Stat 532 Final Project: Modeling an Oil Field Discovery Process

Kenny Flagg

December 4, 2015

Contents

1 Introduction

The locations and sizes of petroleum fields are valuable pieces of information in today's economy. One statistical approach to investigating the amount of oil in a region is to consider the process of finding oil fields as taking a random sample of the fields in a region where larger fields have higher probabilities of being sampled than smaller fields. Sinding-Larsen and Xu [3] used an empirical Bayes model of the oil field discovery process to study the distribution of undiscovered oil reservoirs in the Halten Terrace, off the shore of Norway. The primary research questions were as follows.

1. How many undiscovered oil fields exist in the region?
2. What is the total volume of petroleum present in the undiscovered fields?

The data consist of estimated sizes in millions of cubic meters of all 22 known petroleum reservoirs in the region, ranked in the order in which they were discovered (Figure 1 and Table 1).

The Halten Terrace contains a play, or collection of petroleum fields and prospective petroleum fields that share the same geological circumstances [4, p. 106]. It is comprised of two sub-plays (Figure 2). The western sub-play is overpressured, meaning that oil fields there are under high pressure and so exploratory drilling is considered high-risk. The eastern sub-play is considered to be normally pressured. More detailed geological information is available in Koch and Heum [1].

The discovery process model was previously applied to the eastern sub-play of the Halten Terrace by Sinding-Larsen and Chen [2]. Since then, the western sub-play has been more thoroughly explored and the size estimates for the previously known fields have been revised. The previous analysis was done using maximum likelihood estimation, and it was concluded that it was difficult to find estimates when the total number of fields was unknown. Sinding-Larsen and Xu were interested in repeating the analysis with the more recent data and incorporating expert knowledge about the total number of fields. They fit the discovery process model to a combined dataset where the entire region was treated as a single play, and they also fit the model separately to each sub-play. Their discussion emphasized results from the combined data.

In this project, I evaluated the discovery process model on simulated data using both the empirical Bayes priors of Sinding-Larsen and Xu and my own less-informative priors. I then repeated the analysis on the combined Halten Terrace data comparing both sets of priors.

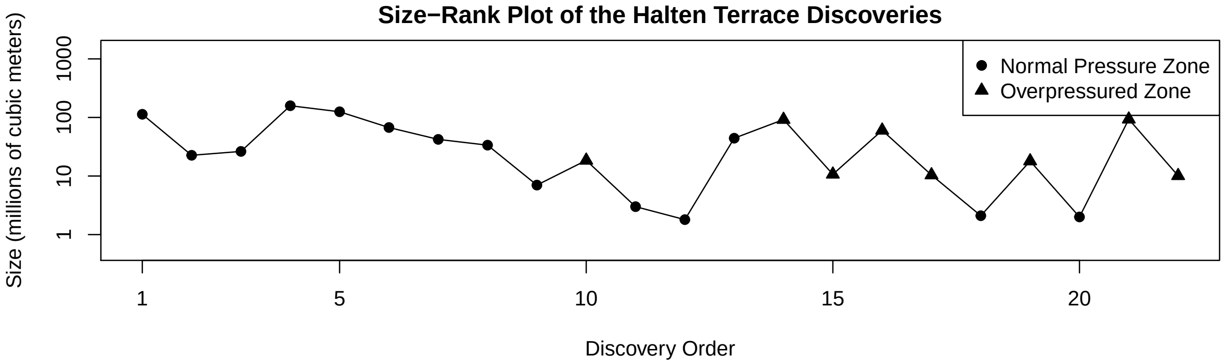

Figure 1: Plot of the sizes versus discovery order of the 22 fields discovered in the Halten Terrance as of 2000.

Field Name	Discovery Order	Size (2000 estimate)	Size (1992 estimate)	Pressure Zone
Midtgard	1	113.20	101.00	Normal Pressure
Tyrihans S	2	22.60	15.50	Normal Pressure
Tyrihans N	3	26.30	18.90	Normal Pressure
Smoerbukk	4	158.50	125.30	Normal Pressure
Heidrun	5	125.10	109.40	Normal Pressure
Smoebukk S	6	67.30	49.70	Normal Pressure
Njord	7	42.20	36.00	Normal Pressure
Mikkel	8	33.63	21.10	Normal Pressure
Trestakk	9	7.00	3.90	Normal Pressure
Alve	10	18.50		Overpressured
6507/8-4	11	3.00	20.40	Normal Pressure
6407/8-2	12	1.80		Normal Pressure
Lavrans	13	44.20		Normal Pressure
Kristin	14	92.00		Overpressured
Ragnfrid	15	10.70		Overpressured
Sharv	16	60.27		Overpressured
Erled	17	10.40		Overpressured
6407/9-9	18	2.10		Normal Pressure
6507/7-6	19	18.00		Overpressured
6407/7-6	20	2.00		Normal Pressure
6506/6-1	21	93.20		Overpressured
Svale	22	10.10		Overpressured

Table 1: Estimated sizes of discovered oil fields in the Halten Terrance.

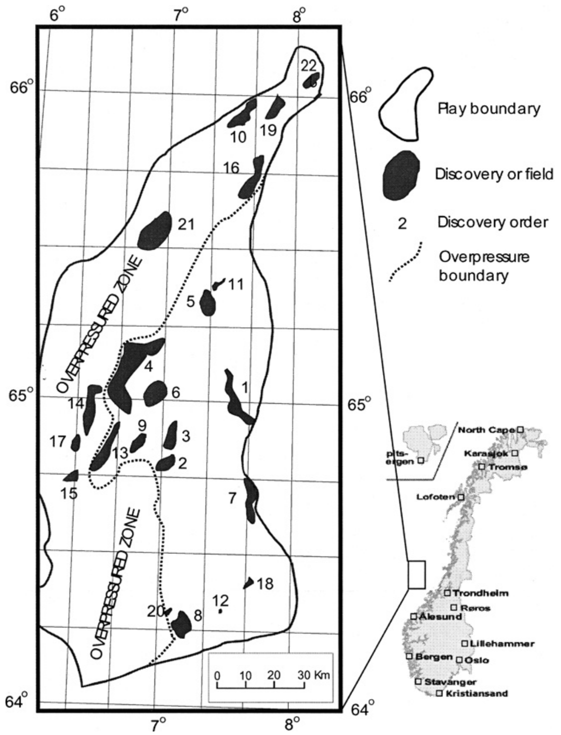

Figure 2: Map of Halten Terrace discoveries and pressure zones from Sinding-Larsen and Xu [3].

2 Model

The discovery process model assumes that the sizes of the oil fields in a play come from a common lognormal population, and that the discovered fields are a random sample drawn without replacement from that population. The probability of including a field in the sample is assumed to be proportional to the field's volume raised to a power.

Let N be the unknown number of fields in the play, and let n be the number of fields that have been discovered. The volumes of the discovered fields are denoted Y_j, $j = 1, \ldots, n$ and the volumes of the undiscovered fields are denoted S_k, $k = n+1, \ldots, N$. These values share the distribution

$$\log(Y_j), \log(S_k) \sim \mathrm{N}(\mu, \sigma^2)$$

where μ and σ^2 are the unknown mean and variance of the log-sizes.

