Empirical Scaling Analyzer:
An Automated System for Empirical Analysis of Performance Scaling

Supplementary Material

Yasha Pushak[1], Zongxu Mu[1] and Holger Hoos[1,2]

[1]Department of Computer Science, University of British Columbia
[2]Computer Science Institute of Leiden University

{ypushak,zongxumu,hoos}@cs.ubc.ca

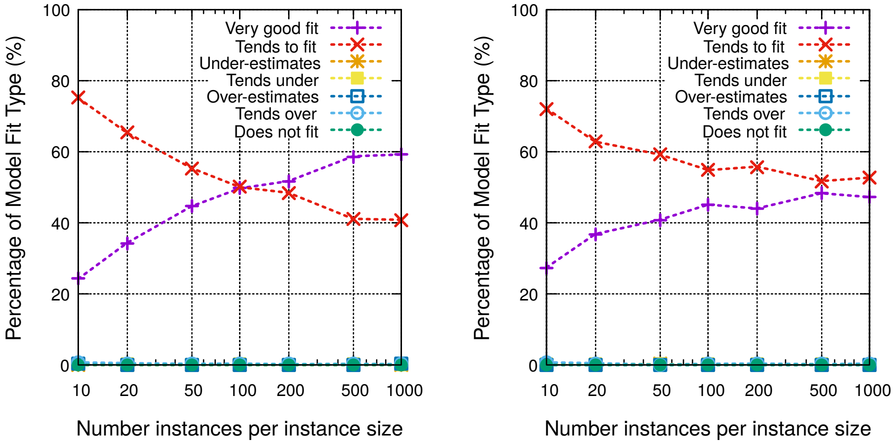
1 Introduction

This document contains a detailed discussion of the stress tests that are summarized in the paper "Empirical Scaling Analyzer: An Automated System for Empirical Analysis of Performance Scaling" [1].

2 ESA Stress Testing

There are many potential factors that could cause ESA to report misleading or incorrect results: for example, it could mistakenly accept the wrong scaling model because of misleading lower-order terms, or because the bootstrap prediction samples are so large that any model appears to fit the data. While the latter case is more benign than the first, it is not immediately clear what steps should be taken to resolve the problem. Most likely, more data is needed of some kind – perhaps running times for more instance sizes, larger challenge sizes, more instances per size, or more independent runs per instance.

We generated running time data sets with varying properties to allow us to examine how the performance of ESA degrades as the scenarios become increasingly difficult. In total, we chose five different properties to vary: the number of instances per instance size, the number of support instance sizes, the number of independent runs per instance, the extrapolation distance and the number of bootstrap samples used in ESA's bootstrap sampling procedure. Clearly, as the total number of support instances are decreased, we expect the variability of the fitted models to increase. However, it is not immediately obvious which property will have the largest impact, and hence how users should allocate their time budget to obtain the most reliable and accurate results. The extrapolation distance is another key property that the users can control. Intuitively, the further we are able to extrapolate with a model while still obtaining consistent results, the more strongly we will trust the scaling of the model. The number of



(a) Poly. model on Poly. data set

(b) Exp. model on Exp. data set

Figure 1: The model fit types reported by ESA's automatic interpretation of the scaling results as the number of instances per instance size is varied.

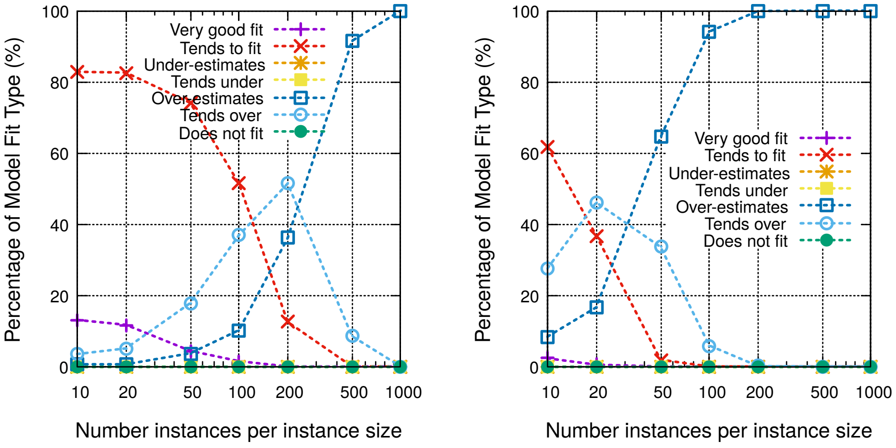
bootstrap samples performed by ESA is another interesting parameter to study since it affects the variance of the bootstrap intervals in a particular run of ESA.

