Algorithms **2011**, *4*, 115-130; doi:10.3390/a4020115

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

Alternatives to the Least Squares Solution to Peelle's Pertinent Puzzle

Tom Burr [1,⋆], **Todd Graves** [1], **Nicolas Hengartner** [2], **Toshihiko Kawano** [3], **Feng Pan** [4] **and Patrick Talou** [3]

[1] Statistical Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; E-Mail: tgraves@lanl.gov

[2] Information Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; E-Mail: nickh@lanl.gov

[3] Nuclear and Particle Physics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; E-Mails: kawano@lanl.gov (T.K.); talou@lanl.gov (P.T.)

[4] Decision Applications, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; E-Mail: fpan@lanl.gov

⋆ Author to whom correspondence should be addressed; E-Mail: tburr@lanl.gov; Tel.: +1-505-665-7865; Fax: +1-505-667-4470.

Received: 28 March 2011; in revised form: 1 June 2011 / Accepted: 8 June 2011 / Published: 23 June 2011

Abstract: Peelle's Pertinent Puzzle (PPP) was described in 1987 in the context of estimating fundamental parameters that arise in nuclear interaction experiments. In PPP, generalized least squares (GLS) parameter estimates fell outside the range of the data, which has raised concerns that GLS is somehow flawed and has led to suggested alternatives to GLS estimators. However, there have been no corresponding performance comparisons among methods, and one suggested approach involving simulated data realizations is statistically incomplete. Here we provide performance comparisons among estimators, introduce approximate Bayesian computation (ABC) using density estimation applied to simulated data realizations to produce an alternative to the incomplete approach, complete the incompletely specified approach, and show that estimation error in the assumed covariance matrix cannot always be ignored.

Keywords: approximate Bayesian computation using density estimation; mean squared error; Peelle's puzzle

---

1. Introduction

Peelle's Pertinent Puzzle (PPP) was introduced in 1987 in the context of estimating fundamental parameters that arise in nuclear interaction experiments [1]. PPP is described briefly below and in more detail in [2]. When PPP occurs, generalized least squares (GLS) behaves reasonably under the model assumptions [2]. However, in PPP, GLS parameter estimates fall outside the range of the data, which has raised concerns that GLS is somehow flawed and has led to suggested alternatives to GLS estimators.

Others [1,3–9,12–15] have investigated alternatives to the GLS approach to PPP, but no numerical performance comparisons to other methods have been reported. And, one ad hoc approach involving simulated data realizations is statistically incomplete [8].

The following section includes additional background for PPP and the GLS approach. Section 3 discusses the ordinary sample mean as a perhaps naive alternative and shows that estimation error in the assumed covariance matrix cannot always be ignored. Sections 4 and 5 develop an alternative involving a modified form of approximate Bayesian computation (ABC). This ABC-based approach uses density estimation applied to realizations of simulated data that follow an error model appropriate for PPP and having various error distributions. Section 6 provides a second Bayesian alternative resulting from completion of the ad hoc approach by specifying a suitable prior probability density function (pdf). Section 7 is a summary.

2. Background

A situation that leads to PPP is as follows. Two experiments aim to estimate a physical constant μ. Experiment one produces a measurement y_1, and experiment two produces y_2. Measurements y_1 and y_2 arise by imperfect conversion of underlying measurements m_1 and m_2. Suppose $m_1 = \mu + \epsilon_{R1}$ and $m_2 = \mu + \epsilon_{R2}$ where ϵ_{R1} is zero-mean random error in m_1 and similarly for ϵ_{R2}. Burr *et al.* [2] prove that if $y_1 = m_1 + \epsilon_S$ and $y_2 = m_2 + \alpha\epsilon_S$, where $\epsilon_S \sim N(0, \sigma_S)$ and α is any positive scale factor other than 1, the covariance matrix Σ of (y_1, y_2) can lie in the PPP region in which the GLS estimate of μ lies outside the range of y_1 and y_2. And, Burr *et al.* [2] give an example for which this error model is appropriate.

Let Σ be the 2-by-2 symmetric covariance matrix for y_1 and y_2 with diagonal entries σ_1^2, σ_2^2, and off-diagonal entry σ_{12}, which denote the variance of y_1, the variance of y_2, and the covariance of y_1 and y_2, respectively. For the case considered here and in Zhao and Perry [4]

$$\Sigma = \begin{pmatrix} 0.1134 & 0.06 \\ 0.06 & 0.0504 \end{pmatrix} \quad (1)$$

and we assume throughout that Σ exactly equals the numerical values given in Equation 1.

If the three errors $\epsilon_{R1}, \epsilon_{R2}$, and ϵ_S have normal (Gaussian) distributions, then the joint pdf of (y_1, y_2) is also normal. If the three errors have non-normal distributions, then the joint pdf of (y_1, y_2) will not be normal and might be difficult to calculate analytically. Therefore, to handle non-normal distributions that might be difficult to calculate analytically, Sections 4 and 5 develop an estimator based on ABC. This ABC-based estimation can be applied with the error model just described resulting from the two experiments producing y_1, y_2 values with a covariance given by Equation 1, and can accommodate non-normal distributions.

It is well known that GLS applied to y_1 and y_2 results in the best linear unbiased estimate (BLUE) $\hat{\mu}$ of μ [16]. Here, "best" means minimum variance and unbiased means that on average (across hypothetical or real realizations of the same experiment), the estimate $\hat{\mu}$ will equal its true value μ. GLS properties are well known and are reviewed in [2]. The GLS estimate for μ arising from the generic model

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mu + e_1 \\ \mu + e_2 \end{pmatrix} \tag{2}$$

with $\Sigma = \mathrm{Cov}(e_1, e_2) = \mathrm{Cov}(y_1, y_2)$ is given by

$$\hat{\mu} \;=\; cG^t\Sigma^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{3}$$

where the scalar $c = (G^t\Sigma^{-1}G)^{-1}$. And the variance σ^2 of the GLS estimate $\hat{\mu}$ is given by

$$\sigma^2 \;=\; \left(G^t\Sigma^{-1}G\right)^{-1} \tag{4}$$

where the sensitivity (or design) matrix G is given by $G^t = (1, 1)$. Putting the covariance of Equation 1 into Equations 3 and 4 and following the example in Burr *et al.* [2] with $y_1 = 1.5$ and $y_2 = 1.0$ gives $\hat{\mu} = c_1y_1 + c_2y_2 = 0.89$ and $\sigma = 0.22$ where $c_1 = -0.22$ and $c_2 = 1.22$. Notice that $c_1 + c_2 = 1$ (so that $\hat{\mu}$ is unbiased) but also that $c_1 < 0$ and that $\hat{\mu}$ is smaller than each of the two measured values.

GLS estimation is guaranteed to produce the BLUE even if the underlying data are not normal. However, if the data are not normal, then the minimum variance unbiased estimator is not necessarily linear in the data. Also, unbiased estimation is not necessarily superior to biased estimation [17].

