Approximate Bayesian Computation in Population Genetics

Mark A. Beaumont,*[,1] Wenyang Zhang[†] and David J. Balding[‡]

*School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AJ, United Kingdom,
†Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom and
‡Department of Epidemiology and Public Health, Imperial College School of Medicine,
St. Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom

Manuscript received March 22, 2002
Accepted for publication October 2, 2002

ABSTRACT

We propose a new method for approximate Bayesian statistical inference on the basis of summary statistics. The method is suited to complex problems that arise in population genetics, extending ideas developed in this setting by earlier authors. Properties of the posterior distribution of a parameter, such as its mean or density curve, are approximated without explicit likelihood calculations. This is achieved by fitting a local-linear regression of simulated parameter values on simulated summary statistics, and then substituting the observed summary statistics into the regression equation. The method combines many of the advantages of Bayesian statistical inference with the computational efficiency of methods based on summary statistics. A key advantage of the method is that the nuisance parameters are automatically integrated out in the simulation step, so that the large numbers of nuisance parameters that arise in population genetics problems can be handled without difficulty. Simulation results indicate computational and statistical efficiency that compares favorably with those of alternative methods previously proposed in the literature. We also compare the relative efficiency of inferences obtained using methods based on summary statistics with those obtained directly from the data using MCMC.

VALID and efficient statistical inferences are often difficult to achieve in population genetics problems, because data sets are large and complex, and because even the simplest models typically have many nuisance parameters, which often include the entire genealogical tree underlying the observations. Until recently, the only feasible approach to statistical inference proceeded by comparing summary statistics with their null distribution under a simplified model. This approach is statistically inefficient and inflexible, the results can be difficult to interpret, and quantitative model comparison is usually not possible.

In recent years, advances in methods of stochastic simulation have begun to permit likelihood-based statistical inference in population genetics problems. In particular, the Bayesian paradigm has many advantages in this setting (Shoemaker *et al.* 1999). In addition to conveying statistical efficiency, Bayesian methods have advantages of interpretation since they provide probability distributions for the unknown(s) of interest, either singly or jointly. Perhaps most importantly, the Bayesian approach resolves, via integration, the theoretical problems caused by the presence of many nuisance parameters. In many scenarios, however, the large number of nuisance parameters means that computational problems continue to limit the practicality of Bayesian methods.

[1]*Corresponding author:* School of Animal and Microbial Sciences, Whiteknights, PO Box 228, Reading RG6 6AJ, UK.
E-mail: m.a.beaumont@reading.ac.uk

Tavaré *et al.* (1997) pioneered a rejection-sampling method for simulating an approximate posterior random sample. Any properties of the posterior distribution, such as 95% intervals, can then be approximated by corresponding properties of the sample. Under the method, a candidate value, ϕ', for the parameter of interest, ϕ, is simulated from its prior distribution. Ideally, the next step would be to accept ϕ' with probability proportional to its likelihood, otherwise ϕ' is rejected. However, because the likelihood is difficult to compute, Tavaré *et al.* (1997) replaced the full data with a summary statistic S and accepted ϕ' with probabilty proportional to $P(S = s|\phi')$. This is implemented by accepting ϕ' if and only if

$$P(S = s|\phi') > cU, \tag{1}$$

where U is a uniform random variable, s denotes the observed value of S, and c is a constant satisfying $c \geq \max_{\phi} P(S = s|\phi)$. As is typical in population genetics, no sufficient statistic was available in the setting considered by Tavaré *et al.*, but they argued that their statistic S (the number of segregating sites) is close to sufficient for their parameter ϕ (the scaled mutation rate).

Although a useful advance, the approach of Tavaré *et al.* (1997) is limited to relatively simple settings in which $P(S = s|\phi)$ can readily be computed and maximized over ϕ. Fu and Li (1997) opened the way to greater generality of this approach by, after simulation of ϕ', replacing the computation of $P(S = s|\phi')$ with a further simulation step: They simulate a data set under

their model and accept ϕ' if the observed summary statistic s matches the simulated value s'. For Fu and Li (1997), the unknown of interest ϕ was the time since the most recent common ancestor of the sample, for which a standard coalescent prior distribution was assumed, and the statistic S was the maximum over haplotypes of the number of nucleotide differences.

This Monte Carlo likelihood approximation was extended by Weiss and von Haeseler (1998) to multiple summary statistics, some with near-continuous distributions, and multiple parameters. Instead of requiring an exact match, they accept ϕ' whenever $||\mathbf{s}' - \mathbf{s}|| \leq \delta$, for some appropriate metric $||\cdot||$ and tolerance δ, where $\mathbf{s}'$ and $\mathbf{s}$ are vectors of summary statistics calculated at, respectively, simulated and observed data sets. Rather than simulate ϕ' from a prior distribution, they employed a grid of ϕ values, which is equivalent to assuming a uniform prior.

Pritchard *et al.* (1999) adopted a rejection-sampling method similar to that of Weiss and von Haeseler (1998) but with simulation from a prior. Their investigation of mutation and demographic parameters, based on a sample of human Y chromosome data, is discussed further below. Similar methods have been adopted by Wall (2000), Tishkoff *et al.* (2001), and Estoup *et al.* (2002). These rejection-sampling methods combine the computational convenience of summary statistics with the advantages of the Bayesian paradigm. A key feature of the approach is that it can handle complex models with many nuisance parameters, provided only that simulation of data under the model is feasible. Moreover, the ratio of acceptances under two models approximates the Bayes factor, and hence quantitative model comparison is possible.

A crucial limitation of the rejection-sampling method is that only a small number of summary statistics can usually be handled. Otherwise, either acceptance rates become prohibitively low or the tolerance δ must be increased, which can distort the approximation, because the $\mathbf{s}'$ are treated equally whenever $||\mathbf{s}' - \mathbf{s}|| \leq \delta$, irrespective of the precise value of $||\mathbf{s}' - \mathbf{s}||$. Here, we introduce two improvements to existing rejection-sampling methods, smooth weighting and regression adjustment, described further below. The key benefit is insensitivity of the approximation to δ. This insensitivity permits increasing the number of summary statistics, thus potentially increasing the information extracted from the data. An additional feature of our study is that it is the first to compare the inferences obtained using summary statistics with those obtained by full-data Markov chain Monte Carlo (MCMC) methods. Given the potential that the summary-statistic methods have for substantially widening the access to and scope of Bayesian inference in population genetics, it is important to illustrate the relative efficiency of both methods.

METHODS

Rejection-based approximate Bayesian inference: Initially we assume that there is a single parameter of interest, ϕ; the case of vector-valued ϕ is discussed below. The basic rejection-sampling algorithm is: (1) choose a summary statistic $\mathbf{S}$ and calculate its value $\mathbf{s}$ for the observed data set; (2) choose a tolerance δ; (3) simulate ϕ' from the prior distribution for ϕ; (4) simulate a genealogical tree under the chosen model, such as a coalescent model (see, *e.g.*, Nordborg 2001); (5) simulate ancestral allelic types at the root of the tree, and then mutation events along the tree to generate a data set at the leaves; (6) compute $\mathbf{s}'$, the value of $\mathbf{S}$ for the simulated data set; (7) if $||\mathbf{s}' - \mathbf{s}|| \leq \delta$, then accept ϕ', otherwise reject; and (8) repeat steps 3 to 7 until k acceptances have been obtained.

