sigma blowup analysis 2, large p

Contents

Experiment Overview

In this experiment, I analyze the behavior of the `covdepGE` algorithm when applied to a large dataset. Specifically, I have observed that in the large p regime, the MAPE-fitted σ^2 values blow up along with the μ values, and so, the purpose of this analysis is to more closely examine this blowup.

I perform 100 trials on data generated as described in the section titled *Data Generation* with $p + 1 = 25$. For each variable, I allowed each CAVI 1,000 iterations to converge during the grid search for optimal π. Additional grid searches were performed until stability in the optimal π value was attained. π was selected by maximizing ELBO over the following grid:

$$\underset{\sim}{\pi} = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$$

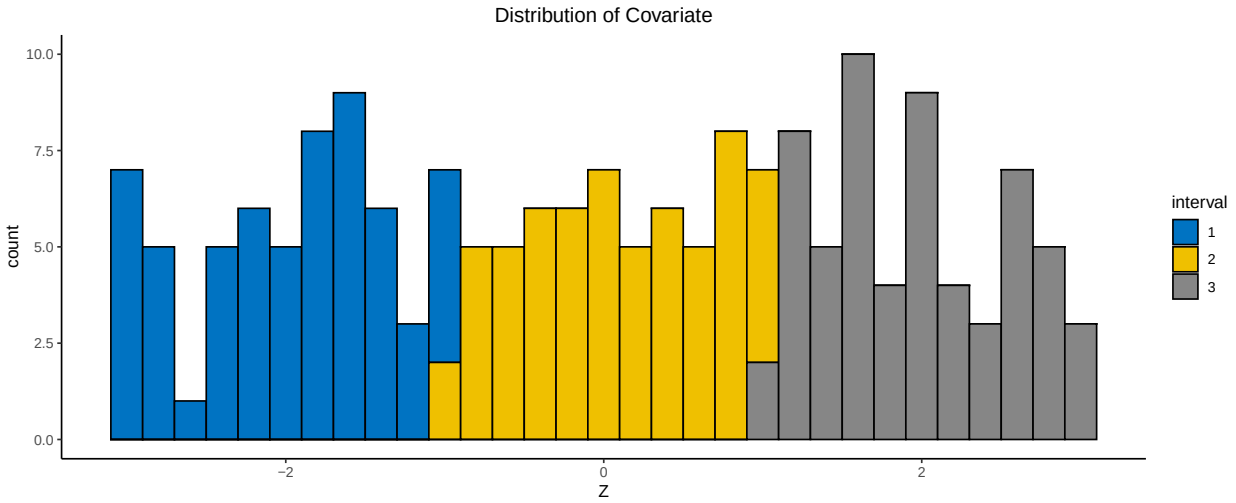
σ^2 and σ_β^2 were both fit to the data using MAPE.

In the *Results* section, I analyze the frequency and patterns of the blowups and the condition numbers associated with the blowups.

Data Generation

Extraneous Covariate

I generated the covariate, Z, as the union of three almost disjoint intervals of equal measure. That is, $Z = Z_1 \cup Z_2 \cup Z_3$ with $Z_1 = (-3, -1), Z_2 = (a, b) = (-1, 1), Z_3 = (1, 3)$. Within each interval, I generated 60 covariate values from a uniform distribution. For example:



Precision Matrix

All of the individuals in interval 1 had the same precision matrix, $\Omega^{(1)}$:

$$\Omega^{(1)}_{i,j} = \begin{cases} 2 & i = j \\ 1 & (i,j) \in \{(1,2),(2,1),(2,3),(3,2)\} \\ 0 & o.w. \end{cases}$$