The main alternative to GLS estimation in this context is estimation that uses the pdf. When viewed as a function of the unknown model parameters (μ in our case) conditional on the observed data (y_1, y_2 in our case), the pdf is referred to as the likelihood. The maximum likelihood (ML) estimate therefore of course depends on the pdf for the errors. For example, the normal distributions are replaced with lognormal distributions, the ML estimate will change. Because the ML approach makes strong use of the assumed error distributions, the ML estimate is sensitive to the assumed error distribution. The ML method has desirable properties, including asymptotically minimum variance as the sample size increases. However, in our example, the sample size is tiny (two), so asymptotic results for ML estimates are not relevant. It still is possible that an ML estimator will be better for non-normal data than GLS. In this paper, "better" is defined as the mean squared error (MSE) of the estimator, which is well known to satisfy MSE = variance + bias^2. In some cases, biased estimators have lower MSE than unbiased estimators because the bias introduced is more than offset by a reduction in variance [17]. The other estimators considered here are two versions of Bayesian estimators, both of which rely on the likelihood as does ML.

3. Estimation error in $\hat{\Sigma}$

In nearly all real situations, Σ is not known exactly, but must be estimated. Suppose the estimate $\hat{\Sigma} = S$, where S is the sample covariance matrix defined as $(n-1)S = \sum_{i=1}^{n}(y_{1i} - \overline{y}_1)(y_{2i} - \overline{y}_2)^t$

(t denotes transpose) based on n pairs of y_1, y_2 values that are auxiliary data pairs separate from the $y_1 = 1.5$ and $y_2 = 1.0$ values. We will consider two examples.

- Example 3.1

  Assume that the distribution of $(n-1)S$ is Wishart with $\nu = n-1$ degrees of freedom, which is exactly true if the y_1 and y_2 pairs are bivariate normal, and approximately true for many other bivariate distributions [18]. Then the root mean squared error (RMSE = $\sqrt{E(\hat{\mu}-\mu)^2}$, where E denotes expected value) in estimating the true mean μ using GLS, is 0.46, 0.32, 0.27, 0.25, 0.23, and 0.22 for $\nu = 2, 3, 4, 5, 10,$ and 30, respectively. These RMSE estimates are based on 10,000 simulations in *R* [19] and are repeatable to ± 0.01. And the RMSE is 0.27 using the simple mean of y_1 and y_2. Recall that if Σ is known exactly, then the RMSE in $\hat{\mu} = 0.22$. Therefore, if $n \leq 5$, the naive sample mean has the same or lower RMSE than does GLS.

  A third estimation option suggested in [20] uses only the variances of $\hat{\Sigma}$, not because y_1 and y_2 are thought to be independent, but because the estimated covariance $\hat{\sigma}_{12}^2$ is thought to be unreliable. For $\nu = 2, 3, 4, 5, 10,$ and 30, option 3 has an RMSE of 0.25, which is the theoretical best RMSE corresponding to the situation in which σ_1 and σ_2 are known exactly. Therefore, there is merit in Schmelling's [20] suggested option 3 in this case, because it has lower RMSE for all values of ν than the simple mean and lower RMSE than GLS for small ν values. Whether option 1 (GLS), option 2 (naive sample mean), or option 3 (use variances only, not covariance) is preferred depends on the true data likelihood and on the true Σ.

- Example 3.2

  Still assuming a normal likelihood but changing $\sigma_1^2, \sigma_2^2,$ and σ_{12}^2 in Σ to 0.1134, 0.02, and 0.0475, respectively (cor(y_1, y_2) = 0.997), options 1, 2, and 3 have RMSE of 0.025, 0.24, and 0.17, respectively, for $\nu = 3$, and 0.04, 0.24, and 0.17, respectively, for $\nu = 2$. This Σ corresponds to a very strong PPP effect with high correlation between y_1 and y_2, so even with large estimation error in $\hat{\Sigma}$, the GLS method has by far the lowest RMSE for all values of ν. We note that as $\nu \to \infty$, the GL $\hat{\mu}$ has $\sigma_{\hat{\mu}} \to 0.017$. As another example, let $\sigma_1^2, \sigma_2^2,$ and σ_{12}^2 in Σ be 0.1134, 0.0104, and 0.02, respectively. Then options 1, 2, and 3 have RMSE of 0.13, 0.20, and 0.12, respectively, for $\nu = 3$, and RMSE of 0.25, 0.20, and 0.13, respectively, for $\nu = 2$. So in this case option 3 is again a good choice for small values of ν. We note that as $\nu \to \infty$, the GLS $\hat{\mu}$ has $\sigma_{\hat{\mu}} \to 0.096$.

4. Approximate Bayesian Computation

Bayesian analysis requires a prior probability $p(\theta)$ for the parameters θ, and a likelihood $f(D|\theta)$ for the data given the parameter values. The result is a posterior probability $\pi(\theta|D) \propto f(D|\theta)p(\theta)$ for the parameters θ. For many priors and likelihoods, calculating the posterior is difficult or impossible, but observations from the posterior can be obtained by Markov chain Monte Carlo (MCMC). MCMC is a Monte Carlo sampling scheme that accepts a candidate value θ' with probability that depends on the relative probability of the current θ value and the candidate θ' value. The goal of MCMC is to simulate many observations from the posterior, and use summary statistics such as posterior means and

standard deviations to characterize the state of knowledge about the parameters. In our examples, θ is the scalar-valued mean which we denote throughout as μ.

ABC has arisen relatively recently in applications for which the likelihood $f(D|\theta)$ is intractable [22,23], but data can be simulated somehow,often relying on detailed simulations of physical processes. In our context, the simulation is very simple, based on realizations of the measurement error pdfs. And if some subset of the three underlying error distributions described in the background section are non-normal, leading to a complicated bivariate distribution for y_1, y_2 that could be difficult to derive analytically, then ABC is attractive. Predating ABC is a nonBayesian ("frequentist") method referred to as "inference for implicit statistical models" [21,24] that was based on density estimation. We use the term ABC to include any Bayesian method in a situation for which an analytical likelihood is not available, but simulations can produce realizations from the likelihood.

In the common version of ABC, because data can be simulated from the pdf, candidate values θ' can be accepted with a probability that depends on how close certain summary statistics such as the first few moments computed from the simulated data having parameter value θ' are to the corresponding summary statistics from the real data. Some sources [24] reserve the term ABC for methods that rely on such summary statistics, so would not use the term ABC with density estimation. Our combination of MCMC with density estimation appears to be a new combination of ideas, though some might not refer to it as ABC. Regardless of the jargon, the method will be clearly delineated below.

