Expectation Propagation for Likelihood-Free Inference

Simon Barthelme, Nicolas Chopin

Expectation-Propagation for Likelihood-Free Inference

Simon Barthelmé[*]
Nicolas Chopin[†]

Abstract

Many models of interest in the natural and social sciences have no closed-form likelihood function, which means that they cannot be treated using the usual techniques of statistical inference. In the case where such models can be efficiently simulated, Bayesian inference is still possible thanks to the Approximate Bayesian Computation (ABC) algorithm. Although many refinements have been suggested, ABC inference is still far from routine. ABC is often excruciatingly slow due to very low acceptance rates. In addition, ABC requires introducing a vector of "summary statistics" $\boldsymbol{s}(\boldsymbol{y})$, the choice of which is relatively arbitrary, and often require some trial and error, making the whole process quite laborious for the user.

We introduce in this work the EP-ABC algorithm, which is an adaptation to the likelihood-free context of the variational approximation algorithm known as Expectation Propagation (Minka, 2001a). The main advantage of EP-ABC is that it is faster by a few orders of magnitude than standard algorithms, while producing an overall approximation error which is typically negligible. A second advantage of EP-ABC is that it replaces the usual global ABC constraint $\|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\boldsymbol{y}^\star)\| \leq \varepsilon$, where $\boldsymbol{s}(\boldsymbol{y}^\star)$ is the vector of summary statistics computed on the whole dataset, by n local constraints of the form $\|s_i(y_i) - s_i(y_i^\star)\| \leq \varepsilon$ that apply separately to each data-point. In particular, it is often possible to take $s_i(y_i) = y_i$, making it possible to do away with summary statistics entirely. In that case, EP-ABC makes it possible to approximate directly the evidence (marginal likelihood) of the model.

Comparisons are performed in three real-world applications which are typical of likelihood-free inference, including one application in neuroscience which is novel, and possibly too challenging for standard ABC techniques.

Key-words: Approximate Bayesian Computation; Composite Likelihood; Expectation Propagation; Likelihood-Free Inference; Quasi-Monte Carlo.

1 Introduction

In natural and social sciences, one finds many examples of probabilistic models whose likelihood function is intractable. This includes most models of noisy biological neural networks (Gerstner and Kistler, 2002), some time series and choice models in Economics (Train, 2003), phylogenetic models in evolutionary Biology (Beaumont, 2010), spatial extremes in Environmental Statistics (Davison et al., 2012), among others. That the likelihood is intractable is unfortunate, because one would still like to perform the usual statistical tasks of parameter inference and model comparison, and the traditional statistical tool-kit assumes that the likelihood function is either directly available or can be made so by introducing latent variables. This explains that researchers have often had to content themselves with semi-quantitative analyses showing that a model could reproduce some aspect of an empirical phenomenon for some values of the parameters; for two representative examples from vision science, see Nuthmann et al. (2010), Brascamp et al. (2006).

A breakthrough was provided by the work of Tavaré et al. (1997), Pritchard et al. (1999) and Beaumont et al. (2002), in the form of the Approximate Bayesian Computation (ABC) algorithm, which enables Bayesian inference in the likelihood-free context. (See also Diggle and Gratton (1984) and Rubin (1984) for early versions of ABC.) Assuming some model $p(\boldsymbol{y}^*|\boldsymbol{\theta})$ for the data $\boldsymbol{y}^\star$, and a prior $p(\boldsymbol{\theta})$ over the parameter $\boldsymbol{\theta} \in \Theta$, the ABC algorithm iterates the following steps:

1. Draw $\boldsymbol{\theta}$ from the prior, $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$.
2. Draw a dataset $\boldsymbol{y}$ from the model conditional on $\boldsymbol{\theta}$, $\boldsymbol{y}|\boldsymbol{\theta} \sim p(\boldsymbol{y}|\boldsymbol{\theta})$.
3. If $d(\boldsymbol{y}, \boldsymbol{y}^\star) \leq \epsilon$, then keep $\boldsymbol{\theta}$, otherwise reject.

and therefore produces samples from the so-called ABC posterior:

[*]Université de Genève. Faculté de Psychologie et de Sciences de l'Education, Unimail CH-1211 Genève 4, Switzerland. simon.barthelme@unige.ch
[†]CREST/ENSAE. 3, Avenue Pierre Larousse 92245 Malakoff CEDEX France. nicolas.chopin@ensae.fr

$$p_\epsilon\left(\boldsymbol{\theta}|\boldsymbol{y}^*\right) \propto p(\boldsymbol{\theta}) \int p(\boldsymbol{y}|\boldsymbol{\theta}) \mathbb{1}_{\{d(\boldsymbol{y},\boldsymbol{y}^\star)\leq\boldsymbol{\epsilon}\}}\, d\boldsymbol{y}. \tag{1.1}$$

The pseudo-distance is usually taken to be $d(\boldsymbol{y},\boldsymbol{y}^\star) = ||\boldsymbol{s}\left(\boldsymbol{y}\right) - \boldsymbol{s}\left(\boldsymbol{y}^*\right)||$, for some norm $||\cdot||$, where $\boldsymbol{s}(\boldsymbol{y})$ is a vector of summary statistics, for example some empirical quantiles or moments of $\boldsymbol{y}$. Unless $\boldsymbol{s}$ is sufficient, the approximation error does not vanish as $\epsilon \to 0$, as $p(\boldsymbol{\theta}|\boldsymbol{s}(\boldsymbol{y}^\star)) \neq p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$. In that respect, the ABC posterior suffers from two levels of approximation: a nonparametric error, akin to the error of kernel density estimation, and where ϵ plays the role of a bandwidth (see e.g. Blum, 2010), and a bias introduced by the summary statistics $\boldsymbol{s}$. The more we include in $\boldsymbol{s}\left(\boldsymbol{y}\right)$, the smaller the bias induced by $\boldsymbol{s}$ should be. On the other hand, as the dimensionality of $\boldsymbol{s}(\boldsymbol{y})$ increases, the lower the acceptance rate will be. We would then have to increase ϵ, which leads to an approximation of lower quality.

Thus, ABC requires in practice some more or less arbitrary compromise between what summary statistics to include and how to set ϵ. To establish that the results of the inference are somewhat robust to these choices, many runs of the algorithm may be required. Although several variants of the original ABC algorithm that aim at increasing acceptance rates exist (e.g. Beaumont et al., 2002), the current state of the matter is that an ABC analysis is very far from routine use because it may take days to tune on real problems. The semi-automatic method for constructing summary statistics recently proposed by Fearnhead and Prangle (2012) seems to alleviate partly these problems, but it still requires at least one, and sometimes several pilot runs.

In this article we introduce EP-ABC, an adaptation of the Expectation Propagation (EP) algorithm (Minka, 2001a; Bishop, 2006, Chap. 10) to the likelihood-free setting. The main advantage of EP-ABC is that it is much faster than previous ABC algorithms: typically, it provides accurate results in a few minutes, whereas a standard ABC algorithm may need several hours, or even days. EP-ABC requires that the data $\boldsymbol{y}^\star$ may be decomposed into n "chunks", $y_1^\star, \ldots, y_n^\star$ (of possibly different dimensionality or support), in such a way that is possible to simulate sequentially the n chunks; i.e. one is able to simulate from $p(y_i^\star|y_{1:i-1}^\star, \boldsymbol{\theta})$ for each i. More precisely, EP-ABC builds an EP approximation of the following type of ABC posterior distributions:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|s_i(y_i) - s_i(y_i^\star)\| \leq \varepsilon\right\}}\, dy_i \right\} \tag{1.2}$$

where $s_i(y_i)$ is a summary statistic specific to chunk i, and with the convention that $p(y_1|y_{1:0}^\star, \boldsymbol{\theta}) = p(y_1|\boldsymbol{\theta})$. Given the way it operates, we shall see that EP-ABC essentially replaces the initial, possibly difficult ABC problem, by n ABC simpler (i.e. lower-dimensional) ABC problems. In particular, it may be easier to construct a set of "local" summary statistics $s_i(y_i)$ for each chunk y_i rather than a global set $\boldsymbol{s}(\boldsymbol{y})$ the whole dataset $\boldsymbol{y}$, in such a way that the probability that $\|s_i(y_i) - s_i(y_i^\star)\| \leq \varepsilon$ is not too small.

Of course, not all ABC models are amenable to the factorisation in (1.2), at least directly. We shall discuss this point more in detail in the paper. At the very least, factorisable models include situations with n repeated experiments, which are not easily tackled through standard ABC, because of the difficulty to define some vector $\boldsymbol{s}(\boldsymbol{y})$ which summarises these n experiments. Bazin et al. (2010) discusses this point, in the context of genetic data collected over a large number of loci, and recommend instead to construct loci-specific summary statistics, and combine results from loci-specific ABC exercises.

In certain problems such that y_i is of low dimension, it may be even possible to take $s_i(y_i) = y_i$. In that case, EP-ABC provides an EP approximation of a summary-less ABC posterior; that is, when $\epsilon \to 0$ (under appropriate measure-theoretic conditions), $p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \to p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$, the true posterior. In such situations, EP-ABC also provides an approximation of the evidence $p(\boldsymbol{y}^\star)$, which is a very useful quantity for model comparison. Dean et al. (2011) show that, in the specific context of hidden Markov models such that the observation density is intractable (as this work was motivated by the filtering ABC algorithm developed in Jasra et al. (2010)), the ABC error of a summary-less ABC algorithm should not be interpreted as a non-parametric error, but rather as a misspecification error, where the misspecified model is the true model, but with observations corrupted with noise. This misspecification implies a bias of order ε. Thus, in addition to being more convenient for the user, as it avoids specifying some summary statistics, summary-less ABC does not seem to suffer from the curse of dimensionality (in the dimension of $\boldsymbol{s}(\boldsymbol{y})$) of standard ABC. Allowing summary-less ABC inference in certain cases seems to be another advantage of EP-ABC.

We start with a generic description of the EP algorithm, in Section 2, and explain in Section 3 how it can be adapted to the likelihood-free setting. We explain in Section 4 how EP-ABC can be made particularly efficient when data-points are IID (independent and identically distributed). Section 5 contains three case studies drawn from finance, population ecology, and vision science. The two first examples are borrowed from already known applications of ABC, and illustrate to which extent EP-ABC may outperform standard ABC techniques in realistic scenarios. To the best of our knowledge, our third example from vision science is a novel application of likelihood-free inference, which, as which shall argue, seems too challenging for standard ABC techniques.

Section 6 discusses how EP-ABC may cope with difficult posteriors through two additional examples. Section 7 discusses possible extensions of EP-ABC; in particular, for non-factorisable likelihoods. In such cases, one may replace the likelihood by some factorisable approximation, such as a composite likelihood. Section 8 concludes.

We use the following notations throughout the paper: bold letters refer to vectors or matrices, e.g. $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and so on. We also use bold face to distinguish complete sets of observations, i.e. $\boldsymbol{y}$ or $\boldsymbol{y}^\star$, from their components, y_i, and $y_i^\star$, $i = 1, \dots n$, although we do not assume that these components are necessarily scalar. For sub-vectors of observations, we use the colon notation: $y_{1:i} = (y_1, \dots, y_i)$. The notation $\| \cdot \|$ refers to a generic norm, and $\| \cdot \|_2$ refers to the Euclidean norm. The Kullback-Leibler divergence between probability densities π and q is denoted as

$$KL(\pi \| q) = \int \pi(\boldsymbol{\theta}) \log \left(\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta}.$$

The letter p always refers to probability densities concerning the model; i.e. $p(\boldsymbol{\theta})$ is the prior, $p(y_1|\boldsymbol{\theta})$ is the likelihood of the first observation, and so on. Transpose of a matrix $\boldsymbol{A}$ is denoted $\boldsymbol{A}^t$, and the diagonal matrix with diagonal elements given by vector $\boldsymbol{V}$ is $\mathrm{diag}(\boldsymbol{V})$.

2 Expectation Propagation

Expectation Propagation (EP, Minka, 2001a) is an algorithm for variational inference, a class of techniques that aim at finding a tractable probability distribution $q(\boldsymbol{\theta})$ that best approximates an intractable target density $\pi(\boldsymbol{\theta})$. One way to formulate this goal is as finding the member of some parametric family $\mathcal{Q}$ that is in some sense closest to $\pi(\boldsymbol{\theta})$, where "closest" is defined by some divergence between probability distributions. A distinctive feature of EP is that it based on the divergence $KL(\pi||q)$, whereas many variational methods (e.g. Variational Bayes, see Chap. 10 of Bishop, 2006) try to minimize $KL(q||\pi)$. For a good discussion of the relative merits of each divergence, see Bishop (2006, p. 468-470) and Minka (2005). As a brief summary, minimizing $KL(q||\pi)$ tends to produce approximation that are too compact; see also Wang and Titterington (2005). The divergence $KL(\pi||q)$ does not suffer from this drawback, and is perhaps more appealing from a statistical point of view (if only because maximum likelihood estimation is based on the corresponding contrast), but on the other hand, minimizing $KL(\pi||q)$ when π is multimodal tends to produce an approximation that covers all the modes, and therefore has too large a support. We shall return to this point.

2.1 Assumptions of Expectation Propagation

EP assumes that the target density $\pi(\boldsymbol{\theta})$ decomposes into a product of simpler factors

$$\pi(\boldsymbol{\theta}) = \frac{1}{Z_\pi} \prod_{i=0}^{n} l_i(\boldsymbol{\theta}) \tag{2.1}$$

and exploits this factorisation in order to construct a sequence of simpler problems. For instance, $l_0(\boldsymbol{\theta})$ may be the prior $p(\boldsymbol{\theta})$, $l_i(\boldsymbol{\theta}) = p(y_i|y_{1:i-1}, \boldsymbol{\theta})$ if $\boldsymbol{y}$ is a set of n observations y_i, with the convention that $p(y_1|y_{1:0}, \boldsymbol{\theta}) = p(y_1|\boldsymbol{\theta})$, and $Z_\pi = p(\boldsymbol{y})$.

EP uses an approximating distribution with a similar structure:

$$q(\boldsymbol{\theta}) \propto \prod_{i=0}^{n} f_i(\boldsymbol{\theta}) \tag{2.2}$$

where the f_i's are known as the "sites". In a spirit close to a coordinate-descent optimization algorithm, each site is updated in turn, while the n other sites are kept fixed.

