ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՍՏԻՏՈՒՏ

Գալստյան Տիգրան Վահագնի

Հատկանիշներ համապատասխանեցնող արտապատկերումների հայտնաբերման խնդրի վիճակագրական և հաշվողական բարդությունը

Ե.13.05 «Մաթեմատիկական մոդելավորում, թվային մեթոդներ և ծրագրերի համալիրներ» մասնագիտությամբ ֆիզիկամաթեմատիկական գիտությունների թեկնածուի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան - 2024

INSTITUTE FOR INFORMATICS AND AUTOMATION PROBLEMS OF NAS RA

Tigran Galstyan

Statistical and Computational Complexity of the Feature Matching Map Detection Problem

SYNOPSIS

of dissertation for the degree of candidate of physical and mathematical sciences specializing in E.13.05 – ”Mathematical modelling, numerical methods and program complexes”

Yerevan - 2024

Ատենախոսության թեման հաստատվել է Հայ-Ռուսական համալսարանում:

Գիտական ղեկավար՝	ֆիզ.-մաթ. գիտ. դոկտոր Վ. Կ. Օհանյան
Պաշտոնական ընդդիմախոսներ՝	ֆիզ.-մաթ. գիտ. դոկտոր Մ. Հեբիրի
	ֆիզ.-մաթ. գիտ. թեկնածու Ա. Ն. Հարությունյան
Առաջատար կազմակերպություն՝	Հայաստանում ֆրանսիական համալսարան

Պաշտպանությունը կայանալու է 2024թ. մայիսի 3-ին, ժ. 15^{00}-ին, Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում գործող ԲՈԿ-ի 037 մասնագիտական խորհրդի նիստում հետևյալ հասցեով՝ 0014 ք. Երևան, Պ. Սևակի փողոց 1:

Ատենախոսությանը կարելի է ծանոթանալ ԻԱՊԻ-ի գրադարանում:

Սեղմագիրն առաքված է 2024 ապրիլի 3-ին:

Մասնագիտական խորհրդի գիտական քարտուղար՝ Մ. Հարությունյան

Dissertation topic was approved at Russian-Armenian University.

Supervisor:	Doctor of phys-math sciences V. K. Ohanyan
Official opponents:	Doctor of phys-math sciences M. Hebiri
	Candidate of phys-math sciences A. N. Harutyunyan
Leading organization:	French University in Armenia

Defense of the thesis will be held at the meeting of the specialized council 037 of SCC (Supreme Certifying Committee) of Armenia at Institute for Informatics and Automation Problems on May 3, 2024 at 15^{00} (1 P. Sevak street, Yerevan 0014).

You can get acquainted with the thesis in the library of the IIAP.

Synopsis was sent on April 3, 2024.

Scientific secretary of specialized council, M. Harutyunyan

General characteristics of the work

Relevance of the theme.

The problem of finding the optimal matching between two point clouds has been extensively investigated both theoretically and experimentally, due to its relevance in various applications, such as computer vision and natural language processing. For instance, in computer vision, matching local descriptors extracted from two images of the same scene is a well-known example of a matching problem, while in natural language processing, the correspondence between vector representations of the same text in different languages is another example.

Permutation estimation and related problems have been recently investigated in different contexts such as statistical seriation, noisy sorting, regression with shuffled data, isotonic regression and matrices, crowd labeling, recovery of general discrete structure, and multitarget tracking.

Feature matching is a problem that has received significant attention in the field of computer vision. One of the main directions aims to accelerate matching algorithms using fast approximate methods, as demonstrated in recent studies. Another direction is to improve the matching quality by improving the quality of descriptors of image keypoints.

Measuring the quality of statistical procedures in hypothesis testing relies on the use of separation rates. Recently, the practice of using separation rates has been adopted in the field of machine learning. While traditionally used in the context of two hypotheses, this approach is also applicable to multiple testing frameworks, including variable selection, and the matching problem being considered in this work.

In the field of single-cell biology research, it is common to collect datasets using similar measurement protocols or experimental conditions but from different batches. When analyzing such datasets, matching similar cells across different batches is a crucial step in correcting technical variations and batch effects. Another common practice is integrating datasets that have overlapping biological information, such as transcriptomic and proteomic data, obtained from different tissues, species, profiling technologies, or experimental conditions. This inte-

gration requires identifying and aligning cells in comparable states across related datasets. Additionally, matching datasets with complementary biological information, such as spatial information of individual cells within a tissue, with non-spatial single-cell datasets can transfer valuable information to different measurement modalities.