2.1 Varying the Number of Instances per Size

We used seven different values for the number of instances per size, chosen using an approximately logarithmic scale. In particular, we used 10, 20, 50, 100, 200, 500 and 1000 instances per instance size.

In Figure 1a we plot the percentage of times that ESA reported a polynomial model fit the polynomial data very well, tended to fit the data, *etc.*, as we vary the number of instances per instance size. These model classes correspond directly to ESA's automatic interpretation of the scaling. In Figure 1b we show similar results for the exponential model on the exponential data set. We can see that even with a very small number of instances ESA is able to identify that the correct scaling model tends to fit the data in nearly all of the cases. We can also see that as the number of instances per instance size increases, the number of times ESA identifies a very good fit tends to increase.

Overall, the results in Figures 1a and 1b are qualitatively very similar. The biggest difference is that the exponential model is reported to fit the data very well less often. We believe that this is a property specific to these particular running time data sets, rather than the models themselves. Our hypothesis is that the size of the confidence intervals for the model predictions is smaller for the exponential models since the running times for the small support instance sizes are larger than those for the corresponding sizes in the polynomial data set. Since being off by a factor of, for example, 2 contributes less to the support RMSE



(a) SqrtExp. model on Poly. data set

(b) Exp. model on Poly. data set.

Figure 2: The model fit types reported by ESA's automatic interpretation of the scaling results as the number of instances per instance size is varied.

of the model if the running times are small, the smaller running times for the polynomial contribute less to the fitted model. As a result, there is less overall uncertainty in the fitted exponential models. This means that the challenge instance size confidence intervals for the model predictions are smaller and hence are less likely to be strongly consistent with the observed data. Throughout most of the following, we omit the results from the experiment with the exponential data set since they are typically very similar to those for the polynomial data set, apart from this one qualitative difference. We note any exceptions and discuss possible causes for their differences.

Unfortunately, while Figures 1a and 1b may suggest that 10 instances per instance size is reasonable for practical use, they alone do not show the full picture. In Figures 2a and 2b we show the percentage of model fit types ESA reported for the square-root exponential and exponential models on the polynomial scaling running time data set. Here we can see that ESA is unable to correctly distinguish the true model scaling without more instances. Indeed, for 10 instances per instance size, the square-root exponential model is reported to fit the data very well 13.1% of the time and to tend to fit the data 82.8% of the time. We even see that the exponential model obtains many false-positives, and is reported to tend to fit the data 61.7% of the time. For our data sets we need 500 or more instances per instance size before ESA is able to correctly recognize that the square-root exponential model is an over-estimate for the true scaling in most of the runs.

Another important metric for assessing the effect of varying the number of instances per instance size is the location and size of the bootstrap confidence intervals for the model predictions. In Figure 3 we show the location and size of

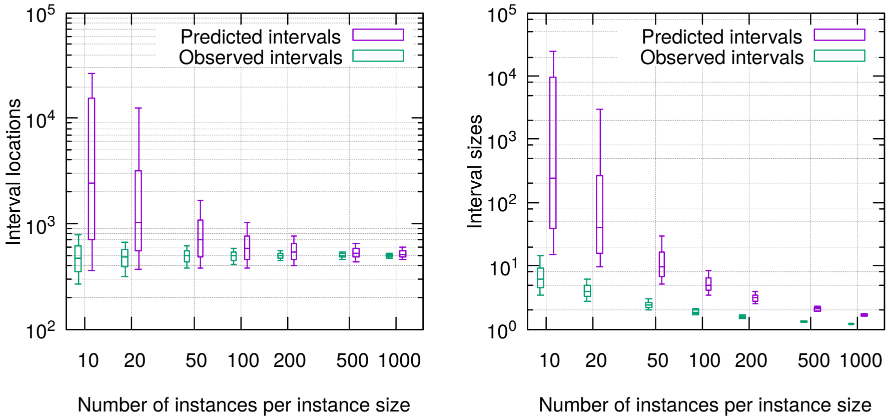
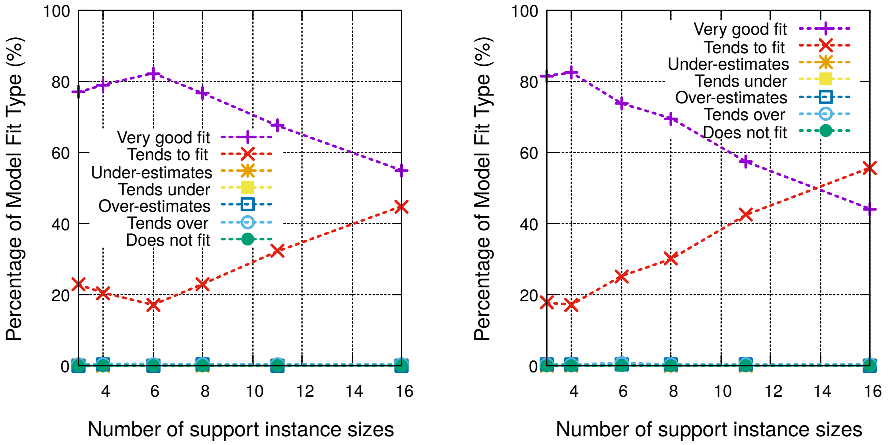