For the sake of conciseness, we focus in this paper on Gaussian sites, expressed under their natural parametrisation: $f_i(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q_i} \boldsymbol{\theta} + \boldsymbol{r}_i^t \boldsymbol{\theta}\right)$, where $\boldsymbol{Q}_i$ and $\boldsymbol{r}_i$ are called from now on the *site parameters*. This generates the following Gaussian approximation q:

$$q(\boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\theta}^t \left(\sum_{i=0}^{n} \boldsymbol{Q}_i \right) \boldsymbol{\theta} + \left(\sum_{i=0}^{n} \boldsymbol{r}_i \right)^t \boldsymbol{\theta} \right\}. \tag{2.3}$$

In addition, we assume that the true prior $p(\boldsymbol{\theta})$ is Gaussian, with natural parameters $\boldsymbol{Q}_0$ and $\boldsymbol{r}_0$. In that case, the site f_0 is kept equal to the prior, and only the sites f_1 to f_n are updated.

We note however that EP may easily accommodate a non-Gaussian prior (by simply treating the prior as additional factor to be approximated), or other types of parametric sites. In fact, the site update described in the following section may be easily adapted to any exponential family; see Seeger (2005).

2.2 Site update

Suppose that (2.2) is the current approximation, and one wishes to update site i. This is done by creating a "hybrid" distribution, obtained by substituting site i with the true likelihood contribution $l_i(\boldsymbol{\theta})$:

$$h\left(\boldsymbol{\theta}\right) \propto q_{-i}\left(\boldsymbol{\theta}\right) l_i(\boldsymbol{\theta}), \quad q_{-i}(\boldsymbol{\theta}) = \prod_{j \neq i} f_j(\boldsymbol{\theta}). \tag{2.4}$$

For Gaussian sites, this leads to:

$$h(\boldsymbol{\theta}) \propto l_i\left(\boldsymbol{\theta}\right) \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q}_{-i}\boldsymbol{\theta} + \boldsymbol{r}_{-i}^t\boldsymbol{\theta}\right), \quad \boldsymbol{Q}_{-i} = \sum_{j\neq i}\boldsymbol{Q}_j, \quad \boldsymbol{r}_{-i} = \sum_{j\neq i}\boldsymbol{r}_j.$$

The new value of site f_i is then obtained by minimising with respect to f_i the Kullback-Leibler pseudo-distance $KL(h_i||q)$ between the hybrid and the Gaussian approximation q (again keeping the f_j's fixed). When Gaussian sites are used, this minimisation is equivalent to taking q to be the Gaussian density with moment parameters that match those of the hybrid distribution

$$\begin{aligned} Z_h &= \int q_{-i}\left(\boldsymbol{\theta}\right) l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \\ \boldsymbol{\mu}_h &= \frac{1}{Z_h}\int \boldsymbol{\theta} q_{-i}\left(\boldsymbol{\theta}\right) l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \\ \boldsymbol{\Sigma}_h &= \frac{1}{Z_h}\int \boldsymbol{\theta}\boldsymbol{\theta}^t q_{-i}\left(\boldsymbol{\theta}\right) l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} - \boldsymbol{\mu}_h\boldsymbol{\mu}_h^t. \end{aligned} \tag{2.5}$$

A key observation at this stage is that the feasibility of EP is essentially dictated by how easily the moments above may be computed. These moments may be interpreted as the moments of a posterior distribution, based on a Gaussian prior, with natural parameters $\boldsymbol{Q}_{-i}$ and $\boldsymbol{r}_{-i}$, and a likelihood consisting of a single factor $l_i(\boldsymbol{\theta})$.

Finally, from these moment parameters, one may recover the natural parameters of q, and deduce the new site parameters for f_i as follows:

$$\boldsymbol{Q}_i \leftarrow \boldsymbol{\Sigma}_h^{-1} - \boldsymbol{Q}_{-i}, \quad \boldsymbol{r}_i \leftarrow \boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h - \boldsymbol{r}_{-i}.$$

EP proceeds by looping over sites, updating each one in turn until convergence is achieved. In well-behaved cases, one observes empirically that a small number of complete sweeps through the sites is sufficient to obtain convergence. However, there is currently no general theory on the convergence of EP.

Appendix A gives an algorithmic description of EP, in the more general case where an exponential family is used for the sites.

2.3 Approximation of the evidence

EP also provides an approximation of the normalising constant Z_π of (2.1), using the same ideas of updating site approximations through moment matching. To that effect, we rewrite the EP approximation with normalising constants for each site (assuming again the prior is Gaussian and does not need to be approximated):

$$q(\boldsymbol{\theta}) = p(\boldsymbol{\theta})\prod_{i=1}^{n}\frac{f_i(\boldsymbol{\theta})}{C_i}, \quad f_i(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{Q}_i\boldsymbol{\theta} + \boldsymbol{r}_i^t\boldsymbol{\theta}\right). \tag{2.6}$$

Then the update of site i proceeds as before, by adjusting C_i, $\boldsymbol{r}_i$ and $\boldsymbol{Q}_i$ through moment matching. Simple calculations (see e.g. Seeger, 2005) lead to the following expressions for the update of C_i:

$$\log\left(C_i\right) = \log\left(Z_h\right) - \Psi(\boldsymbol{r}, \boldsymbol{Q}) + \Psi(\boldsymbol{r}_{-i}, \boldsymbol{Q}_{-i}) \tag{2.7}$$

where Z_h is the normalising constant of the hybrid, as defined in (2.5), $\boldsymbol{r}$, $\boldsymbol{Q}$ (resp. $\boldsymbol{r}_{-i}$, $\boldsymbol{Q}_{-i}$) are the natural parameters of the current Gaussian approximation q (resp. of $q/f_i \propto \prod_{j\neq i} f_j$) and $\Psi(\boldsymbol{r}, \boldsymbol{Q})$ is the log-normalising constant of an unnormalised Gaussian density:

$$\Psi(\boldsymbol{r}, \boldsymbol{Q}) = \log\left\{\int \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\boldsymbol{t}}\boldsymbol{Q}\boldsymbol{\theta} + \boldsymbol{r}^{\boldsymbol{t}}\boldsymbol{\theta}\right)\, d\boldsymbol{\theta}\right\} = -\frac{1}{2}\log|\boldsymbol{Q}/2\pi| + \frac{1}{2}\boldsymbol{r}^t\boldsymbol{Q}\boldsymbol{r}.$$

For each site update, one calculates C_i as defined in (2.7). Then, at the end of the algorithm, one may return the following quantity

$$\sum_{i=1}^{n} \log\left(C_i\right) + \Psi(\boldsymbol{r}, \boldsymbol{Q}) - \Psi(\boldsymbol{r}_0, \boldsymbol{Q}_0)$$

as an approximation to the logarithm of the evidence.

3 EP-ABC: Adapting EP to likelihood-free settings

3.1 Basic principle

As explained in the introduction, our objective is to approximate the following ABC posterior

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|s_i(y_i) - s_i(y_i^\star)\| \leq \varepsilon\right\}} \, dy_i \right\} \tag{3.1}$$

which corresponds to a particular factorisation of the likelihood,

$$p(\boldsymbol{y}^\star|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i^\star|y_{1:i-1}^\star, \boldsymbol{\theta}). \tag{3.2}$$

Note that, in full generality, the $y_i^\star$ may be any type of "chunk" of the observation vector $\boldsymbol{y}^\star$, i.e. the random variables $y_i^\star$ may have a different dimension, or more generally different supports. For simplicity, we assume that the prior $p(\boldsymbol{\theta})$ is Gaussian, with natural parameters $\boldsymbol{Q}_0$ and $\boldsymbol{r}_0$.

One may interpret (3.1) as an artificial posterior distribution, which decomposes into a prior times n likelihood contributions l_i, as in (2.1), with

$$l_i(\boldsymbol{\theta}) = \left\{ \int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|s_i(y_i) - s_i(y_i^\star)\| \leq \varepsilon\right\}} \, dy_i \right\}.$$

We have seen that the feasibility of the EP algorithm is determined by the tractability of the following operation: to compute the two first moments of a pseudo-posterior, made of a Gaussian prior $N_d(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, times a single likelihood contribution $l_i(\boldsymbol{\theta})$. This immediately suggests the following EP-ABC algorithm. We use the EP algorithm, as described in Algorithm 3, and where the moments of such a pseudo-posterior are computed using as described in Algorithm 1, that is, as Monte Carlo estimates, based on simulated pairs $(\boldsymbol{\theta}^{[m]}, y_i^{[m]})$, where $\boldsymbol{\theta}^{[m]} \sim N_d(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, and $y_i^{[m]}|\boldsymbol{\theta}^{[m]} \sim p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}^{[m]})$.

Algorithm 1 Computing the moments of the hybrid distribution in the likelihood-free setting, basic algorithm.

Inputs: ϵ, $\boldsymbol{y}^\star$, i, and the moment parameters $\boldsymbol{\mu}_{-i}$, $\boldsymbol{\Sigma}_{-i}$ of the Gaussian pseudo-prior q_{-i}.

1. Draw M variates $\boldsymbol{\theta}^{[m]}$ from a $N(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$ distribution.
2. For each $\boldsymbol{\theta}^{[m]}$, draw $y_i^{[m]} \sim p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}^{[m]})$.
3. Compute the empirical moments

$$M_{acc} = \sum_{m=1}^{M} \mathbb{1}_{\left\{\|s_i(y_i^{[m]}) - s_i(y_i^\star)\| \leq \varepsilon\right\}}, \quad \widehat{\boldsymbol{\mu}}_h = \frac{\sum_{m=1}^{M} \boldsymbol{\theta}^{[m]} \mathbb{1}_{\left\{\|s_i(y_i^{[m]}) - s_i(y_i^\star)\| \leq \varepsilon\right\}}}{M_{acc}} \tag{3.3}$$

$$\widehat{\boldsymbol{\Sigma}}_h = \frac{\sum_{m=1}^{M} \boldsymbol{\theta}^{[m]} \left\{\boldsymbol{\theta}^{[m]}\right\}^t \mathbb{1}_{\left\{\|s_i(y_i^{[m]}) - s_i(y_i^\star)\| \leq \varepsilon\right\}}}{M_{acc}} - \widehat{\boldsymbol{\mu}}_h \widehat{\boldsymbol{\mu}}_h^t. \tag{3.4}$$

Return $\widehat{Z}_h = M_{acc}/M$, $\widehat{\boldsymbol{\mu}}_h$ and $\widehat{\boldsymbol{\Sigma}}_h$.

Since EP-ABC integrates one data-point at a time, it does not suffer from a curse of dimensionality with respect to n: the rejection rate of Algorithm 1 corresponds to a single constraint $\|s_i(y_i) - s_i(y_i^\star)\| \leq \epsilon$, not n of them, and is therefore

likely to be tolerably small even for small windows ϵ. (Otherwise, in very challenging situations, one has the liberty to replace Algorithm 1 by a more elaborate ABC algorithm.)

The only requirement of EP-ABC is that the factorisation of the likelihood, (3.2), is chosen in such a way that simulating from the model, i.e. $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{\theta})$ can be decomposed into a sequence of steps, where one samples from $p(y_i|y_{1:i-1}, \boldsymbol{\theta})$, for $i = 1, \ldots, n$. We shall see in our examples section, see Section 5, that several important applications of likelihood-free inference fulfil this requirement. We shall also discuss in Section 7 how other likelihood-free situations may be accommodated by the EP-ABC approach.

3.2 Numerical stability

EP-ABC is a stochastic version of EP, a deterministic algorithm, hence some care must be taken to ensure numerical stability. We describe here three strategies towards this aim.

First, to ensure that the stochastic error introduced by each site update does not vary too much in the course of the algorithm, we adapt dynamically M, the number of simulated points, as follows. For a given site update, we sample repetitively M_0 pairs $(\boldsymbol{\theta}^{[m]}, y_i^{[m]})$, as described in Algorithm 1, until the total number of accepted points exceeds some threshold $M_{\min}$. Then we compute the moments (3.3) and (3.4) based on all the accepted pairs.

Second, EP-ABC computes a great deal of Monte Carlo estimates, based on IID (independent and identically distributed) samples, part of which are Gaussian. Thus, it seems worthwhile to implement variance reduction techniques that are specific to the Gaussian distribution. After some investigation, we recommend the following quasi-Monte Carlo approach. We generate a Halton sequence $\boldsymbol{\xi}^{[m]}$ of dimension d, which is a low discrepancy sequence in $[0,1]^d$, and take

$$\boldsymbol{\theta}^{[m]} = \boldsymbol{\mu}_{-i} + \boldsymbol{L}_{-i}\Phi^{-1}\left(\boldsymbol{\xi}^{[m]}\right), \quad \boldsymbol{L}_{-i}\boldsymbol{L}_{-i}^t = \boldsymbol{\Sigma}_{-i}$$

where $\boldsymbol{\mu}_{-i}$, $\boldsymbol{\Sigma}_{-i}$ are the moment parameters corresponding to the natural parameters $\boldsymbol{r}_{-i}$, $\boldsymbol{Q}_{-i}$, $\boldsymbol{L}_{-i}$ is the Cholesky lower triangle of $\boldsymbol{\Sigma}_{-i}$, and Φ^{-1} returns a vector that contains the $N(0,1)$ inverse distribution function of each component of the input vector. We recall briefly that a low discrepancy sequence in $[0,1]^d$ is a deterministic sequence that spreads more evenly over the hyper-cube $[0,1]^d$ than a sample from the uniform distribution would; we refer the readers to e.g. Chap. 3 of Gentle (2003) for a definition of Halton and other low discrepancy sequences, and the theory of quasi-Monte Carlo. Rigorously speaking, this quasi-Monte Carlo version of EP-ABC is a hybrid between Monte Carlo and quasi-Monte Carlo, because the $y_i^{[m]}$ are still generated using standard Monte Carlo. However, we do observe a dramatic improvement when using this quasi-Monte Carlo approach. An additional advantage is that one may save some computational time by generating once and for all a very large sequence of $\Phi^{-1}\left(\boldsymbol{\xi}^{[m]}\right)$ vectors, and store it in memory for all subsequent runs of EP-ABC.

The third measure we may take is to slow down the progression of the algorithm such as to increase stability, by conservatively updating the parameters of the approximation in Step 3 of Algorithm 3, that is, $\boldsymbol{\lambda}_i \leftarrow \alpha(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{-i}) + (1-\alpha)\boldsymbol{\lambda}_i$. Standard EP is the special case with $\alpha = 1$. Updates of this type are suggested in Minka (2004).

