Matching Map Recovery with an Unknown Number of Outliers

Anonymous Author
Anonymous Institution

Abstract

We consider the problem of finding the matching map between two sets of d-dimensional noisy feature-vectors. The distinctive feature of our setting is that we do not assume that all the vectors of the first set have their corresponding vector in the second set. If n and m are the sizes of these two sets, we assume that the matching map that should be recovered is defined on a subset of unknown cardinality $k^* \leq min(n, m)$. We show that, in the high-dimensional setting, if the signal-to-noise ratio is larger than $8(d \log(4nm/\alpha))^{1/4}$, then the true matching map can be recovered with probability $1 - \alpha$. Interestingly, this threshold does not depend on k^* and is the same as the one obtained in prior work in the case of $k = \min(n, m)$. The procedure for which the aforementioned property is proved is obtained by a data-driven selection among candidate mappings $\{\widehat{\pi}_k : k \in [\min(n, m)]\}$. Each $\widehat{\pi}_k$ minimizes the sum of squares of distances between two sets of size k. The resulting optimization problem can be formulated as a minimum-cost flow problem, and thus solved efficiently. Finally, we report the results of numerical experiments on both synthetic and real-world data that illustrate our theoretical results and provide further insight into the properties of the algorithms studied in this work.

1 INTRODUCTION

The problem of finding the best match between two point clouds has been extensively studied, both theoretically and experimentally. The matching problem arises in various applications, for instance in computer vision and natural language processing. In computer vision, finding the correspondence between two sets of local descriptors extracted from two images of the same scene is a well known example of a matching problem. In natural language processing, in particular in machine translation, the correspondence between vector representations of the same text in two different languages is another example of a matching problem. Clearly, in these problems, not all the points have their matching point and one can hardly know in advance how many points have their corresponding matching points. The goal of the present work is to focus on this setting and to gain theoretical understanding on the statistical limitations of the matching problem.

To formulate the problem and to state the main result, let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{X}^\# = (X_1^\#, \ldots, X_m^\#)$ be two sequences of feature vectors of sizes n and m such that $m \geq n \geq 2$. We assume that these sequences are noisy versions of some feature-vectors, *i.e.*,

$$\begin{cases} X_i = \theta_i + \sigma\xi_i \ , \\ X_j^\# = \theta_j^\# + \sigma^\#\xi_j^\#, \end{cases} \qquad i \in [n] \text{ and } j \in [m], \qquad (1)$$

where, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\theta}^\# = (\theta_1^\#, \ldots, \theta_m^\#)$ are two sequences of deterministic vectors from $\mathbb{R}^d$, corresponding to the original feature-vectors. The noise components of $\mathbf{X}$ and $\mathbf{X}^\#$ are two independent sequences of i.i.d. standard Gaussian random vectors. Formally,

$$\xi_1, \ldots, \xi_n, \xi_1^\#, \ldots, \xi_m^\# \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d),$$

where $\mathbf{I}_d$ is the identity matrix of size $d \times d$. 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 $(\boldsymbol{X}_{\pi^*(i)}^\# : i \in S^*)$ *inliers*, while the other vectors from the sequences $\boldsymbol{X}$ and $\boldsymbol{X}^\#$ are considered to be *outliers*. The ultimate goal is to recover π^* based on the observations $\mathbf{X}$ and $\mathbf{X}^\#$ only.

Various versions of this problem have been studied in the literature. Collier and Dalalyan (2013, 2016) considered the outlier-free case with equal sizes of sequences $\boldsymbol{X}$ and $\boldsymbol{X}^\#$ (*i.e.*, $m = n$ and $S^* = [n]$), whereas Galstyan et al. (2021) investigated the case with outliers in one of the sequences only (*i.e.*, $m \geq n$ and $S^* = [n]$). Other variations of the matching problem under Hamming loss have been studied by Chen et al. (2022b); Kunisky and Niles-Weed (2022);

Preliminary work. Under review by AISTATS 2023. Do not distribute.

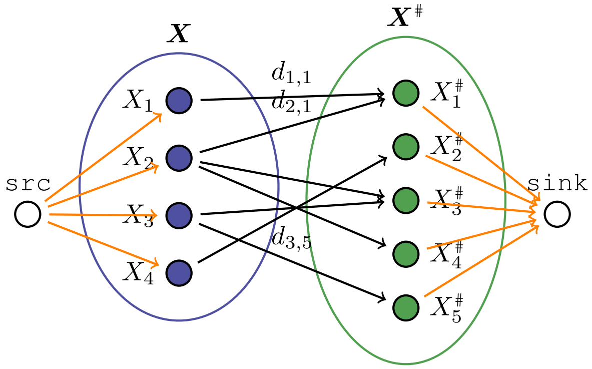

Figure 1: Matching recovery can be cast into a Minimum Cost Flow (MCF) problem. The idea is to augment the bipartite graph with two additional nodes *source* and *sink* and $n+m$ additional edges. The capacities of orange edges should be set to 1, while the cost should be set to 0. Setting the total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

Wang et al. (2022). These papers obtain minimax-optimal separation rates and, in most cases, despite the discrete nature of the matching problem, provide computationally tractable procedures achieving these rates.

When S^* is an arbitrary subset of $[n]$, which is the setting we focus on in this work, one can wonder whether the minimax separation rate is the same as in the case of known S^*. Since the absence of knowledge on S^* brings additional combinatorial complexity to the problem, one can also wonder whether it is still possible to conciliate statistical optimality and computational tractability. We show in this work that the answers to these questions are affirmative.

To explain our result, let us introduce

$$\kappa_{i,j} = \frac{\|\theta_i - \theta^{\#}_j\|_2}{(\sigma^2 + \sigma^{\#2})^{1/2}},$$

which 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 matching is impossible. Therefore, a natural condition for making consistent recovery possible is to assume that the quantity

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

is bounded away from zero. A recovery procedure $\widehat{\pi}$ is considered to be good, if the threshold λ such that $\widehat{\pi}$ recovers π^* with high probability as soon as $\bar{\kappa}_{\text{all}} \geq \lambda$ is as small as possible. It was proved in (Collier and Dalalyan, 2016) that when $k^* = n = m$, one can recover π^* with probability $1-\alpha$ for $\lambda = 4\big\{\big(d\log(4n^2/\alpha)\big)^{1/4} \vee \big(8\log(4n^2/\alpha)\big)^{1/2}\big\}$. Furthermore, it was proved that this threshold is minimax optimal (*i.e.*, optimal in the family of all possible recovery procedures). This implies that there are two regimes. In the low dimensional regime $d \lesssim \log n$, the separation rate is dimension independent. In contrast with this, the separation rate scales roughly as $d^{1/4}$ in the (moderately) high dimensional regime $d \gtrsim \log n$.

is this rate optimal for the arbitrary outlier case?

