HA 3 FMSN50 Ekman & Sundell

Agnes Ekman (ag8720ek-s) & Gustaf Sundell (gu0147su-s)

March 2021

1 The hybrid MCMC sampler

1.1 Marginal posteriors

All marginal posteriors, $f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$, $f(\boldsymbol{\lambda}|\boldsymbol{\theta}, t, \boldsymbol{\tau})$, and $f(\boldsymbol{t}|\theta, \boldsymbol{t}, \boldsymbol{\tau})$, are proportional to the joint distribution function, $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$. This is shown in equations 2, 4, and 5. As exhibited in equation 1, the joint posterior can be divided into a product of the given distributions, found in the instructions for HA3: $f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t})$, $f(\boldsymbol{t})$, $f(\boldsymbol{\lambda}|\theta)$, and $f(\theta)$.

$$\begin{aligned} f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) =& \frac{f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})}{f(\theta, \boldsymbol{\lambda}, \boldsymbol{t})} \times \frac{f(\theta, \boldsymbol{\lambda}, \boldsymbol{t})}{f(\theta, \boldsymbol{\lambda})} \times \frac{f(\theta, \boldsymbol{\lambda})}{f(\theta)} \times f(\theta) = \\ =& f(\boldsymbol{\tau}|\theta, \boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}) f(\boldsymbol{\lambda}|\theta) f(\theta) = \\ =& f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{t}) f(\boldsymbol{\lambda}|\theta) f(\theta) \propto \\ \propto e^{-\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i)} & \prod_{i=1}^{d} \lambda_i^{n_i(\boldsymbol{\tau})} \prod_{i=1}^{d} (t_{i+1} - t_i) \prod_{i=1}^{d} \left(\frac{\theta^2}{\Gamma(2)} \lambda_i e^{-\theta \lambda_i}\right) \frac{\psi^2}{\Gamma(2)} \theta e^{-\psi \theta} \end{aligned} \quad (1)$$

Note that $f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}) = f(\boldsymbol{t})$, is assumed to be a correct simplification, as the prior on $\boldsymbol{t}$ given in the instructions is independent of θ and $\boldsymbol{\lambda}$. In the same manner, the independence of θ in the given density for τ lies behind $f(\boldsymbol{\tau}|\theta, \boldsymbol{\lambda}, \boldsymbol{t}) = f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t})$.

When computing the marginal posterior of θ, $f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$, the parameters λ, $\boldsymbol{t}$, and $\boldsymbol{\tau}$ are viewed as known. Hence, the denominator in equation 2, $f(\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$, is constant, yielding the first proportionality in the same equation. The second proportionality comes from the fact that $f(\boldsymbol{\tau}|\theta, \boldsymbol{\lambda}, \boldsymbol{t})$ and $f(\boldsymbol{t}|\theta, \boldsymbol{\lambda})$ are both constant given t and τ, since again, these parameters are viewed as constant.

$$f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) = \frac{f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})}{f(\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})} \propto f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{t}) f(\boldsymbol{\lambda}|\theta) f(\theta) \propto f(\boldsymbol{\lambda}|\theta) f(\theta) \quad (2)$$

Inserting the given expressions for $f(\boldsymbol{\lambda}|\theta)$ and $f(\theta)$ into equation 2 yields equation 3. Note that, up to a normalizing constant, the marginal posterior of

θ is a Gamma distribution with shape parameter $2d+2$ and rate parameter $\psi + \sum_{i=1}^{d} \lambda_i$.

$$\begin{aligned} f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) &\propto \frac{\psi^2}{\Gamma(2)} \theta e^{-\psi\theta} \prod_{i=1}^{d} (\frac{\theta^2}{\Gamma(2)} \lambda_i e^{-\theta\lambda_i}) \propto \\ &\propto \theta^{2d+1} e^{-\theta(\psi + \sum_{i=1}^{d} \lambda_i)} \propto \\ &\propto Gam(2d+2, \psi + \sum_{i=1}^{d} \lambda_i) \end{aligned} \tag{3}$$

When computing the marginal posterior of λ, the parameters θ, t, and τ are viewed as constant. Consequently, $f(\theta, t, \tau)$, $f(t)$, and $f(\theta)$ are also constant, making the marginal posterior with respect to lambda proportional to a product of Gamma distributions according to equation 4.