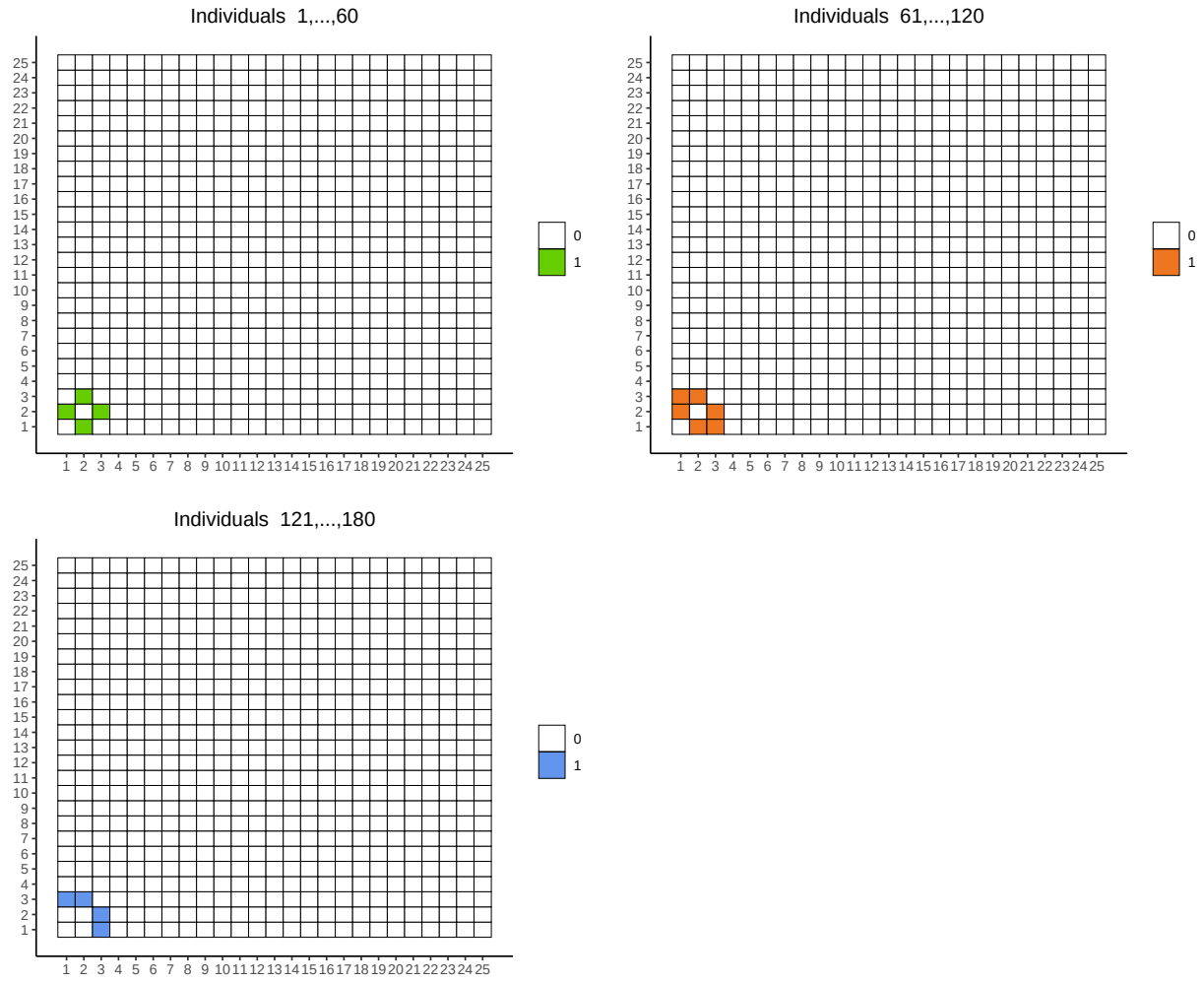
Also, all of the individuals in interval 3 had the same precision matrix, $\Omega^{(3)}$:

$$\Omega^{(3)}_{i,j} = \begin{cases} 2 & i = j \\ 1 & (i,j) \in \{(1,3),(3,1),(2,3),(3,2)\} \\ 0 & o.w. \end{cases}$$

However, the individuals in interval 2 had a precision matrix that was dependent upon Z and (a, b). Let $\beta_0 = -a/(b-a)$ and $\beta_1 = 1/(b-a)$. Then:

$$\Omega^{(2)}_{i,j}(z) = \begin{cases} 2 & i = j \\ 1 & (i,j) \in \{(2,3),(3,2)\} \\ 1 - \beta_0 - \beta_1 z & (i,j) \in \{(1,2),(2,1)\} \\ \beta_0 + \beta_1 z & (i,j) \in \{(1,3),(3,1)\} \\ 0 & o.w. \end{cases}$$

Thus, $\Omega^{(2)}(a) = \Omega^{(1)}$ and $\Omega^{(2)}(b) = \Omega^{(3)}$. That is, an individual on the left or right boundary of Z_2 would have precision matrix $\Omega^{(1)}$ or $\Omega^{(3)}$, respectively. The conditional dependence structures corresponding to each of these precision matrices are visualized below.



Data matrix

Let z_l be the extraneous covariate for the l-th individual. To generate the data matrix for the l-th individual, I took a random sample from $\mathcal{N}(0, \{\Omega_l(z_l)\}^{-1})$, where:

$$\Omega_l(z_l) = \begin{cases} \Omega^{(1)} & z_l \in Z_1 \\ \Omega^{(2)}(z_l) & z_l \in Z_2 \\ \Omega^{(3)} & z_l \in Z_3 \end{cases}$$

Results

Distribution of MAPE fitted σ^2

I first analyze the distribution of the MAPE-fitted σ^2. There are 100 trials, each with 25 variables; for each variable, $n = 180$ weighted regressions are fit. Thus, there are $100 * 25 * 180 = 450,000$ fitted σ^2. Of these, 797 "blew up" - that is, had a value exceeding 15.