Let us set

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

The main contributions of this work are the following.

- For any given $k \in [\min(n,m)]$, we show that the k Least Sum of Squares (k-LSS) procedure, $\widehat{\pi}^{\text{LSS}}_k$, based on maximizing profile likelihood among matching maps between two sets of size k, can be efficiently computed using the Minimum-Cost Flow problem.
- If the value k turns out to be smaller than k^* and $\bar{\kappa}_{\text{all}} \geq \lambda_{n,m,d,\alpha}$, we prove that k-LSS makes no mistake with probability $1-\alpha$.
- We design a model data-driven selection algorithm that adaptively chooses $\widehat{k}$ such that with probability $1-\alpha$, we have $\widehat{k} = k^*$ and $\widehat{\pi}^{\text{LSS}}_{\widehat{k}} = \pi^*$ as soon as $\bar{\kappa}_{\text{all}} \geq (5/4)\lambda_{n,m,d,\alpha}$.

This implies that our data driven algorithm $\widehat{\pi}^{\text{LSS}}_{\widehat{k}}$ achieves the minimax separation rate. More surprisingly, this shows that there is no gap in statistical complexities between the problems of recovering matching maps in outlier-free and outliers-present-on-both-sides settings.

2 OTHER RELATED WORK

In statistical hypotheses testing, the separation rates became key objects for measuring the quality of statistical procedures, see the seminal papers (Burnashev, 1979; Ingster, 1982) as well as the monographs (Ingster and Suslina, 2003; Juditsky and Nemirovski, 2020). Currently, this approach is widely adopted in machine learning literature (Blanchard et al., 2018; Collier, 2012; Ramdas et al., 2016; Wei et al., 2019; Wolfer and Kontorovich, 2020; Xing et al., 2020). Beyond the classical setting of two hypotheses, it can also be applied to multiple hypotheses testing framework, for instance, variable selection (Azaïs and de Castro, 2020; Comminges and Dalalyan, 2012, 2013; Ndaoud and Tsybakov, 2020) or the matching problem considered here.

In computer vision, the feature matching is a well studied problem. One of the main directions is to accelerate matching algorithms, based on fast approximate methods (see e.g. Harwood and Drummond (2016); Jiang et al. (2016); Malkov and Yashunin (2020); Wang et al. (2018)). Another direction is to improve the matching quality by considering alternative local descriptors (Calonder et al., 2010; Chen et al., 2010; Rublee et al., 2011) for given keypoints. The choice of keypoints is considered in Bai et al. (2020); Tian et al. (2020).

The minimum-cost flow problem was first studied in the context of the Hungarian algorithm (Kuhn, 2012) and the assignment problem, which is a special case of minimum cost flow on bipartite graphs with all edges having unit capacity. Generalization of Hungarian algorithm for graphs with arbitrary edge costs guarantees $\mathcal{O}((n+F)m)$ time complexity, where n is the number of nodes in the graph, m is the number of edges and F is the total flow sent through the graph. There have also been other algorithms with similar complexity guarantees (Ahuja et al., 1992; Fulkerson, 1961). Since then many algorithms have been proposed for solving minimum-cost flow problem in strongly polynomial time (Galil and Tardos, 1988; Goldberg and Tarjan, 1989; Orlin, 1993, 1996; Orlin et al., 1993) with the fastest runtime of around $\mathcal{O}(nm)$. Recent advances for solving minimum-cost flow problems have been proposed in Goldberg et al. (2015) and Chen et al. (2022a). The latter proposes an algorithm with an almost-linear computational time.

Permutation estimation and related problems have been recently investigated in different contexts such as statistical seriation (Cai and Ma, 2022; Flammarion et al., 2019; Giraud et al., 2021), noisy sorting (Mao et al., 2018), regression with shuffled data (Pananjady et al., 2017; Slawski and Ben-David, 2019), isotonic regression and matrices (Ma et al., 2020; Mao et al., 2020; Pananjady and Samworth, 2020), crowd labeling (Shah et al., 2021), recovery of general discrete structure (Gao and Zhang, 2019), and multitarget tracking (Chertkov et al., 2010; Kunisky and Niles-Weed, 2022).

3 MAIN THEORETICAL RESULT

This section contains the main theoretical contribution of the present work. 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 injective mappings defined on a subset of $[n]$ of size k. Formally, if we define

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

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

$$\widehat{\pi}_k^{\text{LSS}} \in \arg\min_{\pi \in \mathcal{P}_k} \sum_{i \in S_\pi} \|X_i - X^\#_{\pi(i)}\|_2^2, \tag{3}$$

where S_π denotes the support of function π. In the particular case $k^* = n$, the optimization above is over all the injective mappings from $[n]$ to $[m]$. This coincides with the LSS method from (Galstyan et al., 2021).

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

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

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

1. Compute $\widehat{\Phi}(k)$ and $\widehat{\Phi}(k+1)$.
2. Set $\bar{\sigma}_k^2 = \widehat{\Phi}(k)/(kd)$.
3. If $k = n$ or $\widehat{\Phi}(k+1) - \widehat{\Phi}(k) > \frac{d+\lambda}{1-\gamma}\bar{\sigma}_k^2$, then output $(k, \bar{\sigma}_k, \widehat{\pi}_k^{\text{LSS}})$.
4. Otherwise, increase $k \leftarrow k+1$ and go to Step 1.

In the sequel, we denote by $(\widehat{k}, \bar{\sigma}_{\widehat{k}}, \widehat{\pi}_{\widehat{k}}^{\text{LSS}})$ the output of this procedure. Notice that we start with the value of $k = k_{\min}$, which in 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 1. *Let $\alpha \in (0,1)$ and $\lambda_{n,m,d,\alpha}$ be defined by (2). If $\bar{\kappa}_{\text{all}} > ({}^5\!/\!_4)\,\lambda_{n,m,d,\alpha}$, then the output $(\widehat{k}, \widehat{\pi}_k^{\text{LSS}})$ of the model selection algorithm with parameters $\lambda = \gamma = ({}^1\!/\!_4)\lambda^2_{n,m,d,\alpha}$ satisfies $\mathbf{P}(\widehat{k} = k^*, \widehat{\pi}_{\widehat{k}} = \pi^*) \geq 1 - \alpha$.*

Since the condition on the separation distance $\bar{\kappa}_{\text{all}}$ compared to the case of known k^* is different by only a slightly larger constant, from the perspective of statistical accuracy, the case of unknown k^* is not more difficult than that of the known k^*.

In the sequel, without much loss of generality we assume that the sizes of $\boldsymbol{X}$ and $\boldsymbol{X}^\#$ are equal, *i.e.*, $n = m$. Indeed, in the case $m > n$ one can add $m - n$ points arbitrarily far from the rest of the points to the smaller set $\boldsymbol{X}$, therefore obtaining equal size sets $\boldsymbol{X}^+$ and $\boldsymbol{X}^\#$.

