=1}^{n} p_i^\star$$

where $p^\star = (p_1^\star, \cdots, p_n^\star)$ are the weights obtained as solution of the maximization problem

$$L_{\text{BETEL}}(\theta) = \max_{p^\star} \sum_{i=1}^{n} p_i^\star \log p_i^\star$$

under constraints $p^\star \in [0,1]^n, \sum_{i=1}^n p_i^\star = 1, \sum_{i=1}^n p_i^\star h(y_i, \theta) = 0$, where "BETEL" stands for "Bayesian exponentially tilted likelihood". This method may be called "Bayesian exponentially tilted EL", because it uses the exponential tilting proposed in Efron (1981) and has a Bayesian interpretation. This version of the EL will be used in Section 4.7.

Glenn and Zhao (2007) examined the robustness properties of the estimates arising from the tilted distribution. For example, whereas the root mean squared error (RMSE) of the EL estimator for the mean increases as the non-iidness of the sample increases, the RMSE of the weighted EL estimator remains closer to its theoretical value. Other extensions to standard EL, such as the continuous updating estimator, have also been proposed (Hansen et al., 1996).

In a Bayesian framework, the standard and exponentially tilted likelihoods have been shown to be appropriate for Bayesian inference for a range of set-ups and under certain conditions on the prior, particularly for a prior with sufficiently large variance (Monahan and Boos, 1992; Lazar, 2003; Chaudhuri and Ghosh, 2011).

Notwithstanding these attractions, there are some drawbacks in applying EL. One substantive issue is the formulation of the estimating equations. The number of equations is one issue: there should be at least as many as the dimension of the parameter space, but any more than this (which may be available and desirable from the perspective of model description) has been argued to adversely affect inference (Qin and Lawless, 2001). However, it is suggested by Mengersen et al. (2013) that this concern may not apply in all circumstances, in a Bayesian setup; this is illustrated in the g-and-k example given below.

4.3 Estimation

The most common approach to estimation of the EL is through the method of Lagrange multipliers. In general terms, this method aims to maximise $f(x)$ subject to a (multivariate) constraint $g(x) = 0$. This is achieved by finding $x^* = x^*(\lambda)$ maximising $f(x) = \lambda' g(x)$ such that $g(x') = 0$. Then x^* solves the constrained problem. Considering again the example of estimating $\theta = E[Y]$, the aim is to maximise

$$\log R(p_1, .., p_n) = \sum_{i=1}^n \log(np_i)$$

under the constraints

$$n \sum_{i=1}^n p_i (Y_i - \theta) = 0, \quad 1 - \sum_{i=1}^n p_i = 0 \ .$$

We write

$$G = \sum_{i=1}^n \log(np_i) - n\lambda \sum_{i=1}^n (Yi - \mu) - \gamma(1 - \sum_{i=1}^n p_i)$$

where λ and γ are the Lagrange multipliers. This can be solved to find a unique solution for $\lambda = \lambda(\theta)$.

There is a range of software for computing the EL, particularly targeted towards specific applications. A helpful repository and description of available code is on Art Owen's website. A powerful library available in the R software is the package 'emplik' (Zhou and Yang, 2014).

The underlying computational method is based on the Newton-Lagrange algorithm, whereby the Lagrangian function described above is solved by an application of Newton's method, which iterativly uses a second order Taylor approximation of $f(x)$ to find an optimal value x^* satisfying $f'(x^*) = 0$.

For example, the package `el.test` in the `emplik` library conducts a simple EL ratio test that returns -2 log likelihood ratio (`-2LLR`, which has an approximate chi-squared distribution under the null hypothesis), the associated p-value, the final value of the Lagrange multiplier (`lambda`), the gradient at the maximum (`grad`), the hessian matrix (`hess`), weights on the observations (`wts`) and the number of iterations performed (`nits`).

The following code, provided in the emplik documentation, illustrates two tests on a two-dimensional set of data: (i) $H_0 : \mu_1 = \mu_2$ and (ii) $H_0 : 2\mu_1 - \mu_2 = 0$.

```
# generate data
x <- matrix(c(rnorm(50,mean=1), rnorm(50,mean=2)), ncol=2,nrow=50)
y <- 2*x[,1]-x[,2]
# test hypothesis (i)
el.test(x, mu=c(1,2))
# test hypothesis (ii)
el.test(y, mu=0)
```

In one realisation of this code, the results of the first test were returned after four iterations, with weights ranging between 0.75 and 1.51, and with -2LL=1.50 and a p-value of 0.47 under the assumption that -2LL is approximately chi-square under H_0. The second null hypothesis returned a p-value of 0.22.

Examples of the use of the `emplik` library for survival analysis are given by Zhou (2015). Whereas `el.test` requires uncensored data, the packages developed by Zhou and embedded in the `emplik` library enable estimation of hazard functions, cumulative distribution functions and confidence bands for various types of censored data under a range of survival models.

As an example, the package `em.cen.EM` can be used to test the hypothesis $H_0 : \int g(t)dF(t) = \mu$ versus $H_a : \int g(t)dF(t) \neq \mu$, where $g(t)$ is a user supplied function. For instance, H_0 can be the test about the Kaplan-Meier mean and $g(t) = t$. The myeloma code in the Appendix illustrates this by testing $H_0 : F(10) = 0.2$ in the myeloma dataset incorporated in the `emplik` library. The code also finds the upper and lower confidence limit of a Wilks confidence interval. The output of this analysis provides a value -2LL and a corresponding p-value.

Bayesian EL methods are typically analysed by solving the EL using a Lagrange or similar method, then generating observations from the posterior distribution of the parameters of interest by an MCMC method. A more detailed description of this approach is given in the context of spatial modelling in the next section. An alternative approach, BC_{el}, which employs the `emplik` library to obtain the required likelihood values, is also detailed in a subsequent section.

4.4 EL in practice

The EL approach has been shown to be applicable in a broad range of contexts (Qin and Lawless, 2001). For example, following its formulation for estimation in linear regression (Owen, 1991), it has been extended to nonlinear, generalised, parameric, nonparametric and semiparametric models with and without missing data and censoring, time series models and varying-coefficient models; see the review of Chen and Van Keilegom (2009b) and the references therein. The approach has also been proposed for testing; see again Chen and Van Keilegom (2009b). Einmahl and McKeague (2003) have proposed omnibus tests based on EL for a wide range of hypothesis tests, including symmetry, exponentiality, independence and change of direction. Tests for stochastic ordering using EL have been proposed by El Barmi and McKeague (2013) and El Barmi and McKeague (2015).

Chaudhuri and Ghosh (2011) have proposed an EL approach for small area estimation and have suggested that the approach is also applicable to general random and mixed effects models. As the authors argue, EL overcomes the distributional assumptions of the more dominant parametric models as well as the linearity assumptions of the nonparametric models that have been proposed for this problem. In addition, EL avoids the need for resampling methods like jacknife and bootstrap to obtain mean squared error estimation. The authors' methodology is developed using a multivariate-t prior for the parameter vector θ and both the regular and exponentially tilted formulations for the EL.