Of the σ^2 that did not blow up, the largest value was roughly 13 - these σ^2 are all contained in the first bin of the log-log histogram. Of those that did blow up, the smallest value was on the order of $1e33$.

Note that the spike at the maximum value occurs from imputing 360 `NA` values resulting from `Inf / Inf` as the maximum value.

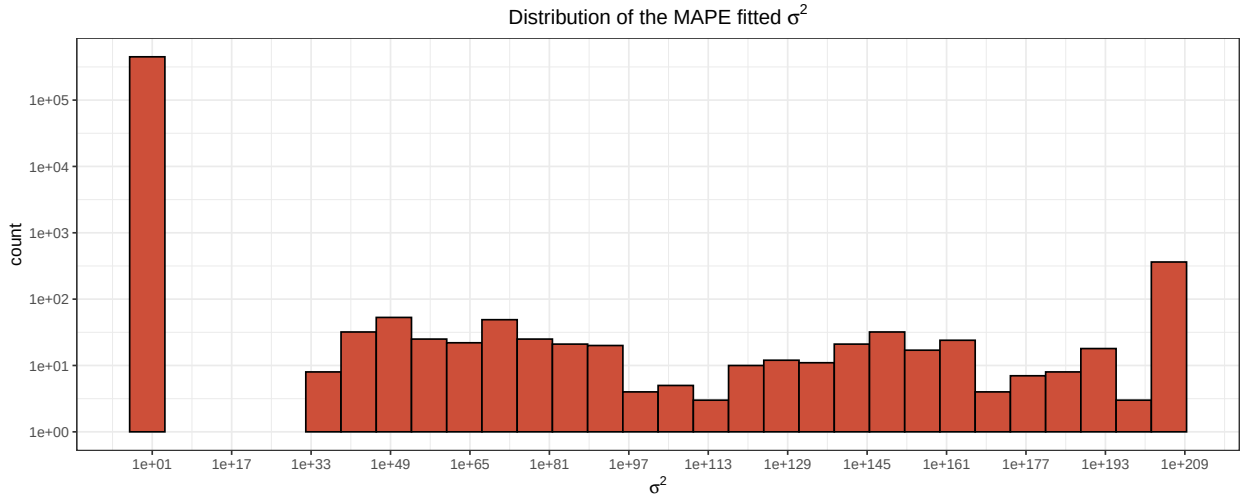
```
## [1] 13.18032
```

```
## [1] 3.858987e+33
```

```
## [1] 450000
```

```
## [1] 797
```