The probability of selecting the sample $(Y_1, \ldots, Y_n)$ from the population $(Y_1, \ldots, Y_n, S_{n+1}, \ldots, S_N)$ is

$$\binom{N}{n} \prod_{i=1}^{n} \left(\frac{Y_i^\beta}{\sum_{j=i}^{n} Y_j^\beta + \sum_{k=n+1}^{N} S_k^\beta} \right)$$

where β is an unknown quantity known as the size-bias parameter.

Let $\log\mathrm{N}(X|\mu, \sigma^2)$ denote the density function of a lognormal random variable X with $E(\log(X)) = \mu$ and $Var(\log(X)) = \sigma^2$. Then the joint distribution of the Y_j, the S_k, and the parameters is

$$\begin{aligned} p(N, \beta, \mu, \sigma^2, S_{n+1}, \ldots, S_N, Y_1, \ldots, Y_n) = p(N, \beta, \mu, \sigma^2) p(S_{n+1}, \ldots, S_N | N, \mu, \sigma^2) \\ \times\, p(Y_1, \ldots, Y_n | N, \beta, \mu, \sigma^2, S_{n+1}, \ldots, S_N) \end{aligned}$$

where $p(N, \beta, \mu, \sigma^2)$ is the joint prior distribution of the parameters,

$$p(S_{n+1}, \ldots, S_N | N, \mu, \sigma^2) = \prod_{k=n+1}^{N} \log\mathrm{N}(S_k | \mu, \sigma^2)$$

is the joint distribution of the unobserved field sizes, and

$$p(Y_1, \ldots, Y_n | N, \beta, \mu, \sigma^2, S_{n+1}, \ldots, S_N) = \binom{N}{n} \prod_{i=1}^{n} \left(\frac{Y_i^\beta}{\sum_{j=i}^{n} Y_j^\beta + \sum_{k=n+1}^{N} S_k^\beta} \log\mathrm{N}(\mu, \sigma^2) \right)$$

is the likelihood.

2.1 Empirical Bayes Priors Used by Sinding-Larsen and Xu

Xu and Sinding-Larsen published a companion paper [5] giving the details of how they chose their prior distributions. They gave independent distributions to N, β, μ, and σ^2. An empirical Bayes approach was used, where the data were used to select informative priors.

The prior distributions of μ and σ^2 were chosen to approximate the sampling distributions of the maximum likelihood estimators. A Normal distribution with mean 2.38 and variance 0.54 was used for μ, and a Gamma distribution with mean 2.89 and variance 2.13 was selected for σ^2 (Figure 3).

The prior distribution for β was derived from the empirical equation

$$b = b_0 \left(1 - e^{-\sigma\beta}\right)$$

where b is the slope of the least-squares line fitting log-size to rank in the discovery order and b_0 is the limit of the slope as β approaches infinity. A Uniform prior was used for b_0. Endpoints of -0.213 and -0.088 were selected by examining the size-rank plot and considering possible slopes of a line connecting the smallest and largest sizes on the plot. The sampling distribution of the least-squares estimator of b was used as the prior for b, which for the Halten Terrace data was estimated as a Normal distribution with mean -0.0925 and variance 0.00193. They advised replacing σ in the equation with the sample standard deviation of the $\log(Y_i)$ values, in this case 1.41. These distributions, and the resulting distribution of β, appear in Figure 4.

The only prior that was not directly based on the data being analyzed was the prior for N. This parameter was modeled hierarchically with a Binomial$(M-n, \pi)$ distribution. The hyperparameter M represents the number of prospective oil fields in the region that could ever be explored, and π is the probability that a prospect actually contains oil. The prior distribution of M was constructed from percentiles elicited from experts. A Beta$(2.3, 4.28)$ distribution was used for π, chosen by examining historical data about the number of prospects being explored at once and the number of those that were found to contain petroleum. The prior distributions of M, π, and N are plotted in Figure 5.

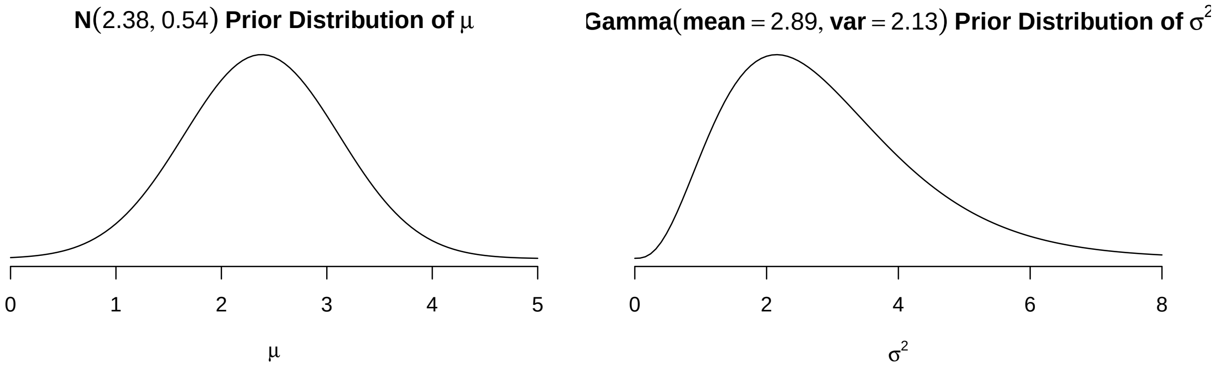

Figure 3: Prior distributions for μ and σ^2 used by Sinding-Larsen and Xu.