$$\begin{aligned} f(\boldsymbol{\lambda}|\boldsymbol{\theta}, t, \boldsymbol{\tau}) &\propto \frac{f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})}{f(\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\tau})} \propto f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{t}) f(\boldsymbol{\lambda}|\theta) f(\theta) \propto \\ &\propto f(\boldsymbol{\tau}|\theta, \boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{\lambda}|\theta) f(\theta) = \\ &= e^{-\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i)} \prod_{i=1}^{d} \lambda_i^{n_i(\tau)} \prod_{i=1}^{d} \lambda_i e^{-\theta\lambda_i} \propto \\ &= \prod_{i=1}^{d} (e^{-\lambda_i (t_{i+1} - t_i)} \lambda_i^{(\theta + t_{i+1} - t_i)}) \propto \\ &\propto \prod_{i=1}^{d} Gam(n_i(\boldsymbol{\tau}) + 2, \theta + t_{i+1} - t_i) \end{aligned} \tag{4}$$

Finally, when computing the marginal posterior with respect to t, the parameters θ, t, and τ are viewed as constant. Consequently, $f(\lambda, t, \tau)$, $f(\lambda|\theta)$, and $f(\theta)$ are also constant, yielding the proportionality relationships in equation 5. The final expression of $f(t|\theta, \lambda, \tau)$, up to a normalizing constant, is not recognized as being proportional to any known distribution.

$$\begin{aligned} f(\boldsymbol{t}|\theta, \boldsymbol{t}, \boldsymbol{\tau}) &\propto \frac{f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})}{f(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\tau})} \propto f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{t}) f(\boldsymbol{\lambda}|\theta) f(\theta) \propto \\ &\propto f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) f(\boldsymbol{t}) = e^{-\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i)} \prod_{i=1}^{d} (t_{i+1} - t_i) \end{aligned} \tag{5}$$

1.2 The algorithm

In order to estimate $\boldsymbol{t}$, θ and $\boldsymbol{\lambda}$, the information given in $\boldsymbol{\tau}$ is to be used, hence looking at the posterior $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}|\boldsymbol{\tau})$. This is once again proportional to equation

1. Gibbs-sampling is applicable to θ and $\boldsymbol{\lambda}$, however Metropolis-Hastings must be used for the breakpoints, $\boldsymbol{t}$. The instructions offered two alternatives for proposals for the MH algorithm, a random walk proposal and an independent proposal, and the former was chosen which is explained by:

For each breakpoint t_i, the candidate is generated by $t_i^* = t_i + \epsilon$, where $\epsilon \sim U(-R, R)$ and $R = \rho(t_{i+1} - t_{i-1})$ and ρ is a parameter that may be tuned to improve the algorithm. Here it may be noted that the proposal kernel is symmetric (since ϵ is uniform), meaning $\frac{r(t_i|t_i^*)}{r(t_i^*|t_i)} = 1,\ \forall\ (t_i, t_i^*) \in (t_1, t_{d+1})$. This means that the expression from slide 16, lecture 9 simplifies to:

$$t_{i+1} = \begin{cases} t_i*, \text{ w.pr. } \alpha(t_i, t_i*) = \min(1, \frac{f(t_i^*)}{f(t_i)}) \\ t_i \text{ otherwise} \end{cases} \tag{6}$$

Where f refers to $f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}, \boldsymbol{\tau})$, the shape of which was found in equation 5, hence regarding θ and $\boldsymbol{\lambda}$ as given. All break-point-vectors were initialized as evenly spaced between the first and last years, i.e. $\boldsymbol{t}^{(j)} = (1658, ..., 1980),\ j = 1, ..., M$. Note here that for all simulations, $M = N + \text{burn-in} = N + N/5$, where the burn in was not used for analysis. For all simulations N was set to 10 000, ensuring we always analyze with sample size N, noting that we simulate 20% more samples. This to ensure correct ordering, and in the hybrid algorithm, it will be necessary to be able to "look forward" towards those break-points not yet updated.

The MH algorithm has f as its stationary distribution, which is well motivated by the proof from lecture 10, slides 9-11, using the lemma presented and proved on slides 7-8 from the same lecture. As the f used in our MH sampler is the posterior on $\boldsymbol{t}$, with all other parameters as given, this is the distribution our algorithm will end up sampling from.

