Bayesian Data Analysis

Analysis of power plant vibration data

# Introduction

For the purpose of this project, we will be doing a Bayesian analysis of data collected from vibration sensors installed on a cooling water pump at a power plant. More specifically, we will use Bayesian Markov Chain Monte Carlo (MCMC) simulation to assess whether the vibration levels plausibly change over time. The MCMC simulation is used to obtain the relevant posterior distributions based on the assumed prior distributions and the collected data. The Bayesian modelling framework Stan was used to implement the models.

## Data

The data we will be analysing stems from a large Danish combined heat and power plant (CHP) running primarily on biomass in the form of woodchips. The data represents vibration levels from a cooling water pump (CWP) at the plant. The pump’s purpose is to move cold water from the sea into a condenser (a large tank), where steam coming from the electricity generating turbine is cooled and thus condensed back into water. For the power plant to produce electricity, the pump must therefore be running. For this reason, it is considered a vital component of the power plant and vibration sensors have therefore been installed in order to allow online monitoring of the condition of the pump.

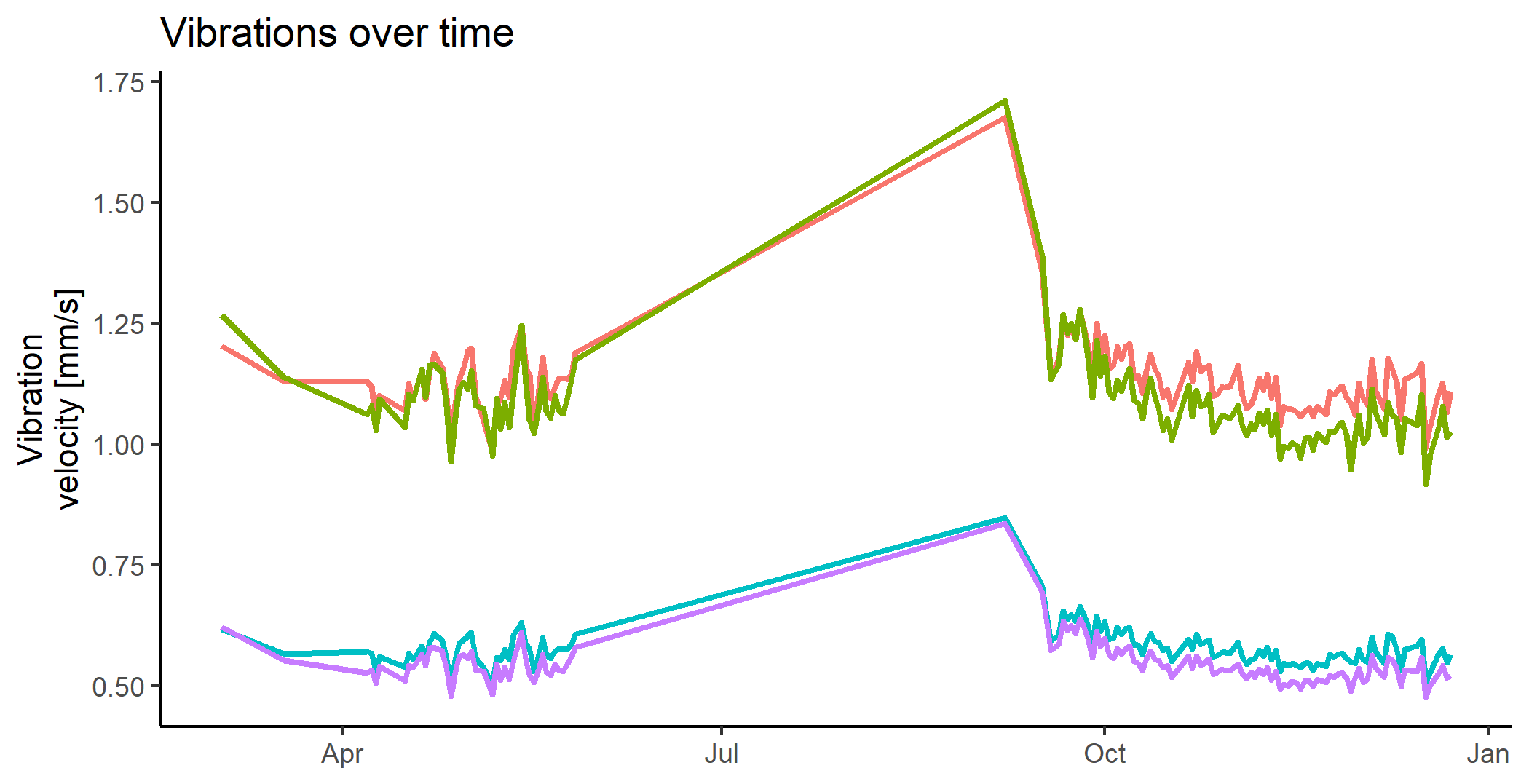
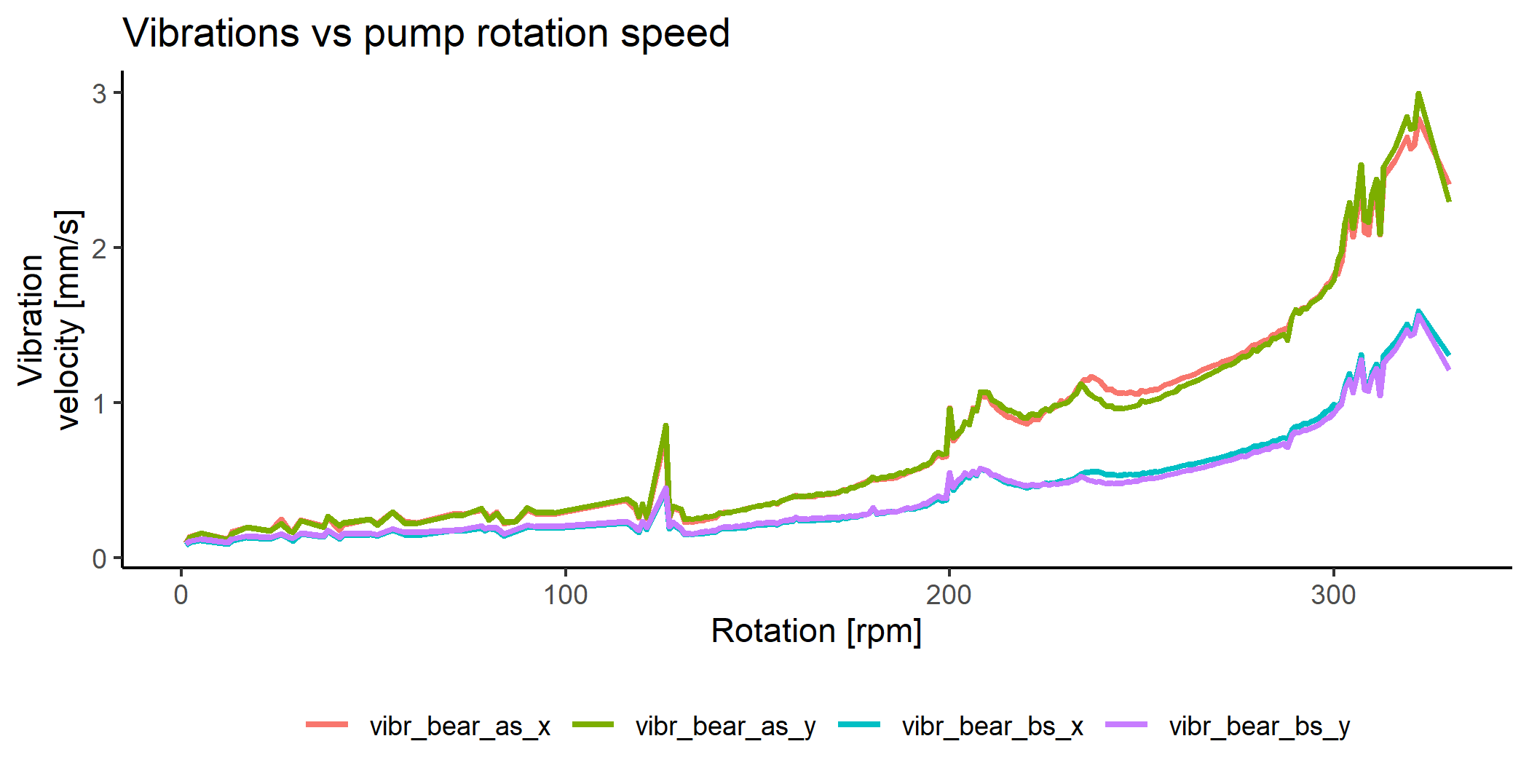
For our purpose we will be using a dataset collected in the period from March to December 2020. The plant operators seemed to notice an increase in the vibration levels in the Spring of this year. We will attempt to quantify this claim as nobody ever found out if the levels truly were on the rise or not, as the plant (and thus also the pump) was shut down and routine overhauled from July until September.