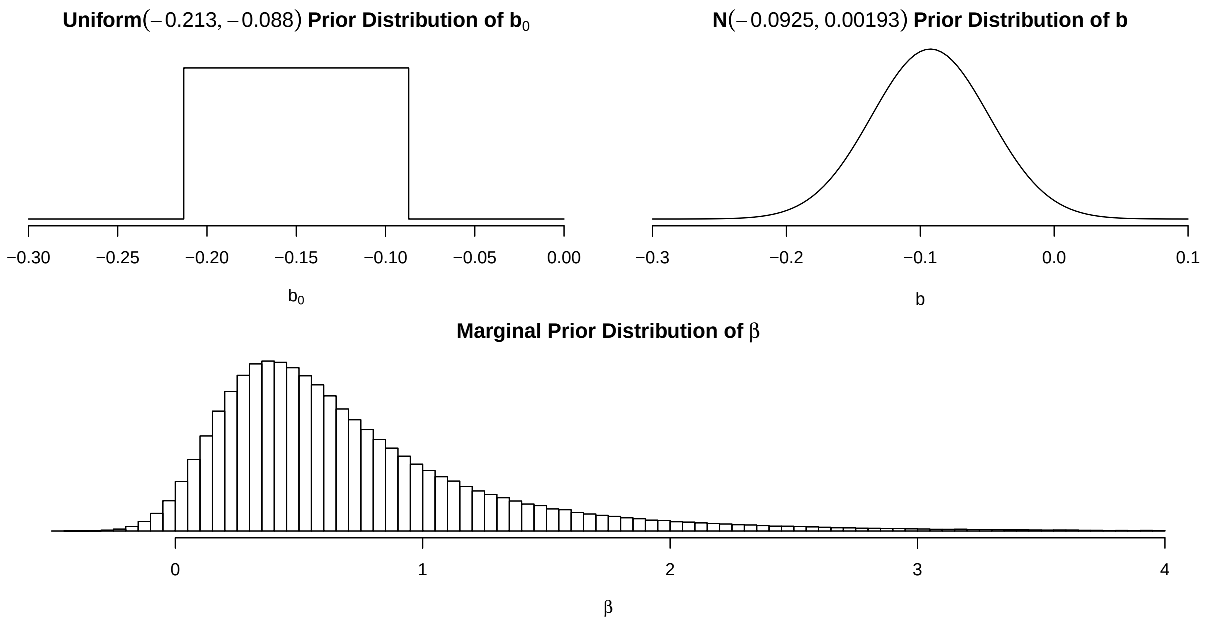

Figure 4: Prior distributions for b_0 and b used by Sinding-Larsen and Xu, and the resulting marginal prior distribution of β.

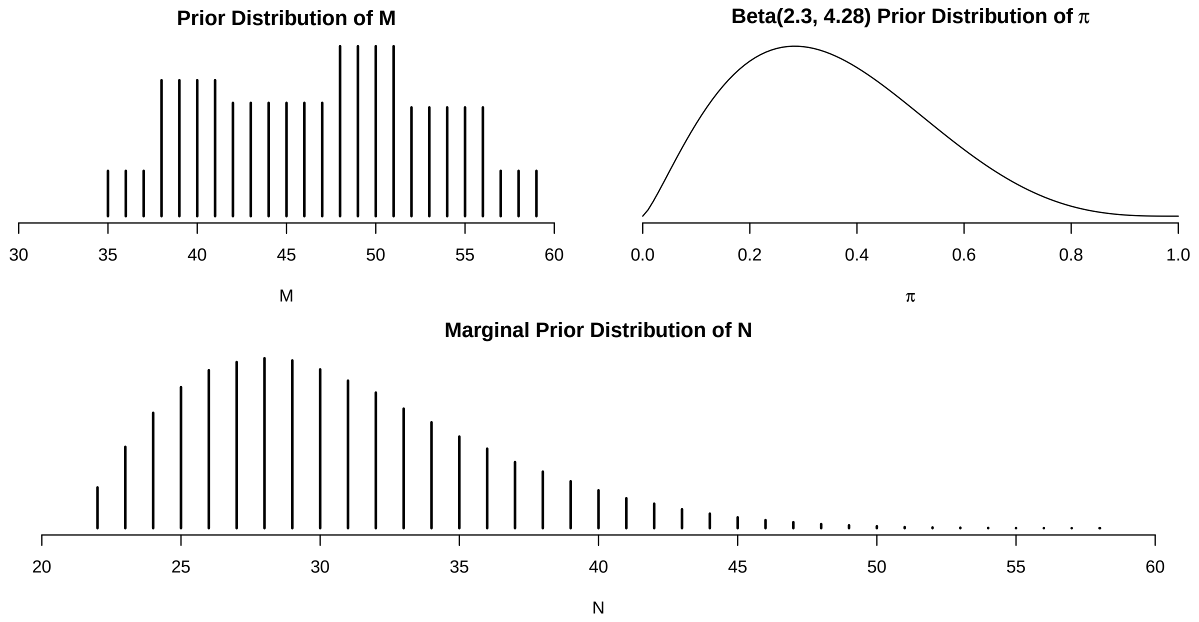

Figure 5: Prior distributions for M and π used by Sinding-Larsen and Xu, and the resulting marginal prior distribution of N.

2.2 Weakly-Informative Priors

I took a fully Bayesian approach, using prior distributions that reflect only the knowledge available to me separately from the data. Where I had no specific knowledge, I chose priors that were meant to have minimal influence on the posterior distribution.

I used weakly-informative priors for μ and σ^2. Oil fields contain volumes on the order of 1 million cubic meters to hundreds of millions of cubic meters. On the natural log scale, 1 and 100 correspond to values of 0 and 4.6, so I used a $\text{Cauchy}(\text{location} = 2, \text{scale} = 1.5)$ prior for μ. This distribution is relatively flat on the interval $(0, 4.6)$ and has 37.1% of its mass outside of that interval, so it will not constrain μ very heavily. If the log-sizes are mostly between 0 and 4.6, then σ^2 is probably less than 4. The Half-Cauchy distribution with scale 3 has 59.0% of its mass below 4 and is rather flat. These distributions are illustrated in Figure 6.

Using the empirical equation to relate β to the slope of a least-squares line seemed like an unnecessarily convoluted procedure. Instead, I opted to place a prior distribution directly on β. To get a sense of what range would be reasonable for β, I simulated several discovery sequences by permuting the Halten Terrace discoveries using several different values of β (Figure 7). Negative values imply that smaller fields tend to be found first. A value of zero represents no relationship between a field's size and discovery order rank. Negative values do not make sense, but I do not want to exclude them if in fact the posterior distribution is centered at zero. Positive β values yield the expected relationship where larger sizes appear earlier in the discovery order. A value of 2 resulted in a strong linear relationship between log-size and rank in the discovery order. We expect variability in size across the discovery order, so β is probably less than 2. To be cautiously vague, I chose a $\text{Uniform}(-1, 2)$ prior distribution. Curiously, the prior used by Sinding-Larsen allows some unrealistically large positive value which I chose to omit (Figure 7, bottom right).

A sample contains little information about N, so the prior distribution must provide constraints. I used the $\text{Binomial}(M - n, \pi)$ distribution since it reflects the way that experts think about the possible number of oil fields. The experts believe the total number of prospects that will be explored to be between 35 and 59, so I used a discrete uniform distribution to set all of these values as equally likely. I chose a $\text{Beta}(1, 1)$ distribution to reflect a lack of knowledge about π. As a result, my prior distribution for N is much more spread out than the prior of Sinding-Larsen and Xu (Figure 8).

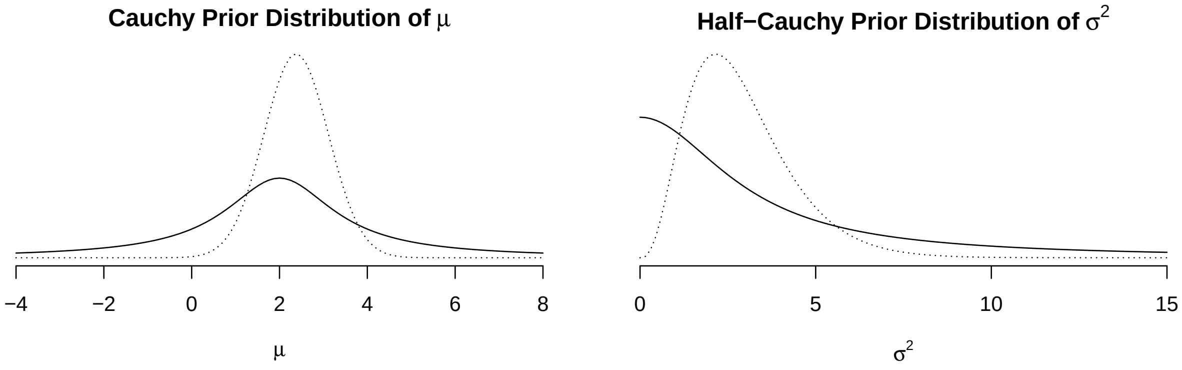

Figure 6: Weakly-informative prior distributions for μ and σ^2. For comparison, the distributions used by Sinding-Larsen and Xu are shown as dotted curves.