A Bayesian EL approach for constructing intervals for the analysis of survey data has been explored by Rao and Wu (2010). This work builds on the EL approaches for complex survey analysis proposed by Chen and Sitter (1999) and Wu and Rao (2006). Rao and Wu (2010) provides a clear exposition of EL methods for sample surveys. The authors set up the problem as one in which N_t denotes the number of units $U = \{1, 2, ..., N_t\}$, in a finite population of size $N = \sum_{t=1}^{T} N_t$, that have the value y_t^*, and n_t denotes the number of units in the sample having this value $y_t^*, t = 1, ..., T$. The sample data are then reduced to a set of so-called scale-loads $(n_1, n_2, ..., n_T)', n_t \geq 0, n = \sum_{t=1}^{T} n_t$. Assuming a negligible sampling fraction, the likelihood can be approximated by using the multinomial distribution with a log likelihood given by

$$l(p) = \sum_{t=1}^{T} n_t \log(p_t)$$

with $p_t = N_t/N$, and the MLE of

$$\bar{Y} = \sum_{t=1}^{T} p_t y^n$$

is the sample mean

$$\bar{y} = \sum_{t=1}^{T} \hat{p}_t y^{n^*}, \quad \hat{p}_t = n_t/n.$$

The authors make the connection with the work of Chen and Sitter (1999) and argue that this 'scale-load' approach is "in the same spirit" as EL as described by Owen (1988).

As described above, survival analysis is another area that lends itself naturally to EL. The popular Kaplan-Meier curve is a nonparametric estimator of the survival function $S(t) =$

$P(T \geq t)$, where T denotes the time to an event. It is conceptually straightforward to see that S can be estimated as a maximum EL estimator. This field has been developed by a number of authors: see, for instance, Wang and Jing (2001) for a general exposition of the survival model, Murphy and van der Vaart (1997) for doubly censored data, Qin and Jing (1994) for Cox modelling using EL, and McKeague and Zhao (2002) for an EL approach to relative survival. The recent text by Zhou (2015) provides an excellent overview of the field as well as new models and computational algorithms, with associated R code to facilitate implementation. A simple illustration of an EL approach to survival analysis is provided in the next section.

Recent years have also seen an increase in popularity of EL for spatial modelling. Chaudhuri and Ghosh (2011) pioneered a Bayesian EL approach for small area estimation. Their model can accommodate continuous and discrete and area- and unit-level data, random and mixed effects, and the original and exponentially tilted empirical likelihoods.

A similar approach has also been proposed recently by Porter et al. (2015a) for this purpose. The so-called semiparametric hierarchical EL (SHEL) model can be applied to irregular lattices and irregularly spaced point-referenced data, and was shown to have improved mean squared prediction error compared with standard parametric analyses in a simulation study, a large community survey and a bird survey. In the SHEL model, EL is employed in an empirical data model, which is combined with a parametric process model that accounts for the spatial dependence through a rank-reduced intrinsic conditional autoregressive (ICAR) prior and, finally, with a model at the highest level of the hierarchy describing the parameter.

A companion paper by the same authors (Porter et al., 2015b) extends this work to a multivariate context, with focus on the Fay-Herriot (FH) model which is a mainstay in small area estimation. The argument is made that this approach encompasses spatial correlation (via the FH model) but avoids the usual restrictive Gaussian distributional assumptions (via EL).

One of the fields in which EL has been prominent is economics and related fields. For example, Riscado (2012) promote the use of EL as a natural framework for estimation of dynamic stochastic general equilibrium (DSGE) models for macroeconomic analysis, since these models represent complex economic systems as a constrained optimisation problem and can be described as a set of moment conditions. The authors favourably compare EL with calibration and ML approaches, since the model parameters have complex correlation structures that hinder calibration and are typically characterised by nonlinear systems of difference equations that have no closed form and hence hinder ML. The likelihood is thus often approximated and then estimated (and maximised) using methods such as the Kalman filter and sequential Monte Carlo. The authors interpret the EL approach as mapping from the set of moment conditions to the stochastic processes of the economic variables, and then performing estimation by inverting that mapping. As discussed above, the importance and very often the difficulty of defining a set of "good" moment conditions, or constraints, is highlighted in this setting.

4.5 The BC_{el} algorithm

A Bayesian EL algorithm was proposed by Mengersen et al. (2013). It was originally designed in the spirit of ABC, in that it avoids computation of the likelihood, but during its

development the authors realised that simulation from the likelihood could also be avoided and replaced with importance sampling. Thus the so-called BC_{el} algorithm generates values $\theta_i, i = 1, \ldots, M$ from the prior distribution for θ and uses the values $w_i = L_{el}(\theta_i|y)$ as (unnormalised) weights in an importance sampling (IS) framework. The basic BC_{el} sampler is given below. Of course, this IS algorithm will not be efficient if the posterior is very different to the prior. Later, we describe a more sophisticated algorithm based on adaptive multiple importance sampling (AMIS, Cornuet et al. (2012)).

Algorithm 2 BC_{el}, Mengersen et al. (2013)

for $i = 1$ to M **do**
  Generate θ_i from the prior distribution $\pi(\cdot)$
  Set the (unnormalised) weight $\omega_i = L_{el}(\theta_i|y)$
end for

4.5.1 Example 1

As a concrete example, consider estimation of the population mean θ based on a sample of observations $y_i, i = 1, ..., n$. In this case, two main decisions are required: the prior on θ and the constraints. The computation of the EL $L_{el}(\theta_i|y)$ can be performed using the `el.test` package in the `emplik` library in `R`, as described earlier in this chapter. In this case, the unnormalised weight ω_i is taken to be equal to the value of the empirical likelihood, which is calculated from the value of -2LLR obtained from the `el.test` function.

Suppose that a sample of size 100 is drawn from an (unknown) distribution $N(10, 1)$ and the aim is to estimate the population mean θ. A $N(-10, 30)$ prior is imposed on θ and a first-moment constraint is chosen, i.e., that the sample mean should equal the population mean. For the analysis, it is decided to run $M = 5000$ iterations, noting that in practice a smaller value of M can be used but care must be taken to check that the weight has not concentrated too strongly on a small number of sampled values of θ. A resampling step can be included to mitigate this, although at a cost of introducing additional variance. In this case, the algorithm becomes:

Algorithm 3 BC_{el} algorithm for Example 1

for $i = 1$ to M **do**
  Generate $\theta_i \sim N(0, 5)$
  Obtain $-2LL$ from `el.test(y,mu=0)`
  Let $\omega_i = \exp(-0.5 \times (-2LL))$
end for
Resample θ with probability ω
Calculate summary statistics of the resampled values of θ

Example R code for this algorithm is given below.

```
data = rnorm(100,10,1)
M = 5000; theta.propose=w=rep(0,M)
```