As for simulating θ and $\boldsymbol{\lambda}$, Gibbs sampling was used. Again, we use the conditional densities $f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$ and $f(\boldsymbol{\lambda}|\theta, \boldsymbol{t}, \boldsymbol{\tau})$, see equations 3 and 4. We begin by initializing $\boldsymbol{\lambda}^{(1)} = (\lambda_1^{(1)}, ..., \lambda_d^{(1)}) = (1, ..., 1)$, i.e. equal initial intensities. Thereafter, the algorithm goes as follows:

for j = 1:M-1 do:

draw $\theta^{(j)} \sim f(\theta|\boldsymbol{\lambda}^{(j)}, \boldsymbol{t}^{(j)}, \boldsymbol{\tau})$

for i = 2:d do:

draw $t_i^{(j+1)} \sim f(\boldsymbol{t}^{(j+1)}|\theta^{(j)}, \boldsymbol{\lambda}^{(j)}, \boldsymbol{\tau})$

draw $\lambda_i^{(j+1)} \sim f(\boldsymbol{\lambda}^{(j+1)}|\theta^{(j)}, \boldsymbol{t}^{(j)}, \boldsymbol{\tau})$

end i-for

draw$\lambda_1^{(j+1)} \sim f(\boldsymbol{\lambda}^{(j+1)}|\theta^{(j)}, \boldsymbol{t}^{(j)}, \boldsymbol{\tau})$

end j-for (7)

Note that, as $\lambda_i^{(j+1)}$ is produced using a Gibbs sampler, it is, more explicitly, sampled from $f(\lambda_i^{(j+1)}|\boldsymbol{\lambda}_{-i}^{(j)}, \theta^{(j)}, \boldsymbol{t}^{(j)}, \boldsymbol{\tau})$, here using notation inspired by lecture 10, slide 14. Note also that $t_i^{(j+1)}$ are drawn separately for each i.

1.3 Behaviour for different numbers of breakpoints

In order to review how the chain behaves for different numbers of breakpoints, the behaviour of the breakpoint chain across iterations was plotted with varying d, see Figure 2. Here, the optimal values of psi and rho were used, see the discussions on parameter choices in sections 1.4-1.5. Note that the number of breakpoints are always $d-1$. In the plots, we have included the endpoints, which by construction do not vary at all across iterations. This is merely to show the plots on the same scale. The same was done using histograms, in Figure 3, but here the endpoints are not as relevant. All plots were made excluding the burn-in. What is apparent looking at the plots is that the breakpoints appear to be reasonably steady for $d \leq 4$, i.e. up to 3 breakpoints. However, for $d \geq 5$, they show a more erratic behaviour.

One important aspect of the analysis of this MCMC sampler is the behavior of lambda. The accident intensity should be significantly different on the two sides of of each breaking point, since that is what defines a breakingpoint in theory. In figure 4, it is exhibited that all intensities are significantly different from each other for $d = 2$ and $d = 3$. In the case of $d = 4$ two intensities overlap. However, these intensities do not come from adjacent intervals since the overlapping data series are series 2 and series 4. This means that they are not the two sides of a breakpoint, but separated by another interval with a different intensity. Hence, there is no violation of the changing intensity in the $d = 4$ case. The same applies for the case of $d = 5$ where the intensities in data series 3 and 5 overlap, being separated by data series 4, and for the case of $d = 6$ where the intesities in data series 2 and 6, and 3 and 5, respectively, overlap. Since none of these pair of overlapping intensities belong to adjacent time intervals, no violation has occured.

1.4 Sensitivity of the posteriors to the choice of hyperparameter ψ

In order to investigate this, the mean and variance of $\boldsymbol{\lambda}$, θ and $\boldsymbol{t}$ were computed, as well as the acceptance rate, for 100 different values of ψ. This is all illustrated in the 7-plot Figure 5.

Beginning by looking the two plots concerning θ, there appears to be a clear structure. Both mean and variance clearly depend on the hyperparameter, which is by construction, see Equation 3, θ follows a Gamma distribution with ψ as part of its rate parameter. As d is kept constant, the only part of the posterior distribution for θ that changes is the rate parameter. The theoretical mean and variance of θ is $\frac{(2d+2)}{\psi+\sum\lambda_i}$ and $\frac{(2d+2)}{(\psi+\sum\lambda_i)^2}$ respectively [1], and it therefore makes sense that both decrease with increasing ψ. Note here that there are two conventions on parametrization of the Gamma distribution, and here the rate parameter is not inverted in the density function.

Now looking at the behaviour of $\boldsymbol{\lambda}$ in Figure 5, it appears there is no very clear structure in the way that it was with θ. The 5 observed intensities appear to have relatively constant means, however displaying some seemingly synchronous spikes. The spikes for certain values of ψ are even more dramatic when looking at the variance of the intensities, here also demonstrating some kind of synchronicity. That is, at leas three of the five intensities appear to spike both in mean and variance for the same seemingly random values of ψ. This could be explained by randomness, seeing as we're running 10 000 simulations 100 times. This means that we are bound to observe some extreme events.