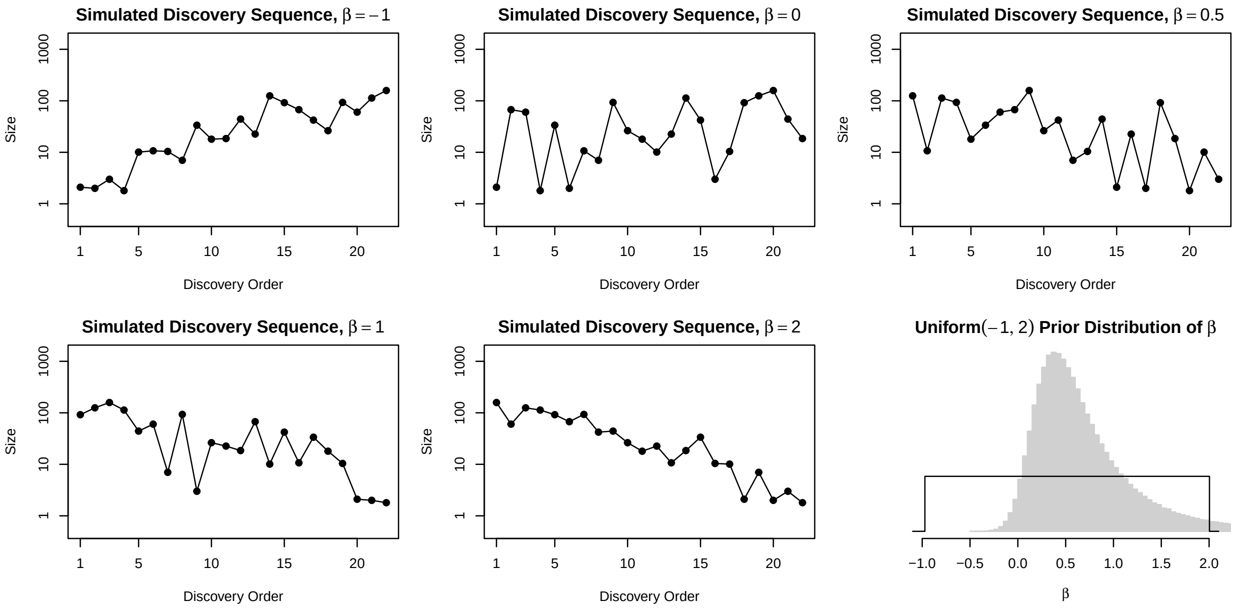

Figure 7: Simulated discovery sequences using the Halten Terrace size estimates and various values of β, and a weakly-informative prior distribution for β. The prior distribution used Sinding-Larsen and Xu is shown in grey.

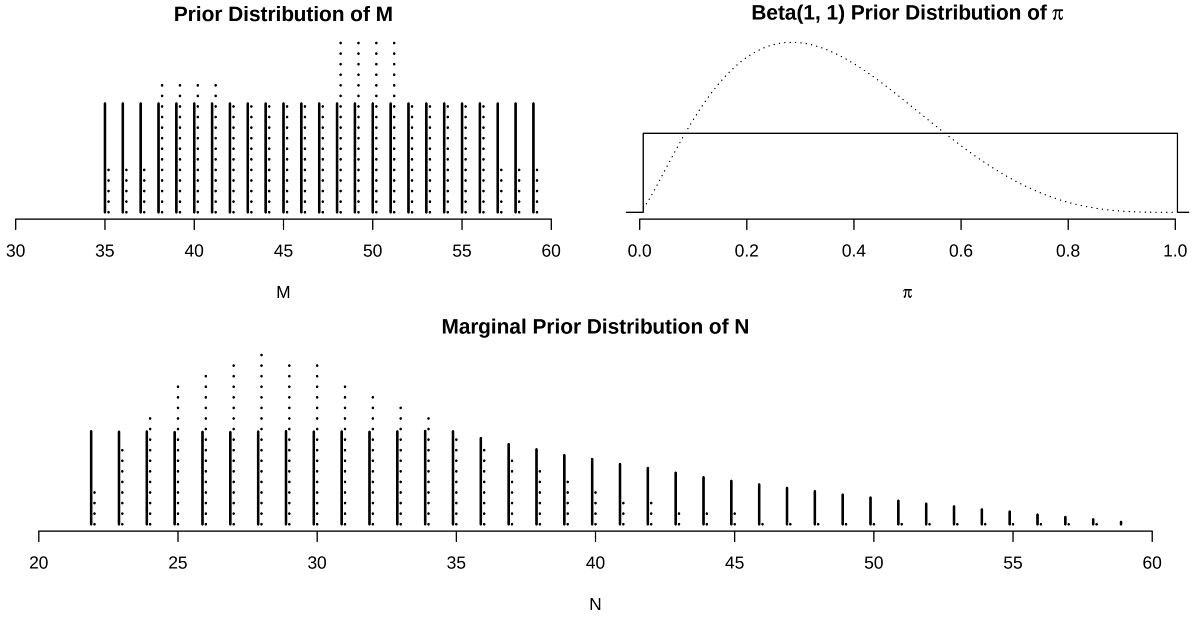

Figure 8: Less-informative prior distributions for M, π, and N. The dotted lines show the priors used by Sinding-Larsen and Xu.

3 Simulation of Artificial Data

The discovery process model assumes the data are a random sample, so simulation from this model was straightforward. For simplicity, I treated the two sub-plays as homogeneous and equal-sized, and assumed they were being explored at the same time so that any discovery was equally likely to come from either sub-play. I set parameter values near the posterior means reported by Sinding-Larsen and Xu for the combined dataset. My chosen values were $N = 32$ total fields, $\beta = 0.45$ as the size-bias parameter, $\mu = 2.6$ as the mean of the log-sizes, and $\sigma^2 = 2.6$ for the variance of the log-sizes. To keep the simulated data as similar as possible to the real data, I kept $n = 22$ as the number of discovered fields.

I first drew 32 log-sizes from a $\mathrm{N}(2.6, 2.6)$ distribution, and then exponentiated those values to get field sizes. Next, I independently labeled each field size as being in the overpressured zone or the normally pressured zone by drawing from the values 1 or 2 with equal probability. I then simulated the discovery order by taking a sample $(i_1, \ldots, i_{32})$, without replacement, from the values $i = 1, \ldots, 32$ using $\text{size}_i^{0.45} / \sum_{j=1}^{32} \text{size}_j^{0.45}$ as the probability of selecting the value i. Finally, I placed the sizes in the order $(\text{size}_{i_1}, \ldots, \text{size}_{i_{32}})$, and set $Y_j = \text{size}_{i_j}$, $j = 1, \ldots, n$ as the discovered fields and $S_k = \text{size}_{i_k}$, $k = n + 1, \ldots, N$ as the undiscovered fields.

