Choosing the Summary Statistics and the Acceptance Rate in Approximate Bayesian Computation

COMPSTAT 2010

Revised version; August 13, 2010

Michael G.B. Blum[1]

Laboratoire TIMC-IMAG, CNRS, UJF Grenoble
Faculté de Mdecine, 38706 La Tronche, France, *michael.blum@imag.fr*

Abstract. Approximate Bayesian Computation encompasses a family of likelihood-free algorithms for performing Bayesian inference in models defined in terms of a generating mechanism. The different algorithms rely on simulations of some summary statistics under the generative model and a rejection criterion that determines if a simulation is rejected or not. In this paper, I incorporate Approximate Bayesian Computation into a local Bayesian regression framework. Using an empirical Bayes approach, we provide a simple criterion for 1) choosing the threshold above which a simulation should be rejected, 2) choosing the subset of informative summary statistics, and 3) choosing if a summary statistic should be log-transformed or not.

Keywords: Approximate Bayesian Computation, evidence approximation, empirical Bayes, Bayesian local regression

1 Introduction

Approximate Bayesian Computation (ABC) encompasses a family of likelihood-free algorithms for performing Bayesian inference (Beaumont et al. (2002)). It originated in population genetics for making inference in coalescent models (Pritchard et al. (1999)). Compared to MCMC algorithms that aim at providing a sample from the *full* posterior distribution $p(\phi|\mathcal{D})$, where ϕ denotes a possibly multi-dimensional parameter and $\mathcal{D}$ denotes the data, ABC targets a *partial* posterior distribution $p(\phi|S)$ where S denotes a p-dimensional summary statistic $S = (S^1, \ldots, S^p)$ typically of lower dimension than the data $\mathcal{D}$. Despite of this approximation inherent to ABC, its ease of implementation have fostered ABC applications in population genetics and evolutionary biology.

1.1 Rejection algorithm

To generate a sample from $p(\phi|S)$, the original ABC rejection algorithm is indeed remarkably simple (Pritchard et al. (1999)):

1. Generate a parameter ϕ according to the prior distribution π;
2. Simulate data $\mathcal{D}'$ according to the model $p(\mathcal{D}'|\phi)$;
3. Compute the summary statistic S' from $\mathcal{D}'$ and accept the simulation if $d(S, S') < \delta$ where d is a distance between the two summary statistics and $\delta > 0$ is a threshold parameter.

It is the user's task to choose a threshold δ. Rather than choosing explicitly a threshold value δ, Beaumont et al. (2002) rather set the percentage of accepted simulations, the acceptance rate p_δ, to a given value. For a total of n simulations, it amounts to setting δ to the p_δ-percent quantile of the distances $d(S_i, S)$, $i = 1 \dots n$. In the following, we choose $d(S, S') = ||S - S'||$ where $|| \cdot - \cdot ||$ denotes the Euclidean distance, and we consider that each summary statistic has been rescaled by a robust estimate of its dispersion (the median absolute deviation).

1.2 Regression adjustment

To weaken the effect of the discrepancy between the observed summary statistic and the accepted ones, Beaumont et al. (2002) proposed two innovations: weighting and regression adjustment. The weighting is a generalization of the acceptance-rejection algorithm in which each simulation is assigned a weight $W_i = K_\delta(||S - S_i||) \propto K(||S - S_i||/\delta)$ where K is a smoothing kernel. Beaumont et al. (2002) considered an Epanechnikov kernel so that simulations with $||S - S'|| > \delta$ are discarded as in the rejection algorithm.

The regression adjustment step involves a local-linear regression in which the least-squares criterion

$$\sum_{i=1}^{n} \{\phi_i - (\beta_0 + (S_i - S)^T \beta_1)\}^2 W_i, \quad \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^p, \tag{1}$$

is minimized. The least-squares estimate is given by

$$\hat{\beta}_{\text{LS}} = (\hat{\beta}^0_{\text{LS}}, \hat{\beta}^1_{\text{LS}}) = (X^T W_\delta X)^{-1} X^T W_\delta \phi, \tag{2}$$

where W_δ is a diagonal matrix in which the i^{th} element is W_i, and

$$X = \begin{pmatrix} 1 & s^1_1 - s^1 & \cdots & s^p_1 - s^p \\ \vdots & \dots & \ddots & \vdots \\ 1 & s^1_n - s^1 & \cdots & s^p_n - s^p \end{pmatrix}, \ \phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}. \tag{3}$$