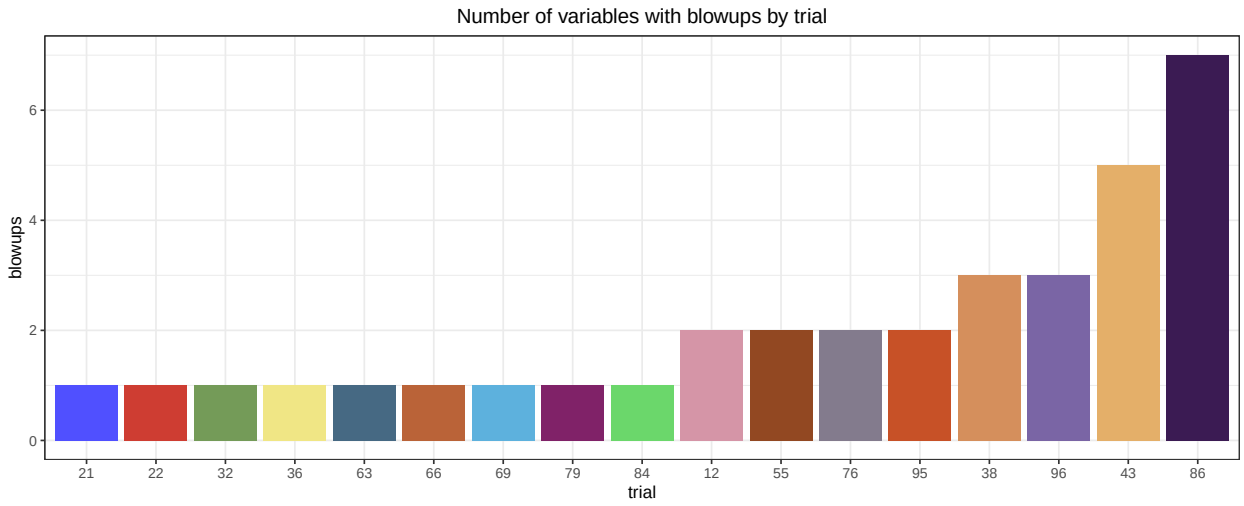
```
## [1] 360
```



Breaking down the blowups

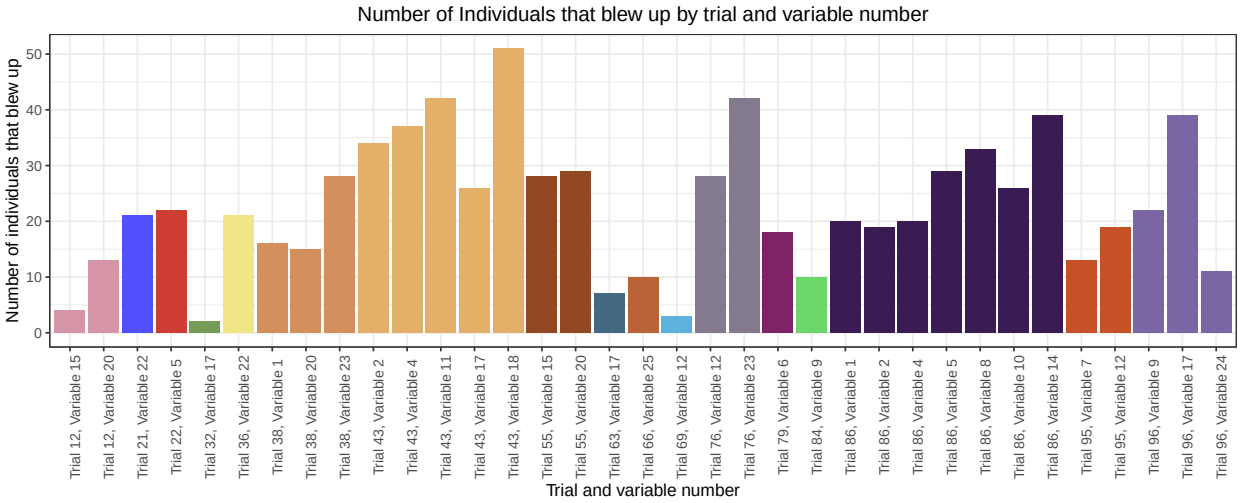
Here, I look at the blowups broken down by trial. Of the 100 trials, 17 had at least one indvidual whose σ^2 value blew up. The maximum number of variables that blew up occured in trial 86, when 7 of the 25 variables had at least 1 individual that blew up. Trials having 0 blowups are omitted from this figure.

```
## [1] 17
```



Breaking this down even further, the following histogram shows the number of individuals per variable per trial that blew up. Variables with 0 blowups are omitted here.

The maximum number of individuals to blow up for a single variable occurred in trial 43 for variable 18, with 51 individuals blowing up. The minimum occurred in trial 32 for variable 17, with just 2 individuals blowing up.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   14.00   21.00   22.77   29.00   51.00
```

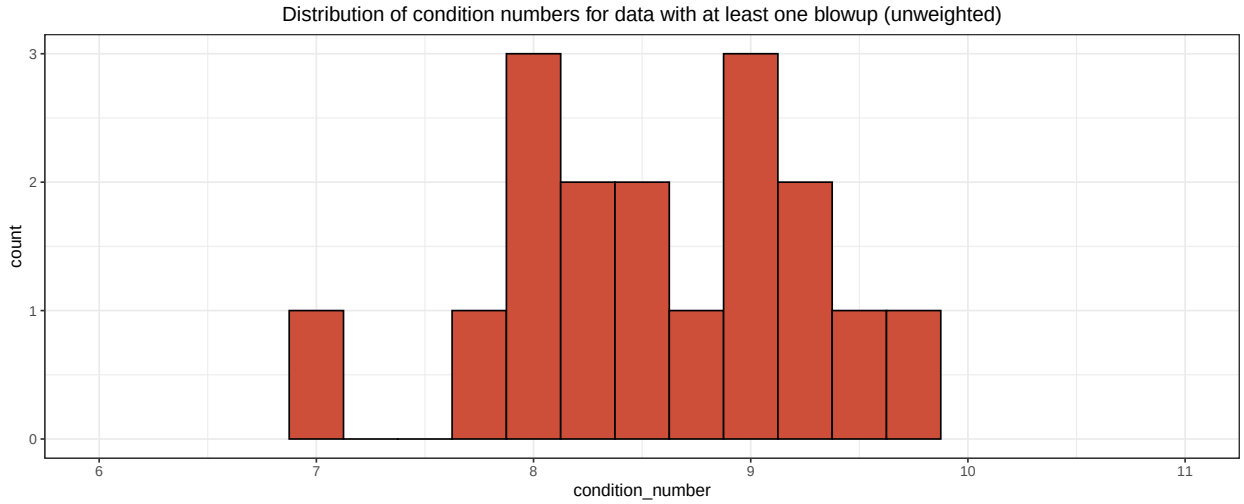
```
## trial32.variable17
##                  5
```