It is evident that in the matching problems mentioned above, not all the points in a dataset have their corresponding matching points in another dataset. It is challenging to predict the exact number of points that will have a match in advance. One of the primary objectives of the current research is to investigate this scenario and develop a comprehensive theoretical understanding of the statistical constraints associated with the matching problem.

The aim of the thesis:

1. Design estimators for matching map detection problem that have an expected error smaller than a prescribed level α under the weakest possible conditions on the nuisance parameter $\boldsymbol{\theta}^{\#}$ and noise level $\boldsymbol{\sigma}^{\#}$.

2. Find the detection boundary in terms of the order of magnitude of $(\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}})$ (4).

3. Introduce a data-driven procedure for estimating the number of inliers for any instance of the matching map detection problem with outliers present in both datasets.

4. Formulate the resulting optimization problem as a graph minimum-cost flow problem and show that it can be solved computationally efficiently.

5. Show that, in the high-dimensional setting, if the signal-to-noise ratio is larger than $5(d\log(4nm/\alpha))^{1/4}$, then the true matching map can be recovered with probability $1-\alpha$.

6. Show that, in the presence of outliers, separation rate for LSL (3) is minimax optimal.

7. Experimentally show that our data-driven procedure for detecting the feature matching map with no additional information before matching achieve similar results to more classical algorithms which were given the true number of inliers as an input.

8. Illustrate achieved results and computational feasibility of proposed algorithms on synthetic and real-world data.

The methods of investigation.

In this thesis we apply methods and techniques obtained on the basis of high-dimensional statistics, probabilistic inequalities, linear programming and related topics. Previous related results also served as a basis of this work.

Scientific innovation.

All results are new and are published in local and international conferences and journals.

Practical and theoretical value.

The results of the work both have theoretical and practical character. The theoretical results are devoted to finding and proving detection boundries of various estimators in different settings of the matching map detection problem. Algorithms studied and proposed in this work have been experimentally proven to work on real-world datasets across various domains (i. e. computer vision, bioinformatics).

Approbation of the results.

The presented results were presented in the scientific seminar at Russian-Armenian University. Some of obtained results were presented in local and international conferences.

Publications.

The main results of this thesis have been published in 3 scientific articles in journals and 1 article in conference. The list of the articles is given at the end of the Synopsis.

The structure and the volume of the thesis.

The thesis consists of introduction, 3 chapters of main results followed by conclusion and discussion, a list of references and 2 appendices. The number of references is 65. The volume of the thesis is 86 pages. The thesis contains 21 figures and 2 tables.

The main results of the thesis

Chapter 1.

First chapter introduces the problem of matching map recovery. In this chapter we formalize the problem, discuss its variations and challenges associated with each problem setting. We also discuss the most simple problem setting already studied in existing literature.

Put formally, this simplest setting goes as follows. We study the problem of matching two sets of equal size $n \geq 2$, $(X_1, \ldots, X_n)$ and $(X_1^\#, \ldots, X_n^\#)$. We assume that observed feature vectors are randomly generated from the following model:

$$\begin{cases} X_i = \theta_i + \sigma_i \xi_i \ , \\ X_i^\# = \theta_i^\# + \sigma_i^\# \xi_i^\#, \end{cases} \qquad i = 1, \ldots, n \tag{1}$$

In this model it is assumed that

- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\theta}^\# = (\theta_1^\#, \ldots, \theta_n^\#)$ are two sequences of vectors from $\mathbb{R}^d$, corresponding to the original features, which are unavailable,
- $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top, \boldsymbol{\sigma}^\# = (\sigma_1^\#, \ldots, \sigma_n^\#)^\top$ are positive real numbers corresponding to the magnitudes of the noise contaminating each feature,
- $\xi_1, \ldots, \xi_n$ and $\xi_1^\#, \ldots, \xi_n^\#$ are two independent sequences of i.i.d. random vectors drawn from the Gaussian distribution with zero mean and identity covariance matrix,
- there exists a bijective mapping $\pi^* : [n] \to [n]$ such that $\theta_i = \theta^\#_{\pi^*(i)}$ for all $i \in [n]$.

The ultimate goal is to detect the feature matching map π^*.

In chapter 1 we also discuss previous related results which served as a foundation for this work.