As for the breakpoints, their means seem completely unaffected by the hyperparameter, and their variances appear to vary with no structure. This is likely to do with the causal distance between the hyperparameter and the breakpoint posterior.

The same reasoning explains why the acceptance rate from the MH algorithm appears like a white noise process with respect to the hyperparameter.

It is not obvious what value of the hyperparameter is to be preferred from the plots. θ in itself is not a variable of interest for the final result, however it is reasoned that low variance is a generally desired quality for the overall stability of the algorithm. For this reason, as the variance of θ decreases with ψ, $\psi = 30$ was deemed to be a suitable value for the parameter, see figure 5.

1.5 Sensitivity of the posteriors and the mixing to the choice of tuning parameter ρ

Looking at the plots in figure 6, neither the mean nor the variance of $\boldsymbol{\lambda}$, θ, and $\boldsymbol{t}$ seem to behave any different depending on the value of ρ. The variation, including the spikes, probably rather comes from the randomness, as discussed above in section 1.4.

[1] Wikipedia page on the Gamma Distribution.

The mixing, however, depends greatly on ρ, which is exhibited in the plot of the acceptance ratio in figure 6. The acceptance ratio rapidly decreases with an increasing rho; being close to 1 when $\rho \approx 0$, and already down to 0.05 when $\rho \approx 0.5$. A benchmark for acceptance ratio, which is known to result in good mixing, is 30%. Looking at the graph, the 30% acceptance ratio is found for $\rho \approx 0.03$.

For this task, $d = 5$ was used, and for this number of breakpoints the optimal ρ with respect to acceptance ratio was specific to $d = 5$. At this point, we did the same algorithm for $d = 2, 3, 4, 5, 6$, to make sure that task 1.c) was solved with a suitable ρ, meaning that it yields an acceptance rate close to 30% for each d.

2 Parametric bootstrap for the 100-year Atlantic wave

The significant wave-heights in the north Atlantic can be modeled as a Gumbel distribution with the distribution function exhibited in equation 8, given in the instructions.

$$F(x; \mu, \beta) = exp\left(-exp\left(-\frac{x-\mu}{\beta}\right)\right), \ x \sim \mathbb{R} \tag{8}$$

2.1 Finding the inverse $F^{-1}(u; \mu, \beta)$

$$F(x; \mu, \beta) = u = exp\left(-exp\left(-\frac{x-\mu}{\beta}\right)\right) \tag{9}$$

$$ln(u) = exp\left(-\frac{x-\mu}{\beta}\right) \tag{10}$$

$$ln(-ln(u)) = -\frac{x-\mu}{\beta} \tag{11}$$

$$\beta ln(-ln(u)) = -x + \mu \tag{12}$$

$$x = \mu - \beta ln(-ln(u)) = F^{-1}(u; \mu, \beta), \ u \in [0, 1] \tag{13}$$

Note that it is theoretically possible for $u = 1$ or $u = 0$. In these cases, x would take the values ∞ or $-\infty$ respectively. However, the propabilities of this occuring are zero.

2.2 Bootstrapped parameter estimates

The estimates of the parameters $\hat{\beta}$ and $\hat{\mu}$ were computed through the use of the given function *est_gumbel* with the data from *atlantic.txt* as input. Then a sample of uniformly distributed numbers in the interval $[0, 1]$ was drawn and ran through the inverse function, derived in section 2.1, together with the parameter

estimates to obtain a simulated set of wave heights. The simulated wave heights were then used as input in the *est_gumbel* function to obtain bootstrapped parameter estimates. This process is also described in equations 14-16.

$$[\hat{\beta}, \hat{\mu}] = est\_gumbel(y) \tag{14}$$

$$y_{boot} = F^{-1}(u; \hat{\beta}, \hat{\mu}),\ u \sim U(0,1) \tag{15}$$

$$[\beta_{boot}, \mu_{boot}] = est\_gumbel(y_{boot}) \tag{16}$$

The errors, $\Delta\beta$ and $\Delta\mu$, were computed through subtracting the vector of bootstrapped estimates, β_{boot} and μ_{boot}, with the initial estimates, $\hat{\beta}$ and $\hat{\mu}$, see equations 17-18.