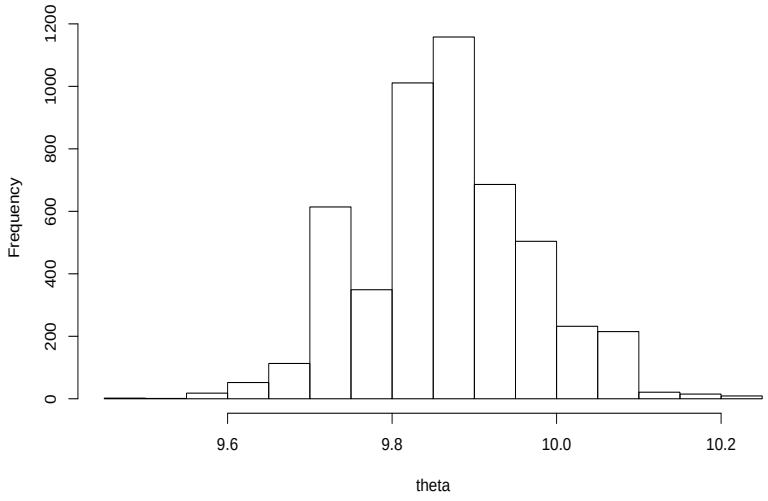


Figure 8: Histogram of draws from the BC_{el} posterior distribution of θ based on data generated from a $\mathrm{N}(\theta = 10, 1)$ distribution

```
for (i in 1:M){
theta.propose[i] = runif(1,-10,30)
el = el.test(data,mu=theta.propose[i])
w[i] = exp(-0.5*(el$'-2LLR'))
}
theta=sample(theta.propose,M,prob=w,replace=TRUE)
mean(theta); sd(theta); quantile(theta,probs=c(0.1,0.9))
hist(theta, main="",xlab="theta")
```

As noted above, the resampling step could be replaced with a weighted mean, standard deviation and quantiles. One realisation of this code provided the following estimates: $\hat{\theta} = 10.08$; s.d.$(\theta) = 0.12$; 95% CrI=$(9.88, 10.21)$. A histogram of the obtained sample of θ is given in Figure 8.

Mengersen et al. (2013) comment on the performance of this algorithm with different constraints, namely one, two and three central moments, $E[(X - \theta)] = 0$, $E[(X - \theta)^2] = 0$, $E[(X - \theta)^3] = 0$. They observe that one and two constraints work well, but three constraints performed more poorly. This was seen to support the general suggestion by Owen (2001) that the number of constraints should be equal to the number of parameters. Interestingly, this was not seen to be the case for the g-and-k distributions, as described in the next example.

A possible measure of the efficiency of the algorithm is the effective sample size (ESS). The ESS is reportedly a measure of the 'equivalent number of independent observations' in a sample, that is, the value that equates the obtained variance of the estimator of interest with the equivalent variance assuming an independent sample. For weighted samples as in EL, the

ESS can be estimated as

$$\text{ESS} = 1/\sum_{i=1}^{M}\{w_i/\sum_{j=1}^{M} w_j\}^2 \ .$$

Kish (1965) argues that this substitution (of the EL for the exact likelihood) can be employed in any algorithm that samples from a posterior distribution. For example, it can be employed in composite likelihoods which are commonly used in areas such as population genetics where the likelihoods are known but complex, and hence computationally difficult. The 'traditional' composite likelihood approach decomposes the target distribution $\pi(\theta)L(\theta|y)$ into several multivariate Student t distributions. In the BC_{el} approach, the EL is used instead. The computation is achieved using AMSI, which can be parallelised on a multi-core computer. The method can also be tailored for some non-i.i.d. problems such as dynamic models with AR structure, although the challenge here is in selecting appropriate constraints; see Mengersen et al. (2013) for details.

4.6 Example 2

We illustrate the use of BC_{el} by expanding on the discussion by Mengersen et al. (2013) of quantile distributions. These distributions are appealing for ABC in general, and BC_{el} in particular: there is typically no closed form expression for the likelihood, so regular algorithms such as MCMC are not immediately applicable; and it is fast and easy to obtain simulations from a quantile function via an inversion algorithm.

There is a body of literature on using ABC for estimation of quantile distributions. Allingham et al. (2009) proposed an ABC-MCMC algorithm in which draws of the parameters of the quantile distribution are based on a Metropolis algorithm with a Gaussian proposal distribution, and are accepted based on the rule $||D - D'|| < h$, where D is the entire set of order statistics, $||\cdot||$ is the Euclidean norm and h is heuristically chosen after inspection of a histogram of $||D - D'||$ obtained from a preliminary run using a very large value of h. Peters and Sisson (2006) also developed an ABC-MCMC algorithm for complex quantile functions. A range of improvements in the MCMC algorithm, selecting low-dimensional summary statistics and methods of choosing h have since been suggested (Prangle, 2011; McVinish, 2012). Sequential Monte Carlo approaches for multivariate extensions of quantile distributions have also been proposed (Drovandi and Pettitt, 2011).

The g-and-k distribution is a popular example of a quantile distribution. This is a transformation of the standard normal distribution function, as follows:

$$Q(z(p);\theta) = a + b\left(1 + c\frac{1-\exp(-gz(p))}{1+\exp(-gz(p))}\right)(1+z(p)^2)^k z(p)$$

where $\theta = (a, b, g, k)$ and c is commonly set fixed at 0.8; see Rayner and MacGillivray (2002). Here p denotes the pth quantile from the g-and-k distribution and $z(p)$ is the corresponding quantile of the standard normal distribution. Thus simulation from the g-and-k distribution requires only the generation of uniform$(0, 1)$ variates.

Figure 9 shows the estimated cdf of a standard normal distribution based on a g-and-k approximation, using the basic BC_{el} procedure described in Algorithm 2. The parameters of the

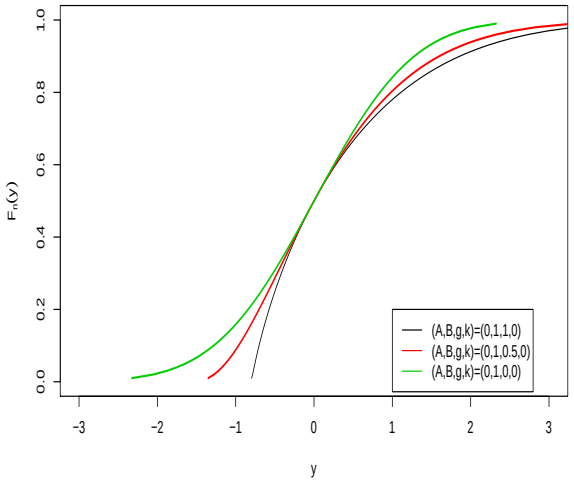


Figure 9: Cumulative distribution functions for three g-and-k distributions.

g-and-k distribution corresponding to a $N(0,1)$ distribution are $\theta = (0,1,0,0)$. The analysis was based on 1000 observations, 100,000 iterations and 10,000 resampled parameter values. The percentiles $(0.1, 0.25, 0.5, 0.75, 0.9)$ were chosen arbitrarily to form the constraints for EL and all parameters were generated from a $U[0,5]$ prior distribution.