Regression adjustment and weighting: Our new algorithm mimics the above up to and including step 6, except that Pritchard *et al.* (1999) used a rectangular acceptance region, whereas after appropriate scaling, for example, to equalize variances, we take $||\cdot||$ to be the Euclidean norm $||\mathbf{s}|| = \sqrt{\sum_{j=1}^{q} s_j^2}$, where $\mathbf{s} \equiv (s_1, \ldots, s_q)$, so that acceptance regions are spheres.

We propose two innovations at step 7: smooth weighting and regression adjustment. In steps 1–6 we have simulated independent pairs $(\phi_i, \mathbf{s}_i)$, $i = 1, 2, \cdots, m$, where each ϕ_i is an independent draw from the prior distribution for ϕ, and the $\mathbf{s}_i$ are simulated values of $\mathbf{S}$ with $\phi = \phi_i$. Under the Bayesian paradigm $P(\phi|\mathbf{S}) = P(\mathbf{S}|\phi)P(\phi)/P(\mathbf{S}) = P(\mathbf{S}, \phi)/P(\mathbf{S})$. The posterior distribution is a conditional density that could be estimated by first estimating the joint density $P(\mathbf{S}, \phi)$ and dividing by an estimate of the marginal density $P(\mathbf{S})$ evaluated at $\mathbf{S} = \mathbf{s}$. The $(\phi_i, \mathbf{s}_i)$ are random draws from the joint density, and the rejection method is just one of many possible methods for estimating the conditional density when $\mathbf{S} = \mathbf{s}$. It is based on the idea that the ϕ_i for which $||\mathbf{s}_i - \mathbf{s}||$ is small form an approximate posterior random sample. Our idea is to improve the approximation by (1) weighting the ϕ_i according to the value of $||\mathbf{s}_i - \mathbf{s}||$ and (2) adjusting the ϕ_i using local-linear regression to weaken the effect of the discrepancy between $\mathbf{s}_i$ and $\mathbf{s}$.

It is convenient to start with standard linear regression, but this is only to explain ideas: Our recommended method uses local-linear regression, described subsequently. Here, we assume that the conditional density that we are trying to estimate can be described by the following regression model for some intercept α and vector of regression coefficients β,

$$\phi_i = \alpha + (\mathbf{s}_i - \mathbf{s})^T\beta + \varepsilon_i, \quad i = 1, \cdots, m, \qquad (2)$$

where the ε_i are uncorrelated with mean zero and common variance. No other assumptions are made about the distribution of the ε_i and hence the ϕ_i. When $\mathbf{s}_i = \mathbf{s}$ the ϕ_i are drawn from the posterior distribution with

mean $E[\phi|\mathbf{S} = \mathbf{s}] = \alpha$. The least-squares estimate of (α, β) minimizes

$$\sum_{i=1}^{m}\{\phi_i - \alpha - (\mathbf{s}_i - \mathbf{s})^T\beta\}^2. \quad (3)$$

The solution is

$$(\hat{\alpha}, \hat{\beta}) = (X^TX)^{-1}X^T\boldsymbol{\theta},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & s_{11} - s_1 & \cdots & s_{1q} - s_q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_{m1} - s_1 & \cdots & s_{mq} - s_q \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_m \end{pmatrix}.$$

It follows from (2) that the ϕ_i^* defined by

$$\phi_i^* = \phi_i - (\mathbf{s}_i - \mathbf{s})^T\hat{\beta}$$

form an approximate random sample from $P(\phi|\mathbf{S} = \mathbf{s})$. This will be exact if the regression model is truly linear and the distributional assumptions given above for the ε_i are met and if the sample is so large that $\hat{\alpha} = \alpha$, $\hat{\beta} = \beta$. Note that $\hat{\alpha}$ is an estimate of the posterior mean $E[\phi|\mathbf{S} = \mathbf{s}]$ and hence can be interpreted as a point estimate of ϕ.

Local-linear regression: The linearity and additivity assumptions underpinning (2) will often be implausible, but may apply locally in the vicinity of $\mathbf{s}$. To implement local-linear regression, we replace the minimization (3) with

$$\sum_{i=1}^{m}\{\phi_i - \alpha - (\mathbf{s}_i - \mathbf{s})^T\beta\}^2K_\delta(\|\mathbf{s}_i - \mathbf{s}\|). \quad (4)$$

The kernel function $K_\delta(t)$ is taken here to be the Epanechnikov kernel,

$$K_\delta(t) = \begin{cases} c\delta^{-1}(1 - (t/\delta)^2), & t \le \delta \\ 0, & t > \delta, \end{cases} \quad (5)$$

where c is a normalizing constant. Other kernel functions could be used, for example, the Gaussian kernel, but (5) is convenient because $K_\delta(t)$ decreases smoothly but steeply to zero as $|t|$ increases, so that few values are assigned small nonzero weights: Such values slow down computations for little gain (see also Fan and Gijbels 1996; Fan and Zhang 1999).

The solution to (4) is

$$(\hat{\alpha}, \hat{\beta}) = (\mathbf{X}^TW\mathbf{X})^{-1}\mathbf{X}^TW\boldsymbol{\theta}, \quad (6)$$

where W is the matrix whose ith diagonal element is $K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)$ while all other elements are zero. The posterior mean estimate is

$$\hat{\alpha} = \frac{\sum_i \phi_i^*K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}{\sum_i K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}. \quad (7)$$

If we had adopted local-constant regression (*i.e.*, $\beta \equiv 0$) and the indicator kernel function

$$I_\delta(t) = \begin{cases} 1, & t \le \delta \\ 0, & t > \delta, \end{cases}$$

in place of (5), then

$$\hat{\alpha} = \frac{\sum_i \phi_iI_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}{\sum_i I_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}, \quad (8)$$

which is the rejection-method estimate. More generally, the rejection method can be viewed as the special case of our local-linear regression approach that uses the indicator kernel and local-constant regression.

The regression approach can be extended to adjust multiple parameters simultaneously, using multivariate regression, in which case β is a matrix and α and ϕ_i are vectors. Examples of the approximation of joint posterior densities for pairs of parameters are given below.

Choice of tolerance, δ: For both rejection and regression methods we set δ to be a quantile, P_δ, of the empirical distribution function of the simulated $\|\mathbf{s}_i - \mathbf{s}\|$. For example, $P_\delta = 0.01$ means that the 1% of simulated $\mathbf{s}_i$ that are closest to $\mathbf{s}$ are assigned a nonzero weight. Choice of δ involves a bias-variance trade-off: Increasing δ reduces variance thanks to a larger sample size for fitting the regression, but also increases bias arising from uncorrected departures from additivity and linearity.