As a simple example to motivate our modified form of ABC, suppose $y_1, y_2, \ldots, y_n$ are independently and identically distributed as $N(\mu, \sigma)$ and it is known that $\sigma = 0.1$. Introductory statistics texts usually prove that the sample mean $\overline{y}$ is the ML estimator (MLE) for μ and more advanced texts investigate properties of MLEs. And, Equation (3) with a diagonal Σ (because in this example the y_i are independent) can be used to show that $\overline{y}$ is the GLS estimator so is also the BLUE and because the data is normal in this simple example (but not in our extended PPP examples below), $\overline{y}$ is also the MVUE. Specify the prior $p(\mu)$ to be $N(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$. Then, because the normal prior for μ is the "conjugate" prior (meaning that the posterior distribution is the same type of distribution as the prior distribution) for the normal likelihood, the posterior $\pi(\theta|D)$ is $N(\mu_{\text{post}}, \sigma^2_{\text{post}})$, where $\mu_{\text{post}} = \frac{0.1^2}{0.1^2+\sigma^2_{\text{prior}}}\mu_{\text{prior}} + \frac{\sigma^2_{\text{prior}}}{0.1^2+\sigma^2_{\text{prior}}}\overline{y}$.

Our modified ABC that relies on density estimation has the following steps:

1. Determine a lower (L) and upper bound (U) for the parameter μ. In this example, having observed $y_1, y_2, \ldots, y_{n_1}$ and knowing $\sigma = 0.1$, we can assume μ lies between say $L = \overline{y} - 5\sigma/n_1$ and $U = \overline{y} + 5\sigma/n_1$.

2. Choose μ values on a grid of equally-spaced values in the range (L, U). For each μ_i value in the grid, simulate n_2 observations from $N(\mu_i, \sigma)$ and use these n_2 observations to estimate the pdf using density estimation [25]. In the toy example, we know the true pdf is $N(\mu, \sigma)$, but in cases below, we might not know the pdf because it involves algebraic combinations of non-normal pdfs.

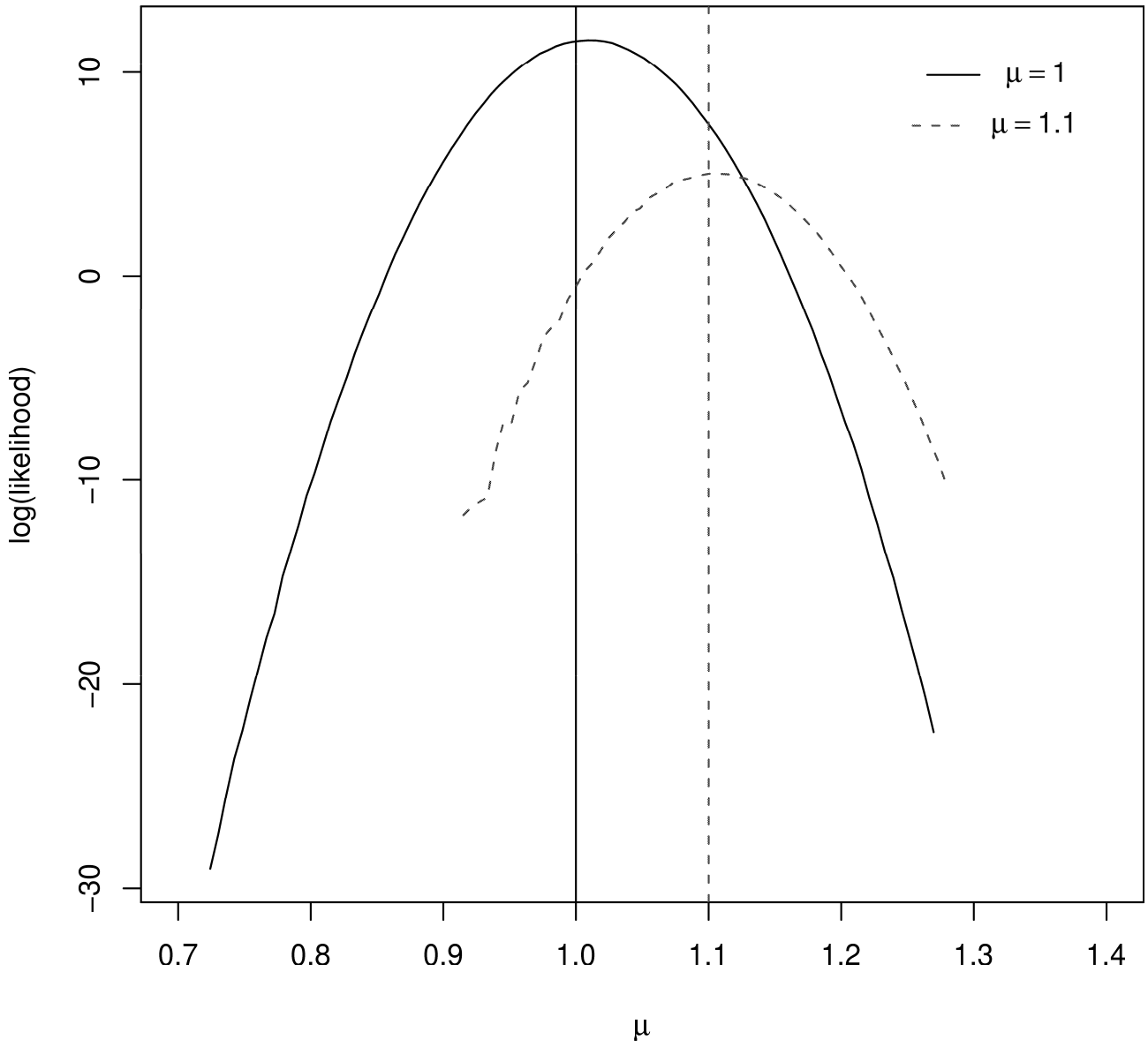
3. The likelihood of each observed y_j can be estimated by interpolation of the estimated pdf for each μ_i value in the grid. The size of the steps in the μ grid can be decreased and n_2 can be increased until estimates of μ stabilize.

4. The overall likelihood is then the product (because the y_j are independent) of the estimated likelihood for each observed y_j.

5. Use MCMC to estimate the posterior probability for μ, and to compute $\hat{\mu}$ as the mean of the posterior. To implement MCMC, we used the `metrop` function in the `mcmc` package in `R` and used the `approx` function to interpolate the estimated pdf on the grid of trial μ values. The Appendix has this "likelihood" function, where the estimated pdf is used in place of an analytical likelihood.

Figure 1 illustrates the resulting log likelihood for $\mu = 1$, $\mu = 1.1$, $n_1 = 10$, and $n_2 = 1000$. Notice that the log likelihood peaks very near the true μ value in both cases.

Figure 1. Example ABC. The solid black curve is for $\mu = 1$ and the dotted red curve is for $\mu = 1.1$.



To check our implementation, we specified the prior $p(\mu)$ to be $N(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$. Then, because the normal prior for μ is the "conjugate" prior (meaning that the posterior distribution is the same type of

distribution as the prior distribution) for the normal likelihood, the posterior $\pi(\theta|D)$ is $N(\mu_{\text{post}}, \sigma^2_{\text{post}})$. The ABC estimate of the posterior mean can then be compared to the mean of the exact posterior $\pi(\theta|D)$ to ensure proper implementation. In this case, our implementation of ABC also leads to $\hat{\mu} \approx \overline{y}$ because we used $\sigma = 0.1$, $\sigma_{\text{prior}} = 10$, and $n_1 = 10$, so the posterior distribution has mean essentially equal to $\overline{y}$. Specifically, in 100 simulations of the entire process of generating $n_1 = 10$ observations, and applying density estimation to estimate the likelihood using $n_2 = 1000$ observations resulted in no statistical difference (based on a paired t-test) between the 100 values of the exact Bayesian $\hat{\mu}$, the ABC estimate of $\hat{\mu}$, and $\hat{\mu} = \overline{y}$. Readers can duplicate our results using the *R* functions listed in the Appendix.