Notice that we optimize the function from (3) over a finite set of injective functions π. For a given value of k, the number of such functions is $k!\binom{n}{k}^2$ making thus performing exhaustive search prohibitively computationally expensive. Instead, we show in Section 5 that this problem can indeed be solved efficiently with complexity $\tilde{\mathcal{O}}(\sqrt{k}\,n^2)$, where the notation $\tilde{\mathcal{O}}$ means up to polylogarithmic factors.

4 INTERMEDIATE RESULTS AND PROOF OF THEOREM 1

This section is devoted to the proof of our main result. Along the way, we establish some intermediate results which are of interest on their own. The proofs of some technical lemmas are deferred to the supplementary material.

4.1 Sub-mapping Recovery by LSS for $k \leq k^*$

The first question we address in this section is under which conditions the LSS estimator $\widehat{\pi}_k^{\mathrm{LSS}}$ from (3) recovers correct matches. Of course, the only way of correctly estimating the true matching is to choose $k = k^*$. However, it turns out that even if we overestimate the number of outliers and choose a value k which is smaller than the true value k^*, with high probability the LSS estimator makes no wrong matches. Naturally, this result, stated in the next theorem, is valid under the condition that the relative signal-to-noise ratio of all incorrect pairs of original features is larger than some threshold.

Theorem 2 (Quality of k-LSS when $k \leq k^*$)**.** *Let $\widehat{S} = \mathbf{supp}(\widehat{\pi})$ for $\widehat{\pi} = \widehat{\pi}_k^{\mathrm{LSS}}$ defined by (3), $\alpha \in (0,1)$ and*

$$\lambda_{n,d,\alpha} = 4\Big(\big(d\log(4n^2/\alpha)\big)^{1/4} \vee \big(8\log(4n^2/\alpha)\big)^{1/2}\Big). \quad (4)$$

If $k \leq k^*$ and the signal-to-noise ratio satisfies the condition $\bar{\kappa}_{\text{all}} \geq \lambda_{n,d,\alpha}$ then, with probability at least $1-\alpha$, the support of the estimator $\widehat{\pi}$ is included in S^* and $\widehat{\pi}$ coincides with π^* on the set $\widehat{S}$. Formally,

$$\mathbf{P}\big(\widehat{S} \subset S^* \text{ and } \widehat{\pi}(i) = \pi^*(i), \forall i \in \widehat{S}\big) \geq 1-\alpha.$$

Proof of Theorem 2. Note that the random vectors

$$\eta_{ij} = \frac{\sigma\xi_i - \sigma^{\#}\xi_j^{\#}}{\sqrt{\sigma^2 + \sigma^{\#2}}}$$

are standard Gaussian and define the following quantities

$$\zeta_1 \triangleq \max_{i \neq j} \frac{|(\theta_i - \theta_j^{\#})^\top \eta_{ij}|}{\|\theta_i - \theta_j^{\#}\|_2},$$
$$\zeta_2 \triangleq d^{-1/2} \max_{i,j} \big|\|\eta_{ij}\|_2^2 - d\big|. \quad (5)$$

For the ease of notations for any matching map π we also define $L(\pi)$ as follows

$$L(\pi) = \sum_{i \in S_\pi} \frac{\|X_i - X_{\pi(i)}^{\#}\|_2^2}{\sigma^2 + \sigma^{\#2}}.$$

We start with two auxiliary lemmas that will be used in other proofs as well. The proofs of these lemmas are deferred to the appendix.

Lemma 1. *Let π be any matching map that can not be obtained as a restriction of π^* on a subset of $[n]$. Let $S_0 \subset S^*$ be an arbitrary set satisfying $|S_0| \leq |S_\pi|$ and $\{i \in S_\pi \cap S^* : \pi(i) = \pi^*(i)\} \subset S_0$ and let π_0 be the restriction of π^* to S_0. On the event $\Omega_0 = \{8\zeta_1 \leq \bar{\kappa}_{\text{all}}\} \cap \{4\sqrt{d}\,\zeta_2 \leq \bar{\kappa}_{\text{all}}^2\}$ we have*

$$L(\pi) - L(\pi_0) \geq (1/4)\bar{\kappa}_{\text{all}}^2 + d(|S_\pi| - |S_0|).$$

Let π be any matching map from $\mathcal{P}_k$ that is not a restriction of π^*. Since $|S_\pi| = k \leq k^*$, there exists necessarily a π_0 as in Lemma 1 such that $|S_0| = |S_\pi|$. For this π_0, we have $L(\pi) - L(\pi_0) \geq (1/4)\bar{\kappa}_{\text{all}}^2 > 0$. This implies that π cannot be a minimizer of $L(\cdot)$ over $\mathcal{P}_k$. As a consequence, on Ω_0, any minimizer of $L(\cdot)$ over $\mathcal{P}_k$ is a restriction of π^*. Therefore, on Ω_0, we have $\widehat{S} \subset S^*$ and $\widehat{\pi}_k = \pi^*|_{\widehat{S}}$. It remains to prove that $\mathbf{P}(\Omega_0) \geq 1-\alpha$.

Lemma 2. *Let $\Omega_{0,x} = \{8\zeta_1 \leq x\} \cup \{4\sqrt{d}\zeta_2 \leq x^2\}$ with ζ_1, ζ_2 defined as in (5). Then, for every $x > 0$, $\mathbf{P}(\Omega_{0,x}^{\complement})$ is upper bounded by*

$$2n^2\Big(\exp\Big\{-\frac{x^2}{128}\Big\} + \exp\Big\{-\frac{x^2}{128d}(x^2 \wedge 4d)\Big\}\Big).$$

We apply Lemma 2 with $x = \bar{\kappa}_{\text{all}}$ to show that $\mathbf{P}(\Omega_0) \geq 1-\alpha$. Clearly, a sufficient condition for the latter is

$$\begin{cases} 2n^2\exp\big\{-\bar{\kappa}_{\text{all}}^2/128\big\} \leq \alpha/2, \\ 2n^2\exp\Big\{-\frac{(\bar{\kappa}_{\text{all}}/16)^2}{d}\big(2\bar{\kappa}_{\text{all}}^2 \wedge 8d\big)\Big\} \leq \alpha/2. \end{cases}$$

This system is equivalent to

$$\bar{\kappa}_{\text{all}} \geq 8\bigg(2\log\frac{4n^2}{\alpha}\bigg)^{1/2} \quad \text{and} \quad \bar{\kappa}_{\text{all}} \geq 4\bigg(\frac{d}{2}\log\frac{4n^2}{\alpha}\bigg)^{1/4}.$$

