Amendments summary

Sayani Gupta 3 February 2022

I thank my examiners for their thorough and constructive comments. The point by point description of changes are below: the examiners' comments are in red and my response is in black.

1 Professor Catherine Hurley

I appreciate all the comments and suggestions, and will incorporate them into future work. Specific comments are addressed as follows:

1.1 Chapter 2

Q1. In Figure 2.1 on the right hand side it should be t/24, t/(24 × 7)

Q2. Last sentence of Section 2.3.4: fix the table reference

Done.

Q3. In Figure 2.4, the outlier glyphs are in my opinion too big and dark and attract too much attention. The log scale for the y-axis should be mentioned in the caption.

Q4. The violin plot in Figure 2.5c is barely recognisable, and comparison of the distributions is challenging. The plot deserves a bit more space.

1.2 Chapter 3

Q1. On page 33 there is a mention of the hakear package, without a reference.

It appears in the abstract of the paper. For the thesis, I have added it since the abstract acts as a summary to the chapter.

Q2. In the Introduction state the methods are for continuous variables.

Added "for a continuous univariate dependent variable" in the last paragraph of Introduction.

Q3. In Figure 3.1, the points are a bit hard to see, so it looks like whiskers extend beyond the data.

Changed to theme_bw() instead of default and outlier.alpha (0.5 ->0.1) and geom_jitter(alpha = 0.04 -> 0.06)

Q4. On page 35, refer to months consistently using short or long names.

No change. Refer as months or month_year depending on the context.

Q5. On page 35 v refers to the number of variables, in Section 3.3.3 it refers to the variable.

Changed. On p35, v changed to p, where p represents the number of variables. This is analogous to many time granularities and their pairwise combinations and v represents the univariate measured variable for which distributions are constructed for different time granularities and their pairs.

Q6. In the Figure 3.2 caption, I do not understand the comment beginning "Difference between the 90th. . .". Also for the sentence "Energy consumption for (a)..", it would clarify to insert the word "median".

No change. Changed to "distribution of energy usage"

Q7. On the top of page 38, it should say "the distribution means are three standard deviations apart".

Done.

Q8. Maybe include a reference for gestalt theory.

To do.

Q9. In Section 3.2.2, I find the references to the null distribution confusing. Surely that is the design in Figure 3.3(a) only?

No change. Null distributions refer to the case when there is no significant distributional differences between categories of one or more granularities.

Q10. In table 3.1 Nc is the number of cyclic granularities, whereas in the text (first sentence of 3.3.2 and 3.3.3) it is m.

Add m to table 3.1 m = number of cyclic granularities to display together

N_c refers to the total number of contextual cyclic granularities. m refers to the number of cyclic granularities we are considering together in the display. For example, contextual cyclic granularities could be $hour_day$, day_week and $month_year$ and we want to visualize any one granularity at a time. So $N_c = 3$ and $m = 1$.

Q11. Add v to table 3.1.

Done. Added v:continuous univariate measured variable

Q12. In the first sentence on the top of page 42, ordered and unordered are mixed up. Have you considered the setting where the facet variable levels are ordered? For the within-facet ordered distances, you could consider a distance measure that respects circular order, or choose the start level appropriate to the display.

first part: Done. second part: To do

Q13. I like Figure 3.4. In (3), the dotted arcs only connect to a1 which might be misleading.

Changed. Caption changed to incorporate that. Within-facet distances are illustrated in Panels 3) (when categories are un-ordered, shown only with respect to a1) and Panel 4) (when categories are ordered)

Q14. In the pairwise distance measure on page 43, is there any adjustment made for varying numbers of observations across the levels of A and B? For example in Figure 3.3(d) if there are few values at xlevel=1 the comparison shown is less interesting

No change. We assume we have enough observations for each level or combination of levels to compute distributions. As long as that holds, no adjustment made for varying number of observations.

Q15. The equations on page 44 could be tidied up. There are extra () on the definition of wpdperm and on the residual definition at the bottom of the page. Use $\times$ for the equations.

Done.

Q16. Is there a practical reason why wpdglm is not working as expected for lower nx and nf ?

To do.

Q17. In the 3.3.1 algorithm, there is m, and then M.

Changed. 1.b) for $i \in 1, 2, \ldots, M$.

Q18. In Table 3.2, I would suggest separating the m = 1 and m = 2 tables.

Not changed.

Q19. The presentation of the material on the simulation study on page 48 could be improved. Where is the notation wpdl,s used? Figure 3.5 shows the m = 2 results only. The text alludes to a simulation involving different underlying distributions, but this is not mentioned again.

first part: How? second part: After NQT, results are pretty similar to an underlying normal distribution.

Q20. In Figure 3.5 the axis tick labels should be smaller. The blue and orange marks are hard to see. Maybe show fewer panels?

To do.

Q21. In Figure 3.6 the axis tick labels should be smaller. The blue and orange are hard to see. The caption should refer to the rug. Maybe show fewer panels?

To do.

Q22. In Figure 3.7 (a) the heatmaps need id labels. The grey color is missing from the legend. Maybe use a different colour in the heatmap for the significant comparisons. State the threshold for significance in the caption. In Table 3.3 the caption should explain the threshold. Maybe use colour instead of stars to indicate significance?

To do.

Q23. The link in the Acknowledgements https://github.com/Sayani07/paper-hakear is not available. Neither are the supplementary materials. I understand this is for the paper version, I mention it for completeness.

Changed. The link works. The repository has now been made public.

1.2.1 Chapter 4

Q1. In Section 4.1, line 5 "method of time series clustering".