To form an approximate sample from $p(\phi|S)$, Beaumont et al. (2002) computed $\phi^*_i = \hat{\beta}^0_{\text{LS}} + \epsilon_i$, where the ϵ_i's denote the empirical residuals of the regression. This translates into the following equation for the regression adjustment

$$\phi^*_i = \phi_i - (S_i - S)^T \hat{\beta}^1_{\text{LS}}. \tag{4}$$

To give an intuition about the benefit arising from the regression adjustment, look at the first and second weighted moments of the ϕ_i^*. The first moment of the ϕ_i^* is equal to the local linear estimate $\hat{\beta}_0$ and therefore provides an estimate of the posterior mean. Compared to the weighted mean of the ϕ_i's obtained with the rejection algorithm (the Nadaraya-Watson estimate in the statistics literature), $\hat{\beta}_0$ is *design adaptive*, i.e. its bias does not depend on the design $p(S)$ (Fan 1992). The second moment of the ϕ_i^* is equal to the second moment of the empirical residuals ϵ_i which is inferior to the total variance of the ϕ_i's. A shrinkage towards $\hat{\beta}_0$ is therefore involved by regression adjustment.

1.3 Potential pitfalls of ABC

As shown by various authors (Joyce and Marjoram (2008), Blum (2010)), the performances of ABC might be hampered by the *curse of dimensionality* when too many summary statistics are included in the analysis. To illustrate the dramatic effect of the curse of dimensionality, we introduce a simple Gaussian example. Assume that we observe a sample of size $N = 50$ in which each individual is a Gaussian random variable of mean μ and variance σ^2. We are interested here in the estimation of the variance parameter σ^2. We assume the following hierarchical prior for μ and σ^2 (Gelman et al. (2004))

$$\sigma^2 \sim \mathrm{Inv}\chi^2(\mathrm{d.f.} = 1) \tag{5}$$

$$\mu \sim \mathcal{N}(0, \sigma^2), \tag{6}$$

where $\mathrm{Inv}\chi^2(\mathrm{d.f.} = \nu)$ denotes the inverse chi-square distribution with ν degrees of freedom, and $\mathcal{N}$ denotes the Gaussian distribution. We consider the following summary statistics

$$(S^1, \ldots, S^5) = (\bar{x}_N, s_N^2, u_1, u_2, u_3), \tag{7}$$

where $\bar{x}_N$ and s_N^2 denotes the empirical mean and variance of the sample, and the u_j, $j = 1, 2, 3$, are three Gaussian summary statistics with mean 0 and variance 1. Of course, the last three summary statistics do not convey any information for estimating μ and σ^2 and are added here for enhancing the curse of dimensionality. As displayed in Figure 1, for an observed empirical variance $s_N^2 = 1.144$ (obtained from the petal lengths of the virginica species in Fisher's iris data), the rejection algorithm ($n = 300$, $p_\delta = 10\%$) retains values of σ^2 as aberrant as 100 and regression adjustment will fail since the accepted points are to widespread for the local linear approximation to hold. As suggested by this example, a methods for selecting the relevant summary statistics is needed.

1.4 Outline of the paper

In this paper, I will provide a criterion for 1) choosing a set of informative summary statistics among the p summary statistics $(S^1, \ldots, S^p)$, 2) choos-

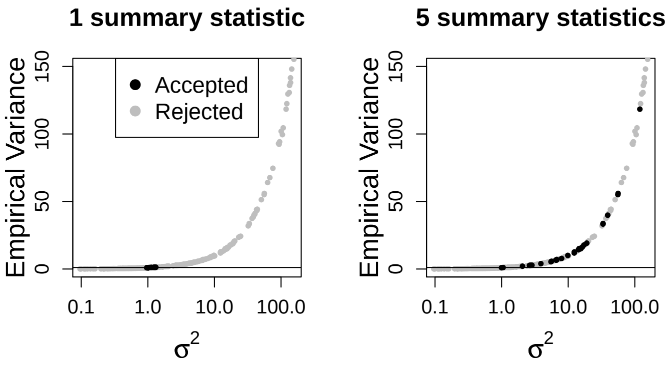

Fig. 1. Rejection algorithm for estimating σ^2 in a Gaussian model. In the left panel, the empirical variance is the single summary statistic in the rejection algorithm whereas in the right panel, we considered the five summary statistics given in equation (7). The horizontal line represents the observed empirical variance $s_N^2 = 1.144$.