Figure 3: The confidence intervals for the model predictions and observations for instance size 4 500 – the largest challenge instance size, as the number of instances per instance size is varied. This data is for the polynomial data set and the intervals are for the fitted polynomial predictions. Plotted on the left are the interval locations and on the right are the interval sizes. The box plot whiskers are the 10^{th} and 90^{th} quantiles. The prediction and observation interval box plots have been shifted slightly right and left, respectively, for visibility.

the prediction and observation bootstrap confidence intervals for instance size 4500 – the largest challenge instance size. In particular, the intervals for the predictions are those of the polynomial model on the polynomial data set. For these figures, we define the location of a bootstrap interval to be the geometric mean of the upper and lower bounds of the interval. We omit the bootstrap interval data for the square-root exponential and exponential model, which tend to show qualitatively similar behaviour, though they are both higher and larger. In some cases, the Levenberg-Marquardt algorithm used by ESA has trouble fitting the exponential model to the polynomial data set, so the default fitting parameters are returned. This happens the most often for small numbers of instances per instance size, which results in bootstrap interval sizes smaller than would otherwise be expected.

From Figure 3 we can see that the size of the confidence intervals for the model predictions is often more than an order of magnitude larger than the observed ones when only 10 instances are used per instance size, in addition, the variability in the size of the intervals is huge. Since we also have very large confidence intervals for the square-root exponential and exponential models predictions, we can see why we observed so many false positives for these models when small numbers of instances per instance size were used. We can also see that the location of the intervals for the predictions tends to be biased towards locations above those of the intervals for the observations. In part, we believe this to be due to an asymmetry in how the fitted models are effected when the



(a) Poly. model on Poly. data set.

(b) Exp. model on Exp. data set.

Figure 4: The model fit types reported by ESA's automatic interpretation of the scaling results as the number of support instance sizes is varied.

largest few instance sizes are outliers. For example, we have observed that if the largest support instance size is an outlier that falls higher than the true (unobserved) running time, then the fit of the model tends to be more affected than if the largest instance size had been an outlier that was a similar amount below true running time (we believe this is because a model with a higher degree can still fit the data well for high outliers, but the models are unable to "bend" downwards, so the fitting procedure is forced to compromise when the outlier is below the data). Nevertheless, increasing the number of instances per instance size decreases the variability of the data, and hence decreases both this bias and the variability in the locations.

2.2 Varying the Number of Support Instance Sizes

In order to keep as many features of the data set constant as possible (such as the extrapolation distance or the range covered by the support sizes) we fixed the location of all of the challenge instance sizes and the largest support instance size, and we forced the smallest support instance size to always be either 500 or 600. Then, to control the number of support instance sizes we varied their "density". For example, when using 8 support instance sizes instead of 16, we used every second support instance size of the ones available from 500, 600, ..., 2000, i.e., we used 600, 800, ..., 2000.

In Figures 4a and 4b we show the percentage of model fit types reported for the polynomial model on the polynomial data and the exponential model on the exponential data. To our initial surprise we found that ESA reported the most very good fits with only 6 support instance sizes for the polynomial data and