In the limit of increasing δ, all $\mathbf{s}_i$ are accepted under the rejection method, and so posterior approximations approach prior values. For practical values of δ, the simulation results below suggest that there can be a notable "bias" of the posterior estimate toward the prior. The local-linear regression method approaches simple linear regression in this limit, and so the accuracy of the results for large δ depends on the adequacy of the linearity and additivity assumptions. In the limit as δ tends to zero, the regression and rejection methods are equivalent. Thus, the relative merits of the two methods hinge on their sensitivity to δ in the vicinity of $\delta = 0$, which is explored via simulation studies below.

Posterior density estimation: The posterior density at a candidate value ϕ_0 for ϕ can be approximated using kernel density estimation applied to the weighted sample,

$$\hat{\pi}(\phi_0|\mathbf{s}) = \frac{\sum_i K_\Delta(\phi_i^* - \phi_0)K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}{\sum_i K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}, \quad (9)$$

where Δ is a density-estimation bandwidth. Once again we employ the Epanechnikov kernel, but note that the role of the density-estimation kernel function is distinct from its regression-weighting role. There is no requirement for the two kernels to have the same functional form. We have chosen to do so here, but note that the regression tolerance δ is usually different from the density-estimation bandwidth Δ.

As noted above, using a local-constant approach with

the indicator kernel leads to the usual rejection-method estimate of the posterior density function:

$$\hat{\pi}(\phi_0|\mathbf{s}) = \frac{\sum_i K_\Delta(\phi_i - \phi_0) I_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}{\sum_i I_\delta(\|\mathbf{s}_i - \mathbf{s}\|)}.$$

Alternative methods, such as local-likelihood-based methods, can be used to estimate densities from the adjusted sample. For all the density estimation in this article we have used the local-likelihood method of Loader (1996).

SIMULATION STUDY

Equation 6 is the standard expression for the fitted value at the intercept in a weighted linear regression and can thus be estimated by any standard method. For the results presented in this article, we used the function lm(), either in the R statistical language (Ihaka and Gentleman 1996; http://cran.r-project.org) or for multivariate response variables in the commercial program Splus. The density-estimation program Locfit is also implemented in R.

Motivating data set and model: In this section we describe a number of simulation-based tests in which the relative performances of the rejection and regression methods are compared. These tests are centered around the models and data set analyzed by Pritchard *et al.* (1999). The data set consists of gene frequencies at eight loci on the Y chromosome, surveyed from 445 males taken from a number of different populations around the world, previously published by Perez-Lezaun *et al.* (1997) and Seielstad *et al.* (1998). Pritchard *et al.* (1999) considered a population growth model similar to that of Weiss and von Haeseler (1998) and Beaumont (1999), in which an ancestral population of constant size N_A chromosomes begins exponential growth t_g generations from the present time, giving a current population size of $N_0 = N_A \exp(rt_g)$, where r is the population growth rate per generation. Pritchard *et al.* (1999) extracted three summary statistics from the data: (1) the mean (across loci) of the variance in repeat numbers; (2) the mean effective heterozygosity (*i.e.*, the probability of two randomly drawn chromosomes differing at a particular locus, averaged across loci); and (3) the number of distinct haplotypes in the sample. They simulated data points under a coalescent model and applied the rejection algorithm described above, keeping only points that were within 10% of the observed values of each summary statistic. Data sets were simulated with a number of mutation models, and they analyzed both the combined data set of 445 chromosomes and the data from each population separately. In our comparisons, we consider only the single-step mutation model (Ohta and Kimura 1973), with no limit on the allele sizes, and the combined data set.

The results from the regression and rejection methods are also compared with those obtained by the MCMC method of Wilson and Balding (1998), which has been expanded to include a model of population growth (Wilson *et al.* 2003; the program BATWING is available at http://www.maths/abdn.ac.uk/~ijw). The growth model is the same as that described above, but with a number of different parameterizations, which are discussed below. Thus we can compare the results from approximate posterior distributions based on summary statistics with those from posterior distributions based on the full data.

Stable population model: The parameter of interest is $\theta = 2N\mu$, where N is the number of chromosomes in the population and μ is the mutation rate. Setting $\theta = 10$, we simulated 100 data sets of 445 chromosomes typed at eight completely linked loci. These data sets were then analyzed with both the regression and the rejection methods, using $P_\delta = 0.00125$, 0.0025, 0.005, 0.01, 0.02, 0.04, 0.08, 0.16 and simulation sizes $k = 2000$, 10,000, and 50,000. For the full-data estimation using MCMC, we ran BATWING for 2×10^6 parameter updates (20 tree updates per parameter update), after a burn-in of 10^5 parameter updates, thinned every 200, to yield 10,000 points. For the regression and rejection methods we assumed rectangular priors on θ of (0, 50) and flat improper priors for the MCMC method. The two priors are comparable because the widths of the rectangular priors are large relative to the posterior density. In particular, although it is likely that the posterior distribution for θ is improper (*i.e.*, has infinite area), to the degree of approximation inherent in MCMC, after burn-in the simulated samples are so far from 50 that in practice there would be no difference in behavior whether a rectangular prior was used or not. The results are illustrated in Figure 1, which shows the relative mean square error (RMSE) of the estimates of θ for $k = 50{,}000$ plotted against the tolerance, P_δ, together with approximate standard errors estimated via a nonparametric bootstrap. The RMSE is calculated as $(1/n)\Sigma^n(\hat{\theta}_i - \theta)^2/\theta^2$. It can be seen that the RMSE of θ estimated by the rejection method diverges rapidly from that estimated by MCMC with increasing values of P_δ, and this contrasts with the regression method in which the divergence is small. We have also repeated the analysis (here and with the growth model discussed below) using the median of the relative absolute errors and obtained very similar results, indicating that the improvement in accuracy does not depend on a few outlying points. Interestingly, although the RMSE of the MCMC method is the smallest of the three methods, the difference between it and that of the regression method is not substantial. For $p_\delta \geq 0.01$ the RMSEs for sample sizes 10,000 and 2000 are very similar. However, for smaller p_δ there is a tendency for the RMSE to begin increasing because of the increased sampling variance. This affects the regression method most strongly, and for $k = 2000$ and $P_\delta = 0.00125$, 0.0025, and 0.005 (sample sizes 3, 5, and 9, respectively)

