Report on PhD thesis of Timothy Hyndman

Contents

1 General comments

The thesis is made up of two separate sections. I only carefully examined the second section, since the first was outside my area of expertise. However, I read through the first part and it was in general very well written and structured in a sound logical manner and appeared to me to be a substantial piece of work; however I cannot speak to its originality or novelty. I will not comment any further on the first part.

The second part, “Efficient optimisation for statistical inference” dealt with two related optimisation problems that arise in statistical inferential procedures involving convolutions where each observation is the sum of a “signal” and a “noise” and we are interested in estimating the distribution of the “signal”:

1. In the case where the distribution of the "noise" is completely known, the nonparametric estimator of the distribution of the "signal" is (under certain regularity conditions) available, is known to be discrete with at most n mass points (n being the number of observations) although in many cases the number is much less than n. Chapter 6 of the thesis examines various questions relating to the number of mass points in the estimate.

2. When the distribution of the "noise" is unknown, but assumed to be symmetric and the distribution of the "signal" is assumed to *not have a symmetric component* (to ensure identifiability) then recent methods developed by Delaigle and Hall (2016) involving *phase functions* are examined. Again, estimates of the distribution of the "signal" are also discrete with a relatively small number of mass points. Chapter 7 starts with an "ideal" estimator proposed in Delaigle and Hall (2016) and discusses various computational issues involved with implementing this, proposing a certain method aimed at overcoming these challenges.

1.1 Chapter 6

This chapter contains a thorough review of existing literature on the analytical properties of nonparametric maximum likelihood estimation of a mixing distribution. It contains a very useful discussion of the pros and cons of penalisation in this setting. Theorem 6.3 is new and interesting. The use of the Karush-Kuhn-Tucker conditions to re-examine the optimisation problem is also interesting and manages to derive a useful new necessary condition for a value θ^* to be a mass point of an m-point maximum likelihood estimate of the mixing distribution. A counter-example to an unproved claim in Lindsay (1995, Theorem 24, third part) is insightful.

The rest of the chapter deals with conditions for $K_{\boldsymbol{x}}$, the number of mass points (regarded as a function of the observation vector $\boldsymbol{x}$). The "flag plots" used to illustrate the partitioning of the sample space induced by different values of $K_{\boldsymbol{x}}$ are innovative and insightful, particularly the $n = 3$ version and serve very well to illustrate the subsequent results in the chapter.

The rest of the chapter collects together some relevant existing results as well as some useful new ones: Theorem 6.9 (with its accompanying Lemmas 6.10 and 6.11) provides new necessary conditions implying a one-component estimate in the case $n = 2$ for a general unimodal symmetric location family of densities; Theorem 6.13 (with its accompanying Lemma 6.14) provides new necessary conditions for the estimate to have m mass points for general

n for the normal location family (with known variance).

The chapter concludes with a simple, elegant proof illustrating the non-consistency of $K_{\boldsymbol{x}}$ as an estimator of the "true" number of mass points when a finite mixture of normals with known variance is actually generating the data. In all the chapter is satisfying and impressive. It provides valuable insights into a difficult statistical optimisation problem and leaves the door open for a few potentially very interesting avenues of research (see section 3 below).

1.2 Chapter 7

This chapter is less theoretical, with more emphasis on computation and implementation of a certain statistical method of deconvolution requiring an optimisation. It opens with a thorough review of certain relevant methods in deconvolution including when the distribution of the "noise" is known, and also when that distribution is unknown, but repeated measurements with the same "signal" but independent "noises" are avaiable. Then the main method under study involving the use of *phase functions* is carefully described.

An "ideal" estimator (suggested by Delaigle and Hall (2016)) is introduced and then two modifications of it are considered, arriving at "Estimator 3". The rest of the chapter deals with numerical implementation of Estimator 3. A graphical illustration of the performance of the resultant estimator is provided as are details of an R package which includes code providing these implementations.

Some interesting parallels between this optimisation problem and that of Chapter 6 are highlighted, including a certain convexity property, suggesting that successful methods in the latter may help with the former. However these are not pursued any further.

1.3 Assessment of the thesis

Four substantial and 6 minor issues are identified in the next section which should be addressed, to the satisfaction of the Chair of Examiners, before the degree is awarded.

2 Issues requiring attention

2.1 Substantial points

2.1.1 Chapter 6: Figure 6.6

There is a problem with Figure 6.6 on page 80. The plot claims to show the "likelihood curve" $(f_\theta(x_1), f_\theta(x_2))$ as θ traverses the parameter space, with $(x_1, x_2) = (0, 0.4)$ and $f_\theta(x) = f(x - \theta)$ where $f(\cdot)$ is "the triangular density with width 1/2". However, the form of this density is

$$f(x) = \begin{cases} 4 - 16|x| & \text{for } |x| \leq 0.25, \\ 0 & \text{otherwise.} \end{cases}$$

In particular the maximal value is 4, not 1 as suggested by the plot. The problem can be remedied by

- mutliplying the markings on the axes by 4 and
- replacing x_2 with 0.2

without changing the qualitative features of the plot.

