Bayesian Analysis of Crack-Seal Veins

Daniel Pliego

18 April, 2025

GitHub repo: https://github.com/pliego02/STAT-447-Project

Introduction

The idea of this project is to do Bayesian analysis on a data set of crack seal veins provided by Dr. Matthew Tarling from the Geological Sciences department at UBC. This data set consists of thickness measurements of consecutive layers or bands measured in micrometers from four different samples of "crack-seal veins".

Crack-seal veins are created beneath Earth's surface through repeated fractures and sealing of these fractures. Geological processes cause rocks to fracture, but due to the high pressure in these environments, the cracks can't stay open. Instead, they are filled with fluids that precipitate crystals, creating distinct layers or bands. Each band represents a separate fracture and sealing event. Studying crack-seal veins is interesting because they preserve a record of tectonic activity over time, which can help us better understand how faults behave and then give us more insight into the behavior of earthquakes.

There have been suggestions that consecutive bands in crack-seal veins do not exhibit correlation, and that their thicknesses follow an exponential distribution (Renard et al., 2005). However, using Monte Carlo simulations, Williams et al. (2022) showed that it is unlikely these events follow an exponential distribution, based on the coefficient of variation (COV), defined as the standard deviation divided by the mean. The aim of this project is therefore to identify better fitting distributions to model the thickness of crack-seal veins, and to use these improved models to fit an AR(1) (autoregressive of order 1) model in order to examine potential correlation between consecutive bands.

Rank and trace plots have been explored for every parameter from every model on this project and they all appear appropriate in that the trace plots of different chains behave similarly and the rank plots appear uniform. However due to spacing none of these plot will be provided on this file. There is a file in the repo which contains all of this plots. I also want to mention that ChatGpt was used in the project to help debug the posterior predictive check plots and the indexes for the AR(1) model

Exponential Likelihood model

The first model I will fit is one with an exponential likelihood. Based on the findings of Williams et al. (2022), we should expect to obtain similar results, where the observed coefficient of variation (COV) falls outside the 95% confidence interval computed using simulated data from the posterior distribution. This would indicate that the posterior does not accurately capture the variability observed in the real data, suggesting that the exponential model may not be an appropriate fit.

The model is of the form:

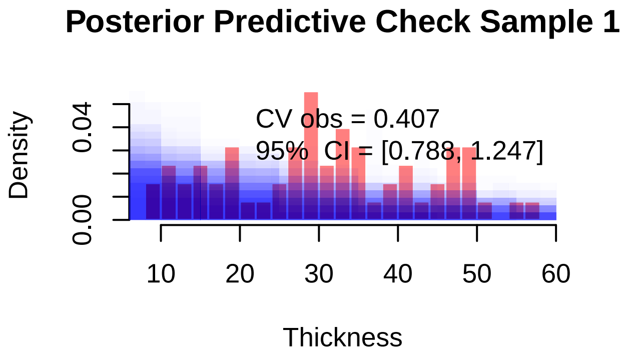
$$\lambda_i \sim \text{Uniform}(0, 1)$$

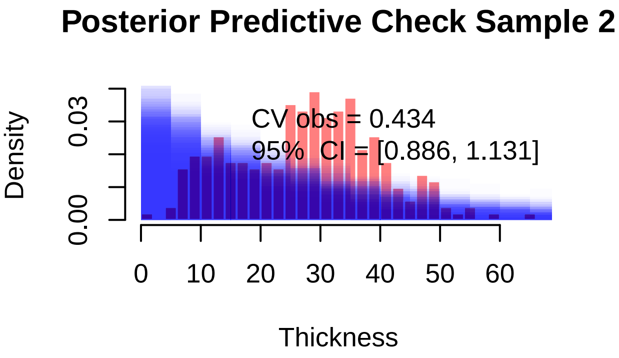
$$X_n \mid \lambda_i \sim \text{Exp}(\lambda_i)$$