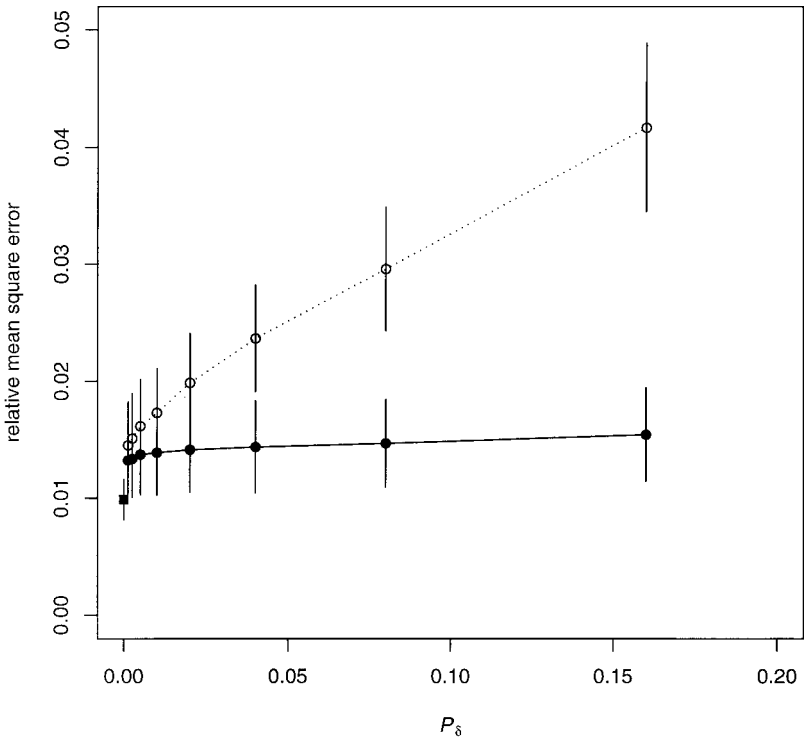

FIGURE 1.—A plot of the RMSE in estimates of θ against a measure of tolerance P_δ, as defined in the text. Estimates using the rejection method are shown as a dotted line and those from the regression method as a solid line. The RMSE for the MCMC method is shown by the solid square at $P_\delta = 0$. Standard errors are shown as vertical bars.

the RMSE of the regression method is worse than that of the rejection method. Most studies will wish to calculate distributional summaries and estimate densities, and therefore generally require sample sizes of ~500, and thus this effect will generally have little practical importance, unless a large number of summary statistics are used, as discussed later.

We have also estimated the RMSE for the heterozygosity-based and variance-based estimators of θ (KING *et al.* 2000, Equations 2–5). For the variance-based estimator the RMSE is 0.46 ± 0.11 and for the heterozygosity-based estimator it is 0.092 ± 0.016. Both estimators have means close to 10 and the cause of the large RMSE is the substantially higher variance compared to the other methods.

Growing population model: Although the model includes four parameters that might be considered estimable—*i.e.*, r, t_g, N_A, and μ in the model of PRITCHARD *et al.* (1999)—only three parameters are identifiable in the likelihood function. Although more parameters can be estimated through the use of informative priors as discussed in TAVARÉ *et al.* (1997) and performed in PRITCHARD *et al.* (1999), for the RMSE analysis we restricted ourselves to a three-parameter model. This leads to a choice of parameterizations (*e.g.*, WEISS and VON HAESELER 1998; BEAUMONT 1999) and, for compatibility with the BATWING program, we used

$$\theta = 2N_A\mu, \quad \omega = rN_A, \quad \kappa = rt_g.$$

One hundred test data sets were simulated using the parameter values: θ = 2.1, ω = 11.25, and κ = 6.75. These values were chosen because they correspond to the μ = 0.0007, t_g = 900, N_A = 1500, r = 0.0075 estimated by PRITCHARD *et al.* (1999). Because of the time taken for the MCMC method to converge, a sample size of 200 was used rather than 445 as in the original data set. For the MCMC simulation we ran 10^7 parameter updates (20 tree updates per parameter update), after a burn-in of 5×10^5 parameter updates, thinned every 1000, to yield 10,000 points. For both the MCMC and the rejection/regression methods we used gamma priors with parameters (shape, scale): θ (4, 1), ω (3, 2), κ (3, 1). Values of P_δ were the same as for the stable population case, and we present results for k = 50,000 in Figure 2. It can be seen that in general the regression method has superior performance to the rejection method and that the MCMC method is more accurate than either. In the case of θ, the relative performance of the different methods is very similar to that in the stable population case, with the RMSE of the rejection method diverging very rapidly from the regression-based value with increasing values of P_δ. For ω, there is a marked improvement in accuracy using the MCMC method. There is appreciable divergence between the regression and rejection methods with P_δ, but it is less marked than for θ. A similar pattern is seen with κ. It can be seen that with low P_δ the RMSE of the regression method begins to rise.

Additional summary statistics: We investigated the mean across loci of the kurtosis in the allele length distribution and the variance of the variances in length among loci (DI RIENZO *et al.* 1998; REICH *et al.* 1999), the mean maximum allele frequency, a multivariate equivalent of kurtosis (*i.e.*, based on the fourth power of