$$\Delta\beta = \beta_{boot} - \hat{\beta} \tag{17}$$

$$\Delta\mu = \mu_{boot} - \hat{\mu} \tag{18}$$

The errors were sorted, and the 2.5% and 97.5% quantiles were identified. Furthermore, the respective biases were estimated as the mean of the respective errors. In order to compute the confidence intervals for the parameters, the biases were subtracted from $\hat{\beta}$ and $\hat{\mu}$ respectively, and the quantiles were subtracted from these quantities, yielding equations 19-20.

$$I_{\beta}^{95\%} = [1.3879,\ 1.5827] \tag{19}$$

$$I_{\mu}^{95\%} = [4.0151,\ 4.2764] \tag{20}$$

2.3 Bootstrapped 100-year wave estimate

To estimate a one-sided confidence interval of the height of the 100-year wave, y, the initial estimate, $\hat{y}$, was computed through the inverse, see equation 21. Then, y estimates were computed based on the bootstrapped estimates of β and μ, see equation . This was repeated once for each parameter pair.

$$\hat{y} = F^{-1}(1 - 1/T; \hat{\beta},\ \hat{\mu}) \tag{21}$$

$$y_{boot} = F^{-1}(1 - 1/T; \beta_{boot},\ \mu_{boot}) \tag{22}$$

$$,\ where\ T = 3 \times 14 \times 12$$

Then, the errors were computed similarly as in section 2.2, see equation 23.

$$\Delta y = y_{boot} - \hat{y} \tag{23}$$

In contrast to section 2.2, the confidence interval in this case is one-sided. This means that all of the relevant uncertainty is gathered on the same side

of the distribution. Conceptually, this comes from only being interested in knowing that the 100-year wave is smaller than a certain value. Hence, when computing the confidence interval, the right α-quantile was used. As in the previous section, the bias was estimated as the mean error and subtracted from the estimate before making the confidence interval presented in equation 24:

$$I_y = [0,\ 17.2934] \tag{24}$$

3 Appendix - figures

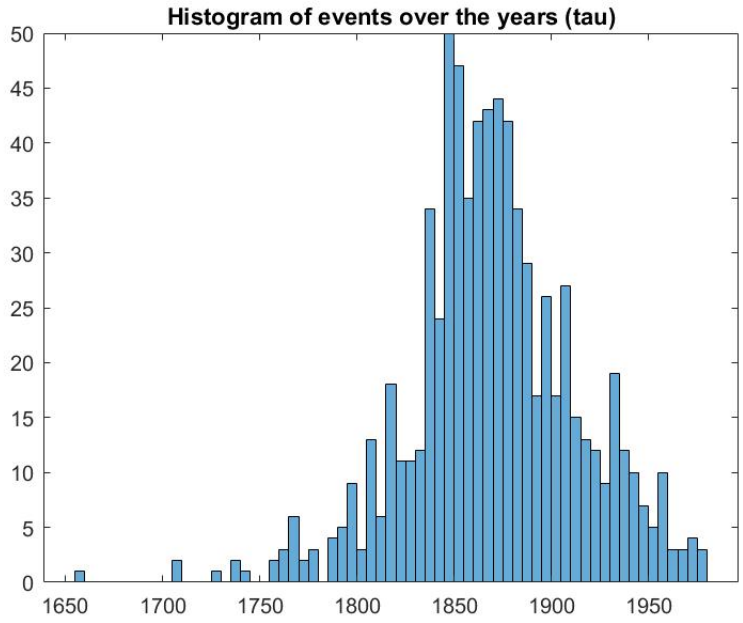

Figure 1: Histogram of coal mine disasters from the given data in τ. Bin width for this histogram was 5 years.

Figure 2: Breakpoint chains for varying d. Note that the "data"-labeling is ordered, signifying 1st, 2nd and so on breakpoint.