The simulated discovery sequence (Figure 9) had a decreasing trend in field size with some local runs of increasing sizes, but no runs were as long as the run of 5 consecutive decreases seen in the first half of the Halten Terrace discovery sequence. Across the simulated discovery sequence, the sizes had similar variability to the Halten Terrace data. However, all of the first 9 Halten Terrace discoveries occurred in the normally pressured zone because that zone was explored first. This characteristic is absent from the simulated data because the two zones were combined into one.

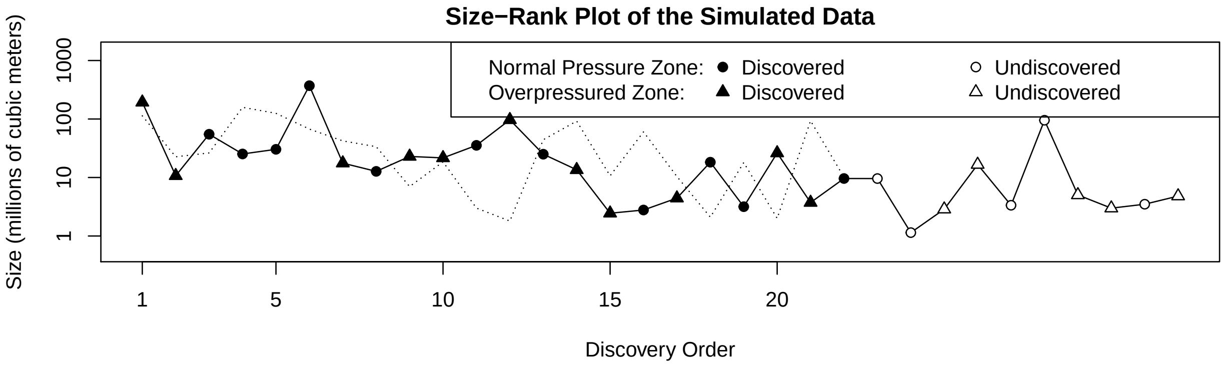

Figure 9: Plot of the sizes versus discovery order for the simulated data. The dotted line segments show the Halten Terrace discovery sequence.

4 Model Fitting and Results

Sinding-Larsen and Xu fit the discovery process model to the combined play, assuming that oil fields in the two sub-plays have a common distribution and that the same exploration and discovery process occurred in both sub-plays. They used a Monte Carlo simulation to obtain 80,000 draws from the joint posterior distribution,

$$p(N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N|Y_1,\ldots,Y_n)$$
$$= \frac{p(N,\beta,\mu,\sigma^2)p(S_{n+1},\ldots,S_N|N,\mu,\sigma^2)p(Y_1,\ldots,Y_n|N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N)}{\sum_N \int p(N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N,Y_1,\ldots,Y_n)d\beta d\mu d\sigma^2}.$$

The companion paper [5] described their Monte Carlo procedure as follows. For $i = 1,\ldots,d$,

1. Draws $N^{(i)}$, $\beta^{(i)}$, $\mu^{(i)}$, and $\sigma^{2(i)}$ from the prior distributions of N, β, μ, and σ^2.
2. Draw $S^{(i)}_{n+1},\ldots,S^{(i)}_{N^{(i)}}$ independently from the lognormal(μ,σ^2) distribution.
3. Compute the likelihood weight, $W_i = p(Y_1,\ldots,Y_n|N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N)$.

The marginal distribution of $Y_1,\ldots,Y_n$,

$$\sum_N \int p(N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N,Y_1,\ldots,Y_n)d\beta d\mu d\sigma^2,$$

is approximated by $\sum_{j=1}^d W_j$ so posterior expectations are computed as

$$E\left(g(N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N)|Y_1,\ldots,Y_n\right) = \frac{\sum_{i=1}^d W_i g(N,\beta,\mu,\sigma^2,S_{n+1},\ldots,S_N)}{\sum_{j=1}^d W_j}.$$

The posterior distributions were summarized with histograms, means, and standard deviations.

I fit the model using this Monte Carlo method. I found that the resulting collection of prior draws and weights was inconvenient to work with, so after taking d prior draws and computing their weights, I obtained a new sample of d posterior draws by sampling with replacement from the prior draws and using $W_i/\sum_{j=1}^d W_j$ as the probability of selecting $\left(N^{(i)},\beta^{(i)},\mu^{(i)},\sigma^{2(i)}\right)$. I took enough samples that, to three significant digits, quantities computed from the resampled draws were identical to those computed from the prior draws by the weighted sum method.

One issue that Sinding-Larsen and Xu did not mention is that the calculation of β requires $b > b_0$. They described independent prior distributions that allow $b \le b_0$. When fitting the model with their priors, I truncated the distribution of b in the following manner. For any i such that $b^{(i)} \le b_0^{(i)}$, I replaced $b^{(i)}$ and $b_0^{(i)}$ with new draws from the prior distributions. I repeated this until $b^{(i)} > b_0^{(i)}$ for all i.

Since this is not an iterative algorithm, convergence is not an issue that one would naturally be concerned with. However, the Monte Carlo draws contained very little information about the posterior distribution. I used histograms of the resampled draws as crude measures of convergence. A histogram that approximates a smooth curve suggests that the draws contain enough information to make posterior inferences. I initially took 100,000 draws from their priors, but the histograms showed spikes where a few values were were sampled much more often than others. I increased the number of draws to 500,000. When using my priors, I took 2 million draws. I would prefer larger samples than this, but the 2 million draws used up nearly all of my computer's available memory.

4.1 Simulated Data

The data were simulated using the reported posterior means of the parameters, so they can be used to investigate whether the model and the Monte Carlo simulation function correctly. If the procedure works as intended, the posterior draws should be centered near the known parameter values. Inconsistency between the posterior distributions and the parameter values indicates problems with either the model or the simulation.

I used the Monte Carlo simulation and resampling to obtain 500,000 posterior draws using the empirical Bayes prior and 2 million posterior draws using the weakly-informative prior. The posterior distributions for μ and σ^2 look similar for both priors, with the main difference being longer tails coming from the weakly-informative priors. The true values $\mu = 2.6$ and $\sigma^2 = 2.6$ are consistent with the posterior distributions (Figure 10).

The posterior distributions of β have similar centers, but the distribution resulting from the empirical Bayes prior is symmetric, while the weakly informative prior yields a posterior with a mild but noticeable left-skew. The empirical prior may have been too constraining on the left. A more distressing observation is that the true value, $\beta = 0.45$, is far to the left in both posteriors (Figure 11, left). It is not far enough into either tail to appear outright inconsistent, but it would be worthwhile to perform additional simulation studies to investigate whether this model produces reliable posterior distributions for β.