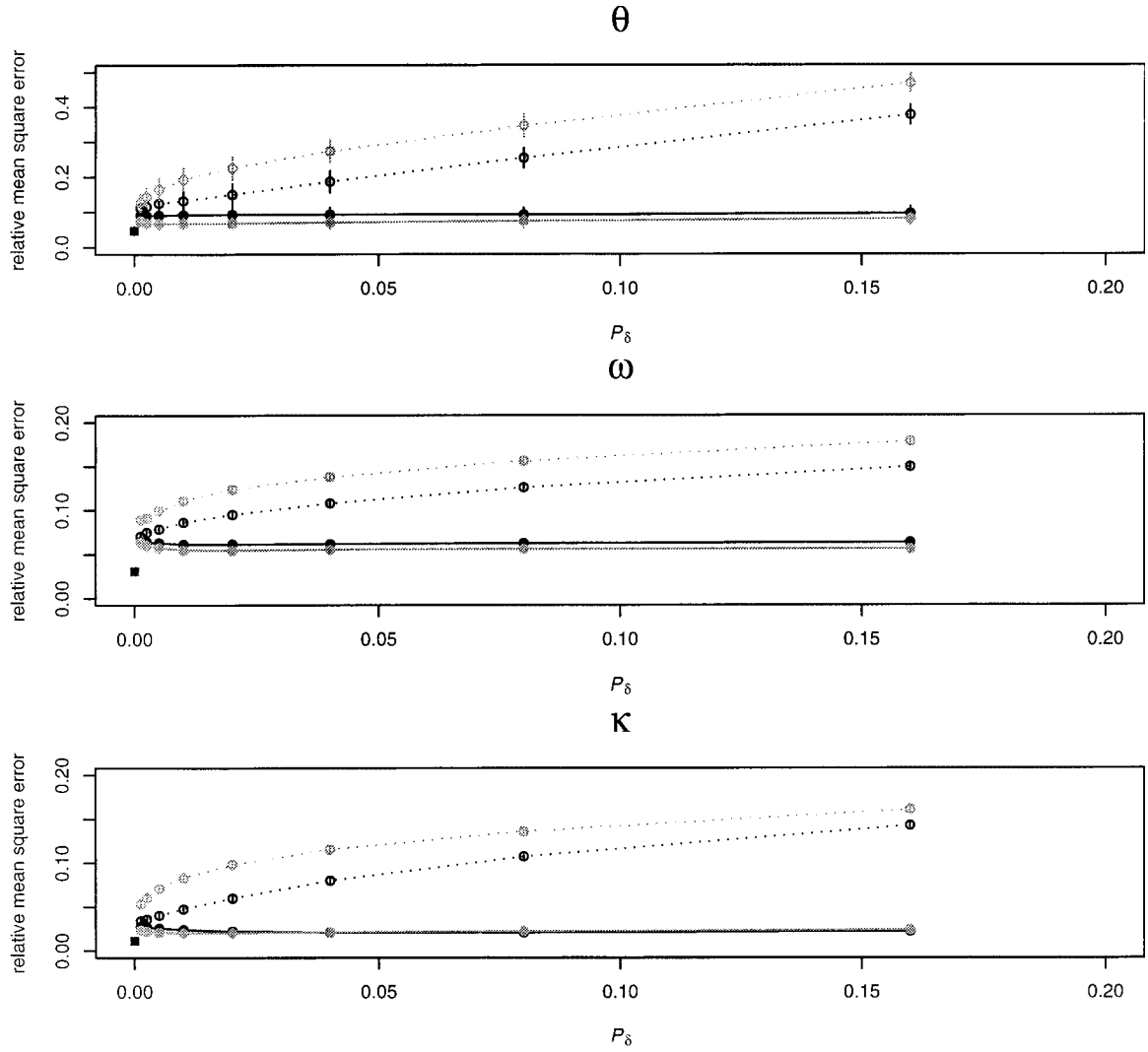

FIGURE 2.—A plot of the RMSE in estimates of θ, ω, and κ against a measure of tolerance P_δ, as defined in the text. Solid lines were obtained using three summary statistics. Shaded lines were obtained using five summary statistics. Other details are as for Figure 1.

the Euclidean distance from the centroid of the lengths measured at each locus for each chromosome), and the measure of linkage disequilibrium, Δ^2 (see, *e.g.*, HUDSON 2001), averaged over all pairs of loci. The rationale for the latter stems from the observation by SLATKIN (1994) that the degree of linkage disequilibrium is reduced in growing populations, even for completely linked loci. None of these statistics individually led to a marked improvement in RMSE. The latter two statistics appeared to lead to a small, and possibly statistically significant, reduction in the RMSE for ω, and the results from using five summary statistics (the original three plus the multivariate kurtosis and Δ^2) are illustrated as shaded lines in Figure 2. The effect of the extra summary statistics is to produce either no change or a small improvement in the RMSE for the regression method for all tolerances, but a substantial worsening for the rejection method (for all tolerances). With five summary statistics, as with three, the RMSE of the regression method begins to rise with small P_δ, as shown in Figure 2, but the lines do not cross. By contrast we found that with $k = 2000$ the RMSE for the regression method began to be notably larger than that for the rejection method at $P_\delta = 0.02$. This is due to the increased variability in the regression estimates with increasing number of summary statistics (the "curse of dimensionality").

For fixed simulation size k, increasing the number of summary statistics requires an increase in the tolerance δ. For the rejection method, the results of Figure 2 indicate that the bias thus introduced outweighs the benefits of the additional information. This is not so for our regression method, because of its insensitivity to δ, although the improvement is small. Inevitably, there will be a tendency for diminishing returns from each extra summary statistic. Overall, however, there seems to be room for additional improvement in the fit using regression-based methods, such that accuracy close to that of full-data methods may be obtained, but

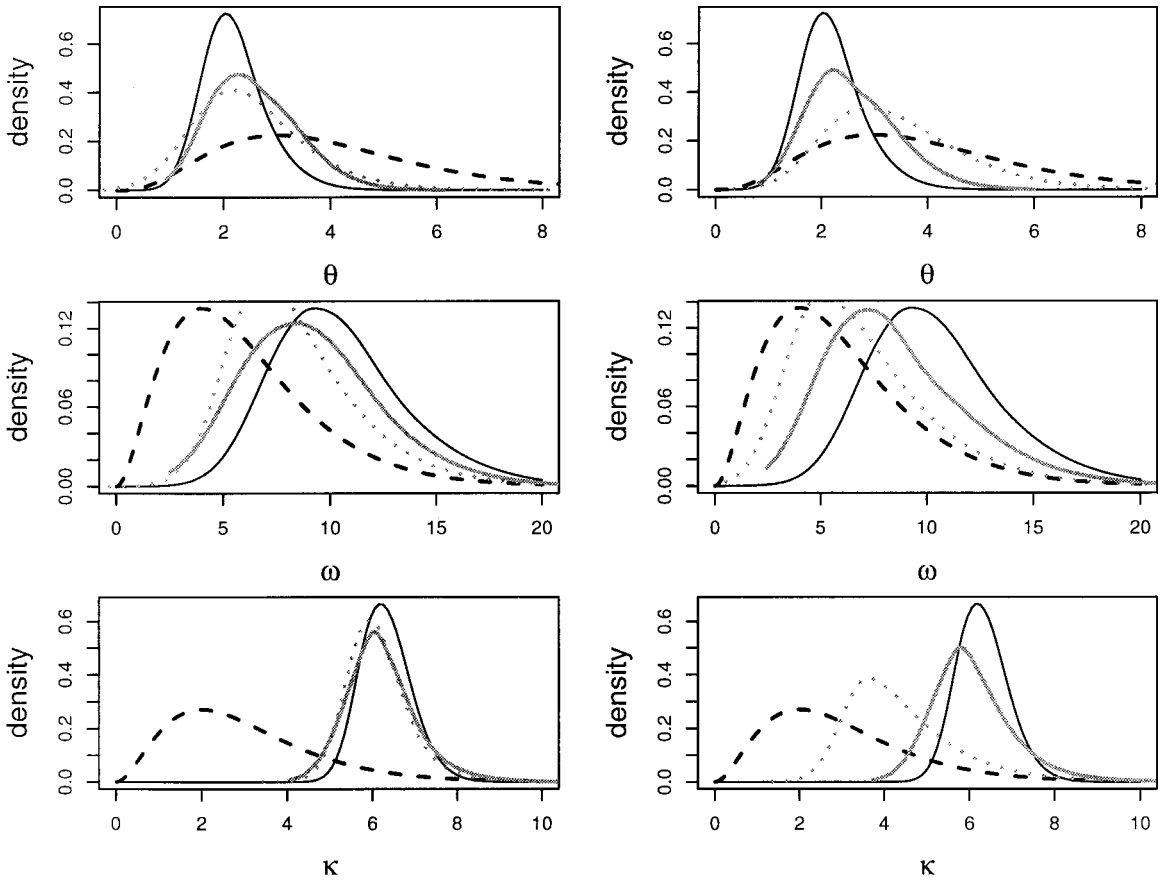

FIGURE 3.—Plots of the posterior densities for θ, ω, and κ estimated by MCMC, regression, and rejection methods. Densities estimated by the regression method are shown on the left, and those estimated by the rejection method are on the right. The posterior density estimated by MCMC is shown as a solid line, and the prior is shown as a dashed line. Posterior densities from the regression/rejection methods are shown as shaded lines ($P_\delta = 0.00125$, shaded; $P_\delta = 0.16$, shaded dotted).

using orders of magnitude fewer computations. The MCMC simulations for the growing population model took 100 processor days on 700 Mhz Pentium 3 processors, whereas the summary statistic analysis took 4 hr for three parameters and 27 hr for five parameters. The latter increase in time does not reflect the consequences of scaling up with more summary statistics, but that the computation of Δ^2 is very time consuming, whereas that for the other summary statistics is generally small compared to the time for generating samples. It should also be noted that in all the simulations described in this article the time spent performing the regression calculations is a negligible proportion of the total time to carry out the coalescent simulations—a few seconds in comparison to several hours. It is conceivable that with a very large number of summary statistics and a large number of points accepted within the tolerance limits, the time spent performing the regression calculations predominates, in which case the rejection method may have an advantage in that more points can be analyzed within a given time.