In our experiments, we found that the two first strategies improve performance very significantly (in the sense of reducing Monte Carlo variability over repeated runs), and that the third strategy is sometimes useful, for example in our reaction time example, see Section 5.4.

3.3 Evidence approximation

In this section, we consider the special case $s_i(y_i) = y_i$, and we normalise the ABC posterior (3.1) as follows:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) = \frac{1}{p_\epsilon(\boldsymbol{y}^\star)} p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \frac{\mathbb{1}_{\left\{\|y_i - y_i^\star\| \leq \varepsilon\right\}}}{v_i(\epsilon)} \, dy_i \right\}, \tag{3.5}$$

where $v_i(\epsilon)$ is the normalising constant of the uniform distribution with respect to the ball of centre $y_i^\star$, radius ϵ, and norm $\|\cdot\|$. For the Euclidean norm, and assuming that the y_i's have the same dimension d_y, one has: $v_i(\epsilon) = v_i(1)\epsilon^{d_y}$, with $v_i(1) = \pi^{d_y/2}/\Gamma(d_y/2+1)$; e.g. $v_i(1) = 2$ if $d_y = 1$, $v_i(1) = \pi$ if $d_y = 2$.

Wilkinson (2008) shows that a standard ABC posterior such as (1.1) can be interpreted as the posterior distribution of a new model, where the summary statistics are corrupted with a uniformly-distributed noise (assuming these summary statistics are sufficient). The expression above indicates that this interpretation also holds for this type of summary-less ABC posterior, except that the artificial model is now such all the random variables y_i are corrupted with noise (conditional on $y_{1:i-1}^\star$).

The expression above also raises an important point regarding the approximation of the evidence. In (3.5), the normalising constant $p_\epsilon(\boldsymbol{y}^\star)$ is the evidence of the corrupted model, which converges to the evidence $p(\boldsymbol{y}^\star)$ of the actual model

as $\epsilon \to 0$. On the other hand, EP-ABC targets (3.1), and, in particular, see Section 2.3, produces an EP approximation of its normalising constant, which is $p_\epsilon(\boldsymbol{y}^\star) \prod_{i=1}^n v_i(\epsilon)$. Thus, one needs to divide this EP approximation by $\prod_{i=1}^n v_i(\epsilon)$ in order recover an approximation of $p_\epsilon(\boldsymbol{y}^\star)$. We found in our simulations that, when properly normalised as we have just described, the approximation of the evidence provided by EP-ABC is particularly accurate, see Section 5. In contrast, standard ABC based on summary statistics cannot provide an approximation of the evidence, as explained in the Introduction.

4 Speeding up EP-ABC in the IID case

Typically, the main computational bottleneck of EP-ABC, or other types of ABC algorithms, is simulating pseudo data-points from the model. In this section, we explain how these simulations may be recycled throughout the iterations in the IID (independent and identically distributed) case, so as to significantly reduce the overall computational cost of EP-ABC.

Our recycling scheme is based on a straightforward importance sampling strategy. Consider an IID model, with likelihood $p(\boldsymbol{y}^\star|\boldsymbol{\theta}) = \prod_{i=1}^n p(y_i^\star|\boldsymbol{\theta})$. Also, for the sake of simplicity, take $s_i(y_i) = y_i$. Assume that, for a certain site i, pairs $(\boldsymbol{\theta}^{[m]}, y^{[m]})$ are generated from $q_{-i}(\boldsymbol{\theta})p(y|\boldsymbol{\theta})$, as described in Algorithm 1. We have removed the subscript i in both $y^{[m]}$ and $p(y|\boldsymbol{\theta})$, to highlight the fact that the generative process of the data-points is the same for all the sites. The next update, for site $i+1$, requires computing moments with respect to $q_{-(i+1)}(\boldsymbol{\theta})p(y|\boldsymbol{\theta})\mathbb{1}\left\{\|y - y_{i+1}^\star\| \leq \epsilon\right\}$. Thus, we may recycle the simulations of the previous site by assigning to each pair $(\boldsymbol{\theta}^{[m]}, y^{[m]})$ the importance sampling weight:

$$w_{i+1}^{[m]} = \frac{q_{-(i+1)}\left(\boldsymbol{\theta}^{[m]}\right)}{q_{-i}\left(\boldsymbol{\theta}^{[m]}\right)} \times \mathbb{1}\left\{\|y^{[m]} - y_{i+1}^\star\| \leq \epsilon\right\}$$

and compute the corresponding weighted averages.

Obviously, this step may also be applied to the subsequent sites, $i+2, i+3, \ldots$, until one reaches a stage when the weighted sample is too degenerated. When this happens, "fresh" simulations may be generated from the current site. Algorithm 2 describes more precisely this recycling strategy. To detect weight degeneracy, we use the standard ESS (Effective Sample Size) criterion of Kong et al. (1994): we regenerate when the ESS is smaller that some threshold $\mathrm{ESS}_{\min}$.

The slower the EP approximation evolves, the less often regenerating the pseudo data-points is necessary, so that as the approximation gradually stabilises, we do not need to draw any new samples any more. Since EP slows down rapidly during the first two passes, most of the computational effort will be devoted to the early phase, and additional passes through the data will come essentially for free.

In non-IID cases several options are still available. For some models the data may come in blocks, each block made of IID data-points (think for example of a linear model with discrete predictors). We can apply the strategy outlined above in a block-wise manner (see the reaction times example, section 5.4). In other models there may be an easy transformation of the samples for data-point i such that they become samples for data-point $j \neq i$, or one may be able to reuse part of the simulation.

5 Case studies

5.1 General methodology

In each scenario, we apply the following approach. In a first step, we run the EP-ABC algorithm. We may run the algorithm several times, to evaluate the Monte Carlo variability of the output, and we may also run it for different values of ε, in order to assess the sensitivity to this approximation parameter. We use the first run to determine how many passes (complete cycles through the n sites) are necessary to reach convergence. A simple way to monitor convergence is to plot the evolution of the expectation (or the smallest eigenvalue of the covariance matrix) of the current Gaussian approximation q along the site updates. Note however that, in our experience, it is quite safe to simply fix the number of complete passes to a small number like 4. Finally, note that in each example, we could take $s_i(y_i) = y_i$, so the point of determining appropriate local summary statistics is not discussed.

In a second step, we run alternative algorithms, that is, either an exact (but typically expensive) MCMC algorithm, or an ABC algorithm, based on some set of summary statistics. The ABC algorithm we implement is a Gaussian random walk version of the MCMC-ABC algorithm of Marjoram et al. (2003). This algorithm targets a standard ABC approximation, i.e. (1.1), that corresponds to a single constraint $\{\|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\boldsymbol{y}^\star)\| \leq \epsilon\}$, for some vector of summary statistics $\boldsymbol{s}$, and some ϵ; the specific choices of $\boldsymbol{s}$ and ϵ are discussed for each application. We calibrate the tuning parameters of these MCMC algorithms using the information provided by the first step: we use as a starting point for the MCMC chain the expectation of the approximated posterior distribution provided by the EP-ABC algorithm, random walk scales are taken to be some fraction of the square root of the approximated posterior variances, and so on. This makes our comparisons

Algorithm 2 Computing the moments of the hybrid distribution in the likelihood-free setting, recycling scheme for IID models.

Inputs: i, ϵ, current weighted sample $(\boldsymbol{\theta}^{[m]}, y^{[m]})_{m=1}^M$, moment parameters $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ (resp. $\boldsymbol{\mu}_{-i}$ and $\boldsymbol{\Sigma}_{-i}$) that correspond to the site where data were re-generated for the last time (resp. that correspond to the Gaussian approximation $\prod_{j\neq i} q_j(\boldsymbol{\theta})$).

1. Compute the importance sampling weights

$$w_i^{[m]} = \frac{N(\boldsymbol{\theta}^{[m]}; \boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})}{N(\boldsymbol{\theta}^{[m]}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})} \times \mathbb{1}_{\left\{\|y^{[m]} - y_i^\star\| \leq \epsilon\right\}}$$

   where $N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the Gaussian $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ probability density evaluated at $\boldsymbol{\theta}$, and the effective sample size:

$$\text{ESS} = \frac{\left(\sum_{m=1}^M w_i^{[m]}\right)^2}{\sum_{m=1}^M \left(w_i^{[m]}\right)^2}.$$

2. If ESS $<$ $\text{ESS}_{\min}$, replace $(\boldsymbol{\theta}^{[m]}, y^{[m]})_{m=1}^M$ by M IID draws from $q_{-i}(\boldsymbol{\theta})p(y|\boldsymbol{\theta})$, set $w_i^{[m]} = \mathbb{1}_{\left\{\|y^{[m]} - y_i^\star\| \leq \epsilon\right\}}$, and $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{-i}$, $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{-i}$.

3. Compute the following importance sampling estimates:

$$\widehat{Z}_h = \frac{1}{M}\sum_{m=1}^M w_i^{[m]}, \quad \widehat{\boldsymbol{\mu}}_h = \frac{\sum_{m=1}^M w_i^{[m]} \boldsymbol{\theta}^{[m]}}{\widehat{Z}_h}$$

   and

$$\widehat{\boldsymbol{\Sigma}}_h = \frac{\sum_{m=1}^M w_i^{[m]} \boldsymbol{\theta}^{[m]} \left\{\boldsymbol{\theta}^{[m]}\right\}^t}{\widehat{Z}_h} - \widehat{\boldsymbol{\mu}}_h \widehat{\boldsymbol{\mu}}_h^t.$$

Return $(\boldsymbol{\theta}^{[m]}, y^{[m]})_{m=1}^M$, $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}}$, $\widehat{Z}_h$, $\widehat{\boldsymbol{\mu}}_h$ and $\widehat{\boldsymbol{\Sigma}}_h$.

particularly unfavourable to EP-ABC. Despite this, we find consistently that the EP-ABC algorithm is faster by several orders of magnitude, and leads to smaller overall approximation errors. We report computational loads both in terms of CPU time (e.g. 30 seconds) and in terms of the number of simulations of replicate data-points y_i. The latter should be typically the bottleneck of the computation.

All the computations were performed on a standard desktop PC in Matlab; programs are available from the first author's web page.

5.2 First example: Alpha-stable Models

Alpha-stable distributions are useful in areas (e.g. Finance) concerned with noise terms that may be skewed, may have heavy tails and an infinite variance. A univariate alpha-stable distribution does not admit a close-form expression for its density, buy may be specified through its characteristic function

$$\Phi_X(t) = \begin{cases} \exp\left[i\delta t - \gamma^\alpha |t|^\alpha \left\{1 + i\beta \left(\tan\frac{\pi\alpha}{2}\right) \text{sgn}(t) \left(|\gamma t|^{1-\alpha} - 1\right)\right\}\right] & \alpha \neq 1 \\ \exp\left[i\delta t - \gamma |t| \left\{1 + i\beta \frac{2}{\pi} \text{sgn}(t) \log|\gamma t|\right\}\right] & \alpha = 1 \end{cases}$$

where α determines the tails, $0 < \alpha \leq 2$, β determines skewness, $-1 < \beta < 1$, and $\gamma > 0$ and δ are respectively scale and location parameters; see Nolan (2012, Chap. 1) for a general introduction to stable distributions.

Peters et al. (2010) consider a model of n i.i.d. observations y_i, $i = 1, \ldots, n$ from a univariate alpha-stable distribution, and propose to use the ABC approach to infer the parameters. Likelihood-free inference is appealing in this context, because sampling from an alpha-stable distribution is fast (using e.g. the algorithm of Chambers et al., 1976), while computing its density is cumbersome.

Trying EP-ABC on this example is particularly interesting for the following reasons: (a) Peters et al. (2010) show that choosing a reasonable set of summary statistics for this problem is difficult, and that several natural choices lead to strong

biases; and (b) since alpha-stable distributions are very heavy-tailed, the posterior distribution may be heavy-tailed as well, which seems a challenging problem for a method based on a Gaussian approximation such as EP-ABC.

Our data consist of $n = 1264$ rescaled log-returns, $y_t = 100 * \log(z_t/z_{t-1})$, computed from daily exchange rates z_t of AUD (Australian Dollar) recorded in GBP (British Pound) between 1 January 2005 and 1 December 2010. (These data are publicly available on the Bank of England web-site.) We take $\boldsymbol{\theta} = \left(\Phi^{-1}(\alpha/2), \Phi^{-1}\left((\beta+1)/2\right), \log\gamma, \delta\right)$ where Φ is the $N(0,1)$ cumulative distribution function, and we set the prior to $N\left(0_4, \mathrm{diag}(1,1,10,10)\right)$. Note however that our results are expressed in terms of the initial parametrisation α, β, γ and δ; i.e. for each parameter we report the approximate marginal posterior distribution obtained through the appropriate variable transform of the Gaussian approximation produced by EP-ABC. We run the EP-ABC algorithm (recycling version, as model is IID, see Section 4), with $\epsilon = 0.1$, $M = 8 \times 10^6$, $\mathrm{ESS}_{\min} = 2 \times 10^4$, and $\|\cdot\|$ set to the Euclidean norm in $\mathbb{R}$ (i.e. the n constraints in (3.1) simplify to $|y_i - y_i^\star| \leq \varepsilon$). Variations over ten runs are negligible. Average CPU time for one run is 39 minutes, and average number of simulated data-points over the course of the algorithm, is 4×10^8.

We first compare these results with the output of an exact random-walk Hastings-Metropolis algorithm, which relies on the evaluation of an alpha-stable probability density function for each data-point (using numerical integration). Because of this, this algorithm is very expensive. We ran the exact algorithm for about 60 hours (2×10^5 iterations). One sees in Figure 5.1 that the difference between EP-ABC and the exact algorithm is negligible.

We then compare these results with those obtained by MCMC-ABC, for the set of summary statistics which performs best among those discussed by Peters et al. (2010, see S_1 in Section 3.1). We run 2×10^7 iterations of this sampler, which leads to about 50 times more simulations from an univariate alpha-stable distribution than in the EP-ABC runs above. Through pilot runs, we decided to set $\epsilon = 0.03$, which seems to be as small as possible, subject to having a reasonable acceptance rate (2×10^{-3}) for this computational budget. In Figure 5.1, we see that the posterior output from this MCMC-ABC exercise is not as good an approximation as the output of EP-ABC. As explained in the previous section, we have set the starting point of the MCMC-ABC chain to the posterior mode. If initialised from some other point, the sampler typically takes a much longer time to reach convergence, because the acceptance rate is significantly lower in regions far from the posterior mode.

