Master thesis opponent's assessment

Title: Estimators of probability density function level-sets and their volumes
Author: Mladen Josivljević

Summary of the content of the thesis

The thesis introduces the concepts of levels sets, minimum volume sets and Voronoi tesselation. Then it investigates the idea of estimation of the minimum volume sets with the help of Voronoi tesselation. Finally it illustrates the problem on synthetic as well as real data.

Topic of the thesis. The topic of the thesis is suitable for master students. Although in the 'guidelines for thesis preparations' mainly level sets are mentioned, dealing in particular with the minimum volume sets is from my point of view acceptable. Further in 'guidelines' it is mentioned that at least two approaches (fully parametric, semiparametric and nonparametric) should be considered and discussed in more detail. Here I am not completely sure if this is fulfilled by considering the suggested method and the SVM method.

Author's contribution. I find the main contribution in formalizing the ideas of using Voronoi tessellation to estimate the volume sets. Further the author has provided some non-trivial (from the computational point of view) numerical experiments.

Mathematical level. The author tries to be mathematical rigorous nevertheless the thesis is rather weak from the mathematical point of view. One of the crucial problems is that the author often does not introduce/explain the symbol (or the mathematical terminology), see for some examples below. While it is acceptable on this study programme that there basically no mathematical derivations, it is pity that at least definitions are often not completely clear due to missing assumptions or cumbersome writing. Moreover from my point of view the mathematical statements (named as Properties in the thesis) are not well integrated in the thesis. These statements stand rather alone than somehow interact with the main message of the thesis.

The use of resources. The resources are **not** correctly cited. This is the problem in particular in Chapter 1. It seems that Figure 1.1 coincides with Figure 1 from Scott and Nowak (2006), but the author does not mention that. Also the references *(Tsybakov, 1997; Ben-David and Lindenbaum, 1997; Cuevas and Rodriguez-Casal, 2003; Steinwart et al., 2005; Vert and Vert, 2005)* seems to copy-pasted from Scott and Nowak (2006). Moreover these references are not included in the bibliography. Similarly Figure 1.2 of the thesis seems to be a print-screen version of Fig. 1 of Garcia et al. (2003). The thesis also lacks information that the statements in Chapter 1.2 are also copied from Garcia et al. (2003).

Another problematic place is the beginning of Chapter 3.1 where about 5 lines have been completely copied (except for one font misprint) from Section 1.3 of Scott and Nowak (2006).

Further the references included in the bibliography are usually incomplete (e.g. in [5] the journal, the volume as well as the pages are missing).

Formal level of writing. Formal level of the thesis is far from being good. There are many misprints, missing punctuation marks (in particular in formulas), conflicting notations, not consistent fonts of the symbols...The level of English could be better even when taking into consideration that the author is not a native speaker.

Question and comments

1. **Examples of unexplained notation**: $\mathcal{B}$ and λ in Definition 3; Img(ξ) in Definition 5, $CH(P)$ in Property 5, ...

2. **Examples of misprints**: $A_{\xi_\alpha^{-1}}$ in Definition 5, Euclidian (on many pages), alterantively (p. 9), ...

3. **Examples of unexplained math. notions**: discontinuities of the first type (p. 5), locally strictly decreasing (p. 5), continuous space (p. 7), collinear (p. 12), convex hull (p. 12), the simple average number (p. 13), the biggest outliers (p. 26), ...

4. P is used as probability measure on the sample space, $P(B)$ as the power set and P is also denoted the set of points $\{p_1, \ldots, p_n\}$.

5. p. **3 Definition 1**: Is there any reason to have the strict inequality $P(G) > \alpha$ instead of $P(G) \geq \alpha$ in the definition?

6. p. **4** Figure 1.1: It should be explained in more detail what is on this figure. And similarly for other figures included in the thesis.

7. p. **4 Definition 2**: I do not understand how the definition works when α is given.

8. p. **4 Definition 3**: It seems that the definition requires the existence of the density function f.

9. p. **4 Definition 4**: Maybe it is because of some typos and unfortunate notation, but this definition does not make sense. For instance it is not clear what is the role of the set B in the definition.

10. p. **9 − 10**: What is the advantage/reason to give the alternative definition of Voronoi diagram?

11. p. **12 Property 4**: As far as I see the points $p_1, \ldots, p_n$ should be assumed to be distinct.

12. p. **14** Chapter 3.1: It is not clear what exactly is the collection of sets $\mathcal{G}$. This is really problem as $\mathcal{G}$ is rather crucial in what follows.

13. p. **15 Definition 10**: I have difficulties to understand the sentence: *Given an independent and identically distributed sample $S = (X_1, \ldots, X_n)$ drawn according to S.*

14. p. **15** *Proof* of **Theorem 1**: I do not agree that μ_a is a countable sum of Lebesgue measures.

15. p. **18**: Do I understand correctly that in the suggested cross-validation procedure the subsample that is left out ('test data') is not used?

16. Chapter 3.3: The description of SVM is rather confusing. It is not at all clear why $\widehat{G}_\alpha$ (given on p. 19) should be an estimator of the minimum volume set as (among others) α is not in the text preceding that formula.

17. p. **20**: I do not understand the sentence: *In practice we more often would often sacrifice a small decrease in $P(G)$ to gain an increase in $\mu(G)$.*

18. Chapter 4: Although more than 11 pages are occupied by figures it is not clear what these figures illustrates. The author does not comment what can the reader see and learn from the figures. Moreover the comparison of the figures is rather difficult as the scaling often differs in figures that are to be compared.

Questions for the defense

Prepare the answers for the comments 5, 9, 10, 12 and 15.

The overall evaluation

Unfortunately the thesis suffers from too many problems that are not compatible with what is expected from a solid mathematical text. That is why I have serious doubts that the thesis meets the standards of the study programme Financial and Insurance Mathematics.

doc. Ing. Marek Omelka, Ph.D.
KPMS MFF UK
August 26, 2024

References

Garcia, J. N., Kutalik, Z., Cho, K.-H., and Wolkenhauer, O. (2003). Level sets and minimum volume sets of probability density functions. *International journal of approximate reasoning*, 34(1):25–47.

Scott, C. and Nowak, R. (2006). Learning minimum volume sets. *Advances in neural information processing systems*, 7.