Figure 1: left: vibrations from the 4 sensors over the course of 2020. Right: vibrations as a function of pump speed. We will be using the measurement coloured in green for the purpose of the analysis.

We have 4 vibration sensors installed, but we will only be using data from one of them, as they are strongly correlated since they measure largely the same thing. The vibrations over time are seen in figure 1.

Furthermore, we only look at vibration levels corresponding to a fixed pump load, as vibration levels will naturally vary with how much work the pump must do. Therefore, we only consider vibrations from observations which had a pump speed in the interval 250 to 270 rounds per minute (rpm). Vibration level as a function of pump speed is seen in figure 2. Lastly, as the measurements are quite noisy, we have computed daily averages. Taking a small number of missing values into account, these steps yield us 14 observations for the seven months of available data: March, April, May, September, October, November, and December of 2020. Analysis

# Analysis

## Models

### Common mean and variance (pooled)

This model is the simplest in that it shares both mean and variance across the months. We do not expect this model to be the best fit as intuitively more realistic models can be thought of. It is given as

Where is the common mean vibration level over the months and is the shared standard deviation. Finally, we assume the observations are normally distributed around the common mean with standard deviation . From figure 1 we can see that the vibration levels seem to fluctuate close to 1, so we fix the mean of the Normal prior on at 1 and restrict it to be positive. The prior standard deviation is fixed at 1, as well. The normal prior on the standard deviation is restricted to positive values as well. These choices correspond to weakly informative priors.

### Monthly mean and common variance

This model is largely the same, however, instead of a shared mean vibration we now separate it out so each month gets its own mean. Otherwise, everything is the same as the previous model. In equations it is given as

*, ,*

### Monthly mean and variance (separate)

We now allow each month to have its own standard deviation as well, with everything else being the same as in the previous model. It is given as

One could argue that this is the most sensible choice of model as it is perhaps the one best reflecting reality, as we would expect each month to deviate slightly from the others. This should be well reflected by this model.

### Monthly mean and variance with hierarchical priors (hierarchical)

Finally, we try a hierarchical model, where we assume a prior distribution on the mean of the existing priors. In equations, this is given as

, ,

, ,

This model should allow for more flexibility in the parameters and a priori we would expect this model to perform well.

# Results

## Convergence diagnostics

To assess convergence of the Markov Chains we look at the values returned from Stan. These are shown for each model in table 1 below. We see that the maximum value attained

by either of the models is close to 1 and well below 1.05, which should give us some guarantee that the chains have converged properly.

|  |  |
| --- | --- |
| **Model** | **Maximal Rhat value** |
| Common mean and variance | 1,008 |
| Monthly mean, common variance | 1,012 |
| Monthly mean and variance | 1,009 |
| Hierarchical | 1,006 |

Table 1: Convergence diagnostic values for each of the four models.

## Model selection

With the selected models in hand and confidence that they have all converged, we now turn to model selection.