Finally, we also use EP-ABC, with the same settings as above, e.g. $\epsilon = 0.1$, in order to approximate the evidence of the model above (-1385.8) and two alternative models, namely a symmetric alpha-stable model, where β is set to 0 (-1383.8), and a Student model (-1383.6), with 3 parameters (scale γ, position δ, degrees of freedom ν, and a Gaussian prior $N(0_3, \mathrm{diag}(10,10,10))$ for $\boldsymbol{\theta} = (\log\nu, \gamma, \delta)$). (Standard deviation over repeated runs is below 0.1.) One sees that there no strong evidence of skewness in the data, and that the Student distribution and a symmetric alpha-stable distribution seem to fit equally well the data. We obtained the same value (-1383.6) for the evidence of the Student model when using the generalised harmonic mean estimator (Gelfand and Dey, 1994) based on a very long chain of an exact MCMC algorithm. For both alpha-stable models, this approach proved to be too expensive to allow for a reliable comparison.

5.3 Second example: Lotka-Volterra models

The stochastic Lotka-Volterra process describes the evolution of two species Y_1 (prey) and Y_2 (predator) through the reaction equations:

$$\begin{aligned} Y_1 &\xrightarrow{r_1} 2Y_1 \\ Y_1 + Y_2 &\xrightarrow{r_2} 2Y_2 \\ Y_2 &\xrightarrow{r_3} \emptyset. \end{aligned}$$

This chemical notation means that, in an interval $[t, t+dt]$, the probability that one prey is replaced by two preys is $r_1 d_t$, and so on. Typically, the observed data $\boldsymbol{y}^\star = (y_1, \ldots, y_n)$ are made of n vectors $y_i^\star = (y_{i,1}^\star, y_{i,2}^\star)$ in $\mathbb{N}^2$, which correspond to the population levels at integer times. We take $\boldsymbol{\theta} = (\log r_1, \log r_2, \log r_3)$. This model is Markov, $p(y_i^\star | y_{1:i-1}^\star, \boldsymbol{\theta}) = p(y_i^\star | y_{i-1}^\star, \boldsymbol{\theta})$ for $i > 1$, and one can efficiently simulate from $p(y_i^\star | y_{i-1}^\star, \boldsymbol{\theta})$ using Gillespie (1977)'s algorithm. On the other hand, the density $p(y_i^\star | y_{i-1}^\star, \boldsymbol{\theta})$ is intractable. This makes this model a clear candidate both for ABC, as noted by Toni et al. (2009), and for EP-ABC. Boys et al. (2008) show that MCMC remains feasible for this model, but in certain scenarios the proposed schemes are particularly inefficient, as noted also by Holenstein (2009, Chap. 4).

Following the aforementioned papers, we consider a simulated dataset, corresponding to rates $r_1 = 0.4$, $r_2 = 0.01$, $r_3 = 0.3$, initial population values $y_{0,1}^\star = 20$, $y_{0,2}^\star = 30$ and $n = 50$; see Figure 5.2. Since the observed data are integer-valued, we use the supremum norm in (3.1), and an integer-valued ϵ; this is equivalent to imposing simultaneously the $2n$ constraints $\left|y_{i,1} - y_{i,1}^\star\right| \leq \epsilon$ and $\left|y_{i,2} - y_{i,2}^\star\right| \leq \epsilon$ in the ABC posterior.

First, we run EP-ABC (standard version) with $M_{\min} = 4000$, and for both $\epsilon = 3$ and $\epsilon = 1$. We find that a single pass over the data is sufficient to reach convergence. For $\epsilon = 3$ (resp. $\epsilon = 1$), CPU time for each run is 2.5 minutes (resp.

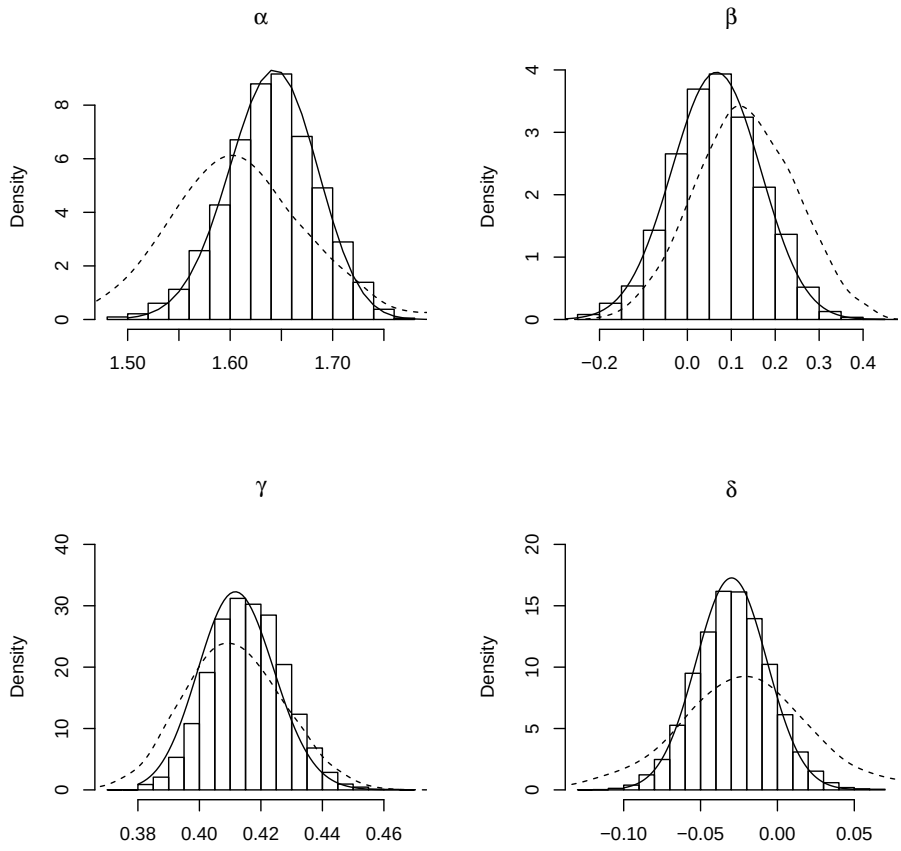


Fig. 5.1: Marginal posterior distributions of α, β, γ and δ for alpha-stable model: MCMC output from the exact algorithm (histograms), approximate posteriors provided by first run of EP-ABC (solid line), kernel density estimates computed from MCMC-ABC sample based on summary statistic proposed by Peters et al. (2010) (dashed line).

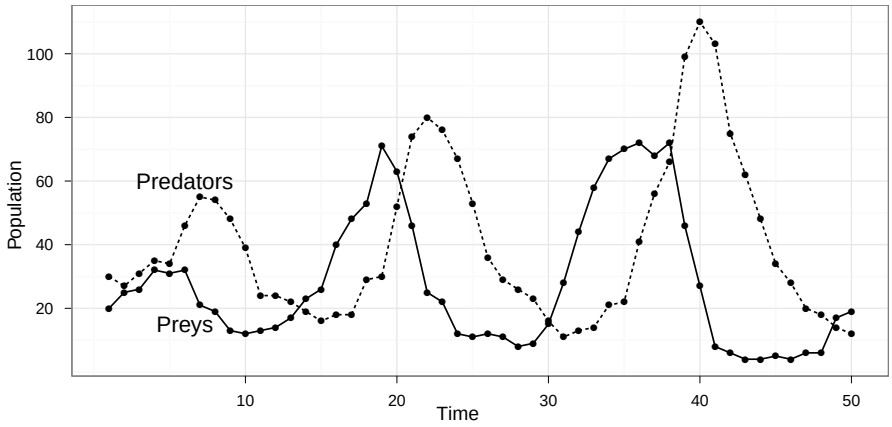


Fig. 5.2: Lotka-Volterra example: simulated dataset

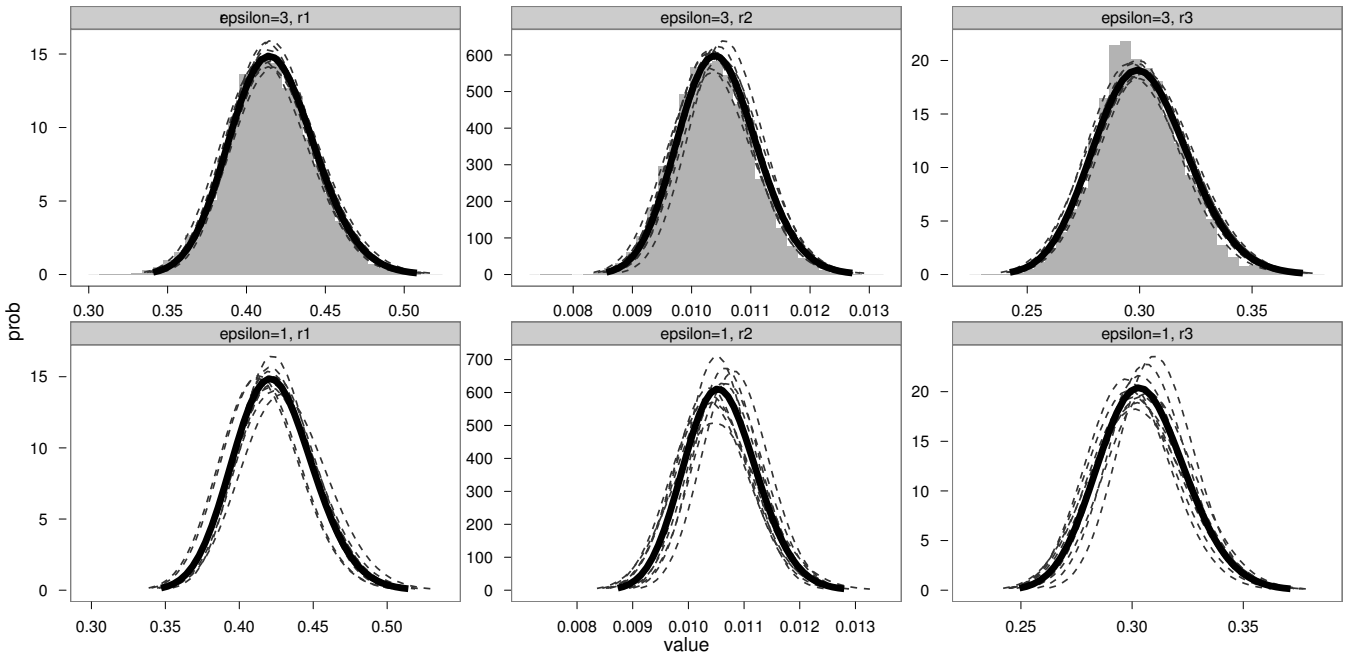


Fig. 5.3: Lotka-Volterra example: marginal posterior densities of rates r_1, r_2, r_3, obtained from PMCMC algorithm (histograms), and from ABC-EP, for $\epsilon = 3$ (top) and $\epsilon = 1$ (bottom); PMCMC results for $\epsilon = 1$ could not be obtained in a reasonable time. The solid lines correspond to the average over 10 runs of the moment parameters of the Gaussian approximation, while the dashed lines correspond to the 10 different runs.

25 minutes), and number of simulated transitions $p(y_i|y^\star_{i-1}, \boldsymbol{\theta})$ is about 10^7 (resp. 9×10^7); marginal posteriors obtained through EP-ABC are reported in Figure 5.3.

When applying ABC to this model, Toni et al. (2009) uses as a pseudo-distance between the actual data $\boldsymbol{y}^\star$ and the simulated data $\boldsymbol{y}$ the sum of of squared errors. In Wilkinson (2008)'s perspective discussed in Section 4, this is equivalent to considering a state-space model where the latent process is the Lotka-Volterra process described above, and the observation process is the same process, but corrupted with Gaussian noise (provided the indicator function $\mathbb{1}\left\{\sum(y_i - y_i^\star)^2 \leq \varepsilon\right\}$ in the ABC posterior is replaced by the kernel function $\exp\left\{-\sum(y_i - y_i^\star)^2/\varepsilon\right\}$). Thus, instead of a standard ABC algorithm, it seems more efficient to resort to a MCMC sampler specifically designed for state-space models in order to simulate from the ABC approximation of Toni et al. (2009). Following Holenstein (2009, Chap. 4), we consider a Gaussian random walk version of the marginal PMCMC sampler. This algorithm is a particular instance of the state of the art PMCMC framework of Andrieu et al. (2010), which is based on the idea of approximating the marginal likelihood of the data by running a particle filter of size N at each iteration of the MCMC sampler. The big advantage of PMCMC in this situation (comparatively to other MCMC approaches for state-space models), is that it does not requires a tractable probability density for the Markov transition of the state-space model.

In Figure 5.3, we report the posterior output obtained from this sampler, run for about 2×10^5 iterations and $N = 1000$ particles (2 days in CPU time, 10^{10} simulated transitions $p(y_i|y^\star_{i-1}, \boldsymbol{\theta})$), with random walk scales set to obtain approximatively a 25% acceptance rate. These plots correspond to $\epsilon = 3$, and a state-space model with an uniformly distributed observation noise. In Figure 5.3, one detects practically no difference between PMCMC and EP-ABC with $\epsilon = 3$ (black lines), although the CPU time of the latter was about 1500 smaller.

The difference between the two EP-ABC approximations (corresponding to $\epsilon = 1$ and $\epsilon = 3$) is a bit more noticeable. Presumably, the EP-ABC approximation corresponding to $\epsilon = 1$ is slightly closer to the true posterior. We did not manage however to obtain reliable results from our PMCMC sampler and $\epsilon = 1$ in a reasonable time.

5.4 Third example: Race models of reaction times

Reaction time models seek to describe the decision behaviour of (human or animal) subjects in a choice task (Luce, 1991; Meyer et al., 1988; Ratcliff, 1978). In the typical experiment, subjects view a stimulus, and must choose an appropriate response. For example, the stimulus might be a set of moving points, and the subject must decide whether the points

move to the left or to the right.

Assuming that the subject may choose between k alternatives, one observes independent pairs, $y_i = (d_i, r_i)$, where $d_i \in \{1, \ldots, k\}$ is the chosen alternative, and $r_i \geq 0$ is the measured reaction time. For convenience, we drop for now the index i in order to describe the random distribution of the pair (d, r).