ing an acceptance rate p_δ, and 3) choosing if a summary statistic should be transformed or not. Here I will consider only log transformation but square root or inverse transformations could also be considered. The first section presents how to compute the $(p+1)$-dimensional parameter β of the local linear regression in a Bayesian fashion. In the context of Bayesian local regression, we define the evidence function that will provide us a rationale criterion for addressing questions 1-3. The second section presents two examples in which we show that the evidence function provides reasonable choices for p_δ, for the selection of the summary statistics, and for the choice of the scale (logarithmic or not) of the summary statistics.

2 Regression adjustment in a Bayesian fashion

2.1 Local Bayesian regression

Carrying out locally-linear regression in a Bayesian fashion has been studied by Hjort (2003). The linear regression model can be written as $\phi_i = \beta^0 + (S_i - S)^T \beta^1 + \epsilon$. The points (S_i, ϕ_i) are weighted by the $W_i = K_\delta(||S_i - S||)/K_\delta(0)$. By contrast to the least-squares estimate, Bayesian local regression is not invariant to rescaling of the W_i's. Here, a weight of 1 is given to a simulation for which S_i matches exactly S and the weights decrease from 1 to 0 as the $||S_i - S||$'s move from 0 to δ.

Here we assume a zero-mean isotropic Gaussian prior such that $\beta = (\beta^0, \beta^1) \sim \mathcal{N}(0, \alpha^{-1} I_{p+1})$, where α is the precision parameter, and I_d is the identity matrix of dimension d. The distribution of the residuals is assumed to be a zero mean Gaussian distribution with variance parameter τ^2.

With standard algebra, we find the posterior distribution of the regression coefficients β (Bishop (2006))

$$\beta \sim \mathcal{N}(\beta_{\mathrm{MAP}}, V), \tag{8}$$

where

$$\beta_{\mathrm{MAP}} = \tau^{-2} V X^T W_\delta \phi \tag{9}$$
$$V^{-1} = (\alpha I_{p+1} + \tau^{-2} X^T W_\delta X). \tag{10}$$

Bayesian regression adjustment in ABC can be performed with the linear adjustment of equation (4) by replacing β^1_{LS} with β^1_{MAP}. By definition of the posterior distribution, we find that β_{MAP} minimizes the regularized least-squares problem considered in ridge regression (Hoerl and Kennard (1970))

$$E(\beta) = \frac{1}{2\tau^2} \sum_{i=1}^{n} (\phi_i - (S_i - S)^T \beta)^2 W_i + \frac{\alpha}{2} \beta^T \beta. \tag{11}$$

As seen from equation (11), Bayesian linear regression shrinks the regression coefficients towards 0 by imposing a penalty on their sizes. The appropriate value for τ^2, α, and p_δ, required for the computation of β_{MAP}, will be determined through the evidence approximation discussed below.

2.2 The evidence approximation

A complete Bayesian treatment of the regression would require to integrate the hyperparameters over some hyperpriors. Here we adopt a different approach in which we determine the value of the hyperparameters, by maximizing the *marginal likelihood*. The marginal likelihood $p(\phi|\tau^2, \alpha, p_\delta)$, called the evidence function in the machine learning literature (MacKay (1992), Bishop (2006)), is obtained by integrating the likelihood over the the regression parameters β

$$p(\phi|\tau^2, \alpha, p_\delta) = \int \left(\mathit{\Pi}_{i=1}^n p(\phi_i|\beta, \tau^2)^{W_i} \right) p(\beta|\alpha)\, d\beta. \tag{12}$$

Finding the value of the hyperparameters by maximizing the evidence is known as *empirical Bayes* in the statistics literature (Gelman et al. (2004)). Here, we do not give the details of the computation of the evidence and refer the reader to Bishop (2006). The log of the evidence is given by

$$\log p(\phi|\tau^2, \alpha, p_\delta) = \frac{p+1}{2} \log \alpha - \frac{N_W}{2} \log \tau^2 - E(\beta_{\mathrm{MAP}}) - \frac{1}{2} \log |V^{-1}| - \frac{N_W}{2} \log 2\pi, \tag{13}$$

where $N_W = \sum W_i$. By maximizing the log of the evidence with respect to α, we find that

$$\alpha = \frac{\gamma}{\beta^T_{\mathrm{MAP}} \beta_{\mathrm{MAP}}}, \tag{14}$$

where γ is the effective number of parameters (of summary statistics here)

$$\gamma = (p+1) - \alpha \mathrm{Tr}(V). \tag{15}$$

Similarly, setting $\delta \log p(\phi|\tau^2, \alpha, p_\delta)/\delta\tau^2 = 0$ gives

$$\tau^2 = \frac{\sum_{i=1}^{n}(\phi_i - (S_i - S)^T\beta)^2 W_i}{N_W - \gamma}. \tag{16}$$