Therefore, if the signal-to-noise ratio satisfies

$$\bar{\kappa}_{\text{all}} \geq 4\Big(\big(d\log(4n^2/\alpha)\big)^{1/4} \vee \big(8\log(4n^2/\alpha)\big)^{1/2}\Big),$$

we have $\mathbf{P}(\Omega_0) \geq 1-\alpha$. □

4.2 Matching Map Recovery for Unknown k^*

If no information on k^* is available, and the goal is recover the entire mapping π^*, one can proceed by model selection. More precisely, one can compute the collection of estimators $\{\widehat{\pi}_k^{\mathrm{LSS}} : k \in [n]\}$ and select one of them using a suitable criterion. To define the selection criterion proposed in this paper, let us remark that

$$\widehat{\Phi}(k) = \min_{\pi \in \mathcal{P}_k} \sum\nolimits_{i \in S_\pi} \|X_i - X_{\pi(i)}^{\#}\|_2^2$$

is an increasing function. The increments of this function for $k \leq k^*$ are not large, since they essentially correspond to the squared norm of a pure noise vector distributed according to a scaled χ^2 distribution with d degrees of freedom. The main idea behind the criterion we propose below is that the increment of $\widehat{\Phi}$ at k^* is significantly larger than the previous ones, and the gap is of the order of $\bar{\kappa}_{\text{all}}^2$. Therefore, if $\bar{\kappa}_{\text{all}}^2$ is larger than the deviations of the χ_d^2 distribution, we are able to detect the value of k^* and to estimate the true matching.

Based on these considerations, for any tolerance level $\alpha \in (0,1)$, we set $\sigma_0^2 = \sigma^2 + \sigma^{\#2}$ and define the estimator[1]

$$\begin{aligned}\widehat{k} = 1 + \max\big\{k \in \{0,\dots,n-1\} : \widehat{\Phi}(k+1) - \widehat{\Phi}(k) \\ \le \sigma_0^2(d + \lambda_{n,d,\alpha}^2/4)\big\}\end{aligned}$$

with $\lambda_{n,d,\alpha}$ as in (4).

Theorem 3 (Model selection accuracy)**.** *Let $\alpha \in (0,1)$. If $\bar{\kappa}_{\text{all}} > \lambda_{n,d,\alpha}$, then it holds that $\mathbf{P}\big(\widehat{k} = k^*$ and $\widehat{\pi}_{\widehat{k}} = \pi^*\big) \ge 1-\alpha$. Therefore, $\lambda_{n,d,\alpha}$ is an upper bound on the separation distance in the case of unknown k^*.*

A remarkable feature put forward by this result is that a data-driven selection of k based on the increments of the test statistics $\widehat{\Phi}$ leads to the recovery of π^*, with high probability, under the same constraint on the separation rate as in the case of known k^*. It is however important to underline that this criterion requires the knowledge of the noise level. Therefore, from the point of view of statistical accuracy, the case of unknown k^* is not more difficult than the case of known k^*, provided the noise level is known.

Proof of Theorem 3. The main parts of the proof will be done in the following two lemmas, the proofs of which are postponed to the appendix. For the known value of σ_0^2 it is more convenient to work with the normalized version of test statistic $\widehat{\Phi}(\cdot)$, denoted by $\widehat{L}(\cdot)$ and defined by

$$\widehat{L}(k) = \min_{\pi \in \mathcal{P}_k} \sum_{i \in S_\pi} \frac{\|X_i - X^{\#}_{\pi(i)}\|_2^2}{\sigma^2 + \sigma^{\#2}} \equiv \frac{\widehat{\Phi}(k)}{\sigma_0^2}.$$

Lemma 3. *On the event, $\Omega_0 = \{8\zeta_1 \le \bar{\kappa}_{\text{all}}; 4\sqrt{d}\,\zeta_2 \le \bar{\kappa}_{\text{all}}^2\}$, we have $\widehat{L}(k^*+1) - \widehat{L}(k^*) \ge d + \bar{\kappa}_{\text{all}}^2/4$.*

Lemma 4. *On the event, $\Omega_0 = \{8\zeta_1 \le \bar{\kappa}_{\text{all}}; 4\sqrt{d}\,\zeta_2 \le \bar{\kappa}_{\text{all}}^2\}$, for every $k < k^*$, we have $\widehat{L}(k+1) - \widehat{L}(k) \le d + \sqrt{d}\,\zeta_2$.*

Lemma 2 implies that the event $\{8\zeta_1 \le \lambda_{n,d,\alpha}\} \cap \{4\sqrt{d}\,\zeta_2 \le \lambda_{n,d,\alpha}^2\}$ has a probability at least $1-\alpha$. This event being included in Ω_0, on this event, we have $\widehat{L}(k+1) - \widehat{L}(k) \le d + \lambda_{n,d,\alpha}^2/4$ for any $k < k^*$, in view of Lemma 4. On the other hand, in view of Lemma 3, on the same event we have $\widehat{L}(k^*+1) - \widehat{L}(k^*) \ge d + \bar{\kappa}_{\text{all}}^2/4 > d + \lambda_{n,d,\alpha}^2/4$. This implies that $\widehat{k} = k^*$. We can infer from this equality that $\widehat{\pi}_{\widehat{k}} = \widehat{\pi}_{k^*}$. In view of Theorem 2, on the same event we have $\widehat{\pi}_{k^*} = \pi^*$. □

4.3 Matching Map Recovery for Unknown k^* and Unknown Noise Level

In the previous subsection we considered the case of unknown k^* with known noise levels σ and $\sigma^{\#}$. Notice that we don't need to estimate parameters $\sigma, \sigma^{\#}$ separately, it is sufficient to estimate only their squared sum, which is denoted by σ_0^2. In the definition of $\widehat{k}$, we use the value of σ_0^2 in the threshold for $\widehat{\Phi}(k+1) - \widehat{\Phi}(k)$. When both k^* and σ_0^2 are unknown, we first estimate σ_0^2 and then plug it in the criterion of selection of k based on the increments of $\widehat{\Phi}(k+1) - \widehat{\Phi}(k)$.

To start with, for any $k \in [n]$ we define "candidate" estimators of σ_0^2 as follows

$$\bar{\sigma}_k^2 = \frac{\widehat{\Phi}(k)}{kd}, \tag{6}$$

The rationale for this definition is that for small values of k, $\widehat{\pi}_k^{\text{LSS}}$ contains only correct matches and, therefore, $\widehat{\Phi}(k)/\sigma_0^2$ is merely a sum of k independent random variables drawn from the χ^2 distribution with d degrees of freedom. Hence, after division by kd, we obtain an estimator of σ_0^2. However, from the perspective testing the values of k, we need to slightly overestimate the noise variance. This is done through the multiplication by the inflation factor $1/(1-\gamma)$.