The posterior distributions of N look very much like the prior distributions (Figure 11, right). The Bayesian model does not seem to offer improvement over the maximum likelihood estimator. It should always be kept in mind that the Y_j contain little information about N, so any inference about N relies primarily on prior information.

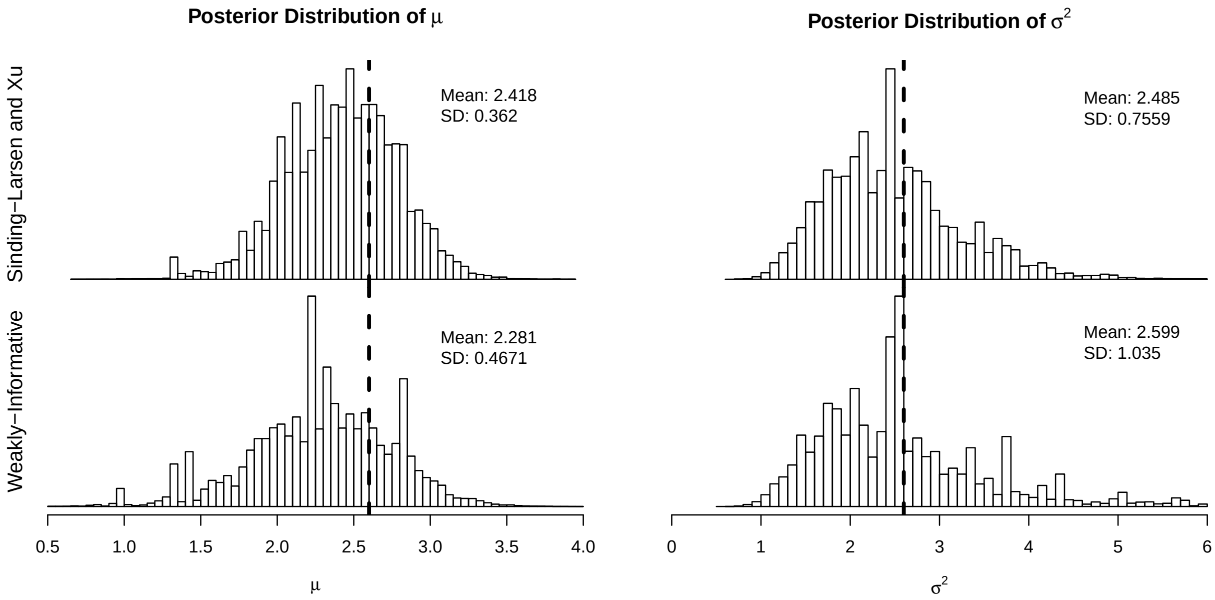

Figure 10: Posterior distributions of μ and σ^2 for the simulated data. The dashed vertical lines mark the values $\mu = 2.6$ and $\sigma^2 = 2.6$ from which the data were simulated.

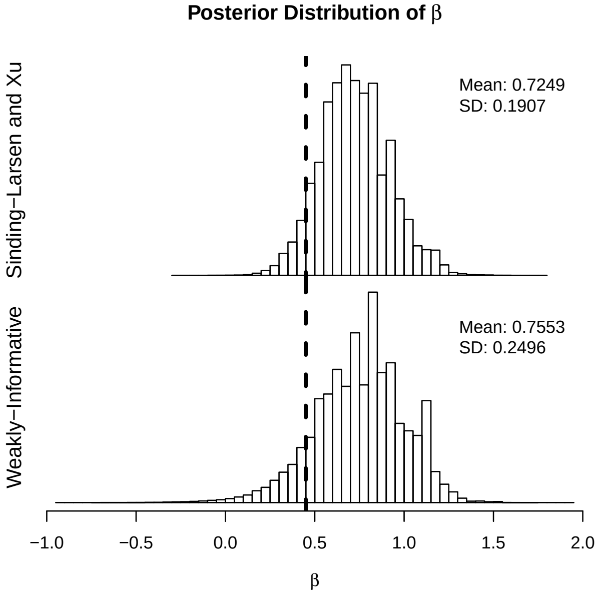

Figure 11: Posterior distributions of β and N for the simulated data. The dashed vertical lines mark the values $\beta = 0.45$ and $N = 32$ from which the data were simulated.

4.2 Halten Terrace Data

To check that I had correctly implemented the model int the way that Sinding-Larsen and Xu described, I repeated the analysis on the Halten Terrace data. I combined all the data into one group, ignoring zone. As with the simulated data, I used the Monte Carlo simulation and resampling to simulate 500,000 posterior draws using the Bayes prior and 2 million posterior draws using the weakly-informative prior.

When using their prior, my posterior distributions were essentially identical to those presented in the paper (Figures 12 and 13, top rows). The weakly-informative prior resulted in posteriors for μ, σ^2, and β with means that were nearly identical to the posterior means from the informative prior (Figure 12 and Figure 13, left). The weakly-informative prior lead to a posterior distribution for β that has a very prominent left-skew, which is not seen in the posterior resulting from the empirical Bayes prior.

As before, the posterior distributions of N strongly resemble the prior distributions (Figure 13, right). Posterior inference about the number of fields is dominated by the prior information, with very little contribution from the data.

Posterior inferences for μ and σ^2 are very similar whether the empirical Bayes prior is used or the weakly informative prior is used. This suggests that the effort of finding the empirical prior was unnecessary. Given the differences between the posterior distributions of β, and the possible discrepancy between β and its posterior from the simulated data, I would be skeptical about inferences for β based on this model.

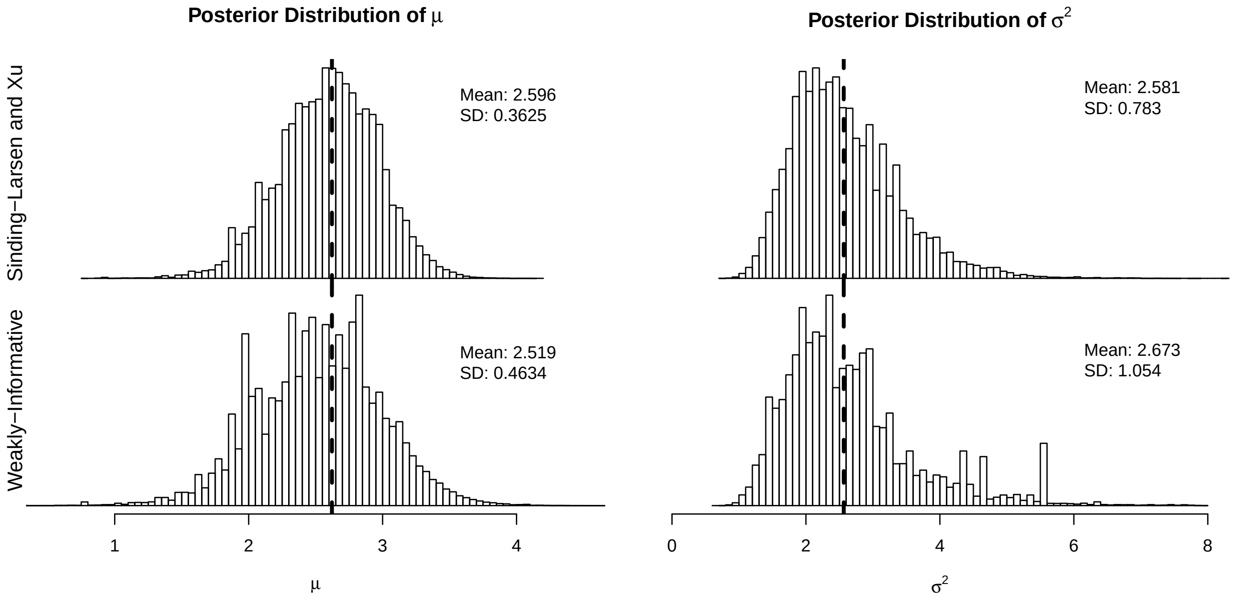

Figure 12: Posterior distributions of μ and σ^2 for the combined Halten Terrace data. The dashed vertical lines mark the values $\mu = 2.6217$ and $\sigma^2 = 2.5643$, the posterior means reported by Sinding-Larsen and Xu.