To our knowledge, this simple use of density estimation to substitute for having an analytical form for the likelihood has not been studied in the ABC literature. However, provided the data can be simulated sufficiently fast to acquire many observations ($n_2 = 1000$ was sufficiently large in this toy problem), and provided the dimension of the data is not too large, density estimation is feasible and effective. In the PPP case, we have only two variables, y_1 and y_2, so certainly 2-dimensional density estimation is feasible. We note here that summary statistics such as moments from the real and simulated data in the MCMC accept/reject steps in typical ABC implementations are fast to compute, but convergence criteria and measures of adequacy are still being developed [23].

5. ABC for PPP

This section again uses the slightly modified form of ABC that relies on density estimation rather than using summary statistics computed from the real and simulated data, but MCMC is applied in the context of PPP.

As mentioned in Section 2, suppose from two experiments $m_1 = \mu + \epsilon_{R1}$ and $m_2 = \mu + \epsilon_{R2}$, where ϵ_{R1} is zero-mean random error in m_1 and similarly for ϵ_{R2}. Then if $y_1 = m_1 + \epsilon_S$ and $y_2 = m_2 + \alpha\epsilon_S$, where $\epsilon_S \sim N(0, \sigma_S)$ and α is any positive scale factor other than 1, the covariance matrix Σ of (y_1, y_2) can lie in the PPP region [2].

To implement our version of ABC, many (y_1, y_2) realizations are simulated from each of many values of μ. Given the observed (y_1, y_2) and its known covariance Σ, we can again set up good L and U values for the grid of candidate μ values.

A key advantage of ABC is that any probability distribution can be easily accommodated for any of ϵ_{R1}, ϵ_{R2}, and ϵ_S. Here are two examples.

- Example 5.1.

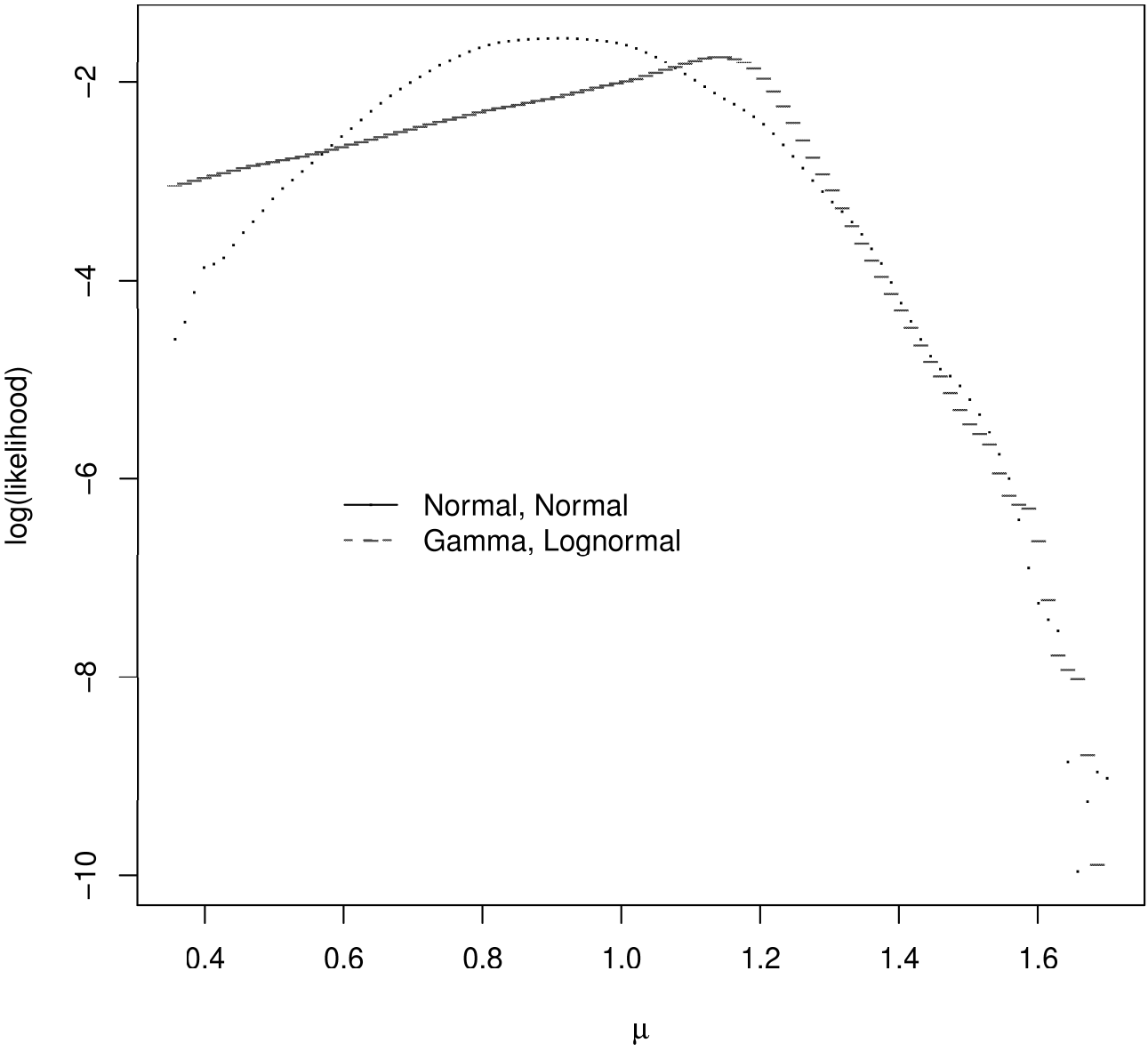
  Assume ϵ_{R1}, ϵ_{R2}, and ϵ_S are all normal, so that y_1, y_2 is bivariate normal. Then as it must, our version of ABC recovers the GLS estimate $\hat{\mu} = 0.89$.

- Example 5.2.

  Assume ϵ_{R1} and ϵ_{R2} each have a centered and scaled lognormal distribution and ϵ_S has a centered and scaled gamma distribution, with parameters chosen so that the covariance matrix Σ of y_1, y_2 is given by Equation 1. Figure 2 plots the log likelihood for μ values in the μ grid for the correct (true) likelihood for y_1, y_2 and assuming a bivariate normal distribution for y_1, y_2. These log likelihoods are meaningfully different and one might expect the corresponding estimators to behave quite

differently. For the single realization of $y_1, y_2 = 1.08, 1.24$, for which Figure 2 was generated, the GLS and ABC estimate assuming a bivariate normal distribution for y_1, y_2 is 1.27. The ABC estimate assuming the correct bivariate distribution for y_1, y_2 is 1.03.