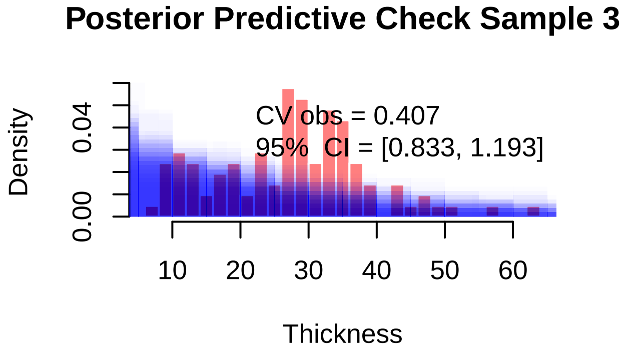
Where X_n is the thickness observed thickness and λ_i is the parameter for each sample i. This will make the stan model generate lambda's for each one of the samples

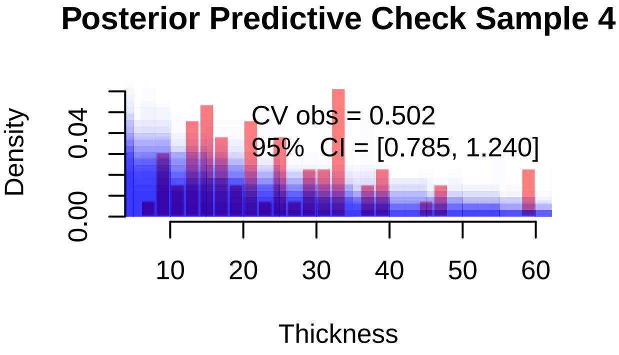
Posterior predictive check for each sample.

The solid red histogram shows the actual observed thicknesses for each sample. Overlaid in translucent blue are 100 histograms of simulated thicknesses each one generated by plugging a draw of the model's parameters (from the joint posterior) into the likelihood. The legend reports the observed coefficient of variation (COV) alongside its 95% posterior predictive interval.


Posterior Predictive Check Sample 1


Posterior Predictive Check Sample 2


Posterior Predictive Check Sample 3


Posterior Predictive Check Sample 4

As expected in all of the samples the observed COV is far away from the 95% interval. We can also see that the posterior distribution with an exponential likelihood for the thickness does not fit the data very well. This is expected and is what was show on the paper.

As said on the introduction I wont provide trace or rank plots on this file. However, they all looked appropriate.