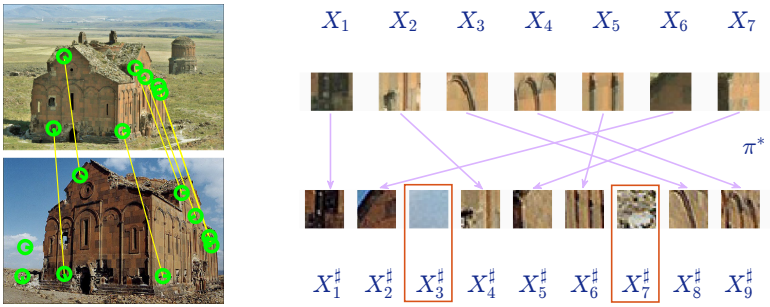

Figure 1: Illustration of the considered framework described in (1). We wish to match a set of 7 patches extracted from the first image to the 9 patches from the second image. The picture on the left shows the locations of patches as well as the true matching map π^* (the yellow lines).

Chapter 2.

Chapter 2 discusses in more detail the setting of matching map detection problem in presence of outliers only in one of the sets and our results achieved in this problem setting.

Formally, in this chapter we discuss the problem of matching vectors from two sets $(X_1, \ldots, X_n)$ and $(X_1^\#, \ldots, X_m^\#)$ with different sizes n and m such that $m \geq n \geq 2$. We assume that vectors are randomly generated from the following model:

$$\begin{cases} X_i = \theta_i + \sigma_i \xi_i \ , \\ X_j^\# = \theta_j^\# + \sigma_j^\# \xi_j^\#, \end{cases} \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, m. \tag{2}$$

In this model all assumptions from (1) hold, the only exception being that here, instead of a bijective mapping π^* our goal is to find an **injective** mapping $\pi^* : [n] \to [m]$, such that $\theta_i = \theta^\#_{\pi^*(i)}, \forall\, i \in [n]$.

Figure 1 illustrates the aforementioned problem setting on image matching application using local descriptors.

The LSL optimizer, one of the main estimators studied in this chapter is defined as follows:

$$\hat{\pi}^{\mathsf{LSL}}_{n,m} \triangleq \operatorname*{arg\,min}_{\pi:[n]\to[m]} \sum_{i=1}^{n} \log \|X_i - X^\#_{\pi(i)}\|^2, \tag{3}$$

Our aim is to develop estimators that can achieve an expected error smaller than a specified threshold α, while imposing minimal restrictions on the nuisance parameter $\boldsymbol{\theta}^{\#}$ and the noise level $\boldsymbol{\sigma}^{\#}$. When dealing with features that are difficult to differentiate, the problem of matching becomes more challenging. To quantify this phenomenon, we introduce two metrics - the normalized separation distance $\bar{\kappa}_{\text{in-in}} = \bar{\kappa}_{\text{in-in}}(\boldsymbol{\theta}^{\#}, \boldsymbol{\sigma}^{\#}, \pi^*)$ and the normalized outlier separation distance $\bar{\kappa}_{\text{in-out}} = \bar{\kappa}_{\text{in-out}}(\boldsymbol{\theta}^{\#}, \boldsymbol{\sigma}^{\#}, \pi^*)$. These metrics measure the ratio of the minimal distance-to-noise between inliers and the minimal distance-to-noise between inliers and outliers, respectively. The specific definitions of these metrics are as follows:

$$\bar{\kappa}_{\text{in-in}} \triangleq \min_{\substack{i,j \notin O_{\pi^*}, \\ j \neq i}} \frac{\|\theta_i^{\#} - \theta_j^{\#}\|}{(\sigma_i^{\#2} + \sigma_j^{\#2})^{1/2}}, \qquad \bar{\kappa}_{\text{in-out}} \triangleq \min_{\substack{i \notin O_{\pi^*}, \\ j \in O_{\pi^*}}} \frac{\|\theta_i^{\#} - \theta_j^{\#}\|}{(\sigma_i^{\#2} + \sigma_j^{\#2})^{1/2}}, \tag{4}$$

where $O_{\pi^*} \triangleq [m] \setminus \text{Im}(\pi^*)$ is the set of indices of outliers. One main result achieved in homoscedastic case, i. e. $\sigma_i = \sigma^{\#}_{\pi^*(i)} \forall\, i \in S$, is formulated below.