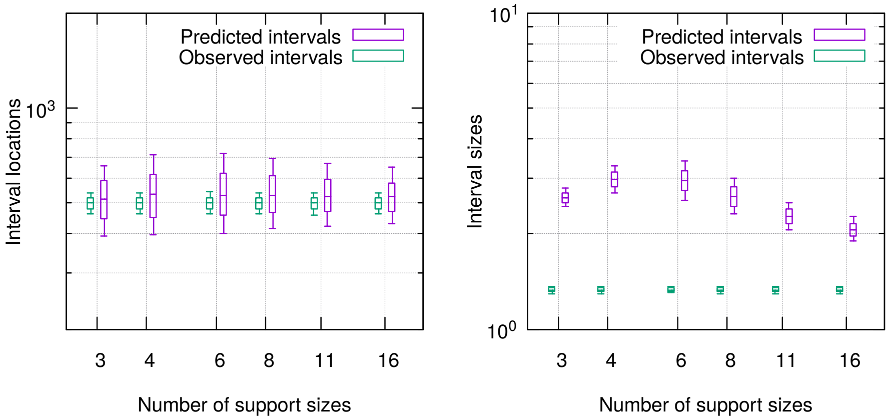
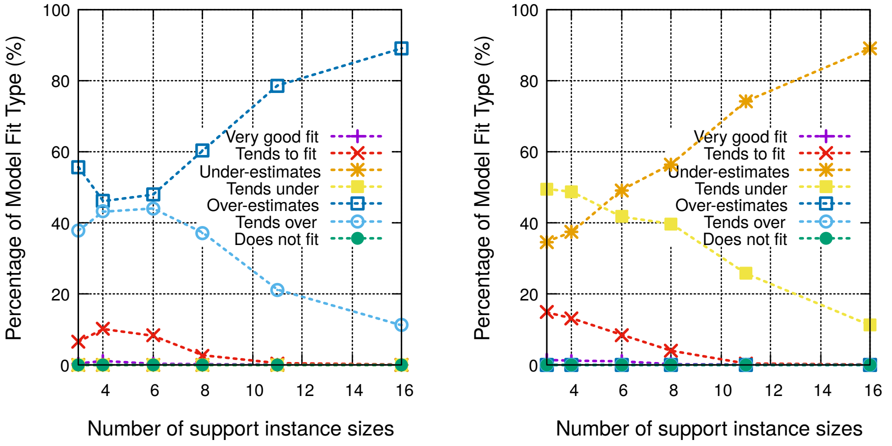


Figure 5: The bootstrap intervals for the model predictions and for the observations for instance size 4 500 – the largest challenge instance size, as the number of support instance sizes is varied. This data is for the polynomial data set and the fitted polynomials.

with only 4 instance sizes for the exponential data. We were also surprised to see that the response was non-monotonic. This observation is most noticeable for the polynomial data set, which starts at 76.8% before increasing to 82.2% and after which it descends to 54.8%. The decrease in the percentage of very good model fits can be easily understood by looking at Figures 5 and 12, which show the location and size of the prediction and observation bootstrap intervals for the polynomial model and data set, and the exponential model and data set, respectively. As expected, we see that in both cases the sizes and locations of the bootstrap intervals for the observations do not change. With the exception of first few data points we see an overall decreasing trend in the size of the bootstrap intervals for the model predictions – which is another unsurprising observation, since we expect the variance in the predictions to decrease as we use more support instance sizes.

An unexpected observation that can be made for Figure 5 is that for 3 support instance sizes the bootstrap intervals for the model predictions are smaller than for 4 support instance sizes, which we believe is what caused the non-monotonic response in Figure 4a. However, because the effect is relatively small, and because we do not recommend to use as few as 6 support instance sizes, we refer the interested reader to Appendix A for a discussion of why we believe we have observed this behaviour.

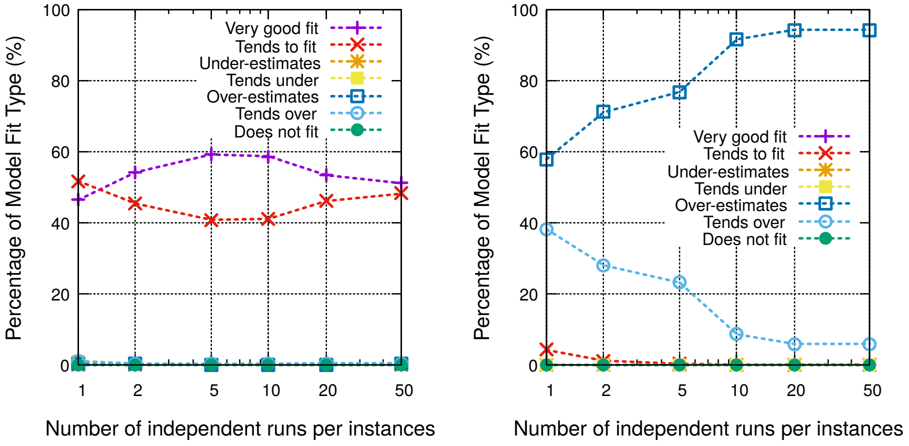
In Figures 6a and 6b we show the percentage of model fit types reported for the square-root exponential models on the two data sets. We omit the results for the remaining two cases since ESA correctly identifies the scaling in 100% of the runs for the exponential model on the polynomial data set, and in nearly 100% of the runs for the polynomial model on the exponential data set. We see from



(a) SqrtExp. model on the Poly. data set. (b) SqrtExp. model on the Exp. data set.