Model 2, Gamma model

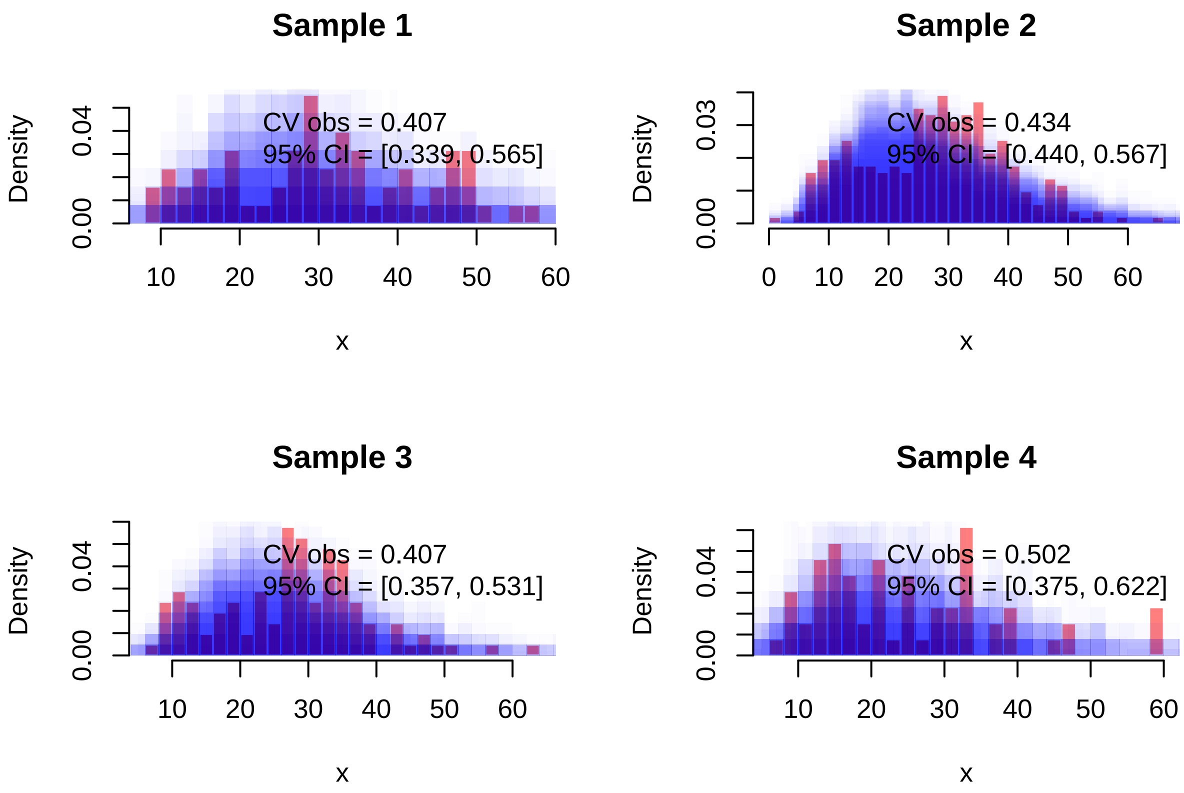
We have seen that the exponential model does not fit the data properly this is likely due to the assumption of equal variance and mean from the exponential distribution. I will implement a Gamma model to relax that assumption.

I will also implement a hierarchical model to help inform the prior.

The model is of the form:

$$\begin{aligned}\mu_\alpha &\sim Exp(1)\\ \theta_\alpha &\sim Exp(1)\\ \mu_\beta &\sim Exp(1)\\ \theta_\beta &\sim Exp(1)\\ \alpha_i \mid \mu_\alpha, \theta_\alpha &\sim Gamma(\mu_\alpha, \theta_\alpha)\\ \beta_i \mid \mu_\beta, \theta_\beta &\sim Gamma(\mu_\beta, \theta_\beta)\\ X_n \mid \alpha_i, \beta_i &\sim Gamma(\alpha_i, \beta_i)\end{aligned}$$

I will provide posterior predictive checks, however due to the hierarchical model, I will perform mixed predictive replication for hierarchical models which are showed how to perform on the stan users guide. This are posterior predictive checks with the difference that the hyper parameters remain fixed



Looking at the plots, the hierarchical gamma model appears to fit the data much better, additionally the CoV measurement is now inside the 95% CI provided by the posterior distribution in all samples but sample 2. This is already a huge improvement compared to the exponential model.