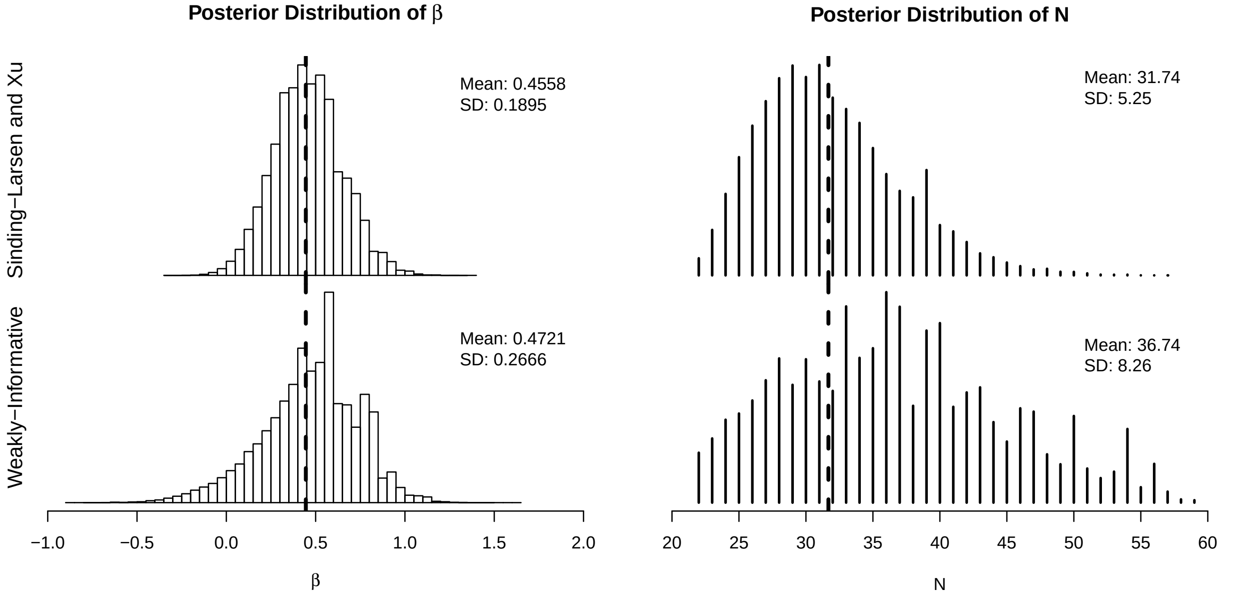

Figure 13: Posterior distributions of β and N for the combined Halten Terrace data. The dashed vertical lines mark the values $\beta = 0.44514$ and $N = 31.6789$, the posterior means reported by Sinding-Larsen and Xu.

5 Posterior Predictive Checks

Two quantities of great practical interest are the volume of the largest undiscovered oil field and the total volume of oil in all undiscovered fields. In reality these are unobserved, but simulation provides an opportunity to evaluate the performance of the discovery process model for predicting these values.

I performed posterior predictive checks in the following manner. I took a simple random sample of 10,000 of the posterior draws. For each draw i, I simulated a predicted discovery sequence $\left(Y_1^{(i)}, \ldots, Y_n^{(i)}, S_{n_1}^{(i)}, \ldots, S_N^{(i)}\right)$ from the parameter values $N^{(i)}$, $\beta^{(i)}$, $\mu^{(i)}$, and $\sigma^{2(i)}$ using the method of Section 3. Then for each predicted sequence, I computed max $\left(S_{n+1}^{(i)}, \ldots, S_N^{(i)}\right)$ and $\sum_{k=n+1}^{N} S_k^{(i)}$.

The model tends to underpredict these quantities (Figure 14). It is apparent the model does a poor job of describing the undiscovered fields. The very long tails in the posterior predictive distributions are also unsettling. Theoretically, the these are proper distributions because the prior distribution and likelihood are proper, and N has an upper bound. The long tails may be due to an inability of the Monte Carlo simulation to remove implausibly large draws, or it could result from the data being unable to reduce uncertainty in N. In any case, this is reason to believe that this model is not useful for its intended purpose of making inference about the amount of undiscovered petroleum.

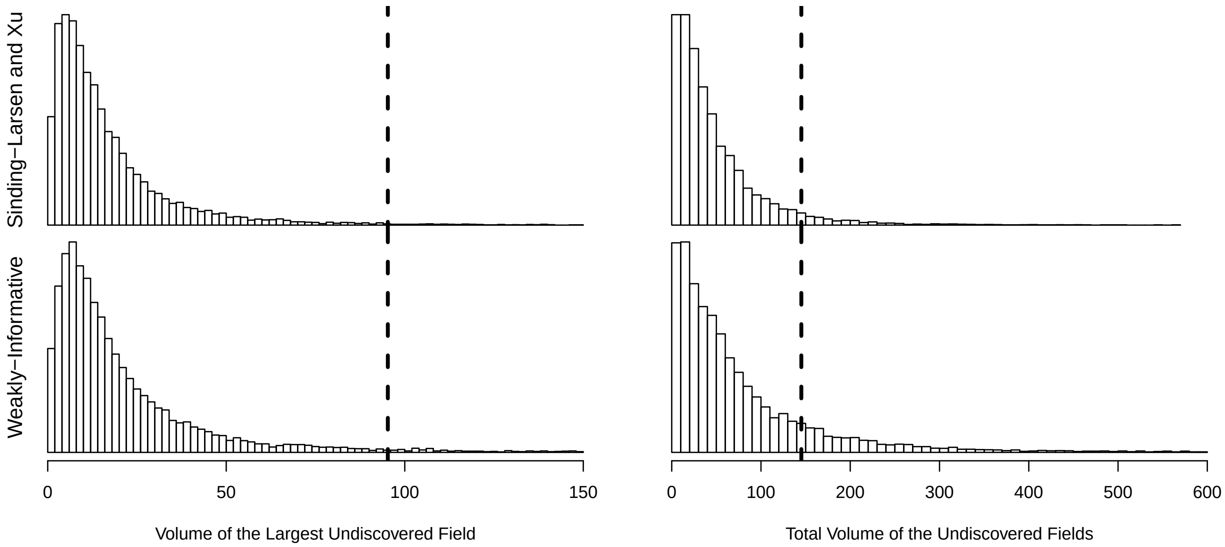

Figure 14: Distributions of $\max(S_{n+1}, \ldots, S_N)$ and $\sum_{k=n+1}^{N} S_k$ from 10,000 posterior predictive datasets. The dashed vertical lines mark the values $\max(S_{n+1}, \ldots, S_N = 95.26)$ and $\sum_{k=n+1}^{N} S_k = 145.2$ from the original simulated data.