Theorem 1 (Upper bound for LSL). *Let $\alpha \in (0, 1/2)$. If the separation distances $\bar{\kappa}_{\text{in-in}}$ and $\bar{\kappa}_{\text{in-out}}$ corresponding to $(\boldsymbol{\theta}^{\#}, \boldsymbol{\sigma}^{\#}, \pi^*)$ and defined by* (4) *satisfy*

$$\min\{\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}}\} \geq \sqrt{2d} + 4\Big\{ \big(2d \log(\frac{4nm}{\alpha})\big)^{1/4} \vee \big(3 \log(\frac{8nm}{\alpha})^{1/2} \Big\} \tag{5}$$

then the LSL estimator (3) *detects the matching map π^* with probability at least $1 - \alpha$, that is*

$$\mathbf{P}_{\boldsymbol{\theta}^{\#}, \boldsymbol{\sigma}^{\#}, \pi^*}(\hat{\pi}^{\text{LSL}}_{n,m} = \pi^*) \geq 1 - \alpha. \tag{6}$$

Experiments on synthetically generated and real-world data are presented to illustrate theoretical findings.

Chapter 3.

In this chapter, we discuss the results achieved for the variation of the matching map detection problem, where both feature vector sets can contain outliers. Formally, we assume that for some $S^* \subset [n]$ of cardinality k^*, there exists an injective mapping $\pi^* : S^* \to [m]$ such that $\theta_i = \theta^{\#}_{\pi^*(i)}$ holds for all $i \in S^*$. We call the observations $(\boldsymbol{X}_i : i \in S^*)$ and $(\boldsymbol{X}^{\#}_{\pi^*(i)} : i \in S^*)$

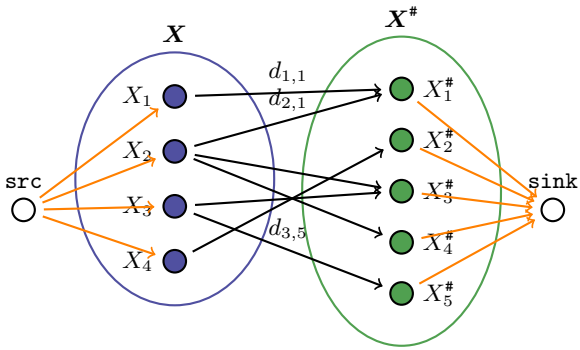

Figure 2: Matching as a Minimum Cost Flow (MCF) problem. The idea is to augment the graph with two nodes, *source* and *sink*, and $n + m$ edges. The capacities of orange edges should be set to 1, while the cost should be set to 0. Setting the total flow sent through the graph to k, the solution of the MCF becomes a matching of size k.

inliers, while the other vectors from the sets $\boldsymbol{X}$ and $\boldsymbol{X}^{\#}$ are considered to be *outliers*. The goal here again is to recover π^* based on the observations $\mathbf{X}$ and $\mathbf{X}^{\#}$ only.

In this section, we introduce a novel procedure to estimate the number of inliers for cases where both sets contain an unknown number of outliers. Our findings indicate that in the high-dimensional setting, the true matching map can be retrieved with a probability of $1 - \alpha$ if the signal-to-noise ratio surpasses a threshold of $5(d \log(4nm/\alpha))^{1/4}$. It is noteworthy that this threshold remains constant and is independent of k^* (the true number of inliers). Our data-driven selection process among candidate mappings $\hat{\pi}_k : k \in [\min(n, m)]$ yielded the aforementioned outcome. Each $\hat{\pi}_k$ minimizes the sum of the squared distances between two sets of size k. The resulting optimization problem can be expressed as a minimum-cost flow problem, thereby enabling efficient resolution. The illustration of the reformulation of the problem as a minimum-cost flow problem is shown on 2. To explain our result, let us introduce

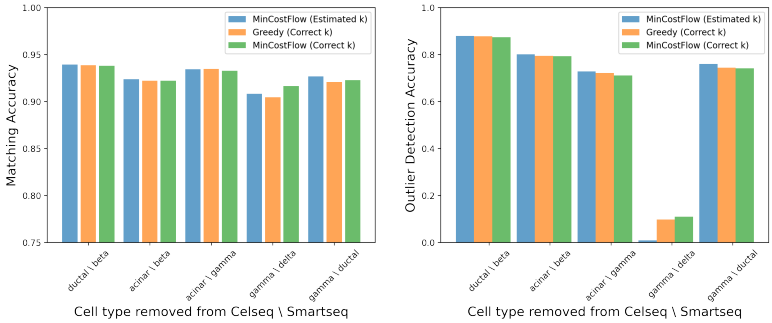

Figure 3: The study compares an algorithm that is unaware of the number of inliers (MinCostFlow estimated k) with algorithms that have the correct number of inliers as input.