Proof of Theorem 1. We will provide the proof only in the high dimensional setting, that is we assume throughout that $d \ge 800\log(2n/\sqrt{\alpha})$. First, we show that for every $k < k^*$ the condition from $\widehat{\Phi}(k+1) - \widehat{\Phi}(k) \le \frac{d+\lambda}{1-\gamma}\bar{\sigma}_k^2$ is satisfied on an event of high probability. Second, we prove that for $k = k^*$ this condition is violated on the same event of high probability. Therefore, the combination of these two results concludes the proof.

Using the first part of the proof of Lemma 4, for all $k < k^*$ on the event $\Omega_0 = \{8\zeta_1 \le \lambda\} \cap \{4\sqrt{d}\zeta_2 \le \lambda^2\}$ we have

$$\begin{aligned}\frac{\widehat{\Phi}(k+1) - \widehat{\Phi}(k)}{\widehat{\Phi}(k)} &= \frac{\sum_{\widehat{S}_{k+1}} \|\eta_{i,\pi^*(i)}\|_2^2 - \sum_{\widehat{S}_k} \|\eta_{i,\pi^*(i)}\|_2^2}{\sum_{i \in \widehat{S}_k} \|\eta_{i,\pi^*(i)}\|_2^2} \\ &\le \frac{d + \sqrt{d}\zeta_2}{kd + k\sqrt{d} \cdot \min_{1 \le i \le n} \frac{\|\eta_{i,\pi^*(i)}\|_2^2 - d}{\sqrt{d}}} \\ &\le \frac{d + \sqrt{d}\zeta_2}{k(d - \sqrt{d}\zeta_2)_+}.\end{aligned}$$

Using the second part of the proof of Lemma 2, we can further upper bound the expression from the last display as follows

$$\frac{\widehat{\Phi}(k+1) - \widehat{\Phi}(k)}{\widehat{\Phi}(k)} \le \frac{d + \lambda^2/4}{k(d - \lambda^2/4)_+}.$$

Now we show that for $k = k^*$ the relative difference of func-

[1] We use the convention $\widehat{\Phi}(0) = 0$.

tion $\widehat{\Phi}(\cdot)$ at points k^*+1 and k^* is large enough. Indeed,

$$\begin{aligned}
&\frac{\widehat{\Phi}(k^*+1)-\widehat{\Phi}(k^*)}{\widehat{\Phi}(k^*)} \\
&= \frac{\sum_{\widehat{S}_{k^*+1}}\|X_i - X^{\#}_{\widehat{\pi}_{k^*+1}(i)}\|_2^2 - \sum_{\widehat{S}_{k^*}}\|X_i - X^{\#}_{\widehat{\pi}_{k^*}(i)}\|_2^2}{\sum_{i\in\widehat{S}_{k^*}}\|X_i - X^{\#}_{\widehat{\pi}_{k}(i)}\|_2^2} \\
&\geq \frac{\sum_{\widehat{S}_{k^*+1}}\|X_i - X^{\#}_{\widehat{\pi}_{k^*+1}(i)}\|_2^2 - \sum_{S^*}\|X_i - X^{\#}_{\pi^*(i)}\|_2^2}{\sum_{i\in S^*}\|X_i - X^{\#}_{\pi^*(i)}\|_2^2} \\
&\geq \frac{\min_{i\neq\pi^*(j)}\|X_i - X^{\#}_j\|_2^2}{\sigma_0^2\sum_{i\in S^*}\|\eta_{i,\pi^*(i)}\|_2^2} \\
&\geq \frac{\bar{\kappa}_{\text{all}}^2 - 2\zeta_1\bar{\kappa}_{\text{all}} + d - \sqrt{d}\zeta_2}{k(d+\sqrt{d}\zeta_2)}.
\end{aligned}$$

Then, on the event Ω_0 we bound the quantities ζ_1 and ζ_2 from the last display along with using the condition on $\bar{\kappa}_{\text{all}}$. One can now check that if $d \geq 800\log(2n/\sqrt{\alpha})$ then $\lambda^2 \leq {}^4\!/_5\, d$, which in turn implies

$$\begin{aligned}
\frac{\widehat{\Phi}(k^*+1)-\widehat{\Phi}(k^*)}{\widehat{\Phi}(k^*)} &\geq \frac{\bar{\kappa}_{\text{all}}^2 - \bar{\kappa}_{\text{all}}\lambda/4 + d - \lambda^2/4}{k(d+\lambda^2/4)} \\
&\geq \frac{d+\lambda^2/4}{kd(1-\lambda^2/4d)}.
\end{aligned}$$

Thus, we have shown that on the event Ω_0 our model selection procedure will select k^*, *i.e.,* $\widehat{k} = k^*$. The last equality implies that $\widehat{\pi}_{\widehat{k}} = \widehat{\pi}_{k^*}$. Moreover, in view of Theorem 2 on the same event Ω_0 we have $\widehat{\pi}_{k^*} = \pi^*$. Finally, using Lemma 2, we get that the event Ω_0 has probability of at least $1-\alpha$. Therefore, the desired result follows. □

4.4 Lower Bounds

We already mentioned that Theorem 2 and Theorem 3 imply that the minimax rate of separation in the problem of recovering π^* is at most of the order of $\lambda_{n,m,d,\alpha}$ defined in (4). An interesting and natural question is whether this rate is optimal. In the literature, the lower bounds for similar models have been proved, see (Collier and Dalalyan, 2016, Theorem 2) for the case of $n = m$ and (Galstyan et al., 2021, Theorem 5) for general rectangular case $m \geq n$. Our statistical model being more general[2] than those considered in these two references, the same lower bound applies to the model considered in this work. Therefore, combining the results of Theorem 3 and (Collier and Dalalyan, 2016, Theorem 2) along with the fact that the separation distance has the same rate in both theorems implies that $\lambda_{n,m,d,\alpha}$ is the optimal rate of separation. Interestingly, it does not depend on k^*.

[2] Indeed, it involves an additional (unknown) parameter k^*.

5 COMPUTATIONAL ASPECTS AND NUMERICAL EXPERIMENTS

In this section, we address computational aspects of the optimization problem from (3). We show that it can be cast into a minimum cost flow problem. The latter is also known as imperfect matching problem and to the best of our knowledge the fastest algorithm with complexity $O(\sqrt{k^*}\, n^2\log(k^*))$ is proposed in (Goldberg et al., 2015). We then report results of numerical experiments conducted both on synthetic and real data.