```
## trial43.variable18
##                 14
```

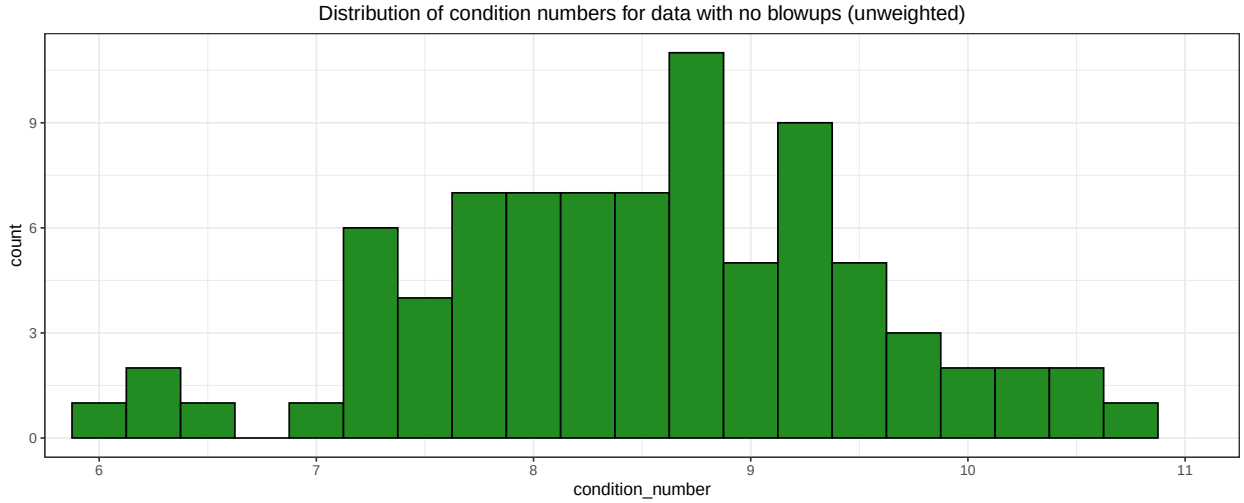
Condition number analysis

Raw data

The following shows the distribution of the condition numbers of the matrices in trials that had at least one individual that blew up, versus those in which no individuals blew up. That these distributions have no striking differences in shape motivates a further analysis of the weighted matrices.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.106   8.096   8.536   8.587   9.103   9.741
```