Equations (14) and (16) are implicit solutions for the hyperparameters since β_{MAP}, V, and γ depend on α and τ^2. For maximizing the log-evidence, we first update β_{MAP} and V with equations (9) and (10), then we update γ using equation (15), and finally update α and τ^2 with equations (14) and (16). This updating scheme is applied in an iterative manner and stopped when the difference between two successive iterations is small enough. Plugging the values of these estimates for α and τ^2 into equation (13), we obtain the log-evidence for the acceptance rate $\log p(\phi|p_\delta)$.

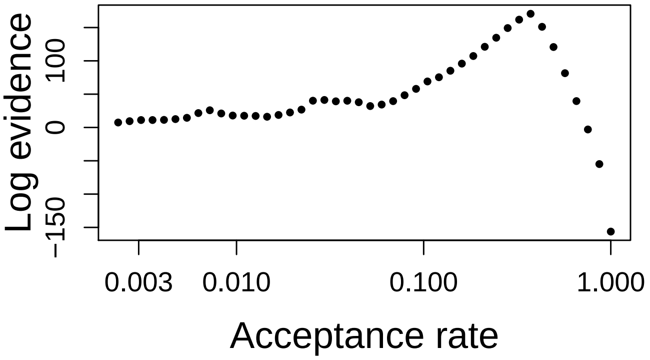

Fig. 2. Log of the evidence as a function of the acceptance rate for the generative model of equation (17). A total of 1,000 simulation is performed and the optimal acceptance rate is found for $p_\delta = 37\%$.

3 The evidence function as an omnibus criterion

3.1 Choosing the acceptance rate

To show that the evidence function provide a good choice for the tolerance rate, we introduce the following toy example. We denote ϕ, the parameter of

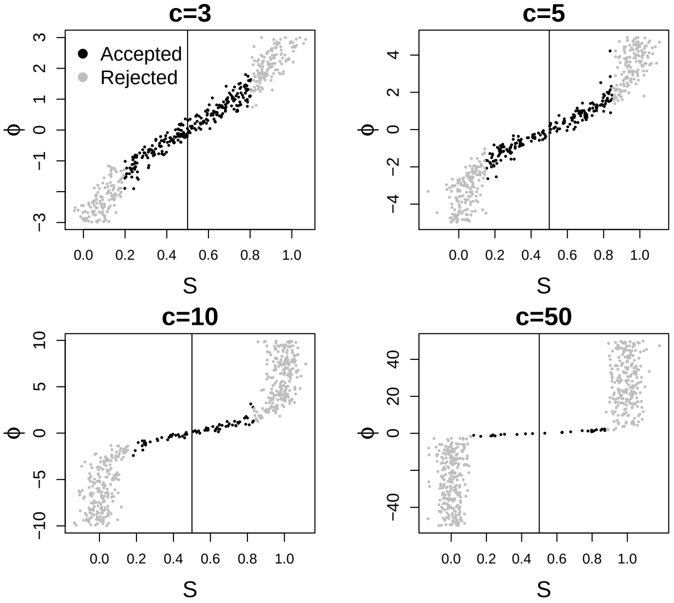

Fig. 3. Plot of the accepted points in the rejection algorithm for four different values of the parameter c. In the four plots, the acceptance rate is chosen by maximizing the evidence function $p(\phi|p_\delta)$.

interest and S the data which is equal here to the summary statistic. The generative model can be described as

$$\phi \sim \mathcal{U}_{-c,c} \quad c \in \mathbb{R},$$
$$S \sim \mathcal{N}\left(\frac{e^\phi}{1+e^\phi}, \sigma^2 = (.05)^2\right), \tag{17}$$

where $\mathcal{U}_{a,b}$ denotes the uniform distribution between a and b. We assume that the observed data is $S = 0.5$. For $c = 5$, Figure 2 displays that the evidence function has a maximum around $p_\delta = 37\%$. As seen in Figure 3, this value of p_δ corresponds to a large enough neighborhood around $S = 0.5$ in which the relationship between S and ϕ is linear. For increasing values of c in equation (17), the width of the neighborhood-around $S = 0.5$-in which the linear approximation holds, decreases. Figure 3 shows that the evidence function does a good job at selecting neighborhoods of decreasing widths in which the relationship between S and ϕ is linear.