The Bayes Factor R code available in the Appendix illustrates the ease with which Bayes Factors (BF) can be computed for g-and-k distributions using EL. The example assumes a true model (Model 1) with $(A,B,g,k) = (0,1,1,0)$ versus two alternatives, $(0,1,0.5,0)$ (Model 2) and $(0,1,0,0)$ (Model 3). Here, all models have zero mean ($A = 0$) and unit variance ($B = 1$) but differ in the degree of skewness, with Model 3 having no skewness ($g = 0$) and hence representing a standard normal distribution. The cumulative distributions functions for these three models are depicted in Figure 9. Two sample sizes of 100 and 500 and five constraints $(0.1, 0.25, 0.5, 0.75, 0.9)$ are considered. The resultant boxplots shown in Figure 10 confirm that Model 1 is preferred over both of the alternative models, with a stronger log BF obtained for the larger sample size as anticipated.

4.6.1 Example 3

Mengersen et al. (2013) also describe a variation on the basic BC_{el} algorithm which employs AMIS in order to improve computational efficiency over plain importance sampling. The so-called BC_{el}-AMIS sampler employs multivariate Student $t_3(\cdot|m,\Sigma)$ distributions (3 degrees of freedom, mean m, covariance matrix Σ) as importance sampling distributions, as described in the following algorithm. The output of this algorithm is a weighted sample $\theta_{t,i}$ of size MT_M.

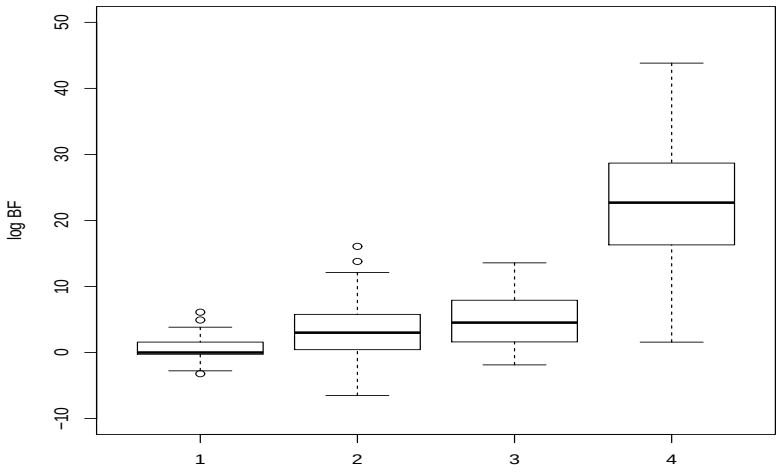


Figure 10: Boxplots of Bayes Factors comparing three g-and-k distributions; data were generated from Model 1. On the x-axis: (1) Model 1 vs Model 2 with sample size $n = 100$, (2) Model 1 vs Model 2 with sample size $n = 500$, (3) Model 1 vs Model 3 with sample size $n = 100$, (4) Model 1 vs Model 3 with sample size $n = 500$.

Algorithm 4 BC_{el}-AMIS

for $i = 1$ to M **do**
  Generate $\theta_{1,i}$ from the prior distribution $\pi(\cdot)$.
  Set the weight $\omega_{1,i} = L_{el}(\theta_{1,i}|y)$.
end for
for $t = 2$ to T_M **do**
  Compute the weighted mean m_t and weighted variance matrix Σ_t of the $\theta_{s,i}(1 \leq s \leq t-1, 1 \leq i \leq M)$.
  Denote by $q_t(\cdot)$ the density of $t_3(\cdot|m_t, \Sigma_t)$.
  for $(i = 1$ to $M)$ **do**
    Generate $\theta_{t,i}$ from $q_t(\cdot)$.
    Set $\omega_{t,i} = \pi(\theta_{t,i})L_{el}(\theta_{t,i}|y)/\Sigma_{s=1}^{t-1} q_s(\theta_{t,i})$.
  end for
  for $(r = 1$ to $t-1)$ **do**
    for $(i = 1$ to $M)$ **do**
      Update the weight of $\theta_{r,i}$ as $\omega_{r,i} = \pi(\theta_{r,i})L_{el}(\theta_{r,i}|y)/\Sigma_{s=1}^{t-1} q_s(\theta_{r,i})$.
    end for
  end for
end for

4.7 Extensions of the BC_{el} algorithm

Since its introduction, the BC_{el} approach has been applied to a range of problems. For example, Cheng et al. (2014) cite the approach as the foundation for their proposed method for estimating the parameters of the extreme value model of Heffernan and Tawn (2004). Through a large simulation study, the method was found to provide good coverage of credible intervals, although one of the parameters needed more informative priors under some more challenging setups.

In a second example, Grazian and Liseo (2017) discuss the use of BC_{el} for copula estimation, whereby the marginal likelihood of the quantity of interest is approximated by the EL.

Copula models are an important tool in multivariate analysis: while a huge literature exist about methods of estimating univariate marginal distributions, the problem of estimating the dependence structure of a multivariate distribution is more complex. Copula models allow for separately working with the univariate marginals and the joint distribution. They are widely used in many applications, including actuarial sciences (Frey and McNeil, 2001), epidemiology (Clayton, 1978), finance (Cherubini et al., 2004), hydrology (Salvadori and De Michele, 2007) among others.

A copula model is a way of representing the joint distribution of a random vector $X = (X_1, \ldots, X_d)$. Given a d-variate cumulative distribution function F which depends on some parameter ψ, it is possible to show (Sklar, 2010) that there always exists a d-variate function $C_\psi : [0,1]^d \to [0,1]$, such that

$$F(x_1, \ldots, x_d; \lambda_1, \ldots, \lambda_d, \psi) = C_\psi(F_1(x_1; \lambda_1), \ldots, F_d(x_d; \lambda_d)),$$

where F_j is the marginal distribution of X_j, indexed by a parameter λ_j, and ψ is a parameter characterizing the joint distribution.

In other terms, the copula C is a distribution function with uniform margins on $[0,1]$ which takes value from the univariate $F_1, F_2, \ldots, F_d$ (which may be of the same form or may differ in terms of the parameters or of the forms) in order to produce the d-variate distribution F. The resulting model is very flexible, because it may utilise different types of marginal distributions and dependence structures.

Many different types of copula functions have been proposed in the literature; see Joe (2015) for a review. An example is the Clayton copula, defined in the general d-dimension case as

$$C(\mathbf{u}) = (u_1^{-\psi} + u_2^{-\psi} + \cdots + u_d^{-\psi} - d + 1)^{-\frac{1}{\psi}}$$

where $\psi \in [-1, \infty) \setminus \{0\}$ is a one-dimensional parameter. The Clayton copula is characterized by lower-tail dependence (that approaches 1 as $\psi \to \infty$) and no upper-tail dependence. A representation of the Clayton copula (obtained through simulation) is available in Figure 11.