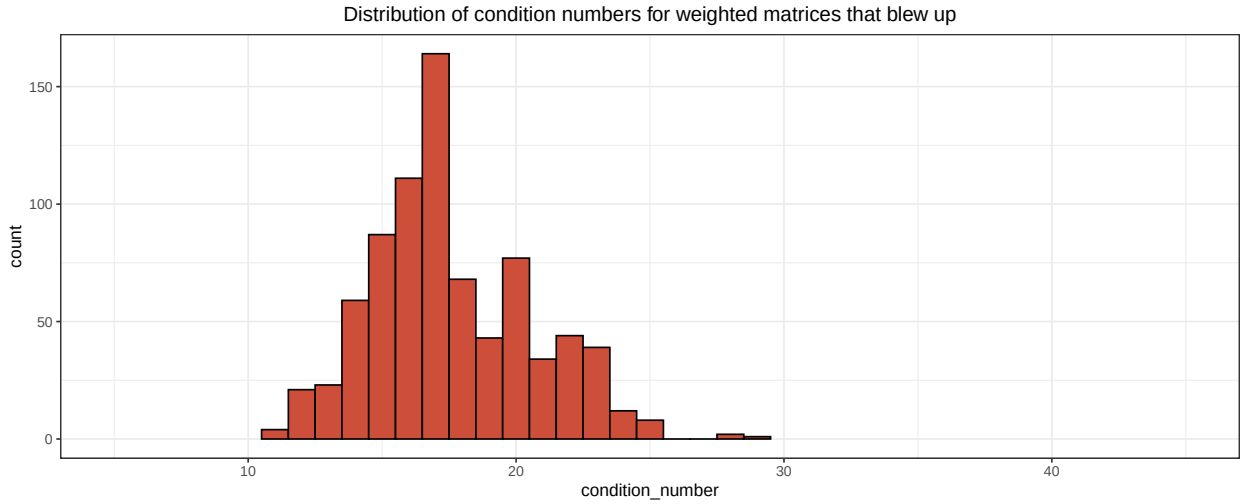



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.051   7.832   8.554   8.517   9.193  10.867
```

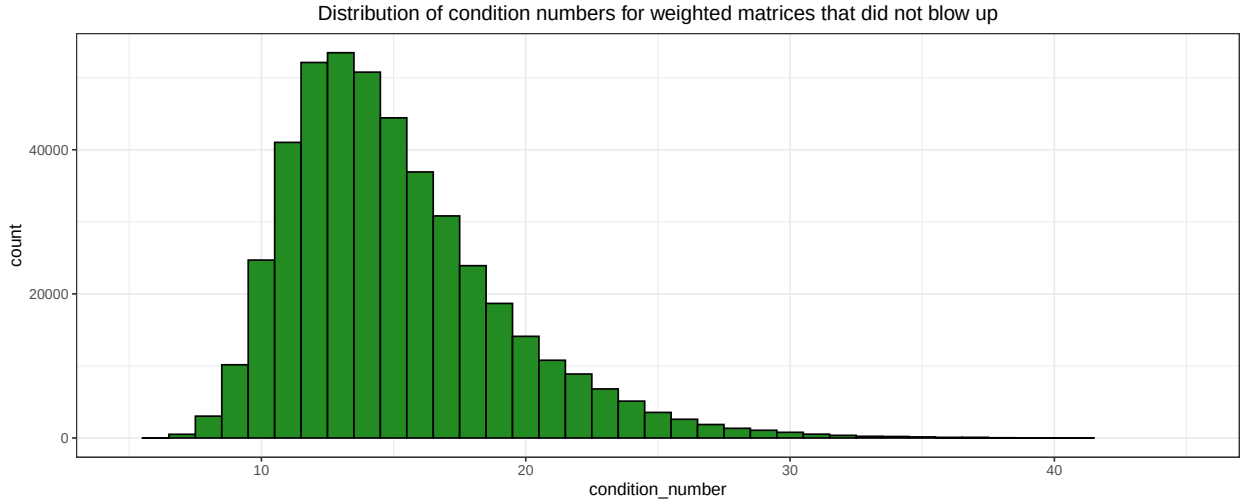
Weighted data

As mentioned in the section analyzing the distribution of the MAPE-fitted σ^2, there were 797 individuals that blew up. Thus, there are 797 unique weighted matrices corresponding to these individuals. In this section, I analyze the condition number of these matrices, as well as the remaining $450,000 - 797$ that did not blow up.

The distribution of the condition numbers for the matrices that blew up demonstrates a greater center than those that did not blow up. Thus, this suggests that the cause of the blowup is associated with the weights, since there was no difference in the condition numbers of the raw data.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.22   15.57   17.08   17.61   19.90   28.56
```