3.2 Choosing the summary statistics

The evidence function can be used to choose a subset of predictor variables in a regression setting. For example, Bishop (2006) used the evidence to select the order of the polynomial in a polynomial regression. Here we show that the evidence function provides a criterion for choosing the set of informative summary statistics in ABC.

Plugging the optimal value for p_δ in equation (13), we obtain the evidence as a function of the set of summary statistics $p(\phi|(S^1, \ldots, S^p))$. To find an optimal subset of summary statistics, we use a standard stepwise approach. We first include the summary statistic S^{j_1} ($j_1 \in \{1, \ldots, p\}$) that gives the largest value of the evidence $p(\phi|S^{j_1})$. We then evaluate the evidence $p(\phi|(S^{j_1}, S^{j_2}))$ ($j_2 \in \{1, \ldots, p\}$) and include a second summary statistics if $\max_{j_2} p(\phi|(S^{j_1}, S^{j_2})) > p(\phi|S^{j_1})$. If a second summary statistics is not included in the optimal subset, the algorithm is stopped. Otherwise, the process is repeated until an optimal subset has been found.

To check the validity of the algorithm, we apply this stepwise procedure to the Gaussian model of Section 1.3 in which there are five different summary statistics. To estimate the posterior distribution of σ^2, we apply the linear correction adjustment of equation (4) to $\log \sigma^2$ and then use the exponential function to return to the original scale. This transformation guarantees that the corrected values will be positive. For each test replicate, we perform $n = 10,000$ simulations of the generative model of Section 1.3 and select an optimal subset of summary statistics with the stepwise procedure. Performing a total of one hundred test replicates, we find that the stepwise procedure always chooses the subset of summary statistics containing the empirical variance only. Figure 4 displays summaries of the posterior distribution obtained with ABC using five summary statistics or with the empirical variance only. As already suggested by Figure 1, the posterior distribution of σ^2 obtained with the five summary statistics is extremely different from the exact posterior distribution (a scaled inverse chi-square distribution, see Gelman et al. (2004)). By contrast, when considering only the empirical variance, we find a good agreement between the true and the estimated posterior.

3.3 Choosing the scale of the summary statistics

Here we show that changing the scale of the summary statistics can have a dramatic effect in ABC. We perform a second experiment in which we replace the empirical variance by the log of the empirical variance in the set of five summary statistics. Performing a total of one hundred test replicates, we find that the stepwise procedure always chooses the subset containing the log of the empirical variance only. However, by contrast to the previous experiment, we find that the posterior distribution of σ^2 obtained with the five summary statistics is in good agreement with the exact posterior distribution (see Figure 5). As usual for regression model, this simple experiment shows that better regression models can be obtained with a good transformation of the predictor variables.

We test here if the evidence function is able to find a good scale for the summary statistics. In one hundred test experiment, we compare $p(\log \sigma^2|s_N^2)$

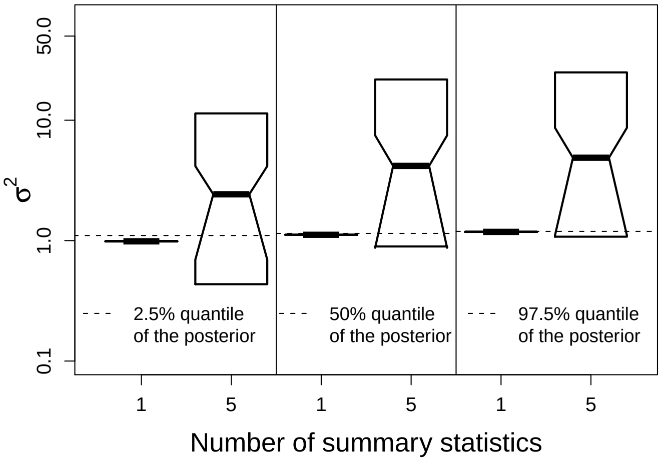

Fig. 4. Boxplots of the 2.5%, 50%, and 97.5% estimated quantiles of the posterior distribution for σ^2. ABC with one summary statistics has been performed with the empirical variance only. A total of 100 runs of ABC has been performed, each of which consisting of $n = 10,000$ simulations.

to $p(\log \sigma^2 | \log(s_N^2))$. We find that the evidence function always selects $\log(s_N^2)$ showing that a good scale for the summary statistics can be found with the evidence function.