Reaction time models assume that the brain processes information progressively, and that a decision is reached when a sufficient amount of information has been accumulated. In the model we use here (a variant of typical models found in e.g. Ratcliff and McKoon, 2008; Bogacz et al., 2007) k parallel integrators represent the evidence $e_1(t), \ldots, e_k(t)$ in favour of each of the k alternatives. The model is illustrated on Figure 5.4. The first accumulator to reach its boundary b_j wins the race and determines which response the subject will make. Each accumulator undergoes a Wiener process with drift:

$$\tau de_j(t) = m_j dt + dW_t^j$$

where the m_j's are the drift parameters, the W_t^j's are k independent Wiener processes; and τ is a fixed time scale, $\tau = 5ms$. The measured reaction time is corrupted by a uniformly-distributed noise r_{nd}, representing the "non-decisional time" (Ratcliff and McKoon, 2008), i.e. the time the subject needs to execute the decision (prepare a motor command, press an answer key, ...). This model is summarised by the following equations:

$$\begin{aligned} r &= r_d + r_{nd}, \quad r_{nd} \sim \mathrm{U}[a, b], \\ r_d &= \min_j \inf_t \{t : e_j(t) = b_j\}, \\ d &= \arg\min_j \inf_t \{t : e_j(t) = b_j\}. \end{aligned}$$

(We fix a and b to $a = 100ms$, $b = 200ms$, credible values from Ratcliff and McKoon (2008))

The model above captures the essential ideas of reaction time modelling, but it remains too basic for experimental data. We now describe several important extensions. First, a better fit is obtained if the boundaries are allowed to vary randomly from trial to trial (as in Ratcliff, 1978): we assume that $b_j = c_j + \tau$, where $\tau \sim N(0, e^s)$, and s is a parameter to be estimated. Second, a mechanism is needed to ensure that the reaction times cannot be too large: we assume that if no boundary has been reached after 1 second, information accumulation stops and the highest accumulator determines the decision. Finally, one needs to account for lapses (Wichmann and Hill, 2001): on certain trials, subjects simply fail to pay attention to the stimuli and respond more or less at random. We account for this phenomenon by having a lapse probability of 5%. In case a lapse occurs, r_d becomes uniformly distributed between 0 and 800 ms and the response is chosen between the alternatives with equal probability. Clearly, although this generalised model remains amenable to simulation, the corresponding likelihood is intractable.

We apply this model to data from an unpublished experiment by M. Maertens (personal communication to Simon Barthelmé). The dataset is made of 1860 observations, obtained from a single human subject, which had to choose between $k = 2$ alternatives: "signal absent" (no light increment was presented), or "signal present" (a light increment was presented), under 15 different experimental conditions: 3 different locations on the screen, and 5 different contrast values. Following common practice in this field, trials with very high or very low reaction times (top and bottom 5%) were excluded from the dataset, because they have a high chance of being outliers (fast guesses, keyboard errors or inattention). The data are shown on Figure 5.5.

From the description above, one sees that five parameters, (m_1, m_2, c_1, c_2, s), are required to describe the random behaviour of a single pair (r_i, d_i), when $k = 2$. To account for the heterogeneity introduced by the varying experimental conditions, we assume that the 2 accumulation rates, m_1, m_2 vary across the 15 experimental conditions, while the 3 parameters related to the boundaries, c_1, c_2 and s, are shared across conditions. The parameter $\boldsymbol{\theta}$ is therefore 33-dimensional.

We note that this model would present a challenge for inference even if the likelihood function was available. It is difficult to assign priors to the parameters, because they do not have a clear physical interpretation, and available prior knowledge (e.g., that reaction times will normally be less than 1 second) does not map easily unto them. Moreover, the model is subject in certain cases to weak identifiability problems. For instance, if one response dominates the dataset, there is little information available beyond the fact that one drift rate is much higher than the other (or one threshold much lower than the other, or both).

We re-parametrised the positive parameters c_1, c_2 as $c_1 = e^{\lambda}$, $c_2 = e^{\lambda+\delta}$, and assigned a $\mathcal{N}(0, 1)$ prior to λ, δ, and s. Taking a $N(0, 5^2)$ prior for these 3 quantities led to similar results. Some experimentation suggested that the drift rates could be constrained to lie between -0.1 and 0.1, because values outside of this interval seem to yield improbable reaction times (too short or too long). We assigned a $[-0.1, 0.1]$ uniform prior for the 30 drift rates, and applied the appropriate transform, i.e. $x \to \Phi^{-1}(0.1 + 5x)$, in order to obtain a $N(0, 1)$ prior for the transformed parameters.

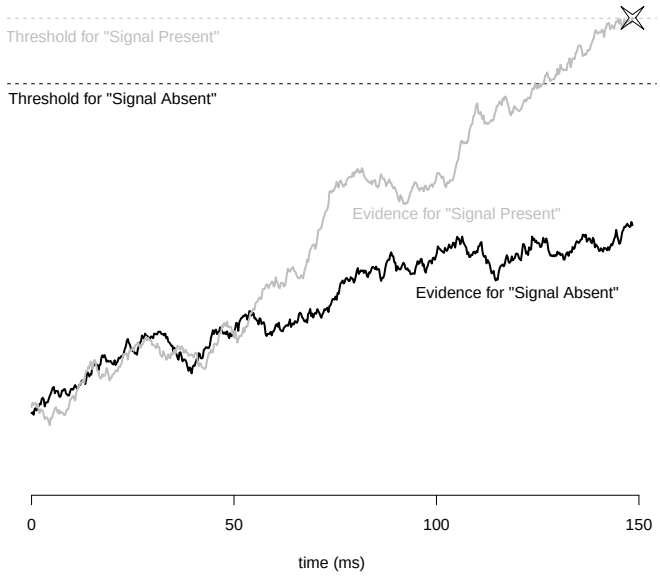


Fig. 5.4: A model of reaction time in a choice task. The subject has to choose between k responses (here, "Signal present" and "Signal absent") and information about each accumulates over time in the form of evidence in favour of one and the other. Because of noise in the neural system "evidence" follows a random walk over time. A decision is reached when the evidence for one option reaches a threshold (dashed lines). The decision time in this example is denoted by the star: here the subject decides for 'B' after about 150ms. The fact that the thresholds are different for "Signal Present" and "Signal Absent" capture decisional bias: in general, for the same level of information, the subject favours option "Signal Absent".

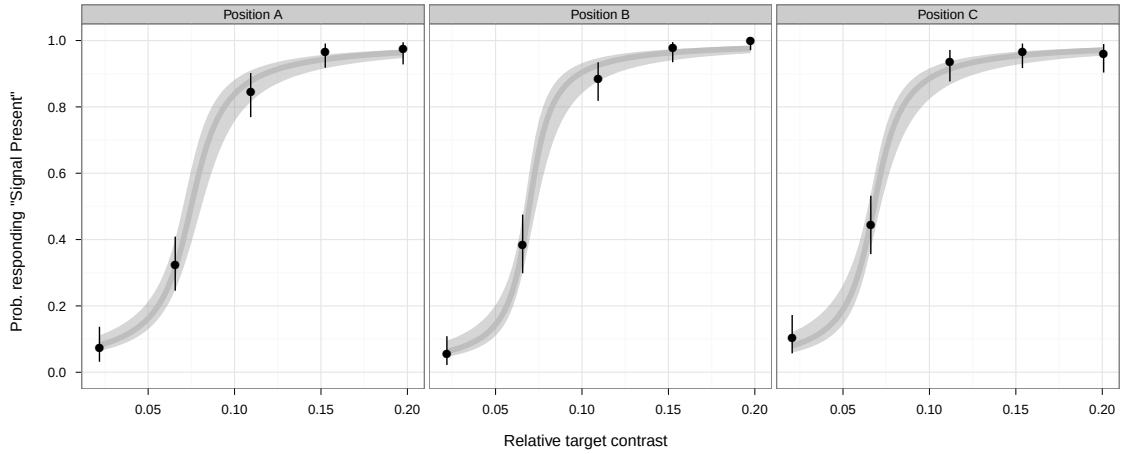
After a few unsuccessful attempts, we believe that this application is out of reach of normal ABC techniques. The main difficulty is the choice of the summary statistics. For instance, if one takes basic summaries (e.g. quartiles) of the distribution of reactions times, under each of the 15 experimental conditions, one ends with a very large vector $\boldsymbol{s}$ of summary statistics. Due to the inherent curse of dimensionality of standard ABC (Blum, 2010), sampling enough datasets, of size 1860, which are sufficiently close (in terms of the pseudo-distance $\|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\boldsymbol{y}^\star)\|$) would require enormous computational effort. Obviously, taking a much smaller set of summary statistics would on the other hand lead to too poor an approximation.

Some adjustments are needed for ABC-EP to work on this problem. First, notice that within a condition the datapoints are IID so that the posterior distribution factorises over IID "blocks". We can therefore employ the recycling technique described in Section 4 to save simulation time, by going through the likelihood sites block-by-block. Second, since datapoints take values in $\{1, 2\} \times \mathbb{R}^+$, we adopt the following set of ABC constraints: $\mathbb{1}\{d_i = d_i^\star\}\,\mathbb{1}\{|\log r_i - \log r_i^\star| \leq \epsilon\}$, where $y_i^\star = (d_i^\star, r_i^\star)$ denotes as usual the actual datapoints. Third, we apply the following two variance-reduction techniques. One stems from the fact that each site likelihood does not depend on the whole 33 parameters but on a subset of size 5. In that case, using simple linear algebra, one can see that it is possible to update only the marginal distribution of the EP approximation with respect to these 5 parameters; see Appendix B for details. The second is a simple Rao-Blackwellisation scheme that uses the fact that the non-decisional component r_{nd} is uniformly distributed, and may therefore be marginalised out when computing the EP update.

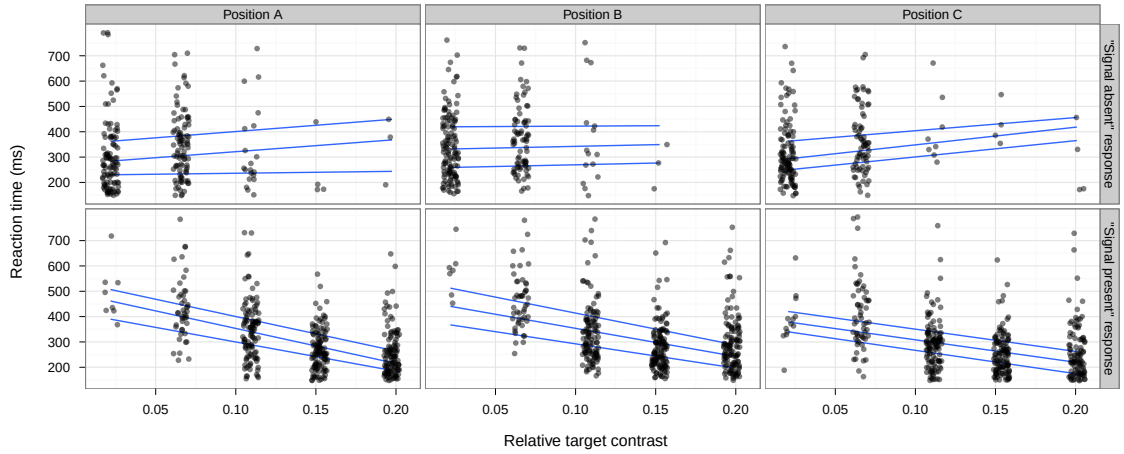
We report below the results obtained from ABC-EP with $\epsilon = 0.16$, $M = 3 \times 10^3$, $\alpha = 0.4$ (see end of Section 3.2), and 3 complete passes over the data; CPU time was about 40 minutes. Results for smaller values of ϵ, e.g. $\epsilon = 0.1$, were mostly similar, but required a larger CPU time.

Since we could not compare the results to those of a standard ABC algorithm, we assess the results through posterior predictive checking. For each of the 15 experimental conditions, we generate 5,000 samples from the predictive density, and compare the simulated data with the real, as follows.

The marginal distribution of responses can be summarised by regressing the probability of response on stimulus contrast, separately for each stimulus position (as on Figure 5.5), and using a binomial generalized linear model (with Cauchit link function). Figure 5.6 compares data and simulations, and shows that the predictive distribution successfully



(a) Probability of answering "Signal present" as a function of relative target contrast in a detection experiment, at three different positions of the target (data from one subject). Filled dots represent raw data, the grey curves are the result of fitting a binomial GLM with Cauchit link function. The light grey band is a 95% pointwise, asymptotic confidence band on the fit obtained from the observed Fisher information. As expected in such experiments, the probability of detecting the target increases with relative target contrast.



(b) Reaction time distributions conditional on target contrast, target position, and response. The semi-transparent dots represent reaction times for individual trials. Horizontal jitter was added to aid visualisation. The lines represent linear quantile regressions for the 30%, 50% and 70% quantiles.

Fig. 5.5: Choice (**a**) and reaction time (**b**) data in a detection experiment by Maertens.

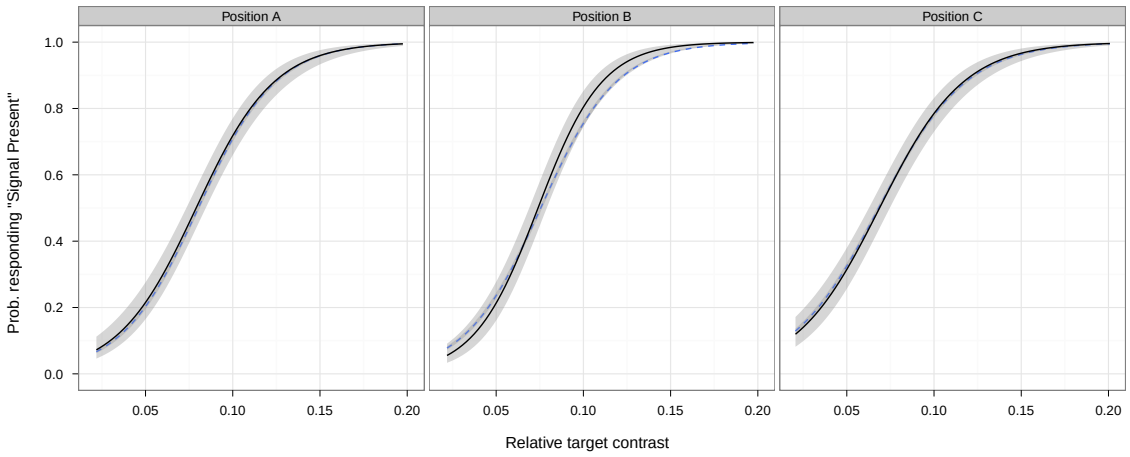


Fig. 5.6: Probability of answering "Signal Present" as a function of contrast: data versus (approximate) posterior predictive distribution. We sampled the posterior predictive distribution and summarised it using a binomial GLM with Cauchit link, in the same way as we summarised the data (see (a) in Figure 5.5). The data are represented as a continuous line, the predictive distribution as dotted. The posterior predictive distribution is very close to the data. The shaded area corresponds to a 95% confidence interval for the fit to the data.

captures the marginal distribution of responses.