### Posterior predictive checking

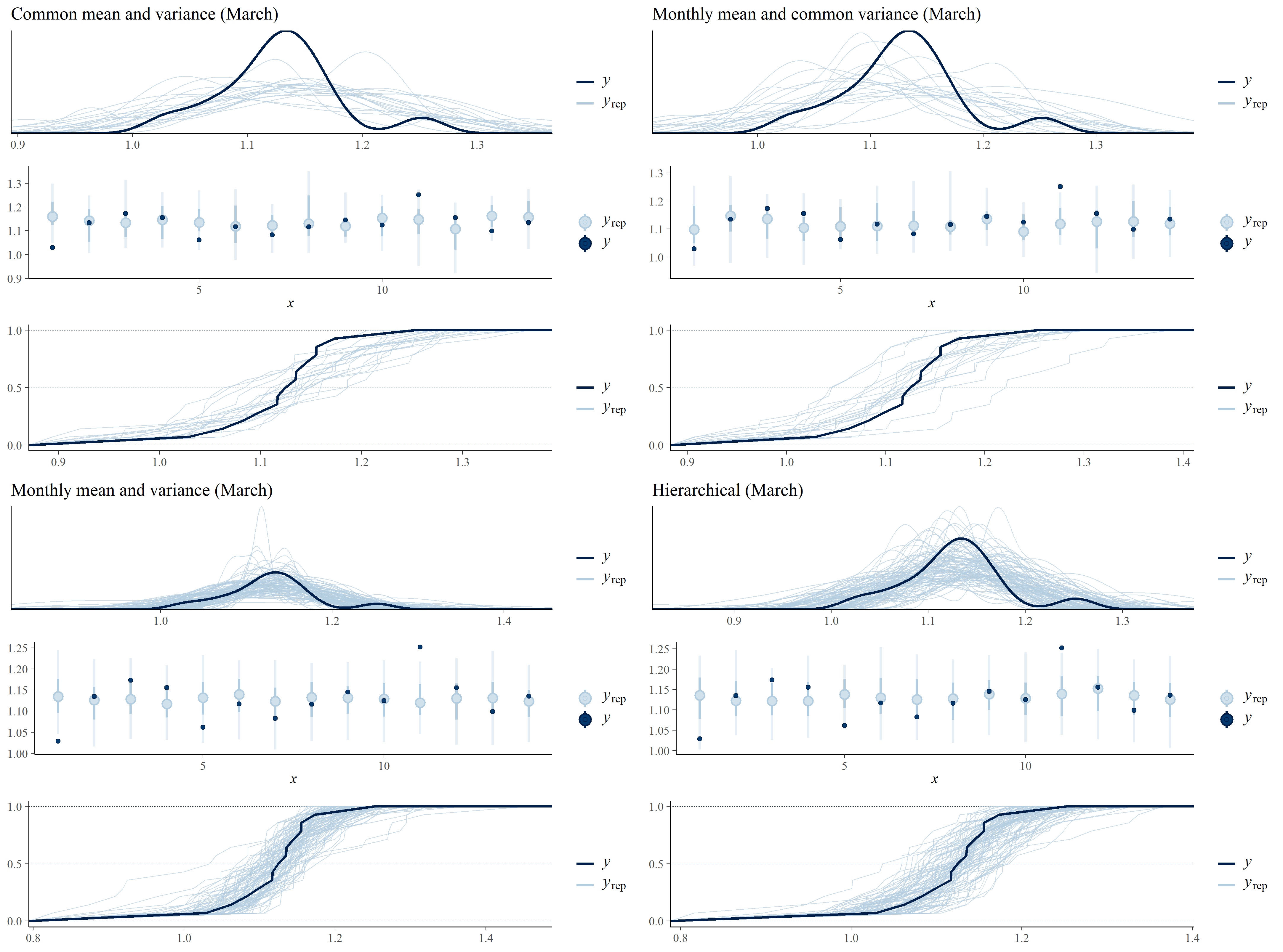
First, we determine which model seems most aligned with reality by drawing random samples from the posterior predictive distributions and comparing these with the observed values. In the case of a good model, they should be well aligned. For the month of March, plots comparing these are shown in figure 2. Judging from the plots it would seem that either the separate or the hierarchical model yields the best fit to the data, so we would expect these two to also score the highest when evaluated with cross-validation in the next section.

Figure 2: posterior predictive checks from all four models. Each model's predictions are shown as from top to bottom as a density, predictive distribution intervals and the empirical cumulative distribution function.