The frequentist standard method of estimating copula models is the "inference from the margins" (IFM) approach (Joe, 2015), i.e. a two-step procedure, where first the marginal distribution functions are separately estimated, either in a parametric or in a nonparametric way (depending on the information available on the marginals) and then the copula function

is estimated. Bayesian alternatives have been explored, nevertheless they are still limited. The reader may refer to Smith (2011) for a review.

In some cases the interest of the analysis is in a function of interest θ of the copula and not in the complete dependence structure; this may be due to a weak information about the type of structure or to the need of a low-dimensional quantification of the dependence. Some typical quantities of interest are, for example, tail dependence indices, Spearman's ρ or Kendall's τ. While tail dependence indices represent, in the bivariate case, the probability that a random variable exceeds a certain threshold given that another random variable has already exceeded that threshold (Großmaß, 2007), Spearman's ρ and Kendall's τ are measures of rank correlation, which are both expressible in terms of the copula C. For example, the Spearman's ρ in the bivariate case is defined as

$$\rho = 12 \int_{[0,1]^2} C(u,v)\, du\, dv - 3 = 12 \int_{[0,1]^2} uv\, dC(u,v) - 3. \tag{3}$$

In this case, Grazian and Liseo (2017), in the same spirit of the IFM method, propose to first estimate the marginal distributions and then study the interest measure of multivariate dependence with an approximate Bayesian approach based on an estimation of the likelihood of θ via EL (the authors use its Bayesian modification described in Schennach (2005) and in Section 4.2). In this way, it is possible to avoid the complete definition of the dependence structure (usually difficult to be determined) and elicite the prior distribution only for the quantity of interest, in order to reduce the bias derived from wrong distributional assumptions. Moreover their Bayesian approach avoids the loss of information of the IFM method and may be proved to be consistent.

The $BCOP$ ("Bayesian computation for copulas") algorithm follows and its final output will then be a posterior sample drawn from an approximation of the posterior distribution of the quantity of interest θ (see Algorithm 5).

This approach presents several advantages with respect to classical approaches to copula estimation. First, it may be applied to a generic dimension d, while in the literature there is a huge difference in terms of consistency results on the proposed estimators between the bivariate and the multivariate case. The authors have applied the $BCOP$ algorithm to a maximum dimension equal to 50 with no loss of precision and with a reasonable computational expense (it has to be noted that the algorithm may be easily parallelised in the first step of estimation of the marginals). Second, the method provides a quantification of the error of estimation, not easily available in the classical approach (see Schmid and Schmidt (2007) for the Spearman's ρ and Schmidt and Stadtmüller (2006) for the tail dependence indices). Third, it avoids the specification of the particular copula function which describes the dependence structure; this is particularly important in absence of information on it, since methods to select the copula function are not yet fully developed.

Since the interest is in small dimension parameter (often only one measure of dependence), the choice of the constraints should be easy; unfortunately, in practical applications these conditions might hold only asymptotically. This is the case, for example, of the Spearman's ρ: its sample counterpart ρ_n is only an asymptotically unbiased estimator of ρ so the moment condition is strictly valid only for large samples.

Algorithm 5 *BCOP* algorithm, Grazian and Liseo (2017)

Given a $n \times d$ data set $x = \{x_1, \ldots, x_n\}'$ and marginal posterior samples $\{\lambda_1^{(s)}, \ldots, \lambda_d^{(s)}\}$ for $s = 1, \cdots, S$

for $s = 1, \ldots, S$ **do**

Use the s-th row of the posterior simulation $\{\lambda_1^{(s)}, \lambda_2^{(s)}, \ldots, \lambda_d^{(s)}\}$ to create a matrix of uniformly distributed data $u_{ij}^{(s)} = F_j(x_{ij}; \lambda_j^{(s)})$

$$u^{(s)} = \begin{pmatrix} u_{11}^{(s)} & u_{12}^{(s)} & \ldots & u_{1d}^{(s)} \\ u_{21}^{(s)} & u_{22}^{(s)} & \ldots & u_{2d}^{(s)} \\ \ldots & \ldots & u_{ij}^{(s)} & \ldots \\ u_{n1}^{(s)} & u_{n2}^{(s)} & \ldots & u_{nd}^{(s)} \end{pmatrix}.$$

end for

Given a prior distribution $\pi(\theta)$ for the quantity of interest ϕ,

for $b = 1, \ldots, B$ **do**

Draw $\theta^{(b)} \sim \pi(\theta)$

for $s = 1, \ldots, S$ **do**

Compute $L_{BEL}(\theta^{(b)}; u^{(s)}) = \omega_{bs}$

Take the average weight $\omega_b = S^{-1} \sum_{s=1}^{S} \omega_{bs}$

end for

end for

Output $(\theta^{(b)}, \omega_b), b = 1, \ldots, B$

Grazian and Liseo (2017) also apply the method to a real data-set based on the study of the dependence among five Italian financial institutes, where the returns are supposed to marginally follow a $GARCH(1,1)$ model with Student's t innovations. They show how it is possible to obtain an approximated posterior distribution of the Spearman's ρ of the financial asset returns of these institutes with Algorithm 5.

As an application, consider the setting of Section 4.6, where five sets of observations are simulated from g-and-k identical but not independent quantile distributions with $a = 0$, $b = 1$, $g = 0.5$ and $k = 0$. The dependence structure is described by a multivariate Clayton copula (McNeil and Nešlehová, 2009) with true unknown multivariate ρ equal to 0.5. There are many ways to extend the bivariate Spearman's ρ defined in (3) to the multivariate case and they are not in general equivalent; nevertheless it is often of interest in many fields of application to describe the dependence with a low-dimensional quantity, for example in the multivariate analysis of financial asset returns where there is the need to express the amount of dependence in a portfolio by a single number. Here, the following is considered

$$\rho = \frac{\int_{[0,1]^d} (C(u) - \Pi(u))\, du}{\int_{[0,1]^d} (M(u) - \Pi(u))\, du} = h(d) \left\{ 2^d \int_{[0,1]^d} C(u) du - 1 \right\}, \tag{4}$$

where $M(u) = \min(u_1, u_2, \ldots, u_d)$ is the upper Fréchet- Hoeffding bound, and $h(d) = (d+1)/\{2^d - (d+1)\}$. For a review of the definitions of the Spearman's ρ in the literature one