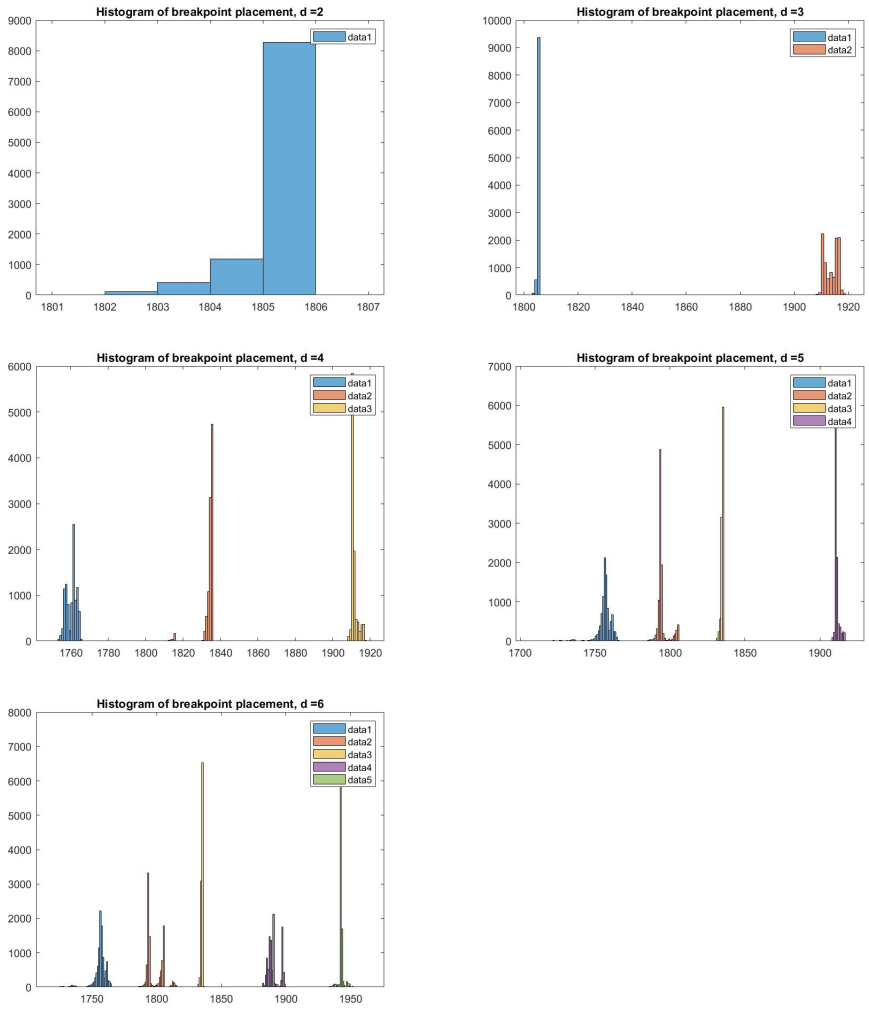

Figure 3: Breakpoint histograms for varying d. Note that the "data"-labeling is ordered, signifying 1st, 2nd and so on breakpoint.