### Cross validation with PSIS-LOO

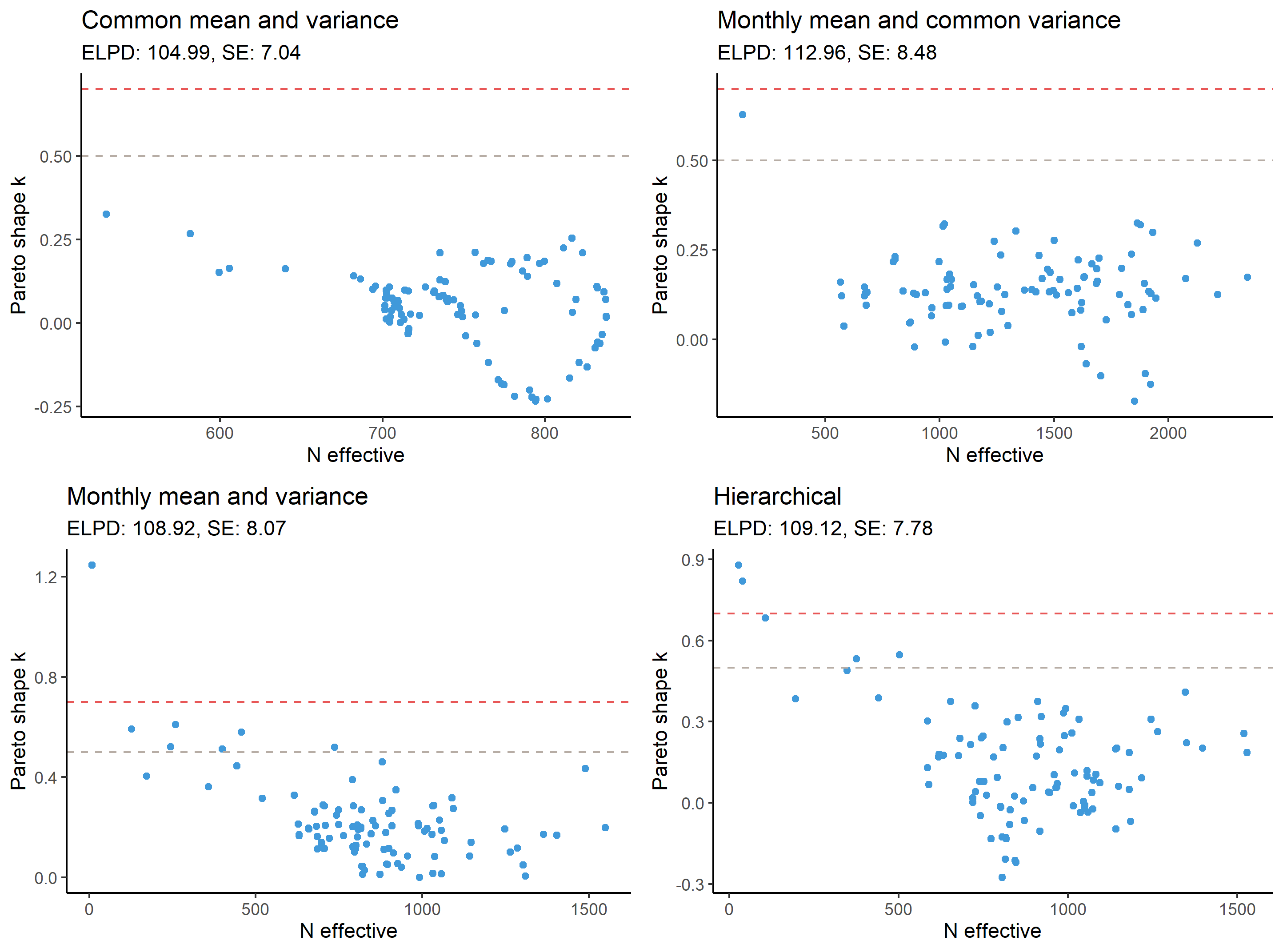
We now turn to cross validation to determine which model is the best fit. From the plot in figure 3, we see the models’ Pareto smoothed importance sampling leave-one-out (PSIS-LOO) cross validation results. The Pareto shape parameter *k* is plotted versus the effective sample size of each sample in the cross validation. Dashed lines at 0.5 and 0.7 indicate the acceptable upper bounds for the shape parameter.

Figure 3: PSIS-LOO diagnostics for all four models. The Pareto shape parameter k is plotted versus the effective sample size for each sample in the cross-validation.

From the plots we see that the models generally have good PSIS-LOO diagnostics. Only few points plot above the shape parameter bounds and most points have high effective sample size. These results indicate, that we should have some trust in the *expected log predictive point density* (ELPD) results obtained from the cross validation. The ELPD differences between the models are shown in table 2, where it is seen that the model with a monthly mean and common variance seems the best fit. However, this is closely followed by the hierarchical model since the difference in ELPD between these two is only -3,8 and the standard error of the difference is 2,9. This indicates that the hierarchical model is likely also a good fit on par with the best model.

|  |  |  |
| --- | --- | --- |
| **Model** | **ELPD difference** | **Standard error of difference** |
| Monthly mean and common variance | 0 | 0 |
| Hierarchical | -3,8 | 2,9 |
| Monthly mean and variance | -4 | 2,6 |
| Common mean and variance | -8 | 5,5 |

Table 2: ELPD differences between the four models. The more negative the ELPD difference, the worse the model. Standard errors on the differences are also shown. The models are colour coded to indicate model performance.