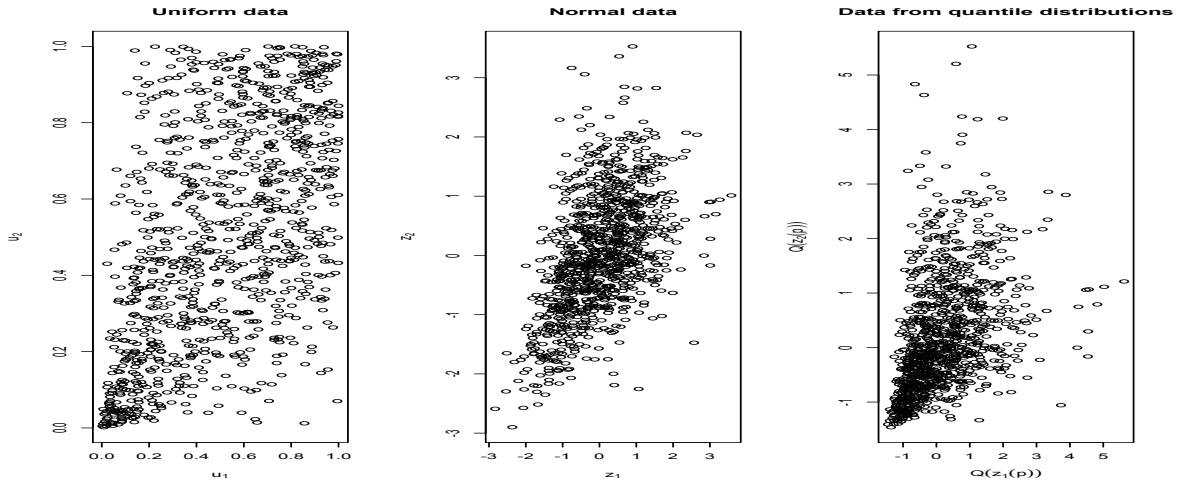


Figure 11: Scatterplots of the first two variables in the generation procedure: data from a Clayton copula with $\rho = 0.5$ (left), transformed to normal data (centre) and, then, to data from a g-and-k distribution with $a = 0$, $b = 1$, $g = 0.5$ and $k = 0$ (right).

may refer to Schmid and Schmidt (2007).

Uniform data have been generated from a multivariate Clayton copula with $\rho = 0.5$ and then inverted in order to obtain data from the corresponding quantile distributions. Figure 11 shows the correlation between the first two sets of observations generated with this procedure.

Figure 12 described the approximation to the posterior distribution of ρ, as defined in (4), obtained via Algorithm 5: it is possible to see that the posterior distribution is concentrated around the true value from which the data have been generated.

The `R` code used is available in the Appendix ("Copula code").

As noted above, one of the key considerations in developing and implementing BC_{el} is the choice of constraints for the EL. This consideration is not particular to BC_{el}, but applies to all EL methods. However, the difference here is that the selected constraints must be also applicable to the ABC context. With this goal in mind, Ruli et al. (2016) advocate the use of scaled composite likelihood score functions as summary statistics in ABC. The scaling takes into account a measure of the relative amount of information provided by the different parameters. They argue that the corresponding ABC procedure is therefore invariant to reparametrisation and accommodates automatically the curvature of the posterior distribution. This approach is argued to be an improvement over that proposed by Pauli et al. (2011) and more fully ABC than the BC_{el} approach.

Acknowledgements

CCD was supported by an Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE160100741). KM was supported by the Australian Research Council. CCD is an Associate Investigator and KM is a Chief Investigator of the Australian Centre

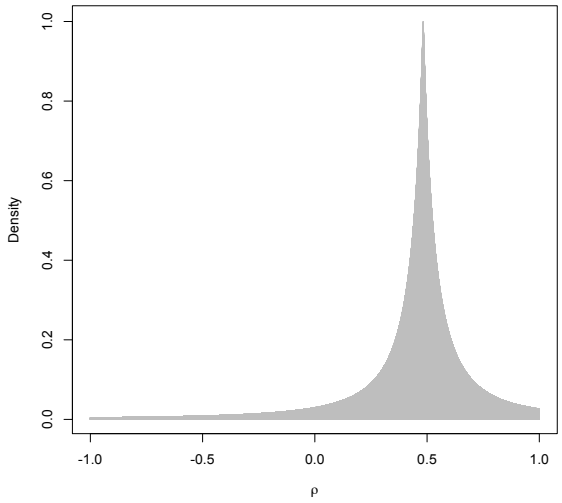


Figure 12: Approximation of the posterior distribution of the Spearman's ρ as defined in (3) for the data described in Figure 11.

of Excellence for Mathematical and Statistical Frontiers (ACEMS).

References

Allingham, D., King, R. A. R., and Mengersen, K. L. (2009). Bayesian estimation of quantile distributionss. *Statistics and Computing*, 19:189–201.

An, Z., South, L. F., Nott, D. J., and Drovandi, C. C. (2016). Accelerating Bayesian synthetic likelihood with the graphical lasso. *https://eprints.qut.edu.au/102263/*.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Blum, M. G. B. (2010). Approximate Bayesian computation: a non-parametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

Brown, V. L., Drake, J. M., Barton, H. D., Stallknecht, D. E., Brown, J. D., and Rohani, P. (2014). Neutrality, cross-immunity and subtype dominance in avian influenza viruses. *PLoS ONE*, 9(2):1–10.

Cai, A. Q., Landman, K. A., and Hughes, B. D. (2007). Multi-scale modeling of a wound-healing cell migration assay. *Journal of Theoretical Biology*, 245(3):576–594.

Chaudhuri, S. and Ghosh, M. (2011). Empirical likelihood for small area estimation. *Biometrika*, 98:473–480.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9:385–406.

Chen, S. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Annals of the Institute for Statistical Mathematics*, 45:621–637.

Chen, S. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis*, 49:24–40.

Chen, S. and Cui, H. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, 93:215–220.

Chen, S. and Gao, J. (2007). An adaptive empirical likelihood test for parametric time series regression models. *Journal of Econometrics*, 141:950–972.

Chen, S. and Quin, J. (2003). Empirical likelihood-based confidence intervals for data with possible zero observations. *Statistics & Probability Letters*, 65:29–37.

Chen, S. and Van Keilegom, I. (2009a). A goodness-of-fit test for parametric and semiparametric models in multiresponse regression. *Bernoulli*, 15:955–976.

Chen, S. and Van Keilegom, I. (2009b). A review on empirical likelihood methods for regression. *Test*, 18:415–447.

Cheng, L., Gilleland, E., Heaton, M., and AghaKouchak, A. (2014). Empirical Bayes estimation for the conditional extreme value model. *Stat*, 3:391–406.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Cornuet, J., MARIN, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997). Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In *Case studies in Bayesian statistics*, pages 37–93. Springer.

Decaestecker, C., Debeir, O., Van Ham, P., and Kiss, R. (2007). Can anti-migratory drugs be screened in vitro? a review of 2D and 3D assays for the quantitative analysis of cell migration. *Medicinal Research Reviews*, 27(2):149–176.

DiCiccio, T., Hall, P., and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Annals of Statistics*, 19:1053–1061.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233.

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, 30(1):72–95.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9:139–172.

Einmahl, J. and McKeague, I. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2):267–290.

El Barmi, H. and McKeague, I. (2013). Empirical likelihood-based tests for stochastic ordering. *Bernoulli*, 19(1):295–307.

El Barmi, H. and McKeague, I. (2015). Testing for uniform stochastic ordering via empirical likelihood. *Annals of the Institute of Statistical Mathematics*, pages 1–22.