5.1 Relation to Minimum Cost Flow Problem

Let $d_{ij} = \|X_i - X^{\#}_j\|_2^2$, for $(i,j)\in[n]\times[m]$, be the squared distances between observed feature-vectors. Consider the following linear program

$$\begin{aligned}
&\text{minimize } \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}w_{ij} \\
&\text{subject to } \ w_{ij}\in\{0,1\}, \quad \forall(i,j)\in[n]\times[m], \qquad (7) \\
&\sum_{i=1}^{n} w_{i\cdot} \leq 1, \quad \sum_{j=1}^{m} w_{\cdot j} \leq 1, \quad \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} = k,
\end{aligned}$$

known in the literature as minimum cost flow problem. Above, the notation $\sum_{i=1}^{n} w_{i\cdot} \leq 1$ means that $\sum_{i=1}^{n} w_{ij} \leq 1$ for all $j\in[m]$, and similar convention is used for $\sum_{j=1}^{m} w_{\cdot j} \leq 1$.

The formulation as an MCF problem is obtained by adding two nodes to the graph, called *source* and *sink* (see Figure 1). The cost of each edge except those adjacent to *source* and *sink* is the corresponding cost from $\{d_{ij}\}_{i,j=1}^{n,m}$. The cost of the rest of the edges is equal to 0. The capacity that can be sent through each edge is equal to 1. The supply of *source* and *sink* are k and $-k$, respectively. The solution of (7) provides the weights $\{w_{ij}\}_{i,j=1}^{n,m}$, from which the matching $\widehat{\pi}_k^{\text{LSS}}$ can be recovered. Indeed, if $w_{ij} = 1$ then X_i and $X^{\#}_j$ are matched. Due to the last constraint from (7) the total matching size (number of w_{ij} that are equal to 1) will be k. In our experiments we used `SimpleMinCostFlow` solver from OR-tools library (Perron and Furnon, 2019).

5.2 Numerical Experiments on Synthetic Data

To support and illustrate our theoretical findings, we performed several experiments on synthetic data. For each experiment, we randomly generated two sets of $d = 100$-dimensional data points of size $n = m = 100$, from which only $k^* = 60$ were inliers (meaning that they have their corresponding match).

The data were generated according to the following procedure. We set $S^* = [k^*]$ and chose an additional parameter τ, used to control $\bar{\kappa}_{\text{all}}$ throughout the experiments. Then,

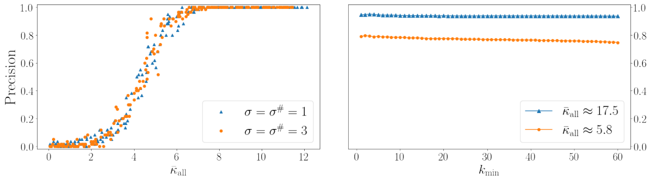

Figure 2: Left plot indicates how the matching precision of $\widehat{\pi}^{\text{LSS}}_{k^*}$ depends on $\bar{\kappa}_{\text{all}}$ for two different noise levels $\sigma = \sigma^{\#} = 1$ (blue triangles) and $\sigma = \sigma^{\#} = 3$ (orange circles). Right plot shows the consistency of accuracy if instead of true value k^* we use smaller value $k_{\min}$, for all $k_{\min} \in \{1, \dots, k^*\}$. Each point on the left plot corresponds to solving the optimization problem from (3) with different $\bar{\kappa}_{\text{all}}$. In the right plot each point corresponds to the averaged value carried over 200 independent trials.

θ and $\theta^{\#}$ were independently sampled from the Gaussian distribution with 0 mean and standard deviation τ. Additionally, for every $i \notin S^*$, we incremented every coordinate of θ_i by τ, *i.e.,* $\theta_i \leftarrow \theta_i + \tau\mathbf{1}$ and, for every $j \notin \text{Im}(\pi^*)$, we incremented every coordinate of $\theta^{\#}_j$ by 2τ. Sequences $\boldsymbol{X}$ and $\boldsymbol{X}^{\#}$ were generated according to (1) with $\pi^*(i) = i$ for $i \in S^*$.

For 200 independent trials, we computed $\bar{\kappa}_{\text{all}}$ and the precision of $\widehat{\pi}^{\text{LSS}}_{k}$ given by (3) for $k = k^*$. The precision is the number of correctly matched inliers divided by $k = k^*$. The results are displayed in the left plot of Figure 2. We see that when $\bar{\kappa}_{\text{all}}$ is large enough, the precision gets close to 1, *i.e.,* $\widehat{\pi}^{\text{LSS}}_{k^*} = \pi^*$. One can also observe that the form of the curves does not depend much on the noise levels $(\sigma, \sigma^{\#})$.

On the right plot we compute the matching precision for $k_{\min} \in [k^*]$. An important observation is that for all values $k \leq k^*$ if $\bar{\kappa}_{\text{all}}$ is large enough the computed matching makes no errors (blue curve), while for lower $\bar{\kappa}_{\text{all}}$ the precision decreases (orange curve). Nevertheless, the accuracy remains consistent over all values $k_{\min} \in [k^*]$, which is aligned with Theorem 2.

We then proceeded with showing that it is possible to adapt to unknown noise levels. The results for different values of $\bar{\sigma}^2_0$ are shown in Fig. 3. For small $\bar{\kappa}_{\text{all}}$, it is impossible to distinguish inliers from outliers, since they are all close to each other. In this scenario, the estimated value $\widehat{k}$ is 100, which means all the points are treated as inliers. Naturally, the estimated value of $\bar{\sigma}^2_0$ is overestimated, since there is a bias due to $\bar{\kappa}_{\text{all}}$. In contrast, for $\bar{\kappa}_{\text{all}}$ large enough, we recover the correct matching size k^* and hence the estimator of $\bar{\sigma}^2_0$ becomes accurate.

Additionally, we show that it is indeed possible to estimate the unknown matching size k^* accurately given $\bar{\kappa}_{\text{all}}$ is large enough. The estimation is carried out according to the scheme presented in Section 4.2, assuming that noise levels σ and $\sigma^{\#}$ are known. We estimate the value of k^* for two different settings noise levels $\sigma = \sigma^{\#} = 1$ and $\sigma = \sigma^{\#} = 2$. The corresponding plots are presented in Figure 4. Observe that the threshold after which the estimation becomes exact does not depend on noise levels.