Figure 6: The model fit types reported by ESA's automatic interpretation of the scaling results as the number of support instance sizes is varied.

Figures 6a and 6b that in a modest percentage of the runs with small numbers of support instance sizes ESA reports that the square-root exponential model tends to fit the data. However, because these percentages are smaller than when we used small numbers of instances per instance size, it may appear that a user of ESA can save time by using less support instance sizes without significant loss. However, we do not believe that this is true. It is also important to compare the relative sizes of the bootstrap intervals for the model predictions. When the number of support instance sizes is reduced to a small number, the size of the intervals only increases by a very small amount relative to the increase that we observed in the previous section (note that we had to use a larger range for the y-axis in Figure 3 to capture the huge variances). When an ESA user sees that the size of the bootstrap intervals for the model predictions are very large they are likely to have little confidence that each model will continue to fit the data as the extrapolation distance is increased. As a result, false positives with very large bootstrap intervals are significantly more benign than false positives relatively small bootstrap intervals. We found that the false positives we observed in these experiments tended to occur when one of the largest two support instance sizes were outliers, and hence all of the models were shifted up or down. In some cases, this resulted in an incorrect model appearing to fit the data reasonably (or very) well. When this happens, because there are so few support instance sizes, there are often no visible signs that ESA has incorrectly classified the scaling of the data. As a result, we recommend to always use as many support instance sizes as possible. In particular, we have found both from these experiments and from our experience using ESA in prior work that it is best to use at least 11 support



(a) Poly. model on Poly. data set.

(b) SqrtExp. model on Poly. data set.

Figure 7: The model fit types reported by ESA's automatic interpretation of the scaling results as the number of independent runs per instance is varied.

instance sizes to obtain robust results. However, depending on the variability in your data you may require more or less support instance sizes.

2.3 Varying the Number of Independent Runs per Instance

We studied a set of 6 values for the number of independent runs per instance: 1, 2, 5, 10, 20 and 50.

In Figures 7a and 7b we show the percentage of model fit types for the polynomial and square-root exponential models on the polynomial data set. We omit the results for the exponential model since it was reported to over-estimate the data in 100% of the runs. Overall, we see that the effect of varying the number of independent runs per instance in Figure 7a is small compared to the response we observed when we varied the number of instances per instance size and the number of support instance sizes. Despite this observation, it is interesting to note that we are again seeing a non-monotonic response in the percentage of very good fits for the polynomial model. To understand what could be causing this response, we need to consider both the variance of the location and size of the bootstrap intervals, shown in Figure 8. In this case we see that the variance of the location of the intervals decreases slightly over the range of 1 to 5 independent runs per instance, after which it remains roughly constant. On the other hand, we see that the size of the intervals tends to decrease over the entire range of values. Since the variance in the interval locations is initially decreasing there is a higher probability that the bootstrap intervals for the predictions will be consistent with the observations; however, as the variance remains constant but the size of the intervals continues to decrease, the probability that they will be strongly consistent decreases.

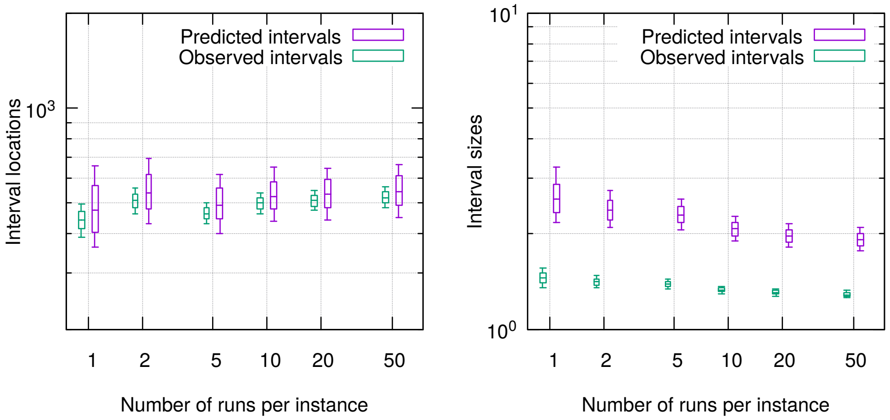
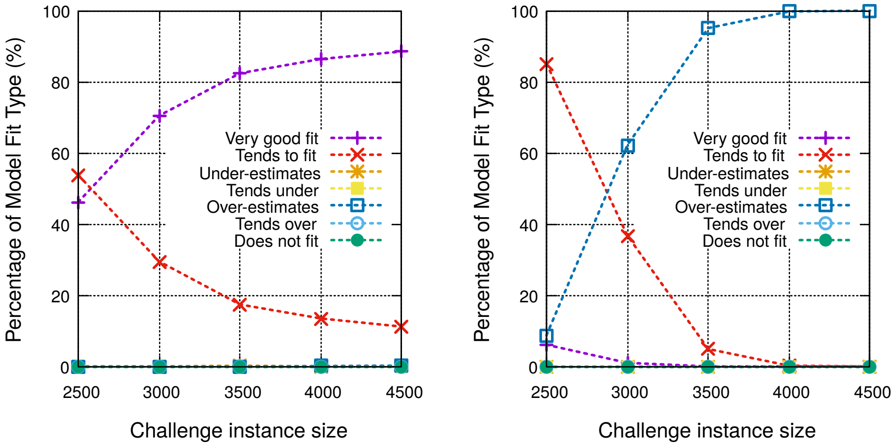