To characterise the distribution of reaction times, we look at means and inter-quantile intervals, conditional on the response and the experimental conditions. The results are presented on Figure 5.7. The predictive distributions capture the location and scale of the reaction time distributions quite well, at least for those conditions with enough data. Such results seem to indicate that, at the very least, the ABC-EP approximate posterior corresponds to a high-probability region of the true posterior.

6 Difficult posteriors

One obvious cause for concern is the behaviour of EP-ABC when confronted with difficult posterior distributions, and in this section we explore a few possible scenarios and suggest methods for detecting and correcting potential problems. Of all potential problematic shapes a posterior distribution could have, the worst is for it to have several modes, and we deal with this question first. Nonmultimodal but otherwise problematic posteriors are discussed later.

6.1 Multimodality

We begin with a toy example of a multimodal posterior. Consider the following IID model: $y_i|\theta \sim N(|\theta|, 1)$, $i = 1, \ldots, n = 50$, and $\theta \sim N(0, 10^2)$. The dataset is obtained by sampling from the model, with $\theta = 2$. The true posterior may be computed exactly, and is plotted in the left hand side of Figure 6.1. It features two symmetric, well separated modes, around -2 and 2. The Gaussian pdf q that minimises $KL(q||\pi)$ where π is the posterior (or, in other words, the Gaussian pdf, with mean equal to posterior expectation, variance equal to the posterior variance) is also represented, as a dashed line, but it cannot be distinguished from the EP-ABC approximation (thick line).

The behaviour of EP-ABC in this case is instructive. If we run the standard EP-ABC algorithm (using the IID version, see Section 4 of the paper, and with $M = 8 \times 10^6$, $ESS_{\min} = 500$), we obtain a result very close to the the thick line in one pass (one cycle over all the observations). However, if the algorithm is run for two passes or more, there is a very high probability that the algorithm stops its execution before completion, because many site updates will produce negative definite contributions $\boldsymbol{Q}_i$. To perform more than one pass, we have to use slow updates (as described in Section 3 of the paper), and set α to a small value. The thick line in Figure 6.1 was obtained with three passes, and slow updates with $\alpha = 0.1$. The right hand side of 6.1 shows the type of plot one may use to assess convergence of EP-ABC: it represents the evolution of the mean of the EP approximation along the iterations; one "iteration" corresponds to one site update, and since we have performed three passes over $n = 50$ sites, there are 150 iterations in total. This plot seems to indicate that the posterior expectation is slowly drifting to the right, and does not stabilise. If we try to run for more passes, we end up again with non-invertible covariance matrices. This behaviour was already described by Minka in his PhD thesis (Minka, 2001b) for the deterministic case.

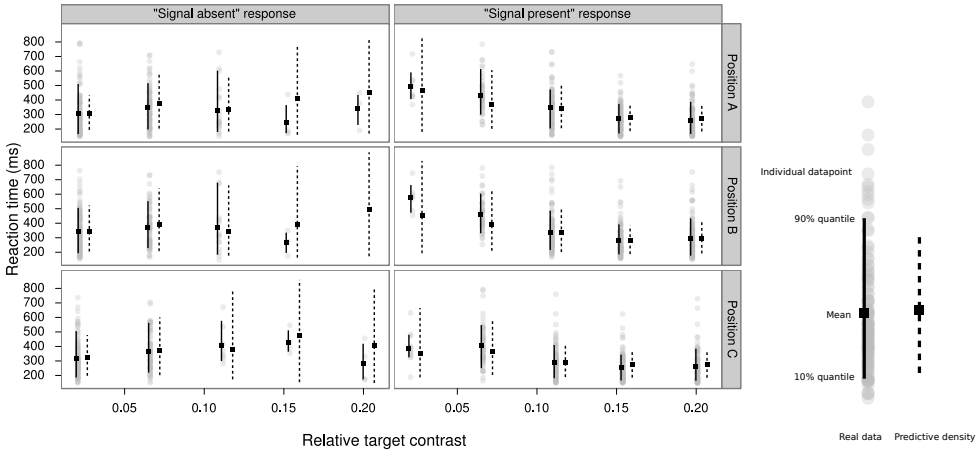


Fig. 5.7: Reaction time distributions conditional on decision: data versus posterior predictive distributions. The reaction times conditioned on contrast, position and response are shown as grey dots and summarised via mean and 10-90% inter-quantile range (continuous lines). The posterior predictive distributions computed from samples are summarised and shown with an offset to the right (dotted lines). The conditional densities are well captured, given sufficient data.

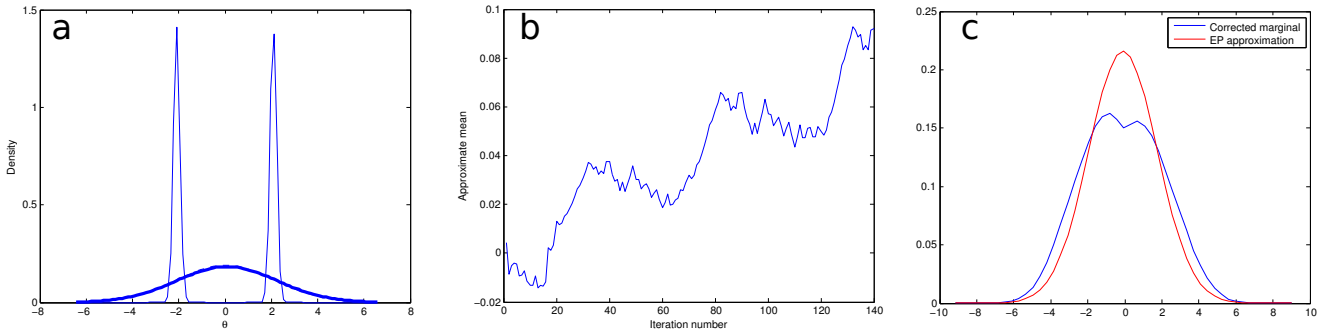


Fig. 6.1: Multi-modal example. **a.** True posterior pdf (thin line) versus EP-ABC approximation (thick line). **b.** mean of the EP-ABC Gaussian approximation during the course of the algorithm (i.e. vs iteration; 3 passes are performed, hence 150 iterations in total) **c.** EP approximation vs. 1st-order PWO correction (see Appendix

The first test to see if the posterior distribution is reasonable is simply to run EP-ABC and see if it fails. There are two possible causes for failures: either one used too few samples and Monte Carlo variance led to negative definite updates, or the model itself is problematic for EP, in which case EP will still fail when using large sample sizes.

Beyond this rather informal test there are more principled things one could do. In Paquet et al. (2009) a very generic class of corrections to EP is introduced, and is described in Appendix 8. Their first-order correction can be obtained relatively cheaply from the hybrid distributions, and can be used to build corrected marginals. In Figure 6.1 we show the first-order correction for our toy multi-modal example: although the corrected marginal is still far from the true posterior, it is clearly bimodal and in this case serves as a warning that the Gaussian approximation is inappropriate (we also applied the same type of correction in our third example, for which no MCMC gold standard is available, but the correction did not modify in any noticeable manner the marginal posterior distributions) In a similar vein, one could use goodness-of-fit tests to see if the n hybrid distributions show large deviations from normality.

Once the problem has been diagnosed, what can one do? The answer is, unfortunately, "not much". If one can determine that multimodality is caused by ambiguity in one or two parameters, it is possible to effectively treat these parameters as hyperparameters. One separates $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_A, \boldsymbol{\theta}_B$ where $\boldsymbol{\theta}_b$ is the set of problematic parameters. Running EP-ABC with $p(\boldsymbol{\theta}_A|\boldsymbol{y}, \boldsymbol{\theta}_B)$ will produce an estimate of $p(\boldsymbol{\theta}_b|\boldsymbol{y})$ which can be used to form an approximation to the marginals in a manner analoguous to the procedure used in INLA (Rue et al., 2009). Although this will require running EP-ABC several times for different values of $\boldsymbol{\theta}_B$, it might still be a lot cheaper than a MCMC procedure.

If no information is available about what the cause of the multimodality is or where the different modes are, our opinion is that all existing ABC methods will have tremendous difficulties. Work might be better invested in injecting more prior information or changing the parameterisation such as to ensure there is only one mode.

6.2 Difficult unimodal posteriors

Compared to the toy example of the previous section, a more realistic scenario is a posterior distribution that is unimodal but otherwise still badly behaved. There are no theoretical results available on convergence in this case, so we evaluated EP's behaviour in a case of a rather nasty posterior distribution, adapted from the model of reaction times described above.

We devised a problem in which we guessed the posterior would be at the minimum very badly scaled and would probably have an inconvenient shape. In the context of the reaction times model described in Section 5.4, imagine that in a choice experiment the subject picked exclusively the second category; so that we have no observed reaction times for the first category. From the point of the model this occurs whenever the threshold for the first category is high enough, compared to accumulation speed, that the second accumulator always wins. This creates complete uncertainty as to the ratio of accumulation speed vs. threshold value for the first accumulator.

We generated a large dataset (1,000 datapoints) based on fixed values of (m_1, m_2, c_1, c_2, s): $m_1 = 10^{-3}$, $m_2 = 0.08$, $c_1 = 20$, $c_2 = 10$, and $s = 0$. In this dataset there are no decisions favouring the first category, and the resulting reaction time distribution is plotted on Fig. 6.2. To make the inference problem still manageable using MCMC-ABC, we limit inference to three parameters: $\log m_1$, $\log m_2$ (log-accumulation rate of the two accumulators) and $\log c_1$ (log-threshold for the first category). The other two parameters are considered known and fixed at their true value. We chose these parameters because we expected the posterior to show quasi-certainty for $\log m_2$ and very high uncertainty for $\log \frac{m1}{c_1}$. To further increase uncertainty, we chose a very vague prior $\boldsymbol{\theta} \sim \mathcal{N}\left(0, 10^2 \times \boldsymbol{I}\right)$.

We ran EP-ABC on this problem using the recycling strategy described in section 3.2 (data are IID). In this case EP-ABC needs large sample sizes to be stable numerically, which is probably due to the very ill-conditioned covariance matrices that arise (itself due to a very poorly-scaled posterior, more on that below). To eliminate any such problems we deliberately chose to use a very high number of samples, 3 million, with a minimum expected sample size of 500. We used an acceptance window of 10ms. We did 3 passes over the data, for a total computing time of 9 minutes on a standard laptop. To check for stability, we ran EP-ABC 10 times.

To check the accuracy of the results, we ran a MCMC-ABC sampler that used as its summary statistics 8 quantiles of the reaction time distribution, linearly spaced from 0.01 to 0.99, and whether the first category was ever chosen. Quantiles were rescaled by a factor of $\frac{1}{200}$. Samples were accepted if the second category was chosen less than 100% of the time, or if the Euclidean distance between $|\boldsymbol{s}(\boldsymbol{y}^\star) - \boldsymbol{s}(\boldsymbol{y})|$ was over $\epsilon = 0.025$. This latter value was found by iteratively adjusting to get an acceptance rate of $1/1000$.

We would like to stress that an appropriate proposal distribution was found using the covariance matrix obtained using ABC-EP (and readjusted later based on short runs). The scaling of the posterior is such that using, e.g., the prior covariance matrix in a proposal distribution is impossible: one ends up using extremely high values of ϵ, and therefore with

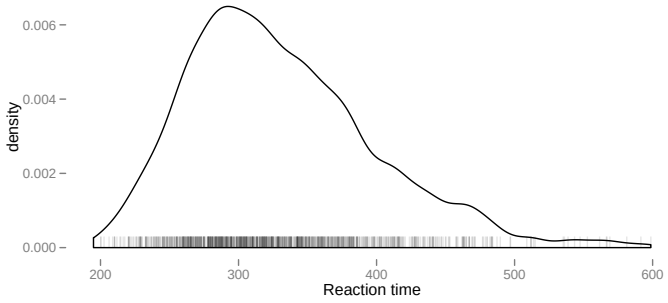


Fig. 6.2: Simulated reaction time data used in creating the difficult posterior of Section 6. Individual reaction times are marked with ticks, and a kernel density estimate is overlaid.

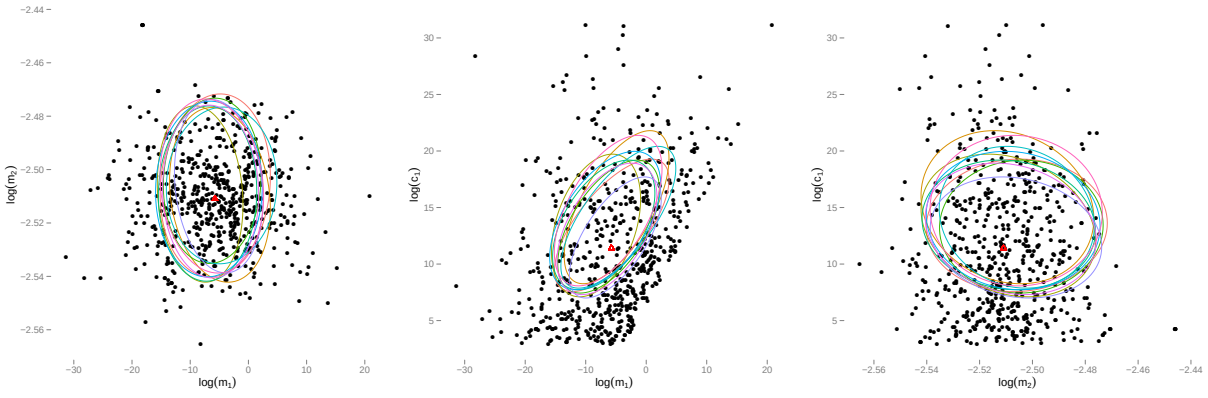


Fig. 6.3: Two-dimensional marginal distributions of the posterior distribution for section 6. In black, MCMC-ABC samples. The colored ellipses represent contours for EP-ABC approximate posteriors (scaled to span 4 times the standard deviation across the diameter). Each ellipse represents the results for a different run of EP-ABC. The posterior distribution exhibits extremely poor scaling (compare the range of $\log m_1$ to that of $\log m_{21}$), and a truncated shape along some dimensions. Red triangles represent posterior mean computed from MCMC-ABC samples.

a very poor approximation of the posterior. As a starting value we used the true parameters. After several adjustments, we ran a MCMC-ABC chain for 3 million iterations, which we trimmed down to 600 samples for visualisation on Fig. 6.3 (auto-correlation in the chain being in any case extremely high).