the following quantities:

$$\kappa_{i,j} = \|\theta_i - \theta_j^{\#}\|_2/(\sigma^2 + \sigma^{\#2})^{1/2}, \tag{7}$$

$$\bar{\kappa}_{\text{all}} = \min_{i\in[n]} \min_{j\in[m]\setminus\{\pi^*(i)\}} \kappa_{i,j} \tag{8}$$

$$\lambda_{n,m,d,\alpha} = 4\Big\{\big(d\log(\tfrac{4nm}{\alpha})\big)^{\frac{1}{4}} \vee \big(8\log(\tfrac{4nm}{\alpha})\big)^{\frac{1}{2}}\Big\}. \tag{9}$$

Here $\bar{\kappa}_{\text{all}}$ is the signal-to-noise ratio of the difference $X_i - X_j^{\#}$ of a pair of feature vectors. Clearly, for matching pairs, this difference vanishes. Furthermore, if $\kappa_{i,j}$ vanishes or is very small for a non-matching pair, then there is an identifiability issue and consistent recovery of underlying true matching is impossible. Therefore, a natural condition for making consistent recovery possible is to assume that the quantity is bounded away from zero.

In order to be able to recover S^* and the matching map π^*, the key ingredient we use is the maximization of the profile likelihood. This corresponds to looking for the least sum of squares (LSS) of errors over all possible injective mappings defined on a subset of $[n]$ of size k. Formally, if we define

$$\mathcal{P}_k := \left\{\pi : S \to [m] \text{ such that } \begin{matrix} S \subset [n], |S| = k, \\ \pi \text{ is injective} \end{matrix}\right\} \tag{10}$$

to be the set of all k-matching maps, we can define the procedure k-LSS as a solution to the

optimization problem

$$\widehat{\pi}_k^{\mathsf{LSS}} \in \operatorname*{arg\,min}_{\pi \in \mathcal{P}_k} \sum_{i \in S_\pi} \|X_i - X^{\#}_{\pi(i)}\|_2^2, \tag{11}$$

where S_π denotes the support of function π.

Let $\hat{\Phi}(k)$ be the error of $\widehat{\pi}_k^{\mathsf{LSS}}$, that is

$$\hat{\Phi}(k) = \min_{\pi \in \mathcal{P}_k} \sum\nolimits_{i \in S_\pi} \|X_i - X^{\#}_{\pi(i)}\|_2^2. \tag{12}$$

For some values of tuning parameters $\lambda > 0$ and $\gamma > 0$, as well as for some $k_{\mathsf{min}} \in [n]$, initialize $k \leftarrow k_{\mathsf{min}}$ and

1. Compute $\hat{\Phi}(k)$ and $\hat{\Phi}(k+1)$.
2. Set $\bar{\sigma}_k^2 = \hat{\Phi}(k)/(kd)$.
3. If $k = n$ or $\hat{\Phi}(k+1) - \hat{\Phi}(k) > \frac{d+\lambda}{1-\gamma}\bar{\sigma}_k^2$,

   then output $(k, \bar{\sigma}_k, \hat{\pi}_k^{\mathsf{LSS}})$.
4. Otherwise, increase $k \leftarrow k+1$ and go to Step 1.

In the sequel, we denote by $(\hat{k}, \bar{\sigma}_{\hat{k}}, \hat{\pi}_{\hat{k}}^{\mathsf{LSS}})$ the output of this procedure. Notice that we start with the value of $k = k_{\mathsf{min}}$, which in the absence of any information on the number of inliers might be set to $k = 1$. However, using a higher value of k_{min} might considerably speed up the procedure and improve its quality.

For appropriately chosen values of γ and λ, as stated in the next theorem, the described procedure outputs the correct values of k^* and π^* with high probability.