Done.

Q2. Page 59, fourth bullet point. I found these sentences confusing.

To do.

Q3. In the material in the bottom of page 59, make it clear from the outset you only have data on energy use, not on property size, location, family size and so on.

To do.

Q4. End of Chapter 4.1: incorrect reference to Section 2.7.

Done.

Q5. I found Figure 4.1 very useful. In each box, would be helpful to put in "or", in places where only one of the steps listed is performed, eq Normal quantile transform or Robust scaling. As there are many steps in the algorithm, it would be helpful to the reader to label the pipeline steps and to refer back to them in the text. The text in the Data pre-processing box does not make sense to me.

To do? (Or not)

Q6. The section describing RS and NQT is confusing. It states RS is applied to each time series separately. Is NQT also applied to each observation separately? How does "it could be useful to standardize it for the selected set of significant granularities prior to computing the distances" relate to the following bullet points?

Q7. Page 64 "D is the Jensen-Shannon distance".

What is the question? I have mentioned that D stands for Jensen-Shannon distances.

Q8. In Table 4.1, the R column should have 250, not 20.

Done.

Q9. The description of the data generation on page 67 and Table 4.2 is confusing. Maybe state the distribution of each vt.

To do.

Q10. Where is μ in Table 4.2?

μ is the difference between means considered for consecutive categories.

Q11. In Section 4.3.3 what are the values of R and T?

Q12. Figure 4.5 caption: MDS summary plots of what?

Q13. Figure 4.7 is missing labels (a), (b) (c).

Q14. Why do you chose to summarize 4 and 5 clusters from JS-NQT when sindex suggests 3 clusters?

Q15. The reference to the stationarity assumption in the Discussion needs clarification.

Q16. Page 81 “can not” should be “cannot”

Done.

Q17. Again, the computational burden could do with more Discussion. At present, if I were to use this method on my data, what kind of sizes are realistic to work with?

2 Prof Juergen Symanzik

2.1 Chapter 4

Q1. There seems to be some contradiction on p. 60. One sentence states: Tureczek and Nielsen (2017) conducted a systematic study of over 2100 peer-reviewed papers on smart meter data analytics. Another sentence states: Time series data, such as smart meter data, are not well-suited to any of the techniques mentioned in Tureczek and Nielsen (2017).

Q2. Use a consistent style and use past tense throughout when you summarize what was previously published.

Done.

Q3. The last cross{reference at the end of the Introduction should be to Section 4.5 (and not Section 2.7).

Done.

Q4. At the end of p. 61, the sentence The flow of the procedures is illustrated in Figure 4.1. should be extended with “and is further described in the following subsections.” Also match subsection headings and headers in Figure 4.1, e.g., Selecting granularities vs. Find significant granularities and more.

first part: Done second part: Discuss (The diagram to be changed as per text or reverse)

Q5. The text on p. 68 speaks of Figure 4.3(b) and Figure 4.3 (right). Either add letters a and b or speak of left and right.

Done.

Q6. Similar to Chapter 3, supplementary material should be specified in more detail, e.g., on p. 68, p. 70, p. 71, and p. 75.

Done. Added links in the supplementary materials section at the end of the chapter.

Q7. In Figure 4.5, it is hard to see which groups are overplotting, in particular for S3. To better reveal this graphically, use different (open) glyphs in addition to different colors. In particular, a + for group 1 and an open circle for group 5 may help to better distinguish the groups (and possibly $\times$, open triangle, and open box to be used for groups 2 to 4).

Discuss.

Q8. Apparently, in Figure 4.6, the vertical axis is not standardized to [0; 1] as it is frequently done for parallel coordinate plots. Therefore, it would make sense to display an actual vertical axis in each of the three graphs.

Discuss.

Q9. Figure 4.7 speaks of (a) and (b) in the caption, but those letters do not appear in the figure. Either add them or refer to left and right. Moreover, in previous chapters, you used orange and green for the quartiles and 10th/90th quantile. Why not using the same colors and quantiles here? If this becomes too confusing for a reader, then at least, indicate in the caption which quantiles are covered by the gray areas.

first part: Done. second part: Discuss

Q10. Be consistent across chapters. For example, in Table 3.3, you use wdwnd. On p. 72 and in Figure 4.7, you use wnwd. Adjust across all chapters. Check whether other abbreviations need to be standardized as well.

Ok. To do in hakear.

Q11. Similar to Figure 4.9, it would help the reader that cluster P-3 is visually represented via three somewhat similar colors for Q-3, Q-4, and Q-5. Changing from red to blue/purple initially hides this information. Also arrange the legend from P-1 to P-3 and Q-1 to Q-5 to make it more obvious that P-1 & Q-1 and P-2 & Q-2 are identic

Discuss.

Q12. It would help the reader to also mention in the caption of Table 4.3 that P-1 & Q-1 and P-2 & Q-2 are identical. I first looked at Figure 4.9 and then at Table 4.3, but if read in the other order, I likely would have been surprised, to see identical values in some of the table rows (without any further explanation).

Done. (but not showing in pdf check again)

Q13. For the thesis, the work is adequate and meaningful. For a journal paper, I would like to see some of the points outlined in the Discussion being addressed. What happens if this method is applied to the data from the 13,000 customers that were introduced earlier on? As a reader of a journal paper, I would rather like to see limitations of the method proposed here, e.g., if it does not scale up then why it most likely does not scale up. Otherwise, it is hard to assess whether it is worthwhile to try this method for one's own data.

Discuss.