2.1.2 Chapter 7: repeated measures condition

The estimator (7.13) is not necessarily consistent without extra assumptions on the distribution of the errors U_{jk}. In particular the Fourier transform $\phi_U(t)$ must be real-valued, so the U_{jk}'s need have a symmetric distribution about zero; see condition (2.2) in Delaigle et al. (2008). This should be mentioned at some point in the lead-up to (7.13); this is an opportune moment to foreshadow the assumption A7 that is made in the following section.

2.1.3 Chapter 7: comparison of MATLAB and R

The comparison of the MATLAB and R implementations of the methods of Chapter 7 is inadequate. Since computational implementation is the focus of Chapter 7, more work is needed to explain the observed differences in performance between the MATLAB and R versions, other than to say "We are unsure as to why out implementation in R tends to produce worse objective values than out implementation in MATLAB. . . .we do not know exactly what is going on". This last phrase should preferably not appear in a PhD thesis, it would be better to identify a few possible causes and investigate them; even if they do not reveal anything it is important to show that such issues are not to blame to assist other future researchers (perhaps this was done but not mentioned?).

There are various aspects one could explore, I list a few below (there are many others):

- Are there any tuning parameters used in either the MATLAB or R implementations? If so, are they set to the same values?
- The R package `NlcOptim` is a third-party package written by an "ordinary R user" (as opposed to a built-in function developed by the R Core Team). It is written entirely in R code, and does not appear to call any compiled C, Fortran or C++ code, unlike the general-purpose optimisation functions "built-in" to R. Can the same be said of the MATLAB implementation? Does a Fortran, C or C++ version of this algorithm exist which could be used in R (see the R extensions manual: `https://cran.r-project.org/doc/manuals/R-exts.pdf` for how to do this).

At least two possible causes should be investigated and reported on.

2.1.4 Incorrect reference

The reference numbered [56] seems to be incomplete and possibly incorrect; I could not find the work it seems to refer to.

2.2 Minor issues

Page 75 I believe that $\sigma_2 = 0.4^2$ should in fact be $\sigma^2 = 0.4^2$; please check.

Page 78 It should be explicitly pointed out that $\mathcal{H}$ and $\mathcal{H}_Q$ are the same, or the definition at (6.32) could be written as

$$\mathcal{H} = \mathcal{H}_Q = \cdots .$$

Page 83 I believe that in inequality (6.65), the θ_j should be θ^*; please check.

Pages 92–3 In the proof of Theorem 6.9, reference is made to results in Lemmas which have not yet been presented e.g. Lemma 6.10. In such cases it would improve readability slightly to insert the word "below" to make clear to someone reading the work linearly that it has not yet been read (this is done at other points in the thesis but not here).

Page 112 At the end of the second paragraph of 7.2.4, it says "... we will explore this further in Section 6.3". It is perhaps the case that the intention was 7.3. Please confirm.

Page 116 It is better to use the term "mass point(s)" when referring to the points of support of a discrete distribution. The term "mass(es)" is ambiguous, is sometimes used for the actual probability/weight supported on the point, as well as for the point itself.

3 Suggestions regarding future research

The points mentioned below are not considered vital for the thesis, but should be borne in mind in any future efforts to publish the research.

- On page 69, as well as referring to the paper labelled [56] (see problems with this reference above), an important reference which sheds important light on the richness of the location-mixture-of-normals family is Magder and Zeger (1996).
- The Theorem 6.13 as written only applies to the location-family of normal distributions with known variance. However, the same arugment could be used to generalise the result to the broader family of one-parameter exponential families of the form (2.2) on page 30 of Lindsay (1995).

References

Aurore Delaigle and Peter Hall. Methodology for non-parametric deconvolution when the error distribution is unknown. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(1):231–252, 2016. ISSN 1369-7412. doi: 10.1111/rssb.12109. URL `https://doi.org/10.1111/rssb.12109`.

Aurore Delaigle, Peter Hall, and Alexander Meister. On deconvolution with repeated measurements. *Ann. Statist.*, 36(2):665–685, 2008. ISSN 0090-5364. doi: 10.1214/009053607000000884. URL `https://doi.org/10.1214/009053607000000884`.

B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, 1995.

Laurence S. Magder and Scott L. Zeger. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Amer. Statist. Assoc.*, 91(435):1141–1151, 1996. ISSN 0162-1459.