Figure 2. Example ABC. The solid black curve is for a bivariate normal. The dotted red curve is for a bivariate non-normal (combination of Gamma and lognormal).



- Example 5.2 continued.

  To investigate whether using the correct likelihood leads to smaller RMSE, we performed 100 simulations for various values of μ in this same example 5.2. In each simulation, one new (y_1, y_2) pair was generated to represent the observed data pair, and then 10,000 observations of (y_1, y_2) pairs were generated to use for density estimation so that the likelihood of each candidate value of μ could be approximated. For example, with $\mu = 1.0$, the RMSE for GLS is 0.22 as expected, and the RMSE for ABC using the correct bivariate distribution is 0.15 (repeatable across sets of 100

simulations to ± 0.01). So, ABC has smaller RMSE than GLS in this case. However, ABC relies much more strongly on the estimated likelihood which relies on the assumed data distributions, so is not expected to be as robust to misspecifying the likelihood as GLS.

Recall that $\hat{\mu}$ via GLS is always outside the range of (y_1, y_2) when PPP occurs, and that some researchers are bothered by this. For the non-normal data just described, even though PPP occurs, we find numerically that $\hat{\mu}$ via ABC sometimes lies outside the range of (y_1, y_2) and sometimes lies within the range of (y_1, y_2).

6. Alternate Bayesian Approach to Complete the Ad hoc Approach

Reference [8] presented an ad hoc procedure to estimate μ as follows.

Using a model of the two experiments that produce y_1, y_2 such as that by [4] in Section 2, generate thousands of a y_1, y_2 pairs but regard them as being μ_1, μ_2 pairs. See below for "justification" in switching from y_1, y_2 to μ_1, μ_2. Because of the strong prior belief that $\mu_1 \approx \mu_2$, restrict attention to bivariate points (y_1, y_2) lying very near the diagonal, defining $\hat{\mu}$ as the value that maximizes the estimated probability density of the scalar-values of either y_1 or y_2 (or both y_1 and y_2) satisfying $|y_1 - y_2| < \epsilon$ for some small ϵ near 0. See Figure 3. As an important aside, recall [2] showed that PPP cannot occur with this error model (Section 2). However, because of how Σ is estimated (using m_1 or m_2 to estimate μ_m), PPP appears to occur with this error model.

In one dimension, the analogous ad hoc procedure for a sample of size n from a normal distribution, is to assume $\mu \sim N(\overline{y}, \sigma^2/n)$ in the case of unknown μ and known variance σ^2 with $\overline{y} = \sum_{i=1}^{n} y_i/n$. The formally correct posterior for μ depends on the likelihood and on the prior, but for a normal likelihood $y \sim N(\mu, \sigma^2)$ and a normal prior $\mu \sim N(\mu_{\text{p}}, \tau^2)$, as the prior variance $\tau^2 \to \infty$, the prior becomes "flat," and the ad hoc procedure is well known to be correct in an asymptotic sense.

Although this ad hoc approach in two dimensions is a recipe that can be followed to produce $\hat{\mu}$, it is statistically incomplete. It can be completed and justified by adopting a two dimensional prior for μ. That is, to regard y_1, y_2 pairs as being μ_1, μ_2 pairs requires that y_1, y_2 pairs be interpreted as being observations from the posterior distribution for μ_1, μ_2 in a Bayesian sense. Second, to formalize the notion of restricting attention to bivariate points (μ_1, μ_2) lying very near the diagonal, we must define a prior probability distribution. Specifically, assume

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N(\mu, \Sigma) \tag{5}$$

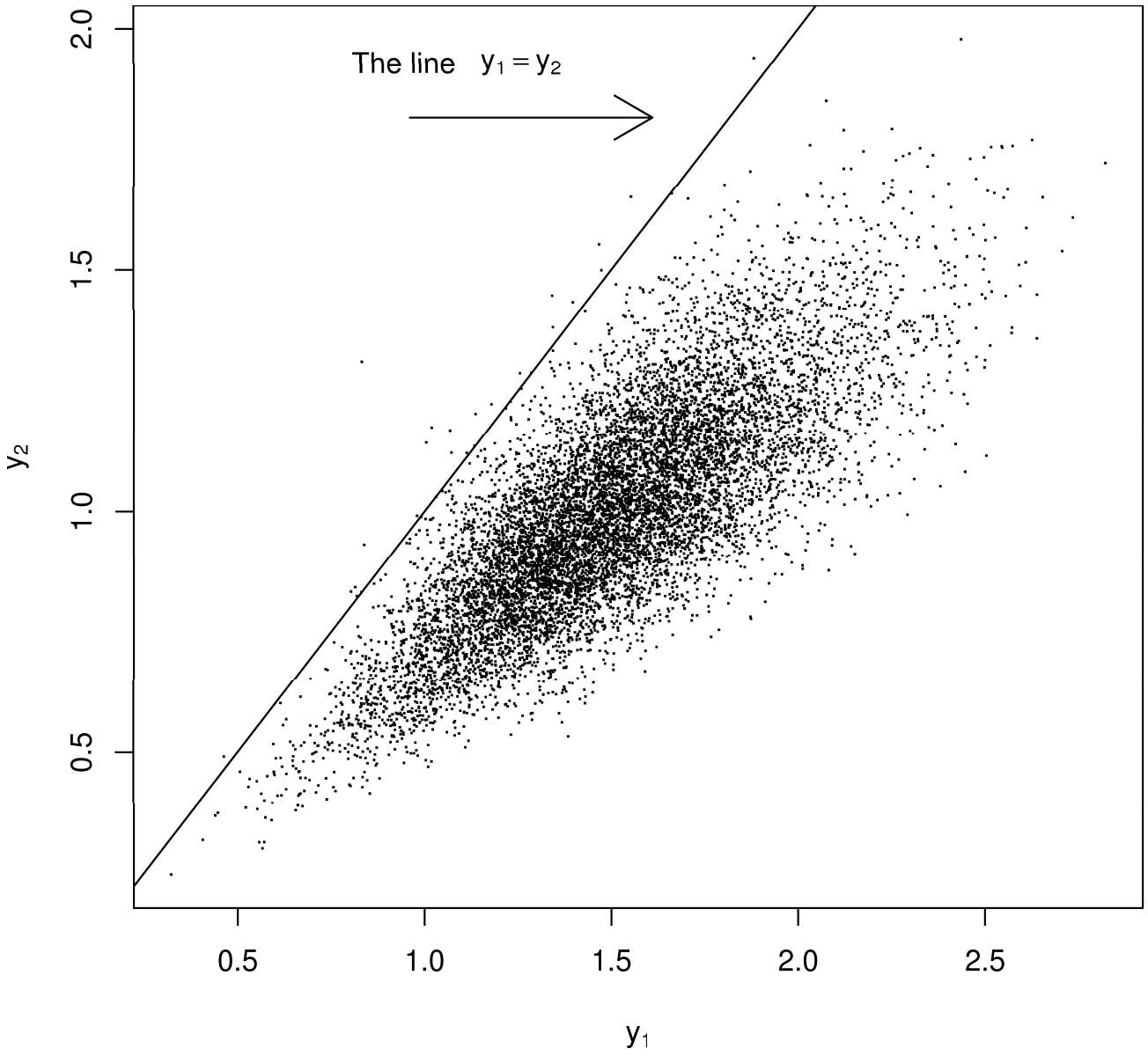
where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Let the prior for μ in Equation 5 be

