Optimal Transport: Algorithms and Applications in Machine Learning

WASP Mini-Course, June 2024

Jia Li, Penn State University, WASP Guest Professor at Umeå University

Course coordinator: Jun Yu, jun.yu@umu.se

1 Introduction

Optimal transport (OT) has been applied across a wide range of fields, including economics, computer vision, computer graphics, and statistics. In this course, we will cover the fundamentals of optimal transport and illustrate its application in clustering analysis, specifically focusing on the development of ensemble clustering methods and approaches for assessing uncertainty. Additionally, we will introduce the concept of Wasserstein barycenters and optimization algorithms for computing these barycenters. We will also address the limitations of OT and the Wasserstein metric in scenarios involving high-dimensional continuous distributions and more complex matching tasks.

Our discussion will include a series of works that introduce a new distance for Gaussian Mixture Models (GMM), namely the Minimized Aggregated Wasserstein (MAW) distance, and develop solutions for solving the corresponding barycenter problem. At last, we will present an extended formulation of OT, termed OT with relaxed marginal constraints (OT-RMC), which offers increased flexibility for tackling matching problems between sets.

Throughout this course, we aim to showcase the broad spectrum of problems that can benefit from the principles of OT. Specifically, we will introduce a suite of tools, implemented as an R package, for analyzing the uncertainty or stability of clustering results. These tools are broadly applicable, as they do not rely on the underlying clustering algorithm. Although OT has traditionally been applied to discrete distributions, we will demonstrate how, through the MAW distance for GMMs, the framework of OT can be extended, with computational efficiency ensured, to encompass high-dimensional continuous distributions.

We plan to cover the following topics.

1. Optimal transport (OT), Wasserstein distance, and Wasserstein barycenter
2. Minimized Aggregated Wasserstein (MAW) distance for Gaussian mixture models (GMM) and hidden Markov models (HMM), MAW barycenter.
3. Applications to clustering: uncertainty assessment, ensemble clustering, cluster alignment, clustering based on multi-modal data. R CRAN package: OTclust.
4. Optimal transport with relaxed marginal constraints: OT-RMC.

2 Prerequisites

Students are expected to have a basic knowledge of linear programming and optimization, and courses in probability theory and mathematical statistics at second cycle level.

3 Schedule

The preliminary time is 17-19 June.

1. Day 1
   (a) Class 1: Topic 1
   (b) Class 2: Topic 2
2. Day 2
   (a) Class 1: Topic 3
   (b) Class 2: Topic 4

4 Assignment

1. Experiment with the R package OTclust. Use it to generate ensemble clustering results and to assess the uncertainty of clusters based on a collection of clustering results.
2. Write a report on the analysis including some background on the datasets used, methods experimented with, comparison of results, etc.
3. Select at least one related reference to read in detail. Recommended papers are listed below but not limited to those.
   - Algorithms for solving Wasserstein barycenters:
     (a) L. Yang, J. Li, D. Sun, and K.-C. Toh, "A Fast globally linearly convergent algorithm for the computation of Wasserstein barycenters," Journal of Machine Learning Research, 22(21):1-37, 2021.
     (b) J. Ye, P. Wu, J. Z. Wang, J. Li, "Fast discrete distribution clustering using Wasserstein Barycenter with sparse support," IEEE Transactions on Signal Processing, 65(9):2317-2332, 2017.
     (c) Y. Zhang, J. Z. Wang, J. Li, "Parallel massive clustering of discrete distributions," ACM Transactions on Multimedia Computing, Communications and Applications, 11(4):1-24, 2015.
     (d) J. Li, J. Z. Wang, "Real-time computerized annotation of pictures," IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(6):985-1002, 2008.
   - Applications to clustering
     (a) L. Zhang, L. Lin, J. Li, "Multi-view clustering by CPS-merge analysis with application to multimodal single-cell data," PLOS Comp. Biology, 19(4):e1011044, 21 pages, 2023.

   (b) L. Zhang, L. Lin, J. Li, "CPS analysis: self-contained validation of biomedical data clustering," Bioinformatics, 36(11):3516-3521, 2020.
   (c) J. Li, B. Seo, L. Lin "Optimal Transport, Mean Partition, and Uncertainty Assessment in Cluster Analysis," Statistical Analysis and Data Mining: The ASA Data Science Journal, 12(5):359-377, 2019.

- Other applications
   (a) X. Zheng, J. Ye, J. Z. Wang, J. Li, "SCOTT: Shape-Location Combined Tracking with Optimal Transport," SIAM Journal on Mathematics of Data Science (SIMODS), 2(2):284-308, 2020.
   (b) J. Li and F. Zhang, "Geometry-sensitive ensemble mean based on Wasserstein barycenters: proof-of-concept on cloud simulations," Journal of Computational and Graphical Statistics, 27(4):785-797, 2018.
- Extensions of OT and Wasserstein metric
   (a) L. Lin, W. Shi, J. Ye, J. Li, "Multi-source single-cell data integration by MAW barycenter for Gaussian mixture models," Biometrics, 79(2):866-877, 2023.
   (b) J. Li, L. Lin "Optimal transport with relaxed marginal constraints," IEEE Access, 9:58142-58160, 2021.
   (c) Y. Chen, J. Ye, J. Li, "Aggregated Wasserstein Distance and State Registration for Hidden Markov Models," IEEE Transaction on Pattern Analysis and Machine Intelligence, 42(9):2133-2147, 2020.

4. Expect to submit in one week.