The performance of RMSE as a measure of accuracy depends on the priors chosen. If the prior mean for a parameter is the same as the value with which the simulations were carried out, then the rejection method will outperform both the MCMC and the regression methods as the tolerance becomes larger. This is because the data will cause the true posterior distribution to fluctuate away from the prior, whereas, in the case of the rejection method, if the tolerance is large all samples will be randomly taken from the prior and have the same mean as the prior and negligible variance in the mean. Similarly, because the posterior distribution estimated by the rejection method tends toward the prior when the tolerance is large, if a prior is chosen such that the mean of the posterior distribution tends to be on one side of the true value (for example, when the likelihood is very skewed) and the prior mean is on the other side of the true value, the RMSE for the rejection method can be seen first to decrease with increasing P_δ and then increase. A comparison of the mean posterior variances might avoid this problem. However, the RMSE is a useful summary because it combines both the effect of variability in the estimates and bias. Therefore we use RMSE to summarize the relative performance of the different methods and have chosen priors that avoid this problem.

Comparison of posterior densities: We estimated margi-

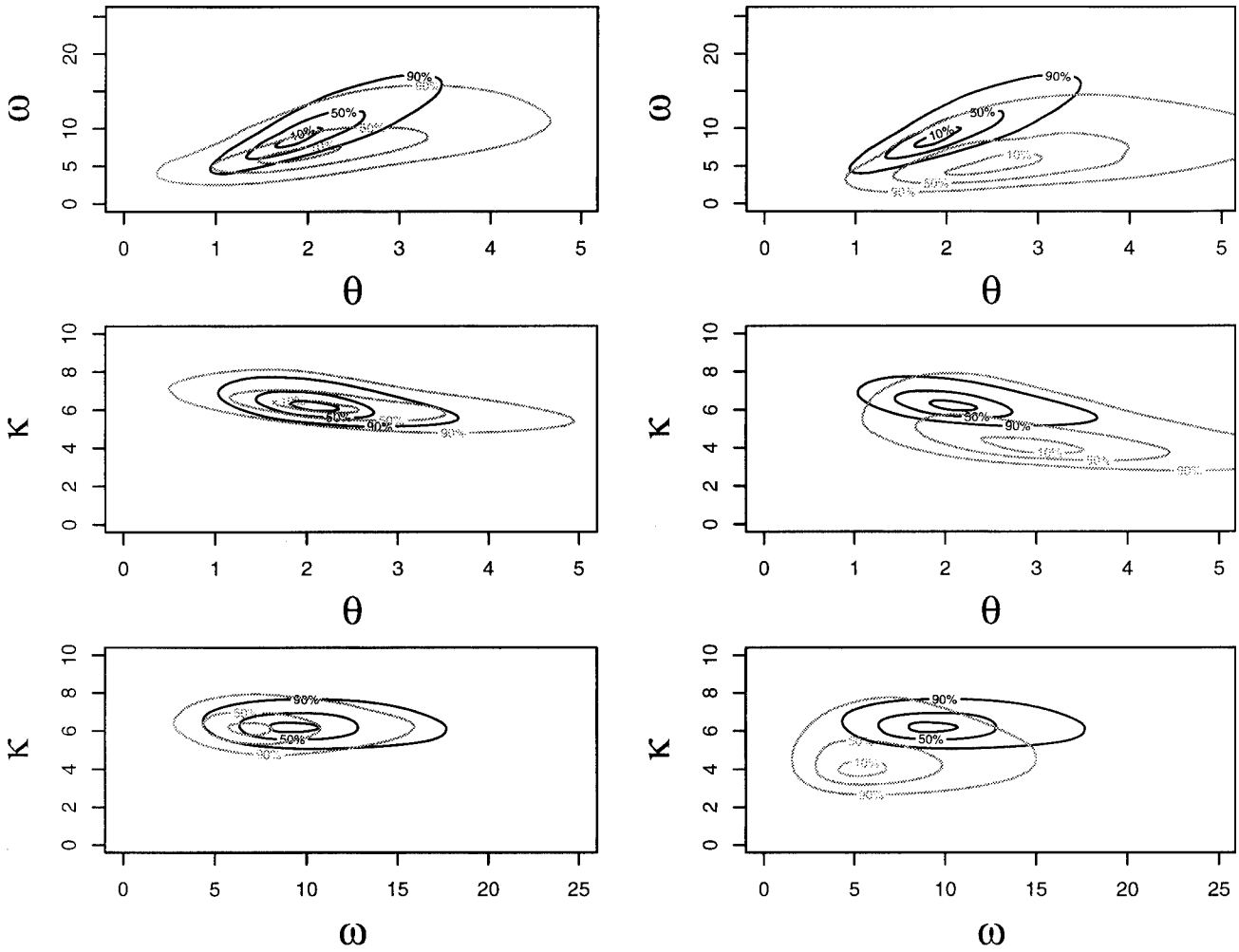

FIGURE 4.—Plots of the joint posterior densities for the three pairs of parameters among θ, ω, and κ, estimated by MCMC, regression, and rejection methods. Densities estimated by the regression method are shown on the left, and those estimated by the rejection method are on the right. A tolerance of 0.08 was used to compare the regression and rejection methods. The 10, 50, and 90% highest posterior density contours are shown. Those estimated by MCMC are shown as solid lines, and those estimated from summary statistics are shown as shaded lines.

nal posterior densities for θ, ω, and κ for one of the 100 data sets simulated under the growing-population model. The regression and rejection methods were applied with two different tolerances ($P_\delta = 0.00125$ and $P_\delta = 0.16$) and $k = 500{,}000$, while for the MCMC method densities were estimated using 10,000 BATWING outputs. The results are presented in Figure 3. It can be seen that there is a general tendency for the posterior density estimated by the rejection method to be closer to the prior than that estimated by the other methods, particularly with $P_\delta = 0.16$. The densities for the two tolerances are generally different with the rejection method and very similar with the regression method. The posterior density from the MCMC method is generally sharper than those from the other methods and is more likely to be centered around the true value.