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N(\mu_{\text{prior}}, \Sigma_{\text{prior}}) \tag{6}$$

Figure 3. Example ABC. The ad hoc "go along the diagonal" approach in [8].



where

$$\mu_{\text{prior}} = \begin{pmatrix} \mu_{\text{p1}} \\ \mu_{\text{p2}} \end{pmatrix} \text{ and } \Sigma_{\text{prior}} = \begin{pmatrix} \sigma^2_{\text{p1}} & \rho\sigma_{\text{p1}}\sigma_{\text{p2}} \\ \rho\sigma_{\text{p1}}\sigma_{\text{p2}} & \sigma^2_{\text{p2}} \end{pmatrix}.$$

The recipe to restrict attention to simulated values satisfying $y_1 \approx y_2$ is ad hoc, but for a given choice of ϵ to define $y_1 \approx y_2$ as satisfying $|y_1 - y_2| < \epsilon$, there is a defensible Bayesian posterior corresponding to a choice for Σ_{prior}. And, given

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{7}$$

the posterior distribution is

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N(\mu_{\text{post}}, \Sigma_{\text{post}}) \tag{8}$$

where

$$\mu_{\text{post}} = (\Sigma^{-1} + \Sigma^{-1}_{\text{prior}})^{-1}(\Sigma^{-1} y + \Sigma^{-1}_{\text{prior}} \mu_{\text{prior}})$$

and

$$\Sigma_{\text{post}} = (\Sigma^{-1} + \Sigma^{-1}_{\text{prior}})^{-1} \tag{9}$$

For example, choosing

$$\Sigma_{\text{prior}} = \begin{pmatrix} 100 & 100(1 - .01^j) \\ 100(1 - .01^j) & 100 \end{pmatrix} \tag{10}$$

results in (μ_1, μ_2) estimates of (1.49, 1.00), (1.08, 0.92), and (0.89, 0.89) for $j = 1, 2,$ and 3 or larger, respectively, for any moderate value of μ_{prior} such as 0.1 to 10. The prior can be chosen such that the resulting estimate of μ_1 is within a small ϵ of the estimate of μ_2. Notice that the (0.89, 0.89) values for the (μ_1, μ_2) estimate agrees with the GLS method for $j \geq 3$ in Equation 11 as claimed for the ad hoc approach of choosing (y_1, y_2) pairs along the diagonal [8].

This Bayesian approach revises the ad hoc recipe and is easily implemented using MCMC for any likelihood. It is sometimes difficult to calculate the likelihood for complicated combinations of distributions involving for example sums and products of random variables (which can arise in modifications of the prescription to generate y_1, y_2 pairs). An appeal of the modified ad hoc approach is that the likelihood does not need to be computed. Instead, one need only be able to generate many y_1, y_2 pairs. This is the same appeal of the ABC approach in our context.

To summarize this section, the ad hoc approach in [8] regards simulated (y_1, y_2) pairs from any distribution as providing an estimate of μ by specifying a small ϵ and computing the mean of all scalar-valued y_1 and y_2 satisfying $|y_1 - y_2| < \epsilon$ (or by finding the value of μ that maximizes the likelihood of the scalar data that arise from concatenating all y_1 and y_2 values that are similar). This approach has a corresponding formally correct Bayesian interpretation, but is easier to implement than the corresponding correct Bayesian implementation with a particular choice of the prior pdf.

For Example 5.1 in Section 5, for which the distributions for ϵ_{R1}, ϵ_{R2}, and ϵ_S are normal, the estimates arising from the Bayesian posterior maximum probability and the posterior mean are 0.89 and 0.89, respectively, both the same as the GLS estimate.

- Example 6.1.

If the distributions for ϵ_{R1}, ϵ_{R2} , and for ϵ_S are centered and scaled product normals, the estimates for all small values of ϵ (ϵ less than approximately 0.01) are 1.10 and 1.09, respectively from the Bayesian posterior maximum probability and the posterior mean.

If the distributions for ϵ_{R1}, ϵ_{R2} are centered and scaled product normal and the distribution for ϵ_S is a centered and scaled gamma, the estimates for all small values of ϵ are 1.33 and 1.18, from the Bayesian posterior maximum probability and the posterior mean, respectively.

The fact that the estimates can lie between y_1 and y_2 appeals in some way.

- Example 6.2

  The ad hoc method can be formally defended by the Bayesian framework just described. Therefore, the ad hoc approach is included among the candidates whose RMSEs were given in Example 5.2 from Section 5. In the second part of Example 5.2, there are 100 simulations of 10,000 observations of (y_1, y_2) pairs (for which the RMSE of GLS and ABC are 0.22 and 0.15, respectively). The the RMSE of the ad hoc approach is 0.24 (repeatable to within ± 0.01).

For Example 6.2, although the RMSEs from low to high (low is good) are ABC, GLS, and alternate Bayes, for this example, we do not make any general performance claims ranking GLS, ABC, and the modified ad hoc ("alternate Bayes") approaches. Instead, the goal is to present statistically defensible approaches and demonstrate that they do perform differently.

Figure 4 illustrates the ad hoc and revised procedures in the case of the normal distribution. The top plot shows the accepted (μ_1, μ_2) pairs in the metrop MCMC (in red), which tightly follow the diagonal line. The bottom plot shows the estimated density of the accepted pairs using the Bayesian strategy and using the ad hoc strategy. Notice that the two densities are not the same. However, in this case, both methods arrive at $\hat{\mu} = 0.89$ as the estimate for μ.

7. Conclusion

There will almost always be estimation error in $\hat{\Sigma}$, and often the measurement errors are non-normal. Therefore, we considered the following three topics: (1) alternatives to GLS when there is estimation error in $\hat{\Sigma}$, (2) approximate Bayesian computation [22] to provide estimators other than GLS that use the estimated likelihood in the case of non-normal error models that make ML estimation difficult, and (3) completion of an incompletely specified ad hoc approach to deal with non-normal likelihoods [8].