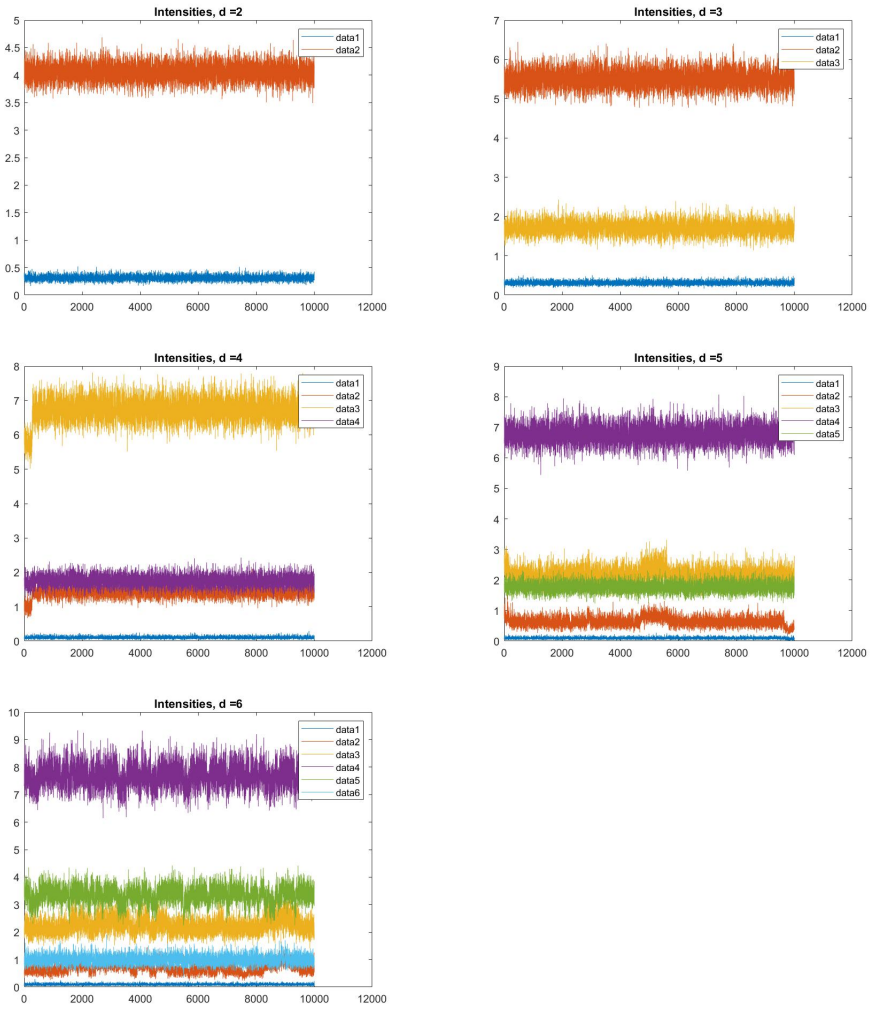

Figure 4: Lambda varying over iterations for different d.Note that the "data"-labeling is ordered, signifying 1st, 2nd and so on intensity.