Figure 8: The prediction and observation bootstrap intervals for instance size 4 500 – the largest challenge instance size, as the number of independent runs per instance is varied. This data is for the polynomial data set and the prediction intervals are for the fitted polynomials.

Despite the relatively small response that we observe for the polynomial model, we do see that ESA has a significantly higher probability of correctly identifying that the square-root exponential model over-estimates the data as the number of runs per instance are increased. This response is unsurprising, since for an incorrect model a decrease in the variance of the intervals' locations and a decrease in size of the intervals both lead to higher probabilities that the intervals for the model predictions will be disjoint from the observations.

What we found the most interesting was the comparison of these results from those we obtained when we decreased the number of instances per instance size. In particular, Consider the decrease from 10 runs per instance to 1 run per instance, compared to the difference from using 500 instances per instance size to 50 instances per size. In both cases, we are decreasing the total number of target algorithm runs by a factor of 10. However, the response in the size of the bootstrap intervals, and hence in ESA's interpretation of the model fit, is drastically different in the two cases. In particular, when 1 run per instance is performed the median size of the bootstrap interval for the polynomial model predictions on the polynomial data set for instance size 4 500 is 2.6, compared to when 50 instances per instance size are used, which resulted in an interval size of 9.9 for the model predictions. The percentage of false positives for the square-root exponential model on the polynomial data set is only 4.1% with 1 run per instance compared to 74.1% with 50 instances per instance size – *i.e.*, there are 70.0% more false positives when reducing the number of instances per size by a factor of 10 than when reducing the number of runs per instance by a factor of 10. With such a striking difference it is clear that if the time required to collect all of the running time data is constrained, then the best option is to



(a) Poly. model on Poly. data set.

(b) SqrtExp. model on Poly. data set.

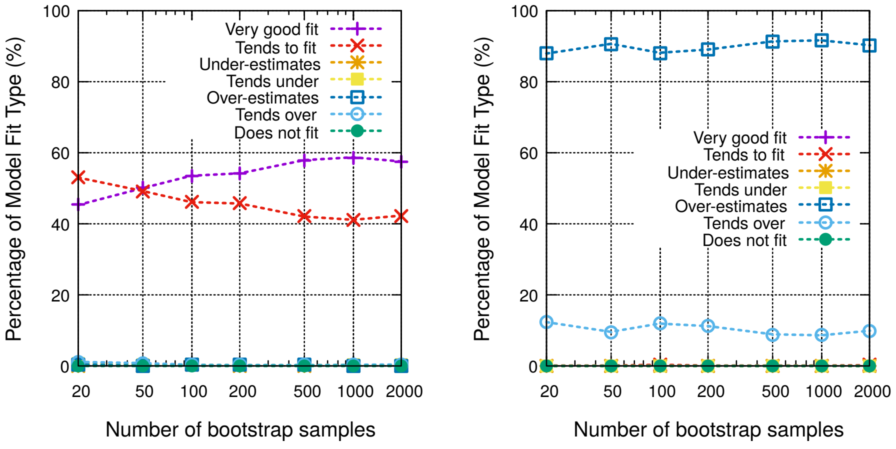
Figure 9: The model fit types reported by ESA's automatic interpretation of the scaling results as the extrapolation distance is varied.

use very few independent runs per instance and to choose to instead use more instances per instance size if they are available.

2.4 Varying the Extrapolation Distance

To isolate the response from varying the extrapolation distance, we used only a single challenge instance size for these experiments. In order to obtain an additional location for the challenge instance size, we also used only 11 support instance sizes, 500, 600, ... 2 000, instead of the 16 that we used in the previous experiments. This allowed us to use five different challenge instance sizes: 2 500, 3 000, ..., 4 500.