MCMC-ABC samples are shown on Fig. 6.3, along with density contours representing EP-ABC results. Although MCMC-ABC and EP-ABC target different densities, it is reasonable to hope that in this case these densities must be fairly close, since we deliberately set the acceptance ratio to be very small, and the data are simple enough for the summary statistics to be adequate. The posterior distribution obtained using MCMC-ABC, as shown on figure 6.3 is pathological, as expected. Scaling is very poor: one parameter varies over a very small range $(-2.56, -2.44)$ while the other two vary over $[-30, 20]$ and $[0, 30]$. The posterior also has a difficult shape and appears truncated along certain dimensions. EP-ABC provides approximations that are located squarely within the posterior distribution, but seems to underestimate posterior variance for the third parameter, $\log c_1$. Given the difficult context, EP-ABC still performs arguably rather well. At the very least it may be used to find an appropriate proposal distribution for MCMC-ABC, which may perform extremely poorly without such help.

7 Extensions

Our description of EP-ABC assumes the prior is Gaussian, and produces a Gaussian approximation. With respect to the latter point, we have already mentioned (see also Appendix A) that EP, and therefore EP-ABC, may propagate more generally approximations from within a given exponential family; say the parametric family of Wishart distributions if the parameter is a covariance matrix. Regarding the prior, even when considering Gaussian approximation, it is possible to accommodate a non Gaussian prior, by treating the prior as an additional site.

Thus, the main constraint regarding the application of EP-ABC is that the likelihood is factorizable, i.e. $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|y_{1:i-1}, \boldsymbol{\theta})$ in such a way that simulating from each factor is feasible.

In this paper, we focused our attention on important applications of likelihood-free inference where the likelihood is trivial to factorise; either because the datapoints are independent, or Markov. But ABC-EP is not limited to these two situations. First, quite a few time series models may be simulated sequentially, even if they are not Markov. For instance, one may apply straightforwardly EP-ABC to a GARCH-stable model (e.g. Liu and Brorsen, 1995; Mittnik et al., 2000; Menn and Rachev, 2005), which is a GARCH model (Bollerslev, 1986) with an alpha stable-distributed innovation. Second, one may obtain a simple factorisation of the likelihood by incorporating latent variables into the vector $\boldsymbol{\theta}$ of unknown parameters. Third, one may replace the true likelihood by a certain approximation which is easy to factorise. The following section discusses and illustrates such an approach, based on the concept of composite likelihood.

7.1 Composite likelihood

Composite likelihood is an umbrella term for a family of techniques based on the idea of replacing an intractable likelihood by a factorisable pseudo-likelihood; see the excellent review of Varin et al. (2011). For the sake of simplicity, we focus on marginal composite likelihood, but we note that other versions of composite likelihood (where for instance the factors are conditional distributions) could also be treated similarly.

In our EP-ABC context, replacing the likelihood by a marginal composite likelihood leads to the following type of ABC posterior

$$p_{\varepsilon}^{CL}(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta}) \prod_{s=1}^{n_s} \int p(\boldsymbol{y}_s|\boldsymbol{\theta})^{\eta_s} \mathbb{1}_{\{\|\boldsymbol{y}_s - \boldsymbol{y}_s^\star\| \leq \varepsilon\}} \, d\boldsymbol{y}_s$$

where $p(\boldsymbol{y}_s|\boldsymbol{\theta})$ is the marginal distribution of some subset $\boldsymbol{y}_s$ of the observations, and η_s is a non-negative weight. There exists some theory on how to choose the weights η_s so that the composite likelihood is close enough to the true likelihood in some sense, see again Varin et al. (2011), but for simplicity we shall take $\eta_s = 1$. Clearly, EP-ABC may be applied straightforwardly to this new ABC posterior, provided one may sample independently from the n_s marginal distributions $p(\boldsymbol{y}_s|\boldsymbol{\theta})$. In fact, the n_s factors may be treated as IID factors, which makes it possible to use the recycling strategy described in Section 4.

To make this idea more concrete, we consider the class of hidden Markov models, where the datapoints y_i are conditionally independent, $y_i|x_i, \boldsymbol{\theta} \sim g_{\boldsymbol{\theta}}(y_i|x_i)$, conditional on latent variable x_i, and the x_i's are Markov: $x_1 \sim \mu_{\boldsymbol{\theta}}(x_1)$, $x_{i+1}|x_{1:i}, \boldsymbol{\theta} \sim f_{\boldsymbol{\theta}}(x_{i+1}|x_i)$ for $i \geq 1$; see Andrieu et al. (2005) for a previous application of composite likelihood to hidden Markov models. We assume assume that the density $g_{\boldsymbol{\theta}}(y_i|x_i)$ is intractable, and that one may sample from it; see e.g. Dean et al. (2011), Calvet and Czellar (2011) for applications of likelihood free inference to this class of intractable hidden Markov models. We assume that $\mu_{\boldsymbol{\theta}}$ is the stationary distribution of the Markov process; hence marginally $x_i|\boldsymbol{\theta} \sim \mu_{\boldsymbol{\theta}}(x_i)$.

Then a particular version of the composite likelihood ABC posterior may be constructed as follows: for some fixed integer $l \geq 2$, take $n_s = \lceil n/l \rceil$ and $\boldsymbol{y}_s = y_{l(s-1)+1:ls}$ if $ls \leq n$, $\boldsymbol{y}_s = y_{l(s-1)+1:n}$ otherwise. One may sample from $p(\boldsymbol{y}_s|\boldsymbol{\theta})$ as

follows: sample $x_{l(s-1)+1}|\boldsymbol{\theta} \sim \mu_\theta(x_{l(s-1)+1})$, then sample recursively $x_{i+1}|x_i, \boldsymbol{\theta} \sim f_{\boldsymbol{\theta}}(x_{i+1}|x_i)$ for $i = l(s-1)+2, \ldots, ls$, and finally sample $y_i|x_i, \boldsymbol{\theta} \sim g_{\boldsymbol{\theta}}(y_i|x_i)$ independently for $i = l(s-1)+1, \ldots, ls$.

As an illustration, we consider the following alpha-stable stochastic volatility model: $x_1 \sim N\left(\mu, \sigma^2/(1-\rho^2)\right)$, $x_{i+1} = \mu + \rho(x_i - \mu) + \sigma u_t$, with $u_t \sim N(0,1)$, and $y_i|x_i, \boldsymbol{\theta}$ is a Stable distribution with the following parameters (using the same parametrisation as in Section 5.2): $\alpha \in (1,2)$ is fixed, $\beta = 0$ (no skewness), $\delta = 0$ (centred at zero), and the scale parameter γ is set to $\exp(x_t/2)$. The parameter vector is therefore $\boldsymbol{\theta} = \left(\mu, \Phi^{-1}((\rho+1)/2), \log\sigma, \Phi^{-1}(\alpha - 1)\right)$, and we use the following prior, $\boldsymbol{\theta} \sim N\left((0, 1.65, 0, 0)^T, \mathrm{diag}\,(100, 0.25, 1, 1)\right)$; for the second and third components (corresponding to ρ and σ), we fitted a Gaussian distribution to the prior suggested by Kim et al. (1998) (after the appropriate transformation), while for the fourth component, the prior is equivalent to $\alpha \sim U[1,2]$.

We simulated a dataset of size $n = 120$ from this model, with true parameters $\mu = 0.35$, $\sigma = 0.2$, $\rho = 0.97$, $\alpha = 1.5$. Note that the high value of ρ creates strong correlations between successive blocks, so it is interesting to see if the marginal composite likelihood remains a reasonable approximation of the true likelihood in this case. We considered several values of l: $l = 2, 3, 4$. The choice of l amounts to a trade-off between the accuracy of the composite likelihood approximation (the larger l, the better), and the computational cost (the larger l, the smaller the probability of the event $\|\boldsymbol{y}_s - \boldsymbol{y}_s^\star\| \leq \varepsilon$). In practice, we observe that it is difficult to take l to be very large, because the probability that $\|\boldsymbol{y}_s - \boldsymbol{y}_s^\star\| \leq \varepsilon$ decrease exponentially in l. For each value of l, we took ϵ to be as small as possible, subject to the running time being about one minute and a half (and roughly about 10^7 draws from an alpha-stable distribution); thus, $\varepsilon = 1$, 1.5, 2.5 for $l = 2$, 3, 4. Fig. 7.1 plots the marginal distributions for each component, obtained by averaging out 10 runs of ABC-EP and applying the appropriate transformation. These marginal densities are to be compared with the output of a PMCMC algorithm that targets the ABC posterior that correspond to the n constraints $|y_i - y_i^\star| \leq \varepsilon$, with $\varepsilon = 1$. (This PMCMC algorithm is a random walk Metropolis sampler in the $\boldsymbol{\theta}$-dimension, which runs the ABC filtering algorithm of Jasra et al. (2010) at each iteration.) The running time of that PMCMC algorithm was three days and six hours (10^5 iterations, $N = 5 \times 10^3$ particles, leading to 6×10^{10} draws from an alpha-stable distribution). Relative to previous examples, the approximation error brought by EP-ABC-CL is more noticeable in this case (especially for ρ), but remains quite reasonable. Presumably the main source of error is the composite likelihood approximation. Note also that the PMCMC output should not considered as a gold standard in this case, as it corresponds to an ABC approximation based on a different set of constraints.

The complexity of EP-ABC-CL in this case is $O(n)$, while the complexity of PMCMC is $O(n^2)$ (Andrieu et al., 2010). Thus it is easy to apply EP-ABC-CL to datasets of size $n = 10^3$, or even 10^4, while this would prove very expensive for PMCMC.

Apart from hidden Markov models, there are several other classes of models that could be treated using the composite likelihood version of EP-ABC. For instance, it is common to use marginal composite likelihood to deal with spatial extremes (see e.g. Davison et al., 2012); however only bivariate marginal distributions are tractable in such a case, whereas with EP-ABC, one could deal with composite likelihood made of larger-order marginals.

8 Conclusion

EP-ABC has several limitations. It requires the likelihood of the model to be factorisable. It produces a parametric approximation of the posterior, which may be a poor approximation in certain cases; e.g. a Gaussian approximation while the posterior is severely multimodal (although one may wonder which ABC method would work under such a scenario.) And the mathematical properties of EP-ABC (i.e. convergence, assessment of the approximation error) are not yet fully understood. Work on EP has started recently (Titterington, 2011), but these preliminary results do not address the issue of the stability of the algorithm when each site update introduces a stochastic error, as in EP-ABC. This is certainly an important (but possibly quite difficult) direction for future research.

As of now, we would like to make two pragmatic remarks. First, EP-ABC is very fast, compared to other ABC methods. We have observed empirically that it produces very accurate results, but the user is free to either use these results directly, or a first step in order to calibrate a second, more expensive step based on a standard ABC approximation. This type of calibration is often critical to obtain decent convergence in standard ABC. Second, EP-ABC greatly reduces the pain of designing summary statistics (as only site-specific summary statistics s_i must be chosen in EP-ABC), and in certain cases make it possible to do away completely with summary statistics (i.e. take $s_i(y_i) = y_i$ at each site). This seems quite convenient as, in real-world applications, one has little intuition and even less mathematical guidance on to why $p(\boldsymbol{\theta}|\boldsymbol{s}(\boldsymbol{y}))$ should be close to $p(\boldsymbol{\theta}|\boldsymbol{y})$ for a given set of summary statistics $\boldsymbol{s}$. One may even argue that the dependence on summary statistics is currently the main limitation of the ABC approach, and that it is essential that this issue is addressed in future research in likelihood-free inference, whether through EP or by other means.

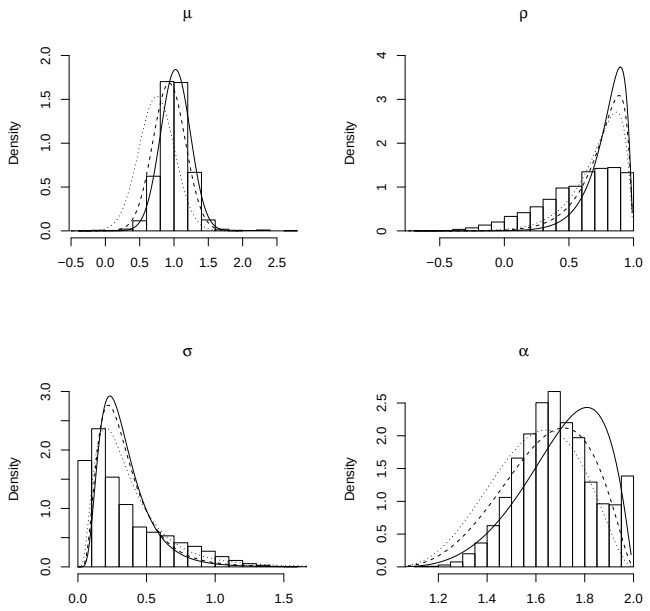


Fig. 7.1: Comparison of EP-ABC approximations based on the composite likelihood for block sizes $L = 2$ ($\varepsilon = 1$, solid line), $L = 3$ ($\varepsilon = 1.5$, dashed line), $L = 4$ ($\varepsilon = 2.5$, dotted line) and output (histograms) from a PMCMC sampler targeting an ABC posterior corresponding to n constraints $|y_i - y_i^\star| < \varepsilon$, and $\varepsilon = 1$.

Algorithm 3 Generic EP for exponential families.