Theorem 2. *Let $\alpha \in (0,1)$ and $\lambda_{n,m,d,\alpha}$ be defined by* (7)*. If $\bar{\kappa}_{\mathsf{all}} > (\frac{5}{4})\,\lambda_{n,m,d,\alpha}$, then the output $(\hat{k}, \hat{\pi}_{\hat{k}}^{\mathsf{LSS}})$ of the model selection algorithm with parameters $\lambda = (\frac{1}{4})\lambda^2_{n,m,d,\alpha}, \gamma = \frac{\lambda}{d}$ satisfies $\mathbf{P}(\hat{\pi}_{\hat{k}}^{\mathsf{LSS}} = \pi^*) \geq 1 - \alpha$.*

Finally, at the end of this chapter, we report the results of our numerical experiments on synthetic and real-world data that serve to illustrate our theoretical findings and offer further insight into the properties of the algorithms studied in this work.

Chapter 4.

Chapter 4 presents studies of the efficacy of recently developed, state-of-the-art entity resolution methods on real-life biomedical datasets. We explore various scenarios for the matching problem, including those without outliers, those with outliers in only one dataset, and those with outliers in both datasets. Subsequently, we conduct an extensive analysis and preprocessing of the biomedical dataset pairs used in our experiments. Our results demonstrate that modern algorithms consistently outperform the original greedy algorithm across all settings. Moreover, we investigate previously proposed procedure that estimates the unknown number of inliers without any supplementary information. We successfully show that algorithms utilizing this estimation technique perform almost as well as those that are provided with the actual number of inliers as input. Figure 3 illustrates some of the results achieved in case of unknown number of inliers, where our proposed algorithm performs as good, if not better, than classical algorithms serving as an oracle baseline, meaning they have additional information of real number of inliers.

List of author's publications

1. Galstyan, T., Minasyan, A., and Dalalyan, A. S. Optimal detection of the feature matching map in presence of noise and outliers. *Electronic Journal of Statistics*, 16(2):5720–5750, 2022.

2. Galstyan, T. Comparison of data matching methods on biomedical datasets. *Vestnik of Russian-Armenian University* , 1(2):46–58, 2022.

3. Galstyan, T. and Minasyan, A. Optimality of the Least Sum of Logarithms in the Problem of Matching Map Recovery in the Presence of Noise and Outliers. *Armenian Journal of Mathematics*, 15(5):1–9, 2023.

4. Minasyan, A., Galstyan, T., Hunanyan, S., and Dalalyan, A. Matching Map Recovery with an Unknown Number of Outliers. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 891–906, 2023.

ԱՄՓՈՓՈՒՄ

Տիգրան Վահագնի Գալստյան

Հատկանիշներ համապատասխանեցնող արտապատկերումների հայտնաբերման խնդրի վիճակագրական և հաշվողական բարդությունը

Բազմաչափ կետերի երկու բազմությունների համապատասխանեցնող արտապատկերման հայտնաբերման խնդիրը լայնորեն ուսումնասիրված է՝ թե՛ տեսականորեն, թե՛ գործնականում: Այդ խնդիրը տարատեսակ կիրառություններ ունի այնպիսի ոլորտներում, ինչպիսիք են համակարգչային տեսողությունը, բիոինֆորմատիկան և բնական լեզվի մշակումը: Համապատասխանեցման խնդրի հայտնի օրինակներից է համակարգչային տեսողության մեջ լոկալ նկարագրիչների երկու բազմությունների միջև համապատասխանության հայտնաբերումը, որը դուրս է բերվում նույն տեսարանի երկու տարբեր նկարներից:

Երբ տվյալների հավաքածուները չեն պարունակում outlier-ներ, այսինքն, երբ երկու համապատասխանող բազմությունները ունեն նույն չափն ու մի բազմության բոլոր կետերը ունեն իրենց համապատասխանը մյուս բազմությունում, համապատասխանեցնող պրոցեդուրայի վիճակագրական հատկությունները ուսումնասիրվել են Կոլիեի և Դալալյանի կողմից: Սակայն, որպես կանոն, վերոնշյալ կիրառություններում ոչ բոլոր կետերը ունեն իրենց համապատասխանը և հաճախ նախապես պարզ չէ համապատասխանող կետերի քանակը: Այս աշխատանքի նպատակն է կենտրոնանալ խնդրի ավելի ընդհանուր դրվածքների վրա և տեսականորեն հետազոտել համապատասխանեցման խնդրի վիճակագրական բարդությունը:

Ատենախոսության մեջ ներմուծվել են համապատասխանեցնող արտապատկերումների այնպիսի մոտարկիչներ, որոնք նախապես տրված (մեկին մոտ) հավանականությամբ համընկնում են ճիշտ համապատասխանեցման

հետ՝ աղմուկի մակարդակի և այլ պարամետրերի հնարավոր ամենաթույլ սահմանափակումների դեպքում: Հատկանիշների համապատասխանեցման դիտարկված բոլոր ընդհանրացումներում ներկայացվել են ալգորիթմներ, որոնք թույլ են տալիս որոշել արտապատկերման չափը (հակադիր բազմությունում համապատասխանն ունեցող կետերի քանակը) հիմնվելով միմիայն դիտարկված տվյալների վրա: Մեր առաջարկած ալգորիթմները պարունակում են օպտիմիզացիայի խնդիրներ, որոնք վերածնակերպվել են որպես արդեն հայտնի գրաֆում ամենաէժան հոսքի հայտնաբերման խնդիր: Ցույց է տրվել խնդիրների համարժեքությունը և լուծման հաշվողական արդյունավետությունը: Որոշվել է նաև ազդանշան/աղմուկ հարաբերության այն օպտիմալ շեմը, որը գերազանցելու դեպքում, ներմուծված ալգորիթմները վերականգնում են ճշգրիտ արտապատկերումը մեծ հավանականությամբ:

Փորձնական ճանապարհով ցույց է տրվել, որ նոր առաջարկված ալգորիթմը ունակ է ավելի ճշգրիտ վերականգնել համապատասխանեցնող արտապատկերումը, նախապես չունենալով համապատասխանեցնող արտապատկերման չափը («ավելորդ» հատկանիշների քանակը), համեմատած դասական մեթոդների հետ, որոնք մոտարկում են արտապատկերումը միայն արտապատկերման չափը նախապես ֆիքսելու դեպքում:

Առաջարկված մեթոդների ճշտությունը և հաշվողական իրագործելիությունը ցուցադրվել են արհեստականորեն գեներացված և իրական տվյալների շտեմարանների վրա:

РЕЗЮМЕ

Галстян Тигран Ваагнович

Статистическая и вычислительная сложность проблемы определения отображений для сопоставления признаков

Проблема поиска наилучшего соответствия между двумя облаками точек тщательно изучалась как теоретически, так и экспериментально. Проблема сопоставления возникает в различных приложениях, например, в компьютерном зрении, биоинформатике и обработке естественного языка. В компьютерном зрении поиск соответствия между двумя наборами локальных дескрипторов, извлеченных из двух изображений одной и той же сцены, является хорошо известным примером вышеупомянутой проблемы.

Когда наборы данных не содержат выбросов, то есть оба совпадающих набора имеют одинаковый размер и все точки имеют соответствующие совпадения в другом наборе данных, оптимальность процедур сопоставления была тщательно изучена с минимаксной статистической точки зрения Коллие и Далаляном. Очевидно, что в вышеупомянутых приложениях не все точки имеют свои точки совмещения, и вряд ли можно знать заранее, сколько точек имеют соответствующие точки совмещения. Цель настоящей работы — сосредоточиться на различных расширенных постановках задачи (т. е. содержащих выбросы) и получить теоретическое понимание статистических ограничений задачи сопоставления.

В диссертации были введены такие аппроксиматоры сопоставления, которые совпадают с правильным сопоставлением с заданной вероятностью (близкой к единице) в случае максимально слабых ограничений на уровень шума и других параметров. Во всех рассмотренных обобщениях сопоставления признаков представлены алгоритмы, позволяющие определять размер отображения (количество точек сопоставления в противоположном множестве) только на основе наблюдаемых данных. Предлагаемые нами алгоритмы содержат задачи оптимизации, которые переформулируются как уже известная задача поиска самого дешевого потока в графе. Показана эквивалентность этих

задач и вычислительная эффективность решения. Также был определен оптимальный порог соотношения сигнал/шум, при превышении которого введенные алгоритмы с высокой вероятностью восстанавливают точное отображение.

Экспериментально показано, что новый предложенный алгоритм способен более точно восстановить сопоставляющее отображение, не имея заранее размера отображения (количества «лишних» признаков), по сравнению с классическими методами, которые аппроксимируют отображение только при знании размера заранее.

Точность и вычислительная осуществимость предложенных методов были продемонстрированы на искусственно созданных и реальных наборах данных.