In Figures 9a and 9b we show the model fit types reported by ESA for the polynomial data set for the polynomial and square-root exponential models, respectively. We omit the results for the exponential model since it was reported to over-estimate the data in 100% of the runs, except for when the challenge instance size is 2 500, in which case it over-estimated the data in 90.8% of the runs and tended to fit the data in the rest. We note that ESA did not report that any of the models tended to over-or under-estimate the data in these experiments since these statements are based on the fraction of challenge instance sizes that are disjoint from the intervals for the observations. In this case, since we had only a single challenge instance size, this fraction was either 0 or 1. Overall, we see that the results from the three figures line up well with our intuition. The farther the extrapolation the higher the probability that ESA will correctly identify the polynomial scaling and the higher the probability that ESA will correctly identify that the exponential and square-root exponential models over-estimate the data.



(a) Poly. model on Poly. data set.

(b) SqrtExp. model on Poly. data set.

Figure 10: The model fit types reported by ESA's automatic interpretation of the scaling results as the number of bootstrap samples is varied.

While this may seem an unsurprising result, it does indicate that the separation of the fitted models grows more quickly than the size of the intervals for the model predictions, otherwise ESA's ability to distinguish between the models would not increase. As a result, increasing the extrapolation distance is one of the best ways to obtain more reliable and statistically significant results with ESA.

2.5 Varying the Number of Bootstrap Samples

We tried seven values for the number of bootstrap samples used by ESA with approximately logarithmic spacing: 20, 50, 100, 200, 500, 1 000 and 2 000. The main effect we observed from varying this parameter was the running time required to run ESA. While this change in running time for ESA was expected, we were surprised by how little the output from ESA changed when a very small number of bootstrap samples was used. In Figures 10a and 10b we show the model fit types reported for the polynomial and square-root exponential models on the polynomial data set. We omit the results for the exponential model since it was reported to over-estimate the data in 100% of the runs of ESA. Overall, we see that the change in the percentage of very good model fits for the polynomial model is only about a 15% increase from 20 bootstrap samples to 2 000. In addition, there is no discernible change in the model fits reported by ESA for the square-root exponential model.

Not surprisingly, we can see from Figure 11 that the locations of the bootstrap intervals tends not to change as the number of bootstrap samples is varied. Similarly, the median size of the bootstrap samples does not change; however, we

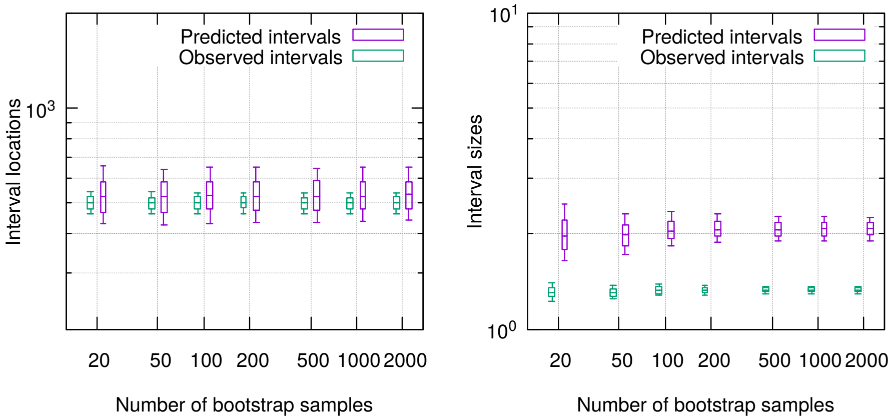


Figure 11: The prediction and observation bootstrap intervals for instance size 4 500 – the largest challenge instance size, as the number of bootstrap samples is varied. This data is for the polynomial data set and the prediction intervals are for the fitted polynomials.

do see a slight increase in the variability of the size of the intervals as the number of bootstrap samples is decreased. We believe that it is this slight increase in variability that is causing the slight decrease in ESA's ability to correctly identify that the polynomial model is a very good fit for the data. We also expect that we would notice a more significant effect in varying the number of bootstrap samples if we were using less running time data. In particular, had we also used a small number of instances per instance size, then we would expect the increase in the variability between runs of ESA to be even more pronounced.