Based on both the posterior predictive checks and the PSIS-LOO we conclude that the hierarchical model is to be preferred for its predictive accuracy judged on the predictive checks. In terms of model simplicity, the monthly mean and common variance model should be preferred.

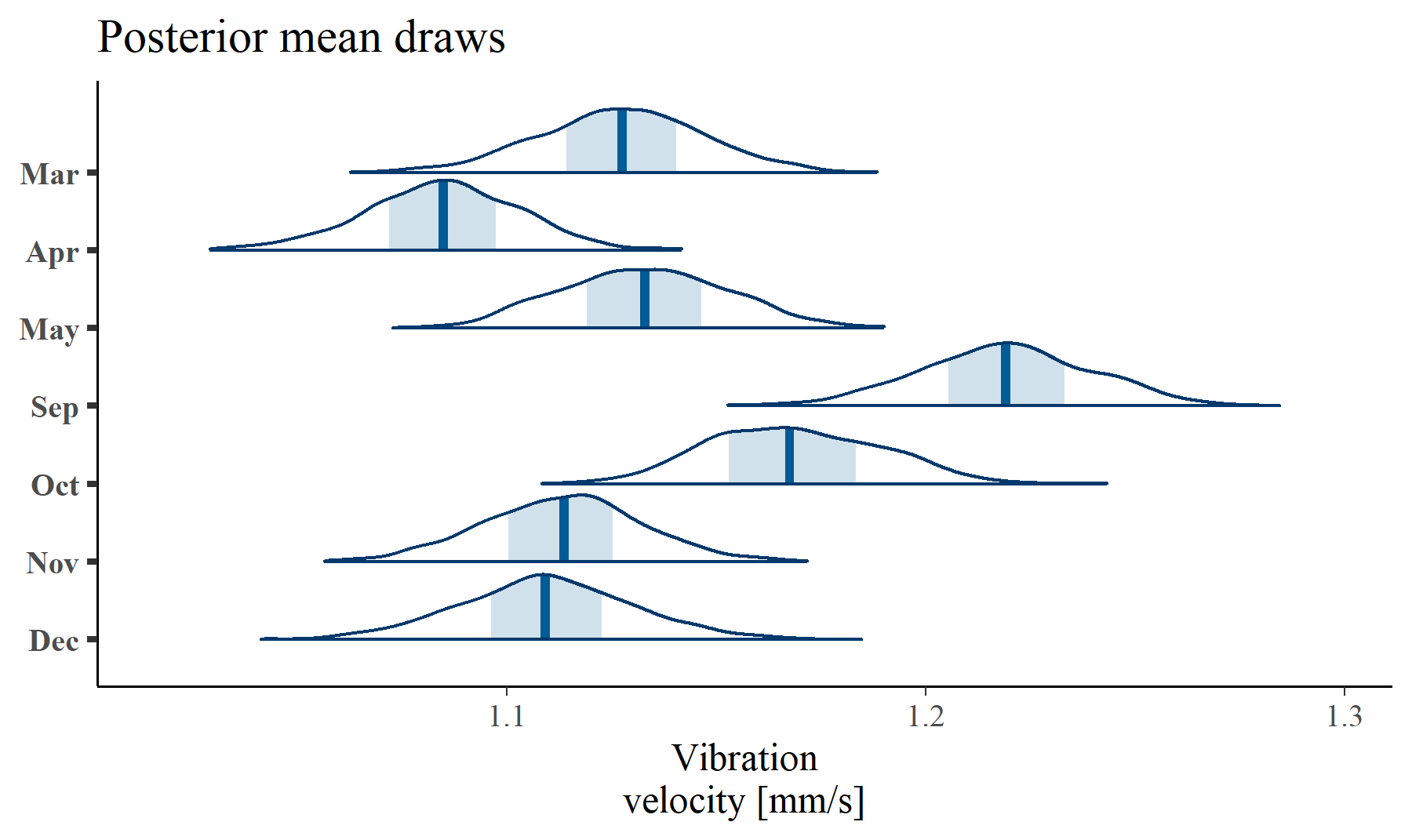
From this model, we can plot posterior draws of the mean value , which is shown in figure 4.

