Mohamed Hebiri
Maître de conférences - HDR
LAMA – Université Gustave Eiffel
E-mail : mohamed.hebiri@univ-eiffel.fr

LAMA

**Report on the manuscript entitled *Statistical and Computational Complexity of the Feature Matching Map Detection Problem* by Tigran Galstyan.**

Tigran Galstyan's Ph.D. dissertation is devoted to the intricate problem of detecting a discrete map that establishes correspondence between two sets of high-dimensional, noisy vectors. This challenge finds its roots in diverse applications such as computer vision and bioinformatics. From a mathematical standpoint, Galstyan formulates the problem as a multiple hypotheses testing issue and thoroughly investigates a highly relevant approach grounded in separation rates.
The utilization of separation rates in statistical hypothesis testing has a rich history, as exemplified by the comprehensive work of Ingster and Suslina (2003). This approach is widely acknowledged as a robust means of assessing the efficacy of testing procedures and determining their optimality. In the realm of vector matching, Dalalyan and Collier (2013, 2016) have previously examined separation rates in a relatively simple setting. Galstyan's Ph.D. thesis elevates this research significantly by offering substantial extensions of existing results. Notably, these extensions address situations that closely mirror real-world applications, thereby enhancing the applicability and relevance of the findings. The breadth and depth of the contributions within this thesis make it a noteworthy advancement in the field.
The dissertation is organized into four distinct parts, supplemented by two appendices housing the primary technical proofs.

The initial part serves as a concise introduction to the challenge of detecting the matching map, providing readers with valuable insights into the intricacies of the problem. Moreover, it offers a thoughtful retrospective on previous work in this domain. While the section is commendably clear in its exposition, there exists an opportunity to enhance its depth. Specifically, the introduction could benefit from a more expansive treatment, possibly incorporating a discussion on the relationship between the problem at hand and the broader landscape of multiple testing problems. Alternatively, a more detailed exploration of concepts such as the separation rate in a two-hypothesis testing problem might further enrich readers' understanding.

Chapter 2 of the dissertation delves into a specific instance of the matching map detection problem, focusing on scenarios where outliers are present in only one of the two sets of vectors. The author starts by examining the case with known noise standard deviations, demonstrating that the least normalized sum of squares estimator effectively identifies the unknown map when the separation distance surpasses a critical threshold. Specifically, this threshold is determined by the expression :

$$(d\log(nm))^{1/4} \vee (\log(nm))^{1/2}.$$

Here, $d$ denotes the dimensionality of the vectors being matched, while $n$ and $m$ represent the sizes of the two sets of vectors. Given that noise standard deviations are seldom known

Laboratoire d'analyse et de mathématiques appliquées – Université Gustave Eiffel
Bâtiment Copernic, 5 boulevard Descartes, Champs-sur-Marne, 77454

in practical applications, the author proceeds to investigate the least sum of logarithms estimator. Notably, this alternative estimator operates without requiring prior knowledge of standard deviations. In this context, the separation rate includes an additional term proportional to $\sqrt{d}$. Intriguingly, the dissertation establishes that this term is an inherent and unavoidable component for any procedure falling within the family of distance-based M-estimators. The chapter also presents findings on scenarios involving mildly varying standard deviations, showcasing an enhanced separation rate. To validate the theoretical results and underscore the computational efficacy of the examined estimators, the author concludes the chapter with a presentation of numerical experiments. The results presented in this chapter have been published in the Electronic Journal of Statistics, which is a Q1 journal in statistics according to Scimago.

Chapter 3 tackles the most intricate scenario, where both sets of vectors may harbor outliers, and the challenge is compounded by the unknown quantity of outliers. To devise efficient estimators under these conditions, Tigran Galstyan and his co-authors focus on a specialized case where all the noise standard deviations share a common value. They propose a novel procedure that combines the least sum of squares concept with model selection, strategically addressing the lack of information regarding the number of inliers. Notably, the authors demonstrate that, somewhat surprisingly, the optimal rate of separation remains consistent with the case explored in Chapter 2, where only one of the two sets contains outliers and the noise standard deviations are known and equal. This stability in optimal separation rates is observed for the least squares of the logarithm as well. The authors substantiate these theoretical findings with numerical experiments, illustrating the efficacy of the proposed methodology in successfully identifying the matching map despite the complicating factors of noise and outliers.

It is noteworthy that the results presented in this section have been disseminated through two published papers, one of which has been featured in the peer-reviewed conference AISTATS—a rank A conference in machine learning and statistics. This external validation underscores the significance and quality of the research outcomes. Finally, Chapter 4 shows that the developed methodology can be applied to biomedical datasets that have been used in some recent work.

In conclusion, I believe this dissertation tackles an interesting problem and introduces new methods with practical applications. These methods rely on advanced mathematical tools from statistics, probability, and optimization. The dissertation is well-organized and easy to follow, making it an enjoyable read. The results obtained are both profound and elegant. For these reasons, I recommend this dissertation for defense.

Champs-sur-Marne, April 25, 2024

Mohamed Hebiri

Laboratoire d'analyse et de mathématiques appliquées – Université Gustave Eiffel
Batiment Copernic, 5 boulevard Descartes, Champs-sur-Marne, 77454