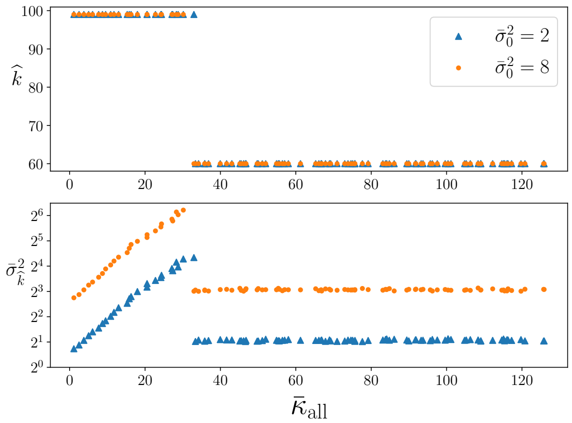

Figure 3: The above plot shows how the estimate of the unknown parameter k^* depends on the value of $\bar{\kappa}_{\text{all}}$ for two different noise levels $\sigma = \sigma^{\#} = 1$ (blue triangles) and $\sigma = \sigma^{\#} = 3$ (orange circles). The below plot shows how the estimator of $\bar{\sigma}^2_0$ computed by (6) depends on $\bar{\kappa}_{\text{all}}$.

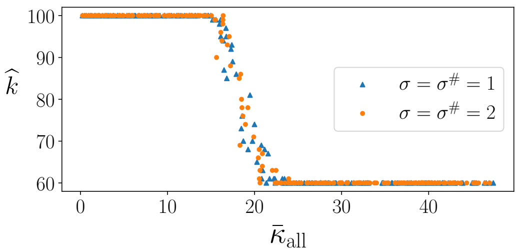

Figure 4: Impact of $\bar{\kappa}_{\text{all}}$ on estimation of k^* for two different known noise levels.

5.3 Choice of k^* for Real Data

Experiments on real data were conducted using IMC-PT 2020 dataset (Jin et al., 2020) consisting of 16 scenes with corresponding image sets and 3D point clouds. In this work we used images from "Reichstag" scene to illustrate how the procedure from Section 4.3 can be used to estimate the matching size. To construct the dataset we randomly chosen 1000 distinct image pairs of the same scene. Then scene point cloud is used to obtain pseudo-ground-truth matching between keypoints on different images of the same scene. We obtained the SIFT descriptors (Lowe, 2004) for all keypoints using Python OpenCV (Itseez, 2015) interface. The histogram shows that there is a spike close to the true value k^*.

Notice that in case of images we first have no information

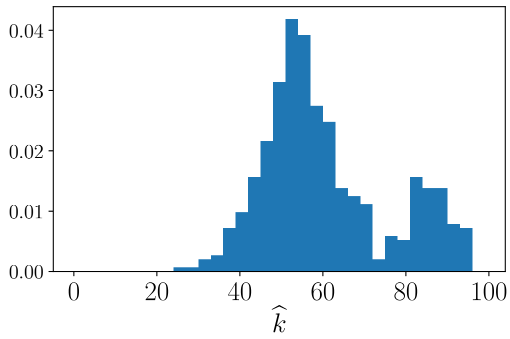

Figure 5: Histogram of $\widehat{k}$ with $\alpha = 0.01$, $d = 128$ (SIFT descriptors), $n = m = 100$ and $k^* = 60$.

about the noise levels σ and $\sigma^\#$ and moreover, the noise levels might not be homoscedastic. Nevertheless, we perform the procedure described in Section 4.3 and show that even in this situation this procedure can be used to estimate k^*. There are several aspects where we can speed-up this procedure. First, using Greedy algorithm to match the feature vectors (as it is done in OpenCV) will allow us at each iteration compute only one distance instead of solving MCF from scratch. Secondly, the stepsize from the second step of the procedure from Section 4.3 can be increased, *i.e.,* by considering the difference $\widehat{\Phi}(k+10) - \widehat{\Phi}(k)$, then with a proper adjustment to the threshold we will obtain a speedup of 10 times.

5.4 Experiments on Biomedical Data

First we replicate the experiment described in Section 5.1 of (Chen et al., 2022b) using our estimator $\widehat{\pi}_k^{\text{LSS}}$ from (3). Here we recover the matching between two datasets collected from human pancreatic islets using two different technologies (CEL-seq2 (Hashimshony et al., 2016) and Smart-seq2 (Picelli et al., 2013)). Both datasets can be found in SeuratData R package (Hao et al., 2021). CEL-seq2 data contains measurements on 34363 RNAs in 2285 cells, Smart-seq2 data contains measurements on 34363 RNAs in 2394 cells. After applying standard pre-processing procedures using Python package scanpy (Wolf et al., 2018), we select 5000 most active RNAs for each dataset. 2808 distinct RNAs appeared in both datasets' top-5000, so we leave out the rest obtaining two datasets of sizes 2808 x 2285 and 2808 x 2394 respectively. Each cell in both datasets has a human annotated type (out of 13 types). As done in Chen et al. (2022b) we also randomly downsample cells to get identical number of cells-per-type in both datasets, eventually getting two datasets of size 2808 x 1935. We proceed to match cells in two datasets using $\widehat{\pi}_k^{\text{LSS}}$. As we don't have the ground truth matching, we calculate the accuracy on cell-type level, meaning that a single match is considered correct, if it matches two cells of the same type. We manage to achieve 97.88% cell-type level accuracy, which is almost identical with th result of (Chen et al., 2022b) (97.93%) without performing any dimensionality reduction techniques and using much simpler approach. Resulting confusion matrix is shown in Figure 6.

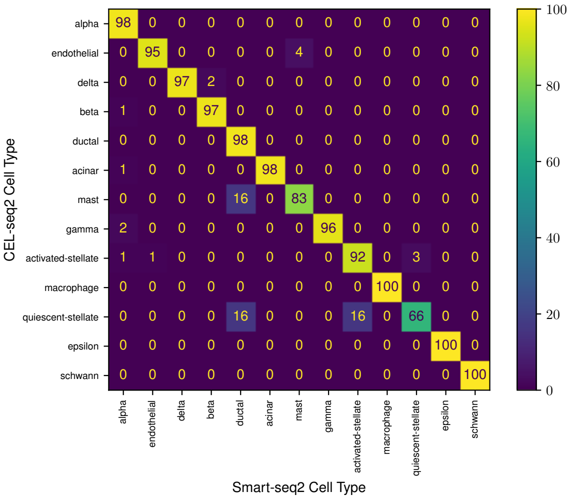

Figure 6: Cell-Type level confusion matrix of $\widehat{\pi}_k^{\text{LSS}}$ on Cel-seq2 and Smart-seq2 datasets.

6 CONCLUSION AND DISCUSSION

We have analyzed the problem of matching map recovery between two sets of feature-vectors, when the number k^* of true matches is unknown. We focused on two practically relevant settings of this problem. Assuming a lower bound k on k^* is available, we proved— under the weakest possible condition on the signal-to-noise ratio—that that the k-LSS procedure makes no mistake with high probability. More precisely, k-LSS provides an estimated map the support of which is included in the support of the true matching map and the values of these two maps coincide on this subset. More importantly, we proposed a procedure for estimating the unknown matching size k^* and proved that it finds the correct value of k^* and the true matching map π^* with high probability. Once again, this holds under the minimal assumption that the signal-to-noise ratio exceeds the minimax rate of separation.