References

1. Pushak, Y., Mu, Z., Hoos, H.H.: Empirical scaling analyzer: An automated system for empirical analysis of performance scaling. Under Review (2019)

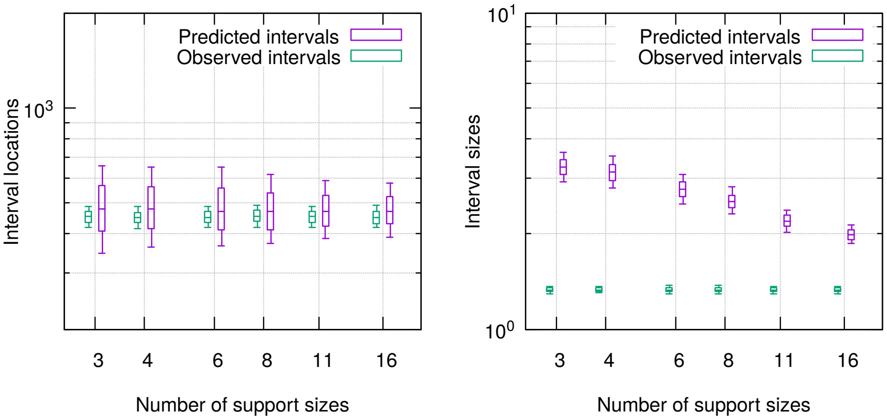


Figure 12: The prediction and observation bootstrap intervals for instance size 4 500 – the largest challenge instance size, as the number of support instance sizes is varied. This data is for the exponential data set and the prediction intervals are for the fitted exponentials.

A Non-Monotonic Responses

Recall that in Figures 4a and 4b we saw a non-monotonic response in the percentage of models that were reported to fit the data very well. In addition, in Figure 5 we saw that the size of the bootstrap intervals for the model predictions for 3 support instance sizes is smaller than for 4. Interestingly, we do not see any such behaviour on the exponential data set in Figure 12. While we have omitted the sizes of the bootstrap intervals for the model predictions for the incorrect models for each data set, we note that they again exhibit qualitatively similar behaviour to the models shown. That is, for the polynomial data set the square-root exponential and exponential models both have smaller bootstrap interval sizes for the model predictions when using 3 support instance sizes instead of 4, and vice versa for the exponential data set. Given this observation, we do not believe that the behaviour we are seeing on the polynomial data set is an artifact that holds true when fitting polynomial models instead of exponential models. Instead, we assume that it must be due to the differences in the data sets themselves – in fact, we believe that the main cause of the difference is the same reason why we see a lower probability that the exponential model is reported to fit the exponential model data very well than the polynomial model is to fit the polynomial data.

We believe that there are three main factors at play that we need to consider: the number of support instance sizes, the magnitude of the observed median running times for the support instance sizes and the location (in terms of instance size) of the support instance sizes. Clearly, as the number of support instance sizes is increased, the uncertainty in the support data decreases and so we can expect this factor to decrease the size of the bootstrap intervals for the model predictions. However, since the running times increase as the problem instance size increases, we know that the fitted models are most sensitive to deviations in the largest couple of support instance sizes – especially when there are a very small number of support instance sizes. In other words, when we only use 3 or 4 support instance sizes, the fitted models are dominated by the running times of the largest two support instance sizes, while the smallest 1 or 2 support instance sizes have only a modest "damping" effect on the fitted models. We also expect that the effect of the smallest couple of instance sizes should be less noticeable for the polynomial data set. This is because the model used to generate the exponential data set grows more slowly over the range of support instance sizes than the polynomial model that was used to generate the polynomial data set. That is to say, in the exponential data set the running times for the smallest and largest support instance sizes are more similar than the corresponding running times in the polynomial data set. This means that the smaller support instance sizes will have a larger effect on the models fitted to the exponential data set than they will for the polynomial data set.

Finally, we need to consider the third factor that we hypothesize is causing the surprising result, i.e., the second largest support instance size is closer (in terms of instance size) to the largest support size in the runs with 4 support instance sizes than in those with only 3. Since these two support instance sizes are closer together, the model is allowed more "wiggle room", which leads to the larger bootstrap intervals for the model predictions. As an analogy, consider the effect of wiggling a pencil around in a short tube that is 2 cm in diameter. If you start with a 2 cm long tube, you will be able to draw only small circles with the end of the pencil. However, as you decrease the length of the tube, the size of the circles you can draw will increase. Our hypothesis is that this behaviour, in combination with the fact that for the polynomial data set the small support instance sizes have less effect on the fitted models, is what is causing the non-monotonic response that we are seeing in the polynomial data set, but only to a small extent in the exponential data set.

We see further evidence that initial increase in the percentage of models that are reported to fit the data very well is due to an increase in uncertainty rather than an increase in precision for the polynomial data set in Figure 6a. In particular, this is because we see that ESA is less confident in reporting that the square-root exponential model is an over-estimate for the data when given 4 or 6 support instance sizes than when given 3. As we would expect from our previous argument, we do not observe similar behaviour for the square-root exponential model on the exponential data set, which is shown in Figure 6b. Instead, in this figure we see the monotonic increase in the percentage of under-estimate fit types that we expect.