Input: a target density $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i=1}^{n} l_i(\boldsymbol{\theta})$.
Initialise $\boldsymbol{\lambda}_0$ to the exponential parameters of the prior p_0, and local site parameters $\boldsymbol{\lambda}_1 \dots \boldsymbol{\lambda}_n = 0$. Set global approximation parameter $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda} = \sum_{i=0}^{n} \boldsymbol{\lambda}_i = \boldsymbol{\lambda}_0$. Loop over sites $i = 1, \dots, n$ until convergence:

1. Create hybrid distribution $h(\boldsymbol{\theta}) \propto q_{-i}(\boldsymbol{\theta})\, l_i(\boldsymbol{\theta})$ by setting $q_{-i}(\boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\lambda}_{-i}^t \boldsymbol{t}(\boldsymbol{\theta})\right)$ with $\boldsymbol{\lambda}_{-i} = \boldsymbol{\lambda} - \boldsymbol{\lambda}_i$.
2. Compute moments $\boldsymbol{\eta}_h$ of hybrid distribution, transform to natural parameters $\boldsymbol{\lambda}_h = \boldsymbol{\lambda}(\boldsymbol{\eta}_h)$.
3. Update site i by setting $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{-i}$, then reset global parameter $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda} = \boldsymbol{\lambda}_h$.

Return moment parameters $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\lambda})$.

Acknowledgements

The authors thank the associate editor, the referees, Pierre Jacob, Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, Sumeetpal S. Singh, Scott Sisson and Darren Wilkinson for helpful comments, and M. Maertens for providing the data used in the third example. The first author acknowledges support from the BMBF (Foerderkennzeichen 01GQ1001B). The second author acknowledges support from the "BigMC" ANR grant ANR-008-BLAN-0218 of the French Ministry of Research.

Appendix A: Algorithmic description of EP

The more general EP algorithm for a generic exponential family is described as Algorithm 3. The sites are therefore of the form $f_i(\boldsymbol{\theta}) = \exp\{\boldsymbol{\lambda}_i^t \boldsymbol{t}(\boldsymbol{\theta}) - \phi(\boldsymbol{\lambda}_i)\}$, so that the global approximation pertains to the same family, i.e. $q(\boldsymbol{\theta}) = \exp\{\boldsymbol{\lambda}^t \boldsymbol{t}(\boldsymbol{\theta}) - \phi(\boldsymbol{\lambda})\}$, $\boldsymbol{\lambda} = \sum_{i=0}^{n} \boldsymbol{\lambda}_i$. For a given exponential family, there exists a smooth invertible mapping $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\eta})$ that transforms the moment parameters $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\lambda}}\{\boldsymbol{t}(\boldsymbol{\theta})\} = \int \boldsymbol{t}(\boldsymbol{\theta}) \exp\{\boldsymbol{\lambda}_i^t \boldsymbol{t}(\boldsymbol{\theta}) - \phi(\boldsymbol{\lambda}_i)\}\, d\boldsymbol{\theta}$, into the natural parameters $\boldsymbol{\eta}$ (see e.g. Schervish, 1995, p. 105). In the particular case where the exponential family is the family of Gaussian distributions of dimension d, as described in the paper, one simply takes $\boldsymbol{\lambda} = (\boldsymbol{r}, \boldsymbol{Q})$, and $\boldsymbol{\eta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In particular, Step 2 of Algorithm 3 corresponds exactly to computing the moments in (2.5).

Appendix B: marginal EP updates

In example 2 we make use of the fact that some sites only depend on a subset of the parameters to obtain more stable updates. We list below some results for multivariate Gaussian families that are essential in deriving these special EP updates.

We generalise the problem slightly to computing the moments of a hybrid $h(\boldsymbol{\theta}) \propto f(\boldsymbol{A\theta})\mathcal{N}(\boldsymbol{\theta};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$, the product of a multivariate Gaussian density and a likelihood which is a function of $\boldsymbol{A\theta}$, where $\boldsymbol{A}$ is a matrix of dimension $k \times m$, $m < k$. When $\boldsymbol{A}$ is a sub-matrix of the identity matrix we have the special case of a likelihood which only depends on a subset of the parameters.

For the normalisation constant Z, we have that:

$$Z = \int f(\boldsymbol{A\theta})\mathcal{N}(\boldsymbol{\theta};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\,\mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{z})\mathcal{N}\left(\boldsymbol{z},\boldsymbol{A\mu}_0,\boldsymbol{A\Sigma}_0\boldsymbol{A^t}\right)\mathrm{d}\boldsymbol{z}$$

where $\boldsymbol{z} = \boldsymbol{A\theta}$. For the expectation of the hybrid:

$$\frac{1}{Z}\int \boldsymbol{\theta} f(\boldsymbol{A\theta})\mathcal{N}(\boldsymbol{\theta};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\,\mathrm{d}\boldsymbol{\theta} = \frac{1}{Z}\int f(\boldsymbol{z})\mathcal{N}\left(\boldsymbol{z},\boldsymbol{A\mu}_0,\boldsymbol{A\Sigma}_0\boldsymbol{A^t}\right)E(\boldsymbol{\theta}|\boldsymbol{z})\mathrm{d}\boldsymbol{z}$$

with $E(\boldsymbol{\theta}|\boldsymbol{z}) = \boldsymbol{V}\boldsymbol{z} + \boldsymbol{b}$, $\boldsymbol{V} = \boldsymbol{\Sigma}_0\boldsymbol{A^t}\left(\boldsymbol{A\Sigma}_0\boldsymbol{A^t}\right)^{-1}$ and $\boldsymbol{b} = \boldsymbol{\mu}_0 - \boldsymbol{\Sigma}_0\boldsymbol{A^t}\left(\boldsymbol{A\Sigma}_0\boldsymbol{A^t}\right)^{-1}\boldsymbol{A\mu}_0$, thus

$$\mathbb{E}_h(\boldsymbol{\theta}) = \boldsymbol{V}\mathbb{E}_h(\boldsymbol{z}) + \boldsymbol{b}$$

where $\mathbb{E}_h(\boldsymbol{z})$ is the expectation of the hybrid. A similar calculation yields an expression for the covariance:

$$Cov_h(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\boldsymbol{A^t}\left(\boldsymbol{A\Sigma}_0\boldsymbol{A^t}\right)^{-1}\boldsymbol{A\Sigma}_0 + \boldsymbol{V}Cov_h(\boldsymbol{z})\boldsymbol{V^t} = (\boldsymbol{I} - \boldsymbol{VA})\boldsymbol{\Sigma}_0 + \boldsymbol{V}Cov_h(\boldsymbol{z})\boldsymbol{V^t} \tag{8.1}$$

These three results yield computational savings and increased stability, because the moments of the hybrid distribution over $\boldsymbol{\theta}$ can be obtained from the moments of the marginal hybrid over $\boldsymbol{z}$, which has lower dimensionality.

Appendix C: Paquet-Winter-Opper corrections

In Paquet et al. (2009) a method for correcting an EP approximation is presented (denoted henceforth the PWO correction). We give below a simpler way of deriving the corrections, and show how to apply them in an ABC context.

Using the factorisations given in (2.1) and (2.2), the PWO correction may derived as

$$\pi(\boldsymbol{\theta}) = q(\boldsymbol{\theta})\prod_{i=1}^{n}\frac{l_i(\boldsymbol{\theta})}{f_i(\boldsymbol{\theta})} \triangleq q(\boldsymbol{\theta})\prod_{i=1}^{n}\{1 + e_i(\boldsymbol{\theta})\} \tag{8.2}$$

where the e_i's are error terms. This product can be expanded in increasing orders of e_i:

$$\pi(\boldsymbol{\theta}) = q(\boldsymbol{\theta})\left\{1 + \sum_{i=1}^{n} e_i(\boldsymbol{\theta}) + \sum_{j<k} e_j(\boldsymbol{\theta})\,e_k(\boldsymbol{\theta}) + \ldots\right\}$$

The PWO corrections are obtained by truncating to a given order. Note that truncation might result in a function that may not be everywhere positive, although in practice the problem did not arise either in the original applications (Paquet et al., 2009), or in ours.

The first-order (un-normalised) correction has a particularly simple form:

$$\begin{aligned} q_c^1(\boldsymbol{\theta}) &\propto q(\boldsymbol{\theta})\left\{1 + \sum_{i=1}^{n} e_i(\boldsymbol{\theta})\right\} \\ &= q(\boldsymbol{\theta}) + \sum_{i=1}^{n} q(\boldsymbol{\theta})\left(\frac{l_i(\boldsymbol{\theta})}{f_i(\boldsymbol{\theta})} - 1\right) \\ &= (n-1)q(\boldsymbol{\theta}) + \sum_{i=1}^{n} q_{-i}(\boldsymbol{\theta})l_i(\boldsymbol{\theta}) \end{aligned}$$

where in the second part of the expression we recognise the *hybrid* distributions $h_i \propto q_{-i}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta})$. The integration constant of $q_c^1()$ can easily be obtained as a by-product of quantities computed during the EP run:

$$\int q_c^1(\boldsymbol{\theta})\, d\boldsymbol{\theta} = (n-1)\, Z_q + \sum_{i=1}^{n} Z_i$$

where the Z_i are the integration constants of the hybrids.

As Paquet et al. (2009) note, the mean and covariance of the first-order approximation q_1 are the same as that of q, provided that EP has converged, which by definition happens when the hybrids of the n sites have the same expectation and covariance matrix, and are equal to respectively the expectation and covariance matrix of the global approximation q.

We can still use the first-order correction for other expectations $\mathbb{E}_{q_c^1}[f(\boldsymbol{\theta})]$: for example, using $f_\alpha(\boldsymbol{\theta}) = \mathbb{I}(\theta_j < \alpha)$ for different values of α leads to an improved estimate of the posterior marginal of θ_j. (This is the strategy we use to obtain the right panel in Figure (6.1).) The first-order correction comes nearly for free in ABC-EP, but higher-orders seem much more expensive to obtain in an ABC context, and are not discussed here.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 72(3):269–342.

Andrieu, C., Doucet, A., and Tadic, V. (2005). On-line parameter estimation in general state-space models. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*, pages 332–337. Ieee.

Bazin, E., Dawson, K., and Beaumont, M. (2010). Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. *Genetics*, 185(2):587–602.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer New York.

Blum, M. G. B. (2010). Approximate Bayesian Computation: A Nonparametric Perspective. *J. Am. Statist. Assoc.*, 105(491):1178–1187.

Bogacz, R., Usher, M., Zhang, J., and McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1655–1670.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327.

Boys, R., Wilkinson, D., and Kirkwood, T. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statist. Comput.*, 18(2):125–135.

Brascamp, J. W., van Ee, R., Noest, A. J., Jacobs, R. H., and van den Berg, A. V. (2006). The time course of binocular rivalry reveals a fundamental role of noise. *Journal of vision*, 6(11):1244–1256.

Calvet, L. and Czellar, V. (2011). State-observation sampling and the econometrics of learning models. *Arxiv preprint arXiv:1105.4519*.

Chambers, J., Mallows, C., and Stuck, B. (1976). A method for simulating stable random variables. *J. Am. Statist. Assoc.*, 71:340–344.

Davison, A., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statist. Science*, 27:161–186.

Dean, T., Singh, S., Jasra, A., and Peters, G. (2011). Parameter estimation for hidden markov models with intractable likelihoods. *Arxiv preprint arXiv:1103.5399*.

Diggle, P. and Gratton, R. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *J. R. Statist. Soc. B*, (74):1–28.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc. B*, 56(3):501–514.

Gentle, J. (2003). *Random number generation and Monte Carlo methods*. Springer-Verlag.

Gerstner, W. and Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 1 edition.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Holenstein, R. (2009). *Particle Markov Chain Monte Carlo*. PhD thesis, University of British Columbia.

Jasra, A., Singh, S., Martin, J., and McCoy, E. (2010). Filtering via approximate bayesian computation. *Statist. Comput.*, pages 1–15.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3):361–393.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputation and Bfayesian missing data problems. *J. Am. Statist. Assoc.*, 89:278–288.

Liu, S. and Brorsen, B. (1995). Maximum likelihood estimation of a GARCH-stable model. *J. Econometrics*, 10(3):273–285.

Luce, D. R. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization (Oxford Psychology Series)*. Oxford University Press, USA.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov Chain Monte Carlo without Likelihoods. 100(26):15324–15328.

Menn, C. and Rachev, S. (2005). A GARCH option pricing model with ialpha-stable innovations. *European journal of operational research*, 163(1):201–209.

Meyer, D. E., Osman, A. M., Irwin, D. E., and Yantis, S. (1988). Modern mental chronometry. *Biological psychology*, 26(1-3):3–67.

Minka, T. (2001a). Expectation Propagation for approximate Bayesian inference. *Proceedings of Uncertainty in Artificial Intelligence*, 17:362–369.

Minka, T. (2001b). *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology.

Minka, T. (2004). Power EP. Technical report, Dep. Statistics, Carnegie Mellon University, Pittsburgh, PA.

Minka, T. (2005). Divergence Measures and Message Passing. Technical report.

Mittnik, S., Paolella, M., and Rachev, S. (2000). Diagnosing and treating the fat tails in financial returns data. *Journal of Empirical Finance*, 7(3-4):389–416.

Nolan, J. P. (2012). *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

Nuthmann, A., Smith, T. J., Engbert, R., and Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, 117(2):382–405.

Paquet, U., Winther, O., and Opper, M. (2009). Perturbation Corrections in Approximate Inference: Mixture Modelling Applications. *J. Machine Learning Research*, 10:1263–1304.

Peters, G., Sisson, S., and Fan, Y. (2010). Likelihood-free Bayesian inference for alpha-stable models. *Comput. Stat. Data Anal.*, (in press).

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, (85):59–108.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *J. Am. Statist. Assoc.*, pages 1151–1172.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B*, 71(2):319–392.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag.

Seeger, M. (2005). Expectation Propagation for Exponential Families. Technical report, Univ. California Berkeley.

Tavaré, S., Balding, D., Griffiths, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505.

Titterington, M. (2011). The EM algorithm, variational approximations, and expectation propagation for mixtures. In Mengersen, K., Robert, C., and Titterington, M., editors, *Mixtures: Estimation and Applications (Chap. 1)*, volume 896. Wiley.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Wang, B. and Titterington, D. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 373–380.

Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception and Psychophysics*, 63(8):1293–1313.

Wilkinson, R. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Arxiv preprint arXiv:0811.3355*.