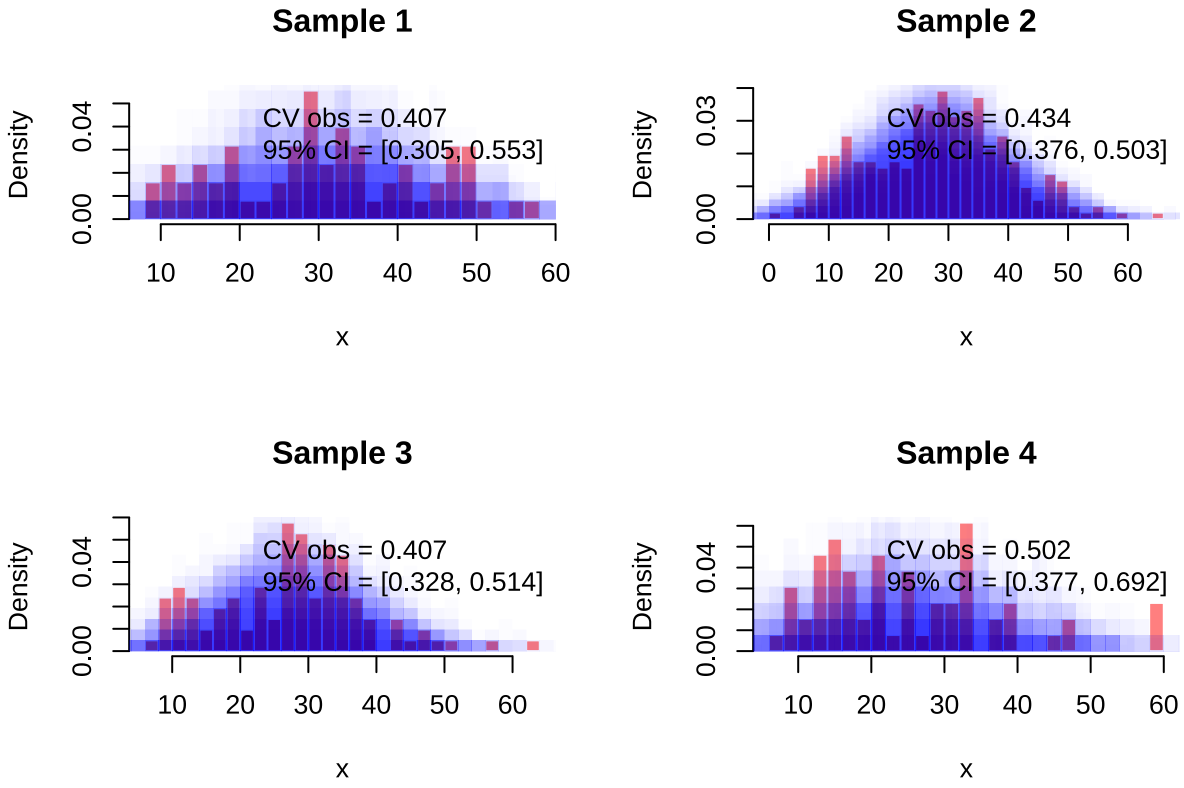
The gamma likelihood seems to miss where the peak of the observed data is. However it seems to model well the right tail of the distribution.

Model 3, Normal model

Finally I want to test a normal distribution model. The normal distribution does not have a support on the positive real numbers as the gamma and exponential distribution do. However the thickness values are on average a few standard deviations away from 0. Which makes it so that fitting a normal distribution should have little to no predictions which are negative. I will also use a hierarchical model to help inform the prior.

model:

$$
\begin{aligned}
\alpha_\mu &\sim Exp(1) \\
\alpha_\theta &\sim Exp(1) \\
\beta_\mu &\sim Exp(1) \\
\beta_\theta &\sim Exp(1) \\
\mu_i \mid \alpha_\mu, \beta_\mu &\sim Gamma(\alpha_\mu, \beta_\mu) \\
\theta_i \mid \alpha_\theta, \beta_\theta &\sim Gamma(\alpha_\theta, \beta_\theta) \\
X_n \mid \mu_i, \theta_i &\sim Normal(\mu_i, \theta_i)
\end{aligned}
$$



By looking at the plots I believe the normal likelihood models the observed data much better. The blue shaded simulated samples appear to be much closer to the observed data. Additionally we can see that now all observed COV are well inside the 95% confidence interval made by the simulated data from the posterior.

The normal likelihood seems to perform the better out of the three distributions even though it has an incorrect support compared to the other 2 distributions.

AR(1) Model Using a Normal Distribution

I will model the AR(1) (autoregressive model of order 1) using a normal distribution. I choose to use a normal distribution as it is the one that best fits the data from the models above. I will also make the model hierarchical to help inform the priors.