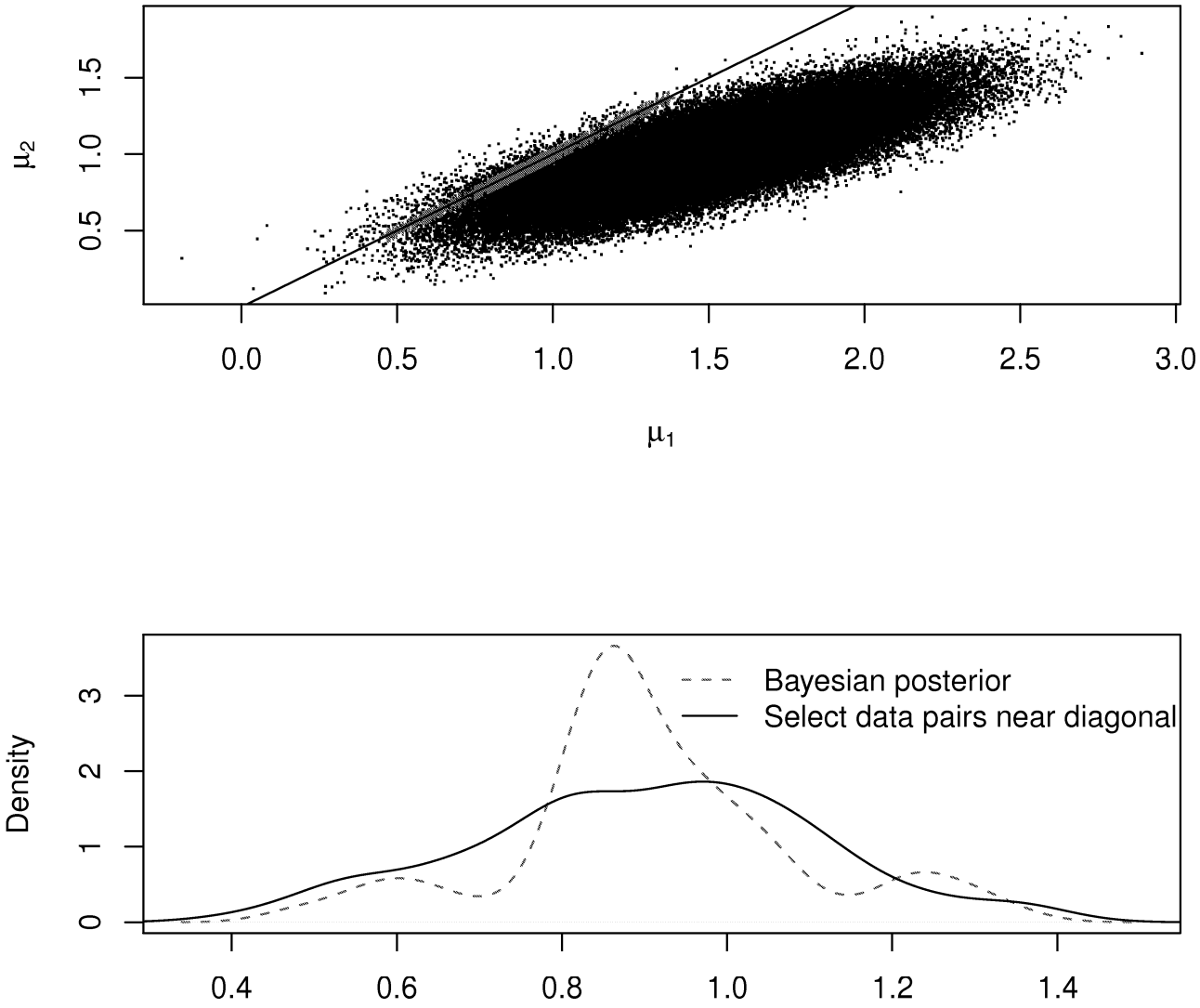
Regarding (1), it was illustrated in the PPP case that weighted estimates do not always outperform equally-weighted estimates when there is estimation error in the weights. We have already noted that estimation of Σ in Equation 1 introduces apparent PPP when PPP does not actually occur.

Regarding (2), ABC based on density estimation to estimate the required likelihood was introduced. We showed that for some likelihoods, Bayesian estimation can outperform GLS, and that for some data realizations, the ABC estimator fell within the range of the simulated x_1, x_2 data pairs. Of course GLS provides a good estimate $\hat{\mu}$ in general because of its well-known BLUE property (and UMVUE if the data is normal) and in particular for the PPP problem if the covariance Σ is well known.

Regarding (3), ad hoc estimators based on incomplete statistical reasoning did produce values that lie within the data range [8]. We modified the incomplete approach by another use of Bayesian reasoning

that involved a two-component prior. We concur that for some data realizations, the ad hoc estimator can lie within the data range, but in the examples considered, the RMSE of this estimator was slightly higher than the RMSE of the GLS or ABC estimators.

Figure 4. Example ABC. The *ad hoc* approach and the completed Bayesian approach using a bivariate prior.



Acknowledgements

We acknowledge the U.S. NNSA, and the Next Generation Safeguards Initiative (NGSI) within the U.S. Department of Energy.

References

1. Peelle, R. *Peelle's Pertinent Puzzle*; Oak Ridge National Laboratory: Oak Ridge, TN, USA, 1987.
2. Burr, T.; Kawano, T.; Talou, P.; Hengartner, N.; Pan, P. Defense of the Least Squares Solution to Peelle's Pertinent Puzzle. *Algorithms* **2011**, *4*, 28–39.