Everitt, R. G., Johansen, A. M., E., R., and Evdemon-Hogan, M. (2017). Bayesian model comparison with un-normalised likelihoods. *Statistics and Computing*, 27(2):403422.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

Fasiolo, M., Pya, N., and Wood, S. (2016). A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. *Statistical Science*, 31(1):96–118.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Frey, R. and McNeil, A. (2001). Correlation and dependence properties in risk management: Properties and pitfalls. *Department of Mathematics, ETH, Zurich RiskLab*.

Ghurye, S. G. and Olkin, I. (1969). Unbiased estimation of some multivariate probability densities and related functions. *The Annals of Mathematical Statistics*, 40(4):1261–1271.

Glenn, N. and Zhao, Y. (2007). Weighted empirical likelihood estimates and their robustness properties. *Computational Statistics and Data Analysis*, 51:5130–5141.

Grazian, C. and Liseo, B. (2017). Approximate bayesian inference in semiparametric copula models. *Bayesian Analysis*, 12(4):991–1016.

Grendar, M. and Judge, G. (2007). A Bayesian large deviations probabilistic interpretation and justification of empirical likelihood. Technical report, Department of Agricultural and Resource Economics, University of California Berkeley.

Grendar, M. and Judge, G. (2009). Asymptotic equivalence of empirical likelihood and Bayesian MAP. *The Annals of Statistics*, 37:2445–2457.

Großmaß, T. (2007). Copulae and tail dependence.

Haardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, United Kingdom.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical. Review*, 58:109–127.

Hansen, L., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics*, 14:262–280.

Hartig, F., Dislich, C., Wiegand, T., and Huth, A. (2014). Technical note: approximate Bayesian parametrization of a process-based tropical forest model. *Biogeosciences*, 11:1261–1272.

Hartley, H. and Rao, J. (1968). A new estimation theory for sample surveys. *Biometrika*, 55:547–557.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Joe, H. (2015). *Dependence modeling with copulas*, volume 134. CRC Press, Boca Raton, FL.

Johnston, S. T., Simpson, M. J., McElwain, D. L. S., Binder, B. J., and Ross, J. V. (2014). Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open Biology*, 4(9):140097.

Kish, L. (1965). *Survey Sampling*. Wiley, New York.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice. Technical report, Cowles Foundation for Research in Economics, Yale University.

Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika*, 90:319–326.

Lian, H. (2012). Empirical likelihood confidence intervals for nonparametric functional data analysis. *Journal of Statistical Planning and Inference*, 142:1669–1677.

McKeague, I. and Zhao, Y. (2002). Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Statistics and Probability Letters*, 60:405–415.

McNeil, A. J. and Nešlehová, J. (2009). Multivariate archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *The Annals of Statistics*, 37:3059–3097.

McVinish, R. (2012). Improving ABC for quantile distributions. *Statistics and Computing*, 22(6):1199–1207.

Meeds, E. and Welling, M. (2014). GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In Zhang, N. L. and Tian, J., editors, *Uncertainty in Artificial Intelligence Proceedings of the Thirtieth Conference*, pages 593–602.

Mengersen, K., Pudlo, P., and Robert, C. (2013). Approximate Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Science*, 110:1321–1326.

Monahan, J. and Boos, D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278.

Moores, M. T., Drovandi, C. C., Mengersen, K. L., and Robert, C. P. (2015). Pre-processing for approximate Bayesian computation in image analysis. *Statistics and Computing*, 25(1):23–33.

Murphy, S. and van der Vaart, W. (1997). Semiparametric likelihood ratio inference. *Annals of Statistics*, 25(4):1471–1509.

Ong, V. M. H., J., N. D., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2017). Likelihood-free inference in high dimensions with synthetic likelihood. *https://eprints.qut.edu.au/112213/*.

Ong, V. M.-H., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018). Variational Bayes with synthetic likelihood. *Statistics and Computing*, 28(4):971988.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.

Owen, A. (1991). Empirical likelihood for linear models. *Annals of Statistics*, 19:1725–1747.

Owen, A. (2001). *Empirical Likelihood*. Chapman-Hall/CRC, New York.

Pauli, F. and Adimara, G. (2010). Bayesian inference with a pairwise likelihood: an approach based on empirical likelihood. *Proceedings of the 45th Scientific Meeting of the Italian Statistical Society, 53*.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, 21:149–164.

Peters, G. W. and Sisson, S. A. (2006). Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, 1(3):27–50.

Pham, K. C., Nott, D. J., and Chaudhuri, S. (2014). A note on approximating ABC-MCMC using flexible classifiers. *Stat*, 3(1):218–227.

Porter, A., Holan, S., and Wikle, C. (2015a). Bayesian semiparametric hierarchical empirical likelihood spatial models. *Journal of Statistical Planning and Inference*, 165:78–90.

Porter, A., Holan, S., and Wikle, C. (2015b). Multivariate spatial hierarchical bayesian empirical likelihood methods for small area estimation. *STAT*, 4(1):108–116.

Prangle, D. (2011). Summary statistics and sequential methods for approximate Bayesian computation. Technical report, Lancaster University, U.K.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *To appear in Journal of Computational and Graphical Statistics*.

Qin, J. and Jing, B. (1994). Empirical likelihood for cox regression model under random censorship. *Annals of Statistics*, 22:300–325.

Qin, J. and Lawless, J. (2001). Empirical likelihood and general estimating equations. *Communications in Statistics - Simulation and Computation*, 30:79–90.

Rao, J. and Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72:533–544.

Rasuga, G. (2006). Bayesian likelihoods for moment condition models. Technical report, University of California, Irvine.

Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Riscado, S. (2012). *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments (Advances in Econometrics, Volume 28)*, chapter On the Estimation of Dynamic Stochastic General Equilibrium Models: An Empirical Likelihood Approach. Emerald Group Publishing Limited.

Rochet, P. (2012). Bayesian interpretation of generalized empirical likelihood by maximum entropy. Technical report.

Rubin, D. (1981). Bayesian bootstrap. *Annals of Statistics*, 9:130–134.

Ruli, E., Sartori, N., and Ventura, L. (2016). Approximate Bayesian computation using composite score functions. *Statistics and Computing*, 26(3):679692.

Salvadori, G. and De Michele, C. (2007). On the use of copulas in hydrology: theory and practice. *Journal of Hydrologic Engineering*, 12(4):369–380.

Schennach, S. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92:31–46.

Schmid, F. and Schmidt, R. (2007). Multivariate extensions of spearman's rho and related statistics. *Statistics and Probability Letters*, 77(4):407–416.

Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33(2):307–335.

Sklar, M. (2010). Fonctions de répartition a n dimensions et leurs marges. With an introduction by Denis Bosq.

Smith, Jr., A. A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(S1):S63–S84.

Smith, M. S. (2011). Bayesian approaches to copula modelling. *Damien, PP Dellaportas, NG Polson, DA Stephens,(2013), Bayesian Theory and Applications, OUP*, pages 336–358.

Treloar, K. K., Simpson, M. J., Haridas, P., Manton, K. J., Leavesley, D. I., McElwain, D. L. S., and Baker, R. E. (2013). Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies. *BMC Systems Biology*, 7(1):137.