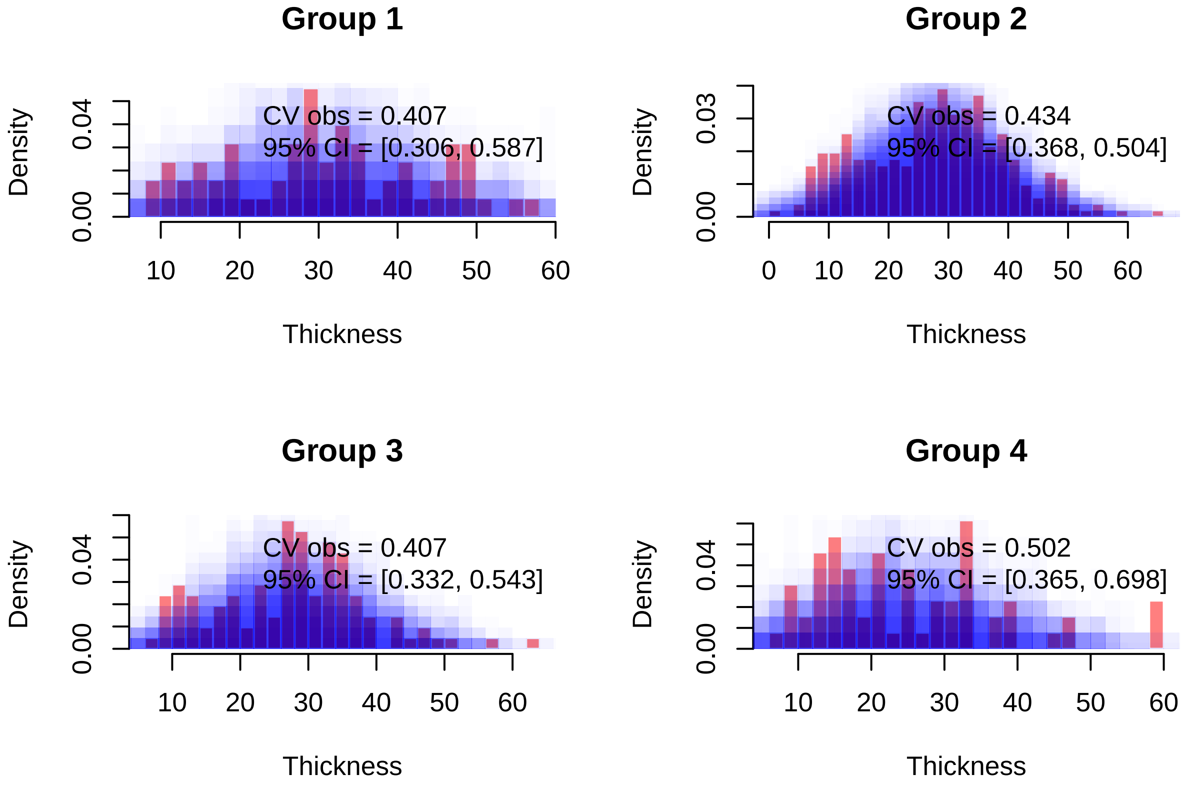
I will make 2 different models. One will have the AR(1) coefficient shared between all 4 samples and the second one will fit a different AR(1) coefficient to each of the different samples. I will only provide the posterior predictive check plots for the model with a shared AR(1) parameter.

model with shared AR(1) parameter ϕ between samples:

$$
\begin{aligned}
\alpha_\mu &\sim Exp(1)\\
\beta_\mu &\sim Exp(1)\\
\alpha_\theta &\sim Exp(1)\\
\beta_\theta &\sim Exp(1)\\
\mu_i &\sim Gamma(\alpha_\mu, \beta_\mu)\\
\theta_i &\sim Gamma(\alpha_\theta, \beta_\theta)\\
\phi &\sim Normal(0, 0.5)\\
X_{f_i} &\sim Normal(\mu_i, \theta_i) \quad \text{(first observation)}\\
X_n &\sim Normal(\mu_i + \phi\,(X_{n-1} - \mu_i),\ \theta_i)
\end{aligned}
$$

The model with parameter AR(1) ϕ_i for each sample is only different in that it has:

$$
\begin{aligned}
\phi_i &\sim Normal(0, 0.5) \quad \text{(for sample i)}\\
X_n &\sim Normal(\mu_i + \phi_i\,(X_{n-1} - \mu_i),\ \theta_i)
\end{aligned}
$$



The posterior predictive checks all suggest the posterior seems appropriate.