The true posterior distribution based on summary statistics need not be very similar to that of the full data, estimated by MCMC—*i.e.*, it could potentially have a different location as well as a broader variance. In comparing the regression- and rejection-based posterior densities there is no independent "true" posterior density based on summary statistics with which to compare the results in Figure 3. However for $P_\delta = 0.00125$ the posterior densities obtained by both the rejection and regression methods are very similar. If it is assumed that the regression method accelerates the rate of convergence to the "true" posterior distribution on the basis of the summary statistics, then similarity of the regression- and rejection-based distributions may indicate convergence, although there is always the possibility that both will continue to change slowly with decreasing P_δ.

To estimate joint posterior densities for pairs of parameters, bivariate local-linear regression adjustments were carried out using lm() in Splus, which can handle multivariate response variables. Figure 4 shows the 10, 50, and 90% highest posterior density (HPD) contours for pairs of parameters considered jointly. It can be seen

TABLE 1

Comparison of estimates using regression and rejection methods

		Pritchard *et al.* (1999)		Regression		Rejection		
		Original	Replicate	$P_\delta = 2\%$	$P_\delta = 16\%$	$P_\delta = 2\%$	$P_\delta = 16\%$	Prior
μ	Mean	7	7.2	6.7	6.8	7.1	7.5	8
$\times 10^{-4}$	95% CI	4–12	3.5–12	3.7–12	3.6–13	3.5–12	3.6–13	4–14
r	Mean	75	75	100	93	82	67	50
$\times 10^{-4}$	95% CI	22–209	23–210	52–290	47–270	24–220	10–210	10–180
t_g	Mean	900	900	750	900	900	1000	1000
	95% CI	300–2150	320–2100	272–2100	320–3200	300–2100	200–2700	25.5–3700
N_A	Mean	1.5	1.5	1.5	1.3	1.3	2.9	36
$\times 10^3$	95% CI	0.1–4.9	0.14–4.4	0.57–6	1–14	0.096–4.6	0.099–11	0.098–250

Means and 95% equal-tailed credible intervals of the marginal posterior distributions of μ, r, t_g, and N_A estimated from the 445 Y chromosomes described in Pritchard *et al.* (1999). The prior values have been recomputed by us and differ slightly from the authors' original values. CI, credible interval.

that both summary-statistic-based methods tend to have wider joint posterior densities than the MCMC-based method, particularly for (θ, ω). However, the densities estimated by the regression method are much closer to those estimated by MCMC than those estimated by the rejection method.

ANALYSIS OF HUMAN Y CHROMOSOME DATA

Pritchard *et al.* (1999) estimated posterior densities for the four "natural" parameters of the growing population model: μ, r, t_g, and N_A. The BATWING program does not use this parameterization, and we use an alternative, similar parameterization described below for comparing posterior distributions from the rejection and regression methods with the full-data likelihoods. However, despite not having the full-data likelihood as a benchmark it is still useful to compare the performances of the regression and rejection methods with these four parameters, and the results are shown in Table 1. We present results for tolerance $P_\delta = 0.02$ and $P_\delta = 0.16$. The results for the rejection and regression methods are based on 100,000 simulations. We summarize the posterior distributions by the mean and 95% equal-tail probability intervals, as in Pritchard *et al.* (1999). In addition, we repeated the results of Pritchard *et al.* (1999), using their definition of tolerance (10% of the observed value) with 10^6 simulations, in which ~1600 points were accepted.

The main pattern evident in Table 1 is that all the methods give broadly similar answers irrespective of the tolerance. This appears to be because there is very little information in the data on much of the parameter space. For example, the priors and posteriors for μ are almost identical, and therefore it is not surprising that even with $P_\delta = 0.16$ in the rejection method, where the acceptance rate is 100-fold greater than in the study by Pritchard *et al.* (1999), there is very little difference in the summaries of the posterior distributions. We are able to replicate the results of Pritchard *et al.* (1999) for their definition of tolerance, and generally there is similarity between their results and those from the rejection method using our definition of tolerance, at least for the narrower tolerances. As with the results above, there is a tendency for the estimates from the rejection method to move closer to the prior with increasing P_δ, whereas this effect is not so strong with the regression method. The results from the regression method with $P_\delta = 0.02$ tend to be more different from those obtained using the method of Pritchard *et al.* (1999) than the results from the rejection method, and it is tempting to conclude that the regression-based results are closer to the true $p(\theta|\mathbf{s})$, despite the 12.5-fold increased intensity of sampling with the method of Pritchard *et al.* (1999).

To compare the results from a four-parameter model, similar to that used by Pritchard *et al.* (1999), using full-data posterior distributions estimated with BATWING as a benchmark, we have used the following parameterization (and priors): μ (gamma: shape $-$ 10, scale -0.00008), N_A (lognormal: mean log 8.5, SD log 2), r (exponential: 0.005), and $\beta = t_g/N_A$ (gamma: shape -2, scale -1). In 5 individuals we detected changes in microsatellite lengths that were fractions of the repeat length, and we therefore used only 440 individuals in the MCMC estimation. The new summary statistics calculated from this group were variance in allele length, 1.123; heterozygosity, 0.635; and number of distinct haplotypes, 312. For the MCMC simulation we ran 10^7 parameter updates (40 tree updates per parameter update), after a burn-in of 5×10^5 parameter updates, thinned every 1000, to yield 10,000 points. The results of this analysis are shown in Figure 5 and reflect those in Table 1. The full-data posterior distributions are generally very broad and similar to the priors, with the exception of values of N_A, r, and β close to 0, which are clearly rejected by the data. Not surprisingly, therefore, the rejection method generally appears to perform well.