6 Discussion

This analysis was hindered by an extremely inefficient method of finding the posterior distribution, and by a pervasive frequentist philosophy that ignored some of the features of Bayesian Statistics which would have been most helpful in this situation. Greater familiarity with Bayesian computation and Bayesian models and inference would have resulted in a more valid analysis.

6.1 Gibbs Sampler

The Monte Carlo method can easily generate many draws, but because the draws come from the prior distribution, most of the draws are in the tails of the posterior distribution. Almost all of the posterior information comes from a very small number of draws with large likelihood values. In order to see how much information was really contained in the Monte Carlo samples, I considered the prior distribution as a proposal distribution and performed rejection sampling. In the best case, when the Halten Terrace data were used with the empirical prior, 84 of 500,000 draws were accepted. For the simulated data with the vague prior, only 30 out of 2,000,000 draws were accepted! This is an extreme waste of computing effort.

Alternatively, it should be possible to use a Gibbs sampler to draw directly from the posterior. This may be difficult or impossible to implement in standard software, but a statistician familiar with Bayesian computation methods could construct a bespoke Gibbs sampler for the discovery process model. Below, I present the beginnings of such an algorithm.

The first step is to find the complete conditional distributions of the parameters. For μ and σ^2, these are

$$p(\mu|N, \beta, \sigma^2, S_{n+1}, S_N, Y_1, Y_n) \propto p(\mu) \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{k=n+1}^{N}(\log S_k - \mu)^2 + \sum_{j=1}^{n}(\log Y_j - \mu)^2\right)\right)$$

and

$$p(\sigma^2|N, \beta, \mu, S_{n+1}, S_N, Y_1, Y_n) \propto \frac{p(\sigma^2)}{(\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{k=n+1}^{N}(\log S_k - \mu)^2 + \sum_{j=1}^{n}(\log Y_j - \mu)^2\right)\right).$$

A Normal prior for μ and an inverse-Gamma prior for σ^2 would be conditionally conjugate. If other priors are used, Normal and inverse-Gamma distributions could be used as proposal distributions for Metropolis-Hastings sampling.

For β, the complete conditional distribution is

$$p(\beta|N, \mu, \sigma^2, S_{n+1}, S_N, Y_1, Y_n) \propto p(\beta) \prod_{i=1}^{n}\left(\frac{Y_i^\beta}{\sum_{j=i}^{n} Y_j^\beta + \sum_{k=n+1}^{N} S_k^\beta}\right).$$

Depending on the prior, it may be possible to find an efficient Metropolis-Hastings proposal distribution. In other situations, I would start by using the prior distribution as the proposal distribution.

The complete conditional distribution of N is

$$p(N|\beta, \mu, \sigma^2, S_{n+1}, S_N, Y_1, Y_n) \\ \propto p(N) \prod_{k=n+1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2} S_k} e^{-\frac{1}{2\sigma^2}(\log S_k - \mu)^2} \right) \binom{N}{n} \prod_{i=1}^{n} \left(\frac{1}{\sum_{j=i}^{n} Y_j^{\beta} + \sum_{k=n+1}^{N} S_k^{\beta}} \right)$$

so, for simplicity, I would use the prior as the proposal distribution. Since the data have little influence on the posterior distribution of N, I would not expect this proposal distribution to cause inefficiency of the algorithm. If μ and σ^2 are sampled efficiently, I expect the proposal distribution for β to be the most important factor in the efficiency of the sampler. However, the Gibbs sampler should provide an accurate description of the posterior distribution in fewer draws than Sinding-Larsen and Xu's Monte Carlo procedure.

6.2 Other Comments

Sinding-Larsen and Xu used methods that are ostensibly Bayesian, but presented them in a way that mimicked a frequentist analysis. Most of the posterior distributions were summarized only by histograms, posterior means, and posterior standard deviations. No posterior intervals were provided, and posterior probabilities were reported only for a few quantities of especially high interest to researchers. Additionally, the phrase "Bayesian estimate" appears numerous times throughout the paper in reference to a posterior mean. This suggests an underappreciation of the flexibility that is possible when making inference from a posterior distribution, or an attempt to write for an audience that was not expected to understand a Bayesian analysis.

They also fit two different models and then combined inferences inappropriately. The bulk of their analysis was based on fitting the model to the combined dataset. This was unjustified because the two sub-plays were subject to different discovery processes. They acknowledged this difference by then fitting the model separately to each sub-play and reporting that the posterior means for the parameters differed between the sub-plays. However, inferences about the site as a whole were based on the combined data rather than averages of the separate results. Instead, they could have considered a more versatile hierarchical model where each sub-play would have its own set of parameters, and prior information about the differences between the sub-plays could be incorporated.

7 Conclusion

My simulation study provides evidence that the discovery process model poorly describes the unobserved field sizes and the size-bias parameter. I have also shown that the complicated process of estimating an empirical prior is unnecessary because very similar results can be obtained from a weakly informative prior. Additionally, the Monte Carlo simulation proposed by Sinding-Larsen and Xu is extremely inefficient. I recommend that future analyses of petroleum discovery data from multiple sub-plays should consider using a hierarchical model fit by Gibbs sampling.

8 References

[1] Koch, J. O., and Heum, O. R., 1995, Exploration trends of the Halten Terrace, offshore mid-Norway: The potential role of mechanical compaction, pressure transfer and stress, in Hanslien, S., ed., *Petroleum exploration and exploitation in Norway*: Norwegian Petroleum Society Spec. Publ. 4, p. 105-114.

[2] Sinding-Larsen, R., and Chen, Z., 1996, Cross-validation of re-source estimates from discovery process modeling and volumetric accumulation modeling: Example from the Lower and Middle Jurassic play of the Halten Terrance, offshore Norway, *in* Dore, A. G., and Sinding-Larsen, R., eds., *Quantitative prediction and evaluation of petroleum resources*, Elsevier, p. 105-114.

[3] Sinding-Larsen, R., and Xu, J., 2005, Bayesian Discovery Process Modeling of the Lower and Middle Jurassic Play of the Halten Terrace, Offshore Norway, as Compared with the Previous Modeling: Natural Resources Research, v. 14, no. 3, p. 235-248.

[4] Stoneley, R., 1995, *Introduction to Petroleum Exploration for Non-Geologists*, Oxford University Press.

[5] Xu, J., and Sinding-Larsen, R., 2005, How to choose priors for Bayesian estimation of the discovery process model: Natural Resources Research, v. 14, no. 3, p. 211-233.