Figure 4: Draws from the posterior of the mean parameter for each month. The median is shown as a vertical line in the middle of each distribution.

From the plot there does not seem to be a noteworthy difference in the mean vibration levels during Spring. However, we do notice that September seems very high compared to all the others. Probabilities that each month is higher than the other are shown in table 3 in the appendix. From this table it is apparent that the probability that September was higher than any of the other months was very close to 1 for all months. However, there is no clear evidence that the Spring months in general tended to have higher vibration levels.

The reason for the high levels in September is due to the pump being completely taken apart and overhauled during the summer. When it was started back up in September, new parts had been installed, which typically causes instability in the first months until the parts have adjusted, either by mechanical changes or by operators adjusting them. This period can thus be viewed as a “burn-in” period, which is a usual phenomenon with large machinery.

## Prior sensitivity analysis

For sensitivity analysis, we choose to proceed with the monthly mean and common variance model, as it yielded a good cross validation result and conceptually is the simplest model. Recall that this model assumed a normal distribution on the variance. This might not be the most realistic assumption. For this reason, we will look at alternative priors, specifically Chi square and log normal. That is, the prior on the variance parameter will then be

and ,

where the prior parameters have been selected based on visual inspection of the empirical standard deviation.

|  |  |  |
| --- | --- | --- |
| **Model** | **ELPD difference** | **Standard error difference** |
| Monthly mean and common variance, **chi square prior** | 0 | 0 |
| Monthly mean and common variance, **log normal prior** | -0,2 | 0,1 |
| Monthly mean and common variance | -0,3 | 0,2 |

Table 3: PSIS-LOO cross validation results from the prior sensitivity analysis.

Due to space constraints we will just state the results of the prior sensitivity analysis using PSIS-LOO cross validation. These results are seen in table 3, from which it is apparent that there is not much difference between the original model and those with alternative priors, even though the chi square prior model comes out with the highest ELPD. The standard errors make this result quite insignificant, although, we should probably go with the chi square prior. Intuitively, this makes sense, as this a quite typical and reasonable distribution for the variance to have.

# Conclusion

We looked at the vibrations from the cooling water pumps from a large Danish power plant, with the initial motivation that we were told that the Spring months were supposed to have unusually high vibration levels. Based on the Bayesian analysis using Hamiltonian Monte Carlo, we did not find any reason to believe that this was actually the case. We did find, however, that the month of September showed higher level of vibrations, but this was simply due to parts being replaced during the summer, causing a burn-in period for the pump, before stability was regained due to adjusted parts.

# Appendix

## Probability table

The table below represents probabilities from the monthly means and common variance model. It shows the probability of a row-month being higher than a given column-month. For example, the probability that the mean vibration level in March was higher than April is 0,94. Also, the probability that September was higher than any of the other months is seen to be very close to 1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | March | April | May | September | October | November | December |
| Is higher than | March |  | 0,94 | 0,43 | 0,001 | 0,07 | 0,69 | 0,72 |
| April | 0,05 |  | 0,04 | 0 | 0,001 | 0,15 | 0,17 |
| May | 0,57 | 0,96 |  | 0 | 0,11 | 0,75 | 0,8 |
| September | 1 | 1 | 1 |  | 0,96 | 1 | 1 |
| October | 0,93 | 1 | 0,89 | 0,04 |  | 0,98 | 0,98 |
| November | 0,3 | 0,85 | 0,25 | 0 | 0,02 |  | 0,55 |
| December | 0,27 | 0,83 | 0,2 | 0 | 0,02 | 0,45 |  |

Table 4: Cross-tabulated probabilities under the monthly mean, common variance model.

## Code

The project was done in R version 4.1.0 together with the Bayesian modelling framework Stan via the RStan R package.

The code and data to execute the analysis is split into a number of files for better structure. This makes it less convenient to simply put a full transcript of the code here. For this reason, the code and everything related to the project can be found in my Github repo: <https://github.com/hviidhenrik/BDA_project>