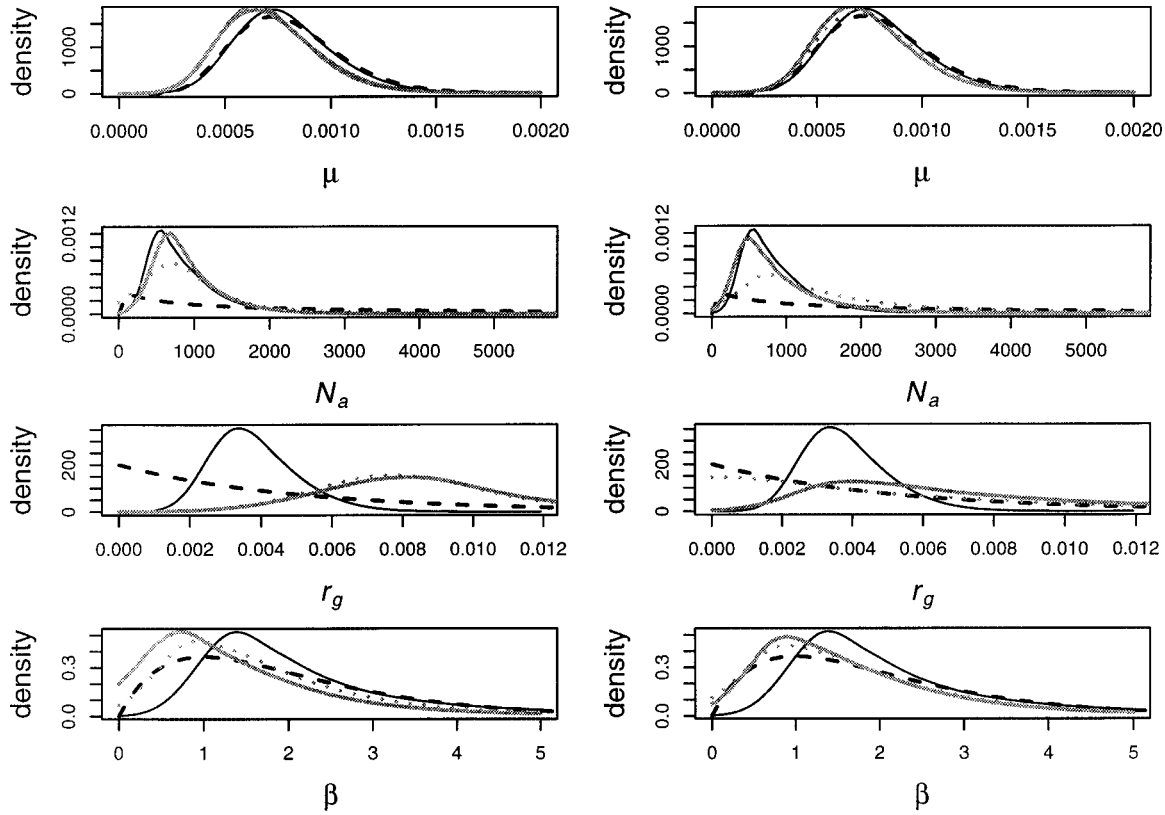

FIGURE 5.—Plots of the posterior densities for μ, N_A, r_g, and β estimated by MCMC, regression, and rejection methods. Details are as for Figure 3.

For N_A and r, where the prior is more different from the posterior, the usual pattern is observed whereby the distributions estimated by the rejection method move toward the prior for large P_δ, and this effect is weaker in the regression method. For μ and β, where the posteriors and priors are very similar, no such effect is observed, and in fact, for β the regression-based density appears to change more with P_δ than the rejection-based density. It should be borne in mind that although a minimum of 1000 independent points are used in the density estimates, there will still be some sampling error in the estimates (and also, of course, sampling error associated with the regression itself). A further point to note is that with β, where the posterior density is close to 0, the regression method gives some negative values, which have been truncated in Figure 5. This could be avoided by the use of transformations or a generalized linear model in the regression.

As with the results in Figure 3 the rejection-based densities tend to be similar to the regression-based densities for $P_\delta = 0.02$, suggesting a degree of convergence. The density for r is the most different. However, tests with $P_\delta = 0.002$ indicate that the rejection-based density does indeed converge to that estimated by the regression method. The results suggest that the posterior distributions for r and β, given the summary statistics, are notably different from the full-data posterior distributions.

CONCLUSIONS

There are two principal advantages of our approach over the rejection method: Simulated ϕ' are assigned a weight that decreases with $||\mathbf{s}' - \mathbf{s}||$, and local-linear regression corrects for the difference between $E[\phi|\mathbf{S} = \mathbf{s}_i]$ and $E[\phi|\mathbf{S} = \mathbf{s}]$. We have illustrated with examples, where we have the full-data posterior distributions, that this innovation leads to substantially improved accuracy over earlier methods. We also illustrate the relative accuracy of MCMC-based and summary-statistic-based methods for inferring past population growth. It can be seen that the MCMC-based method is consistently superior to the summary-statistic-based methods and highlights that it is well worth making the effort to obtain full-data inferences if possible. However, undoubtedly there are advantages to the use of summary statistics, both in the ease of implementation and in the time taken to obtain results, and it appears to be a viable initial approach for applying Bayesian methods to some population genetic problems. Because of the curse of dimensionality there are limitations to the number of summary statistics that can be handled with a reasonable number of simulations (otherwise the problem will approach that of MCMC in computational time). It remains to be seen which population genetic problems can be easily summarized by a small enough number of summary statistics for this approach to be competitive with MCMC. Further

research is needed to find a more rigorous way for choosing summary statistics, including the use of orthogonalization and "projection-pursuit" methods. A problem that needs to be considered is the potential for the regression method to adjust the simulated points so that they fall outside the support of the prior. For example, if the posterior density is large at 0 negative values can be generated. This can be addressed by the use of transformations. In addition, improved regression methods, or other methods of conditional-density estimation, may overcome this problem and allow wider tolerances to be used, thereby decreasing the number of simulations that are needed and increasing the number of summary statistics that can be accommodated. Overall, however, the general approach presented here should allow for a greatly expanded use of approximate Bayesian methods in population genetic analysis.

We are grateful to Claire Calmet and two anonymous referees for their helpful comments on the manuscript. This work was supported by BBSRC/EPSRC grant E13397 awarded to M.A.B. and D.J.B.

LITERATURE CITED

Beaumont, M. A., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013–2029.

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill *et al.*, 1998 Heterogeneity of microsatellite mutations within and between loci and implications for human demographic histories. Genetics **148:** 1269–1284.

Estoup, A., I. J. Wilson, C. Sullivan, J.-M. Cornuet and C. Moritz, 2002 Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. Genetics **159:** 1671–1687.

Fan, J., and I. Gijbels, 1996 *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Fan, J., and W. Zhang, 1999 Statistical estimation in varying coefficient models. Ann. Stat. **27:** 1491–1518.

Fu, Y.-X., and W.-H. Li, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. **14:** 195–199.

Hudson, R. R., 2001 Linkage disequilibrium and recombination, pp. 309–324 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Ihaka, R., and R. Gentleman, 1996 R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5:** 299–314.

King, J. P., M. Kimmel and R. Chakraborty, 2000 A power analysis of microsatellite-based statistics for inferring past population growth. Mol. Biol. Evol. **17**: 1859–1868.

Loader, C. R., 1996 Local likelihood density estimation. Ann. Stat. **24:** 1602–1618.

Nordborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Perez-Lezaun, A., F. Calafell, M. Seielstad, E. Mateu, D. Comas *et al.*, 1997 Population genetics of Y-chromosome short tandem repeats in humans. J. Mol. Evol. **45:** 265–270.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

Reich, D. E., M. W. Feldman and D. B. Goldstein, 1999 Statistical properties of two tests that use multilocus data sets to detect population expansions. Mol. Biol. Evol. **16:** 453–466.

Seielstad, M. T., E. Minch and L. L. C. Cavalli-Sforza, 1998 Genetic evidence for a higher female migration rate in humans. Nat. Genet. **20:** 278–280.

Shoemaker, J. S., I. S. Painter and B. S. Weir, 1999 Bayesian statistics in genetics. Trends Genet. **15:** 354–358.

Slatkin, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331–336.

Tavaré, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293:** 455–462.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wilson, I. J., M. E. Weale and D. J. Balding, 2003 Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities. J. R. Stat. Soc. A (in press).

Communicating editor: W. Stephan