For an AR(1) model like the one we fitted above, the auto correlation coefficient at lag 1 is equal to the AR coefficient.

$$\rho(1) = \phi$$

Therefore the resulting posterior $\rho(1)$ values for each of the samples along with a 95% confidence interval are:

Table 1: 95% Confidence Intervals and Autocorrelation Significance per Sample

Sample	mean	2.5%	97.5%	bound	significant
1	0.147	0.009	0.362	0.252	FALSE
2	0.253	0.129	0.368	0.125	TRUE
3	0.110	0.006	0.275	0.196	FALSE
4	0.293	0.069	0.531	0.248	FALSE

Table 2: 95% Confidence Intervals and Autocorrelation Significance Shared

Sample	mean	2.5%	97.5%	size	bound	significant
all	0.206	0.112	0.294	487	0.091	TRUE

Values outside the interval $(-\frac{2}{\sqrt{n}}, \frac{2}{\sqrt{n}})$ are considered significant when determining whether the data deviates from a white noise process. In the table, “bound” refers to the value of $\frac{2}{\sqrt{n}}$ for each sample.

We can see that the model with a single AR coefficient shared across all samples suggests there is correlation between consecutive bands, since the posterior 95% interval lies entirely outside the bound. However, it’s important to note that sample 2 is the only one out of the four that clearly showed signs of autocorrelation, and it also happens to be the largest sample as it contains about four times as many observations as the others. This suggests that sample 2 may indeed have some autocorrelation. However, due to the limited data and the individual results from each sample, I’m unsure whether all samples exhibit the same behavior.

Results

As expected, and consistent with the findings of Williams et al. (2022), the exponential distribution does not model the thickness of crack-seal veins particularly well. In this project, I tested both gamma and normal distributions as alternatives, with the normal distribution providing the best fit among the three. Based on this result, I used a normal likelihood in an AR(1) model to explore the autocorrelation behavior between consecutive bands.

The analysis showed that sample 2 exhibited significant autocorrelation at lag 1, suggesting that there is a correlation in band thickness from one event to the next in that sample. When fitting an AR(1) model with a shared coefficient across all samples, the results indicated significant autocorrelation between consecutive bands overall. However, when examining each sample individually, it became clear that not all samples exhibit the same behavior. Due to the limited size of the dataset, I remain uncertain whether all samples follow the same correlation structure as sample 2.

Discussion

This analysis of crack-seal veins is based on only four samples. To obtain more reliable results, the procedures outlined above should be tested on a larger number of samples, ideally with more observations per sample. In addition to expanding the dataset, I think it would be very interesting to apply Bayesian analysis of crack-seal veins in the frequency domain. This idea has been mentioned by Renard et al. (2005), but I believe it deserves further exploration using Bayesian methods.

References

Posterior and prior predictive checks. (n.d.). Stan Docs. https://mc-stan.org/docs/stan-users-guide/posterior-predictive-checks.html

Renard, F., Andréani, M., Boullier, A., & Labaume, P. (2005). Crack-seal patterns: records of uncorrelated stress release variations in crustal rocks. Geological Society London Special Publications, 243(1), 67–79. https://doi.org/10.1144/gsl.sp.2005.243.01.07

Williams, R. T., & Kirkpatrick, J. D. (2022). Are low-frequency earthquake moments area- or slip-limited? A rock record examination. Geophysical Research Letters, 49, e2021GL095759. https://doi.org/10.1029/2021GL095759