Vo, B. N., Drovandi, C. C., Pettitt, A. N., and Pettet, G. J. (2015a). Melanoma cell colony expansion parameters revealed by approximate Bayesian computation. *PLOS Computational Biology*, 11(12):e1004635.

Vo, B. N., Drovandi, C. C., Pettitt, A. N., and Simpson, M. J. (2015b). Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences*, 263:133–142.

Wang, Q. and Jing, B. (2001). Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*, 53:517–527.

Wilkinson, R. (2014). Accelerating ABC methods using Gaussian processes. *Journal of Machine Learning Research*, 33:1015–1023.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1107.

Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34:359–375.

Zhou, M. (2015). *Empirical likelihood method in survival analysis*. Chapman and Hall/CRC, London.

Zhou, M. and Yang, Y. (2014). *emplik: Empirical likelihood ratio for censored/truncated data*. R package version 0.9-9-6.

Appendix

Myeloma code

```
data(myeloma)
survtimes <- myeloma[,1]    # survival times
censtatus <- myeloma[,2]    # vital status (0=alive, 1=dead)
myfun1 <- function(t){ as.numeric(t <= 10) }
el.cen.EM(fun=myfun1, x=survtimes, d=censtatus, mu=0.2)
```

Bayes factor code

```
# test Model 1 (A,B,g,k)=(0,1,1,0) [skew]
# versus Model 2 (0,1,0.5,0) [less skew] and
# Model 3 (0,1,0,0) [standard normal]
# Compare B12=el_1/el_2 and B13=el1/el3
# Two sample sizes: n=100, 1000
library(emplik)
# set qc; traditionally set at 0.8
qc=0.8
# specify the models of interest; qp1 is the 'true' model
qp1=c(0,1,1,0) ;  qp2=c(0,1,0.5,0) ; qp3=c(0,1,0,0)
# specify the quantiles for each model
refp=c(0.1,0.25,0.5,0.75,0.9)
simq1=qp1[1]+qp1[2]*(1+qc*((1-exp(-qp1[3]*refp))/
(1+exp(qp1[3]*refp))))*((1+refp^2)^qp1[4])*refp
simq2=qp2[1]+qp2[2]*(1+qc*((1-exp(-qp2[3]*refp))/
(1+exp(qp2[3]*refp))))*((1+refp^2)^qp2[4])*refp
simq3=qp3[1]+qp3[2]*(1+qc*((1-exp(-qp3[3]*refp))/
(1+exp(qp3[3]*refp))))*((1+refp^2)^qp3[4])*refp
# set sample size
nob=c(100, 500)  # no. observations
lennob=length(nob)
nrep=100        # replicates of BF12
# set up matrices and vectors
BF12=logBF12=BF13=logBF13=matrix(0,nrep,lennob)
th1=th2=th3=rep(0,nrep)
# compute BF using el.test based on true parameters for M1 vs M2, M3
for (nk in 1:lennob){
dth1=dth2=dth3=matrix(0,nrow=nob[nk],ncol=length(refp))
for (repk in 1:nrep){
# generate reference data
zp=qnorm(runif(nob[nk]))
dob=qp1[1]+qp1[2]*(1+qc*((1-exp(-qp1[3]*zp))/
(1+exp(-qp1[3]*zp))))*((1+zp^2)^qp1[4])*zp
```

```
for (k in 1:nob[nk]){
for (j in 1:length(refp)){
dth1[k,j] = (dob[k]<simq1[j])*1
dth2[k,j] = (dob[k]<simq2[j])*1
dth3[k,j] = (dob[k]<simq3[j])*1
}}
th1=el.test(dth1,mu=refp)
th2=el.test(dth2,mu=refp)
th3=el.test(dth3,mu=refp)
thll1=th1$'-2LLR' ;  thll2=th2$'-2LLR' ;  thll3=th3$'-2LLR'
logBF12[repk,nk] = -0.5*(thll1 - thll2)
logBF13[repk,nk] = -0.5*(thll1 - thll3)
BF12[repk,nk] = exp(logBF12[repk,nk])
BF13[repk,nk] = exp(logBF13[repk,nk])
}}  #end of repk, nk
par(mfrow=c(2,2))
logBF123=cbind(logBF12,logBF13)
boxplot(logBF123[,1:4],ylim=c(-10,50),
xlab="1=M1 v M2,n=100; 2=M1 v M2,n=500; 3=M1 v M3, n=100; 4=M1 v M3, n=500",
ylab="log BF")
```

Copula code

```
### Function to generate from a quantile function

quantile.fun=function(z,A,B,g,k,c=0.8)
{
val = A + B * ( 1 + c * (1-exp(-g*z)) / (1+exp(-g*z)) ) *
( 1+z^2 )^k * z
return(val)
}

### Simulations from the copula with a fixed Spearman's rho

# Generation from the copula
library(copula)
cc=claytonCopula(d=5,param=1.076)
uu=rCopula(1000,cc)

# Generation from the normal
z=matrix(NA,nrow=1000,ncol=5)
for(i in 1:5)
{
z[,i]=qnorm(uu[,i])
}
```

```
# Generation from the quantile distribution
quant.sim=matrix(NA,nrow=1000,ncol=5)
for(i in 1:5)
{
quant.sim[,i]=quantile.fun(z[,i],A=0,B=1,g=0.5,k=0)
}

#### (Nonparametric) estimation of the marginals

n=1000
F.hat=matrix(NA,nrow=1000,ncol=5)
for(i in 1:5)
{
for(j in 1:1000)
{
F.hat[j,i]=sum(quant.sim[,i]<quant.sim[j,i])/n
}
}

#### BCOP for the Spearman's rho

n=dim(F.hat)[1]
d=dim(F.hat)[2]
S=10^5

# Ranks
U.hat=matrix(NA,ncol=d,nrow=n)
for(i in 1:d)
{
U.hat[,i]=rank(F.hat[,i])/n
}
VV1=apply(1-U.hat,1,prod)
VV2=apply(U.hat,1,prod)

# Frequentist estimate
const=(d+1)/(2^d-(d+1))
estim1=const*(2^d/n*sum(VV1)-1)

# BCOP

rho=runif(S, -1,1)
omega=rep(0,S)

for (s in 1:S)
{
```

```
est=estim1 - rho[s]
omega1[s]<-exp(-EL(est)$elr)
}

rho.sim=cbind(rho, omega)

plot(rho.sim[,1],rho.sim[,2],type="h",
xlab=expression(rho),ylab="Density",main="",col="grey")

par(mfrow=c(1,3))
plot(uu[,1],uu[,2],xlab=expression(u[1]),ylab=expression(u[2]),
main="Uniform data")
plot(z[,1],z[,2],xlab=expression(z[1]),ylab=expression(z[2]),
main="Normal data")
plot(quant.sim[,1],quant.sim[,2],xlab=expression(Q(z[1](p))),
ylab=expression(Q(z[2](p))),main="Data from quantile distributions")
```