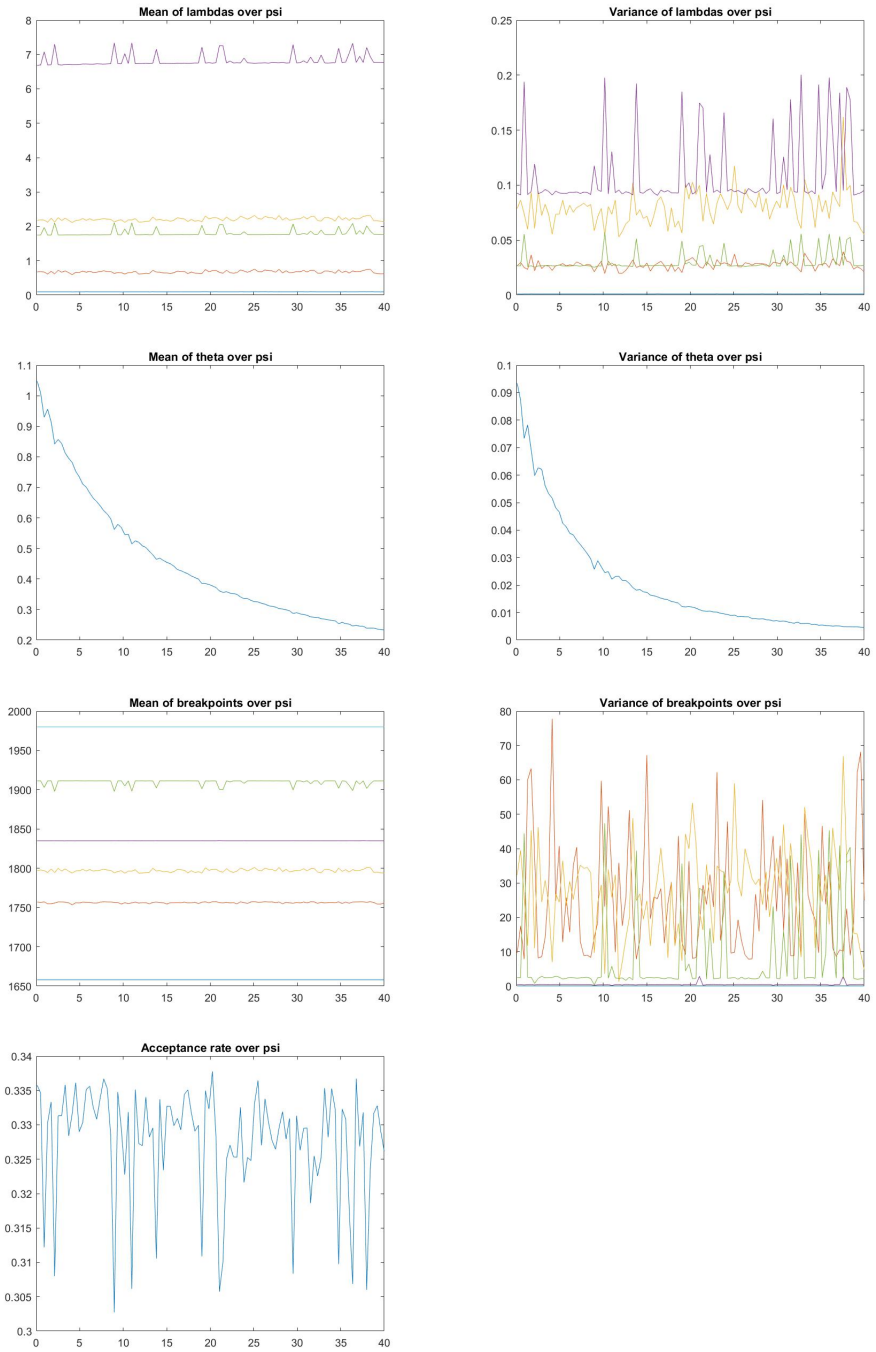

Figure 5: Various plots with varying ψ

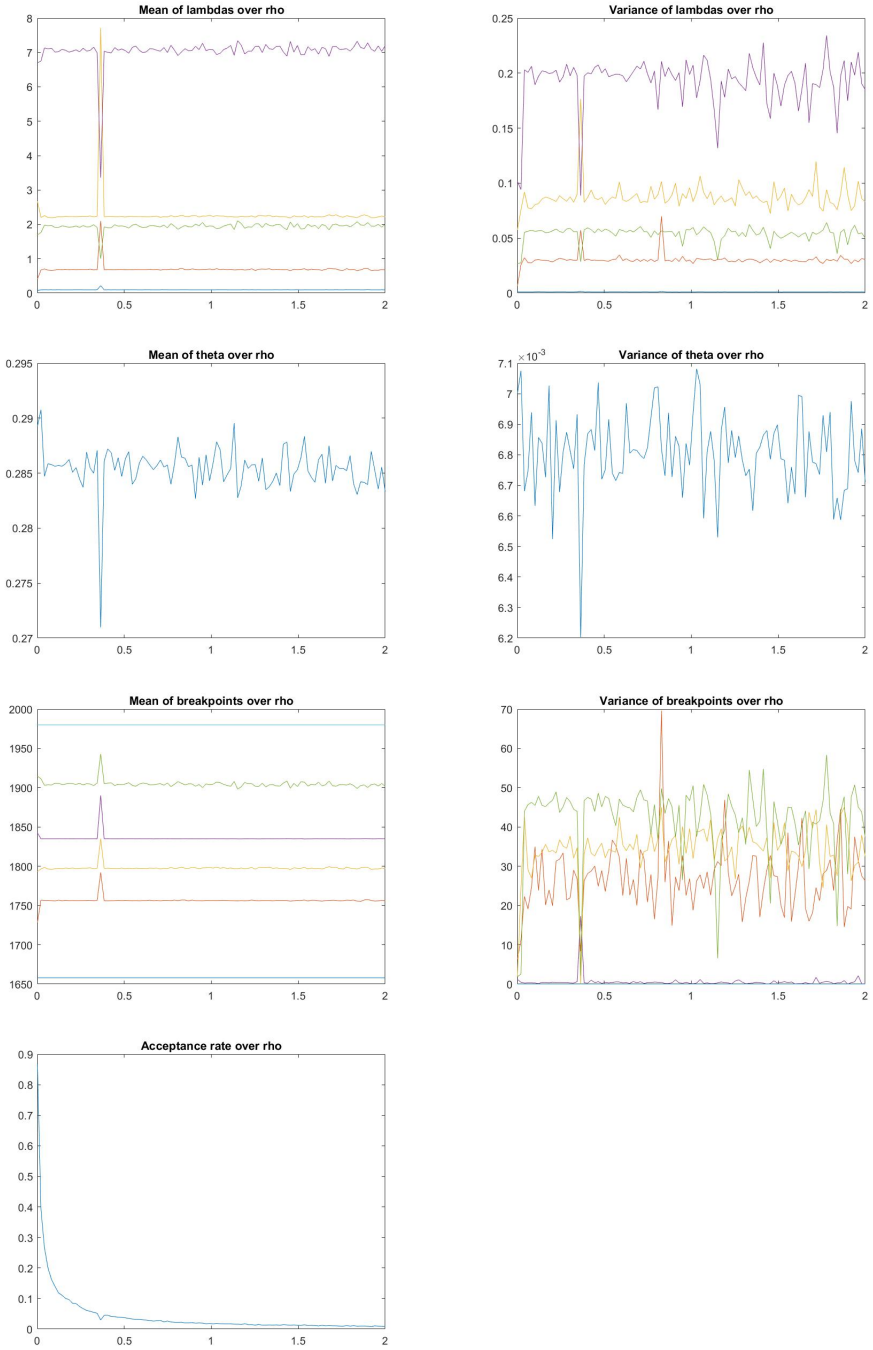

Figure 6: Various plots with varying ρ.