3.4 Using the evidence without regression adjustment

If the standard rejection algorithm of Section 1.1 is considered without any regression adjustment, it is also possible to use the evidence function. The local Bayesian framework is now $\phi_i = \beta_0 + \epsilon$ in which each points (S_i, ϕ_i) is weighted by $W_i = K_\delta(||S_i - S||)/K_\delta(0)$. Assuming that the prior for β_0 is $\mathcal{N}(0, \alpha)$, we find for the evidence function

$$\log p(\phi|\tau^2, \alpha, p_\delta) = \frac{1}{2}\log\alpha - \frac{N_W}{2}\log\tau^2 - E(\beta_{0,\mathrm{MAP}}) - \frac{1}{2}\log|\alpha + \tau^{-2}N_W| - \frac{N_W}{2}\log 2\pi, \tag{18}$$

where

$$\beta_{0,\mathrm{MAP}} = \frac{\tau^{-2}}{\alpha + \tau^{-2}N_W}\sum_{i=1}^{n} W_i\phi_i \tag{19}$$

$$E(\beta_0) = \frac{1}{2\tau^2}\sum_{i=1}^{n} W_i(\phi_i - \beta_0)^2 + \frac{\alpha}{2}{\beta_0}^2. \tag{20}$$

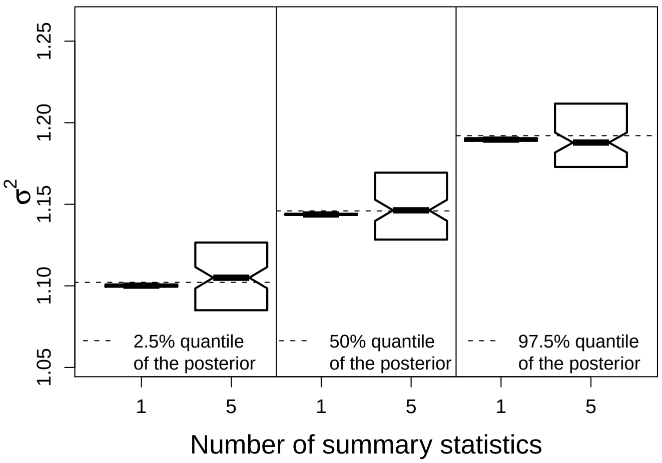

Fig. 5. Boxplots of the 2.5%, 50%, and 97.5% estimated quantiles of the posterior distribution for σ^2. In the ABC algorithms the empirical variance has been log-transformed.

References

BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002): Approximate Bayesian computation in population genetics. *Genetics 162: 2025–2035.*

BISHOP, C. M. (2006): *Pattern recognition and machine learning.* Springer

BLUM, M.G.B. (2010) Approximate Bayesian Computation: a nonparametric perspective. *Journal of the American Statistical Association, to appear.*

FAN, J. (1992): Design-adaptive nonparametric regression. *Journal of the American Statistical Association 87, 998-1004.*

GELMAN, A., CARLIN J. B., STERN H. S. and RUBIN D. B. (2004): *Bayesian Data Analysis.* Chapman & Hall/CRC.

HJORT, N. L. (2003): Topics in nonparametric Bayesian statistics (with discussion). In: P. J. Green, N. L. Hjort and S. Richardson (Eds.): *Highly Structured Stochastic Systems.* Oxford University Press, 455–487.

HOERL, A. E. and KENNARD, R. (1970): Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12, 55-67.*

JOYCE, P. and MARJORAM, P. (2008): Approximately Sufficient Statistics and Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology 7, 26.*

MACKAY, D. J. C. (1992): Bayesian interpolation. *Neural Computation 4, 415-447.*

PRITCHARD, J. K., SEIELSTAD, M. T., PEREZ-LEZAUN, A. and FELDMAN, M. W. (1999): Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16, 1791–1798.*