3. *International Evaluation of Neutron Cross Section Standards*; International Atomic Energy Agency: Vienna, Austria, 2007.
4. Zhao, Z.; Perey, R. *The Covariance Matrix of Derived Quantities and Their Combination*; ORNL/TM-12106; Oak Ridge National Laboratory: Oak Ridge, TN, USA, 1992.
5. Chiba, S.; Smith, D. *A Suggested Procedure for Resolving an Anomaly in Least-Squares Data Analysis Known as Peelle's Pertinent Puzzle and the General Implications for Nuclear Data Evaluation*. ANL/NDM-121; Argonne National Laboratory: Argonne, IL, USA, 1991.
6. Chiba, S.; Smith, D. Some Comments on Peelle's Pertinent Puzzle, JAERI-M 94-068. In *Proceeding of Symptoms on Nuclear Data*; Japan Atomic Energy Research Institute: Tokai, Japan, 1994; p. 5.
7. Chiba, S.; Smith D. Impacts of Data Transformation on Least-Squares Solutions and Their Significance in Data Analysis and Evaluation. *J. Nucl. Sci. Technol.* **1994**, *31*, 770–781.
8. Hanson, K.; Kawano, T.; Talou, P. Probabilistic Interpretation of Peelle's Pertinent Puzzle. In *Proceeding of International Conference of Nuclear Data for Science and Technology*; Santa Fe, NM, USA, 26 September–1 October 2004; Haight, R.C., Chadwick, M.B., Kawano, T., Talou, P., Eds.; American Institute of Physics: College Park, MD, USA, 2005.
9. Sivia, D.; Data Analysis—A Dialogue With The Data. In *Advanced Mathematical and Computational Tools in Metrology VII*; Ciarlini, P., Filipe, E., Forbes, A.B., Pavese, F., Perruchet, C., Siebert, B.R.L., Eds.; World Scientific Publishing Co.: Singapore, Singapore, 2006; pp. 108–118.
10. Jones, C.; Finn, J.; Hengartner, N. Regression with Strongly Correlated Data. *J. Multivar. Anal.* **2008**, *99*, 2136–2153.
11. Finn, J.; Jones, C.; Hengartner, N. *Strong Nonlinear Correlations, Conditional Entropy, and Perfect Estimation*; AIP Conference Proceedings 954; American Institute of Physics: College Park, MD, USA, 2007.
12. Kawano, T.; Matsunobu, H.; Murata, T.; Zukeran, A.; Nakajima, Y.; Kawai, M.; Iwamoto, O.; Shibata, K.; Nakagawa, T.; Ohsawa, T.; Baba, M.; Yoshida, T. Simultaneous Evaluation of Fission Cross Sections of Uranium and Plutonium Isotopes for JENDL-3.3. *J. Nucl. Sci. Technol.* **2000**, *37*, 327–334.
13. Kawano, T.; Hanson, K.; Talou, P.; Chadwick, M.; Frankle, S.; Little, R. Evaluation and Propagation of the Pu^{239} Fission Cross-Section Uncertainties Using a Monte Carlo Technique. *Nucl. Sci. Eng.* **2006**, *153*, 1–7.
14. Oh, S. Monte Carlo Method for the Estimates in a Model Calculation. In *Proceedings of Korean Nuclear Society Conference*; Yongpyung, Korea, october 2004.
15. Oh, S.; Seo, C. *Box-Cox Transformation for Resolving Peelle's Pertinent Puzzle in Curve Fitting*. PHYSOR-2004; The Physics of Fuel Cycles and Advanced Nuclear Systems: Global Developments, Chicago, IL, USA, April 25–29, 2004.
16. Christensen, R. *Plane Answers to Complex Questions, The Theory of Linear Models*; Springer: New York, NY, USA, 1999; pp. 23–25.
17. Burr, T.; Frey, H. Biased Regression: The Case for Cautious Application. *Techometrics* **2005**, *47*, 284–296.
18. Johnson, R.; Wichern, D. *Applied Multivariate Statistical Analysis*; Prentice Hall: Upper Saddle River, NJ, USA, 1988.

19. R Foundation for Statistical Computing. R: A Language and Environment for Statistical Coputing, 2004. Avalable online: www.R-project.org (accessed on 17 June 2011).
20. Schmelling, M. Averaging Correlated Data. *Phys. Scr.* **1995**, *51*, 676–679.
21. Diggle, P.; Gratton, R. Monte Carlo Methods of Inference for Implicit Statistical Models, *J. R. Stat. Soc. B* **1984**, *46*, 193–227.
22. Marjoram, P.; Molitor, J.; Plagnol, V.; Tavare, S. Markov Chain Monte Carlo Without Likelihoods. *Proc. Nat. Acad. Sci. USA* **2003**, *100*, 15324–15328.
23. Plagnol, V.; Tavare, S. Approximate Bayesian Computation and MCMC. In *Proceedings of Monte Carlo and Quasi-Monte Carlo Methods*; National University of Singapore, Singapore, Singapore, 2002; pp. 99–114.
24. Blum, M. Approximate Bayesian Computation: A Nonparametric Perspective, *J. Am. Stat. Assoc.* **2010**, *105*, 1178–1187.
25. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2001.

Appendix

This appendix lists `R` functions.

Remark 1: The `mcmc` function in `R`, requires a log "likelihood" function.
Remark 2: Our ABC implementation uses the estimated pdf instead of an analytical likelihood.

```
mutrue = 1; sigma = .1
mu.grid = seq(mutrue-3*sigma,mutrue+3*sigma,length=100)
llik1 = numeric(length(mu.grid))
for(i in 1:length(mu.grid)) {
xtemp = rnorm(n=10^5,mean=mu.grid[i],sd=sigma)
temp.density = density(xtemp,n=1000)
llik1[i] = sum(log10(approx(temp.density$x,temp.density$y,xout=mu.grid)$y))
}
plot(mu.grid[5:95],llik1[5:95],xlab=expression(mu),ylab="log(likelihood)",type="l")
abline(v=1)
```

```
llik1.fun = function(mu,llik=llik1,mu.grid.use=mu.grid,prior.mean=0,prior.sd=10) {
llik.temp = approx(x=mu.grid.use,y=llik,xout=mu)$y
prior = log(dnorm(mu,mean=prior.mean,sd=prior.sd))
llik.temp = llik.temp + prior
if(is.na(llik.temp)) llik.temp = -10^10
if(llik.temp == -Inf) llik.temp = -10^10
llik.temp
}
```

```
library(mcmc) out1 = metrop(obj=llik1.fun,llik=llik1,initial=1,blen=10,scale=0.1,nbatch=1000)
plot(out1$batch) # basic diagnostic plot to check mcmc convergence
```

```
mean(out1$batch) # Result: approximately 1.0, the correct answer.

#product normal case
llik2.fun = function(mu,llik=llik2,mu.grid.use=mu.grid,prior.mean=0,prior.sd=10) {
llik.temp = approx(x=mu.grid.use,y=llik,xout=mu)$y
prior = log(dgamma(mu,shape=prior.mean/prior.sd,rate=1/prior.sd)) # true mean is nonzero
llik.temp = llik.temp + prior
if(is.na(llik.temp)) llik.temp = -10^10
if(llik.temp == -Inf) llik.temp = -10^10
llik.temp
}

x = c(1.5,1)
library(MASS) # kde2d for 2-dim kernel density estimation
mu.grid = seq(0.3,1.7,length=100)
n = 10^5 # observations from simulation for each value of mu in grid of mu values
llik5a = numeric(length(mu.grid))
nvar1 = 0.1134; nvar2 = 0.0504; ncov12 = 0.06
alpha = .7
svar = ncov12/alpha
rvar1 = nvar1 - svar
rvar2 = nvar2 - alpha^2 * svar
for(i in 1:length(mu.grid))
mu = mu.grid[i]
tempS = svar^.5 * rnorm(n) * rnorm(n)
x1temp = mu + tempS +rvar1^.5 * rnorm(n) * rnorm(n)
x2temp = mu + alpha * tempS + rvar2^.5 * rnorm(n) * rnorm(n)
temp.density = kde2d(x=x1temp,y=x2temp,lims=c(x[1],x[1],x[2],x[2]),n=1)
llik5a[i] = log(temp.density$z)
llik5a[llik5a==-Inf] = min(llik5a[llik5a != -Inf])
tempy = lokerns(x=mu.grid,y=llik5a,x.out=mu.grid)
llik5a = tempy$est
llik2[llik1a==-Inf] = min(llik2[llik2 != -Inf])
tempy = lokerns(x=mu.grid,y=llik1a,x.out=mu.grid)
llik2 = tempy$est
out2 = metrop(obj=llik2.fun,initial=1,llik=llik2,blen=10,scale=.3,nbatch=1000,prior.mean=1,prior.sd=1)
mean(out2$batch) # Result: approximately 0.89.
```