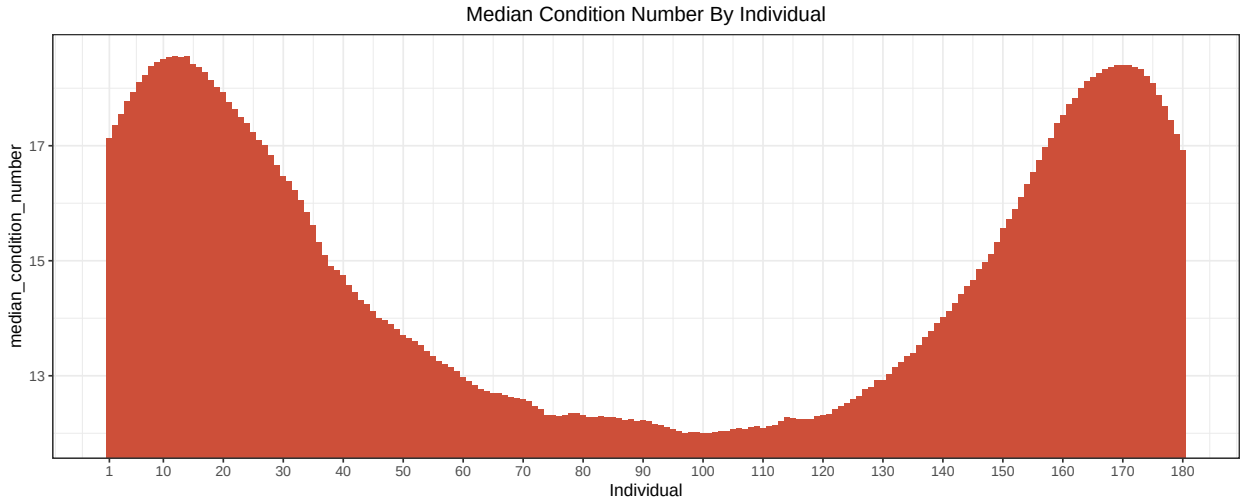



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.824  12.136  14.275  15.038  17.120  40.918
```

Individual breakdown

Since it appears that the increased condition numbers are associated with individual-specific weights, I now look at the blowups on an individual level.

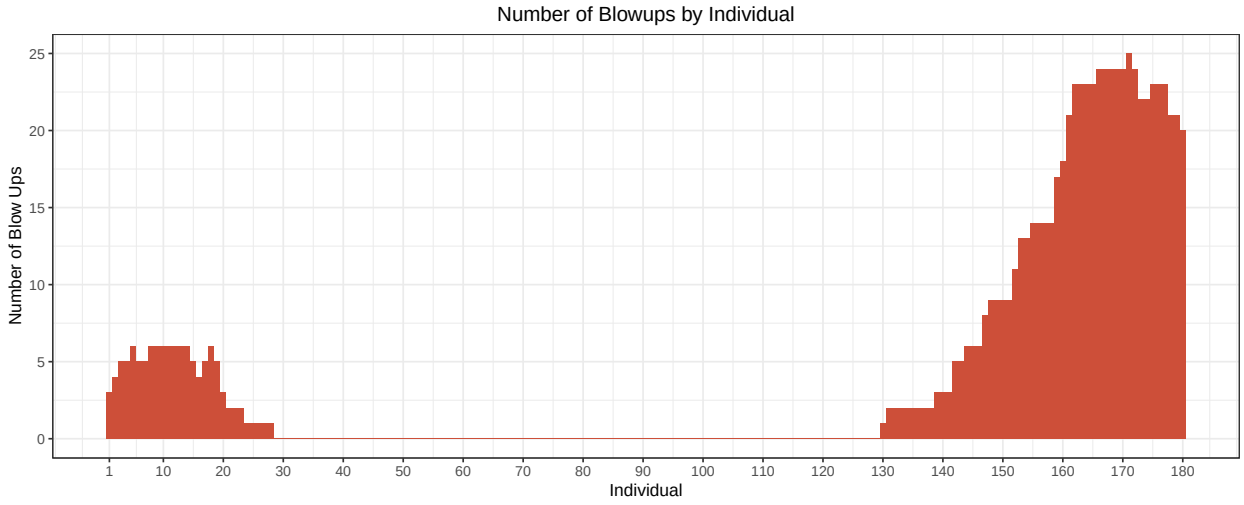
Since each individual has $100 * 25 = 2,500$ unique weighted matrices, I calculate the condition number for each of these matrices and then display the distribution of the median condition number for each individual. As can be seen, individuals with the lowest and highest indices (which corresponds to being on the extremes of the covariate interval) have the greatest condition numbers.



I now display the counts of the blowups for each individual. Note that each individual had $100 * 25 = 2,500$ opportunities to blow up; the greatest number of blowups was achieved by individual 171, with 25.

From the previous and following figures, it is clear that as median condition number increases, so does the tendency to blow up.

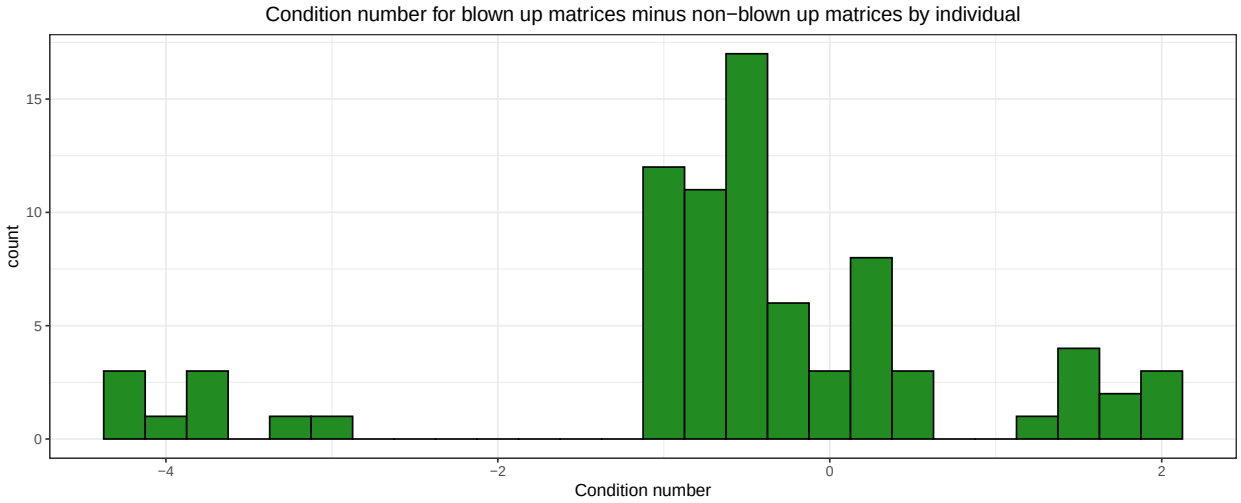
```
## [1] 79
```



Blown-up individual breakdown

I finally restrict my attention to just the individuals that blew up at least one time. There are 79 such individuals. Across the 2,500 opportunities for a given individual to blow up, I first calculate the mean condition number for the matrices corresponding to the times when this individual did blow up, and then compare this to the mean condition number for when this individual did not blow up.

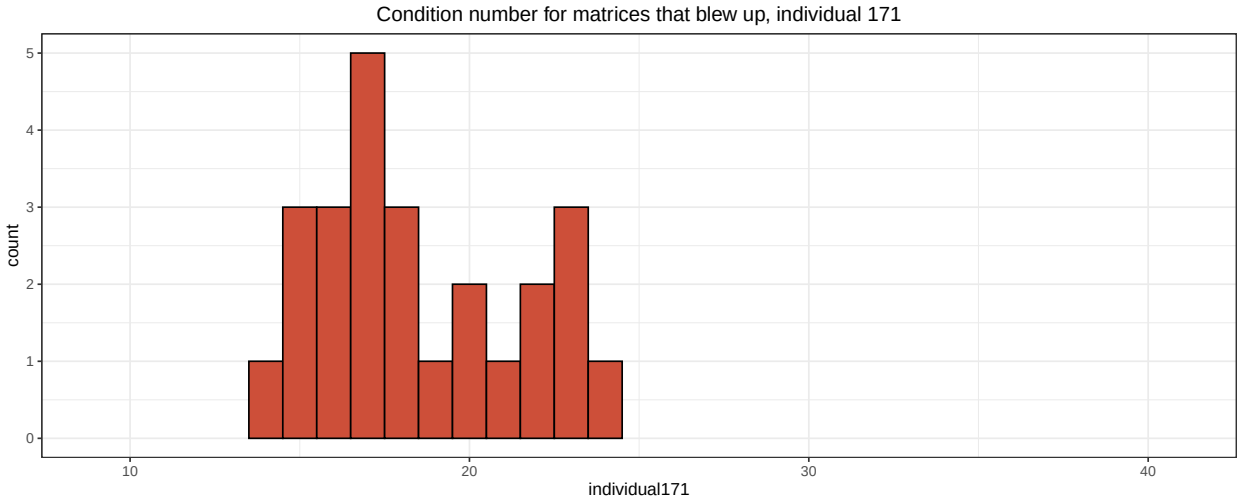
I then take the pairwise difference for each of these individuals; that is, for a given individual, I subtract the mean blown-up condition number from the mean non-blown up condition number. I would expect this difference to be positive on average, however, this is not the trend that I observed, suggesting that there is more at play than just the condition number of the weighted matrices.

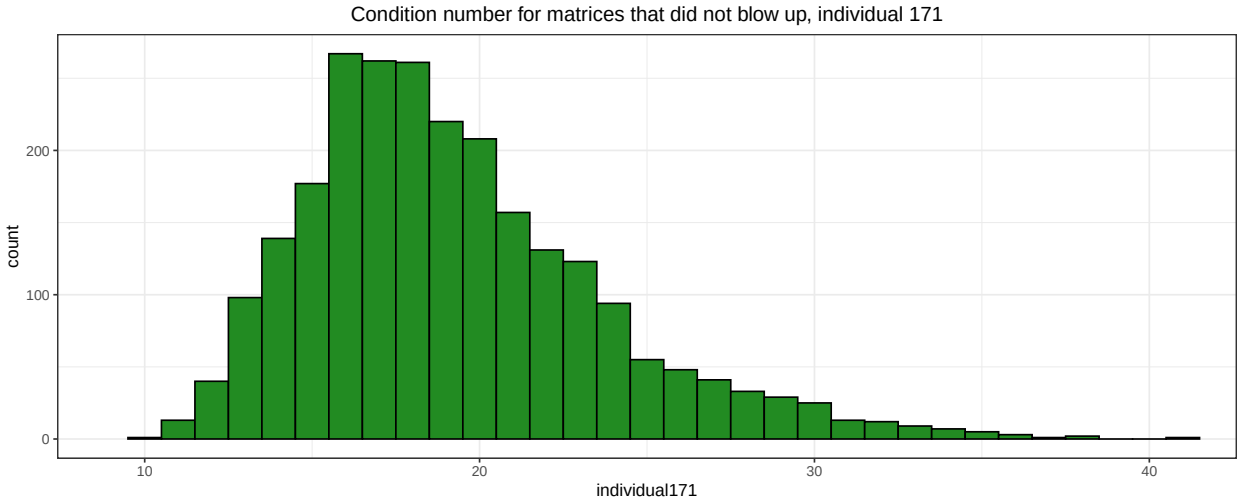



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.2868 -0.9588 -0.5004 -0.5684  0.2371  2.0717
```

Individual 171

I finally display the distribution of the blown-up condition numbers versus the non-blown up condition numbers for a single individual. I selected individual 171, since this individual blew up more than any other indvidual - in the 2,500 opportunities to blow up, this individual blew up a total of 25 times.




Condition number for matrices that did not blow up, individual 171
count
0
100
200
10
20
30
40
individual171