Interestingly, our results demonstrate that the minimax rate of separation does not depend on k^* and, more surprisingly, that the absence of the knowledge of k^* has no impact on the minimax rate. These rates are attained by computationally tractable algorithms solving the minimum cost flow problem.

Our results are limited to Gaussian noise and to noise levels that are equal across observations. Furthermore, we only tackled the recovery problem, leaving the problem of estimation to future work.

References

Ravindra K. Ahuja, Andrew V. Goldberg, James B. Orlin, and Robert Endre Tarjan. Finding minimum-cost flows by double scaling. *Math. Program.*, 53:243–266, 1992. doi: 10.1007/BF01585705. URL https://doi.org/10.1007/BF01585705.

J. M. Azaïs and Y. de Castro. Multiple testing and variable selection along least angle regression's path, 2020.

Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Gilles Blanchard, Alexandra Carpentier, and Maurilio Gutzeit. Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electron. J. Stat.*, 12(2): 3713–3735, 2018. ISSN 1935-7524/e.

M. V. Burnashev. On the minimax detection of an inaccurately known signal in a white Gaussian noise background. *Theory Probab. Appl.*, 24:107–119, 1979. ISSN 0040-585X; 1095-7219/e.

T Tony Cai and Rong Ma. Matrix reordering for noisy disordered matrices: Optimality and computationally efficient algorithms. *arXiv preprint arXiv:2201.06438*, 2022.

Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 778–792, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15561-1.

Jian Chen, Jie Tian, Noah Lee, Jian Zheng, R. Theodore Smith, and Andrew F. Laine. A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Transactions on Biomedical Engineering*, 57(7):1707–1718, 2010. doi: 10.1109/TBME.2010.2042169.

Li Chen, Rasmus Kyng, Yang P Liu, Richard Peng, Maximilian Probst Gutenberg, and Sushant Sachdeva. Maximum flow and minimum-cost flow in almost-linear time. *arXiv preprint arXiv:2203.00671*, 2022a.

Shuxiao Chen, Sizun Jiang, Zongming Ma, Garry P Nolan, and Bokai Zhu. One-way matching of datasets with low rank signals. *arXiv preprint arXiv:2204.13858*, 2022b.

M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, L. Zdeborová, and Boris I. Shraiman. Inference in particle tracking experiments by passing messages between images. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17):7663–7668, 2010. ISSN 00278424. URL http://www.jstor.org/stable/25665416.

Olivier Collier. Minimax hypothesis testing for curve registration. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pages 236–245. JMLR.org, 2012.

Olivier Collier and Arnak S. Dalalyan. Permutation estimation and minimax rates of identifiability. *Journal of Machine Learning Research*, W & CP 31 (AI-STATS 2013):10–19, 2013.

Olivier Collier and Arnak S Dalalyan. Minimax rates in permutation estimation for feature matching. *The Journal of Machine Learning Research*, 17(1):162–192, 2016.

Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Stat.*, 40(5):2667–2696, 2012. ISSN 0090-5364; 2168-8966/e.

Laëtitia Comminges and Arnak S. Dalalyan. Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression. *Electron. J. Stat.*, 7:146–190, 2013. ISSN 1935-7524/e.

Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1): 623–653, 2019.

Delbert R Fulkerson. An out-of-kilter method for minimal-cost flow problems. *Journal of the Society for Industrial and Applied Mathematics*, 9(1):18–27, 1961.

Zvi Galil and Éva Tardos. An $o(n^2(m + n\log n)\log n)$ min-cost flow algorithm. *J. ACM*, 35(2):374–386, 1988. ISSN 0004-5411.

Tigran Galstyan, Arshak Minasyan, and Arnak Dalalyan. Optimal detection of the feature matching map in presence of noise and outliers. 2021. doi: 10.48550/ARXIV.2106.07044. URL https://arxiv.org/abs/2106.07044.

Chao Gao and Anderson Y Zhang. Iterative algorithm for discrete structure recovery. *arXiv preprint arXiv:1911.01018*, 2019.

Christophe Giraud, Yann Issartel, and Nicolas Verzelen. Localization in 1d non-parametric latent space models from pairwise affinities. *arXiv preprint arXiv:2108.03098*, 2021.

Andrew V. Goldberg and Robert E. Tarjan. Finding minimum-cost circulations by canceling negative cycles. *J. ACM*, 36(4):873–886, 1989.

A.V. Goldberg, H. Kaplan, S. Hed, and Robert Tarjan. Minimum cost flows in graphs with unit capacities. *Leibniz International Proceedings in Informatics, LIPIcs*, 30:406–419, 02 2015. doi: 10.4230/LIPIcs.STACS.2015.406.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael

Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021. doi: 10.1016/j.cell.2021.04.048. URL https://doi.org/10.1016/j.cell.2021.04.048.

Ben Harwood and Tom Drummond. Fanng: Fast approximate nearest neighbour graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5713–5722, 2016. doi: 10.1109/CVPR.2016.616.

Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron De Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, et al. Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome biology*, 17(1):1–7, 2016.

Yu. I. Ingster. Minimax nonparametric detection of signals in white Gaussian noise. *Probl. Inf. Transm.*, 18:130–140, 1982.

Yu. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003.

Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015.

Zhansheng Jiang, Lingxi Xie, Xiaotie Deng, Weiwei Xu, and Jingdong Wang. Fast nearest neighbor search in the hamming space. In Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu, editors, *MultiMedia Modeling*, pages 325–336, Cham, 2016. Springer International Publishing. ISBN 978-3-319-27671-7.

Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 2020.

Anatoli Juditsky and Arkadin Nemirovski. *Statistical inference via convex optimization*. Princeton, NJ: Princeton University Press, 2020. ISBN 978-0-691-19729-6/hbk; 978-0-691-20031-6/ebook.

Harold W. Kuhn. A tale of three eras: The discovery and rediscovery of the hungarian method. *Eur. J. Oper. Res.*, 219(3):641–651, 2012.

Dmitriy Kunisky and Jonathan Niles-Weed. Strong recovery of geometric planted matchings. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 834–876. SIAM, 2022.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Rong Ma, T Tony Cai, and Hongzhe Li. Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, pages 1–15, 2020.

Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory*, pages 821–847. PMLR, 2018.

Cheng Mao, Ashwin Pananjady, and Martin J Wainwright. Towards optimal estimation of bivariate isotonic matrices with unknown permutations. *Annals of Statistics*, 48(6): 3183–3205, 2020.

Mohamed Ndaoud and Alexandre B. Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory*, 66(4):2517–2532, 2020. ISSN 0018-9448.

James B. Orlin. A faster strongly polynomial minimum cost flow algorithm. *Oper. Res.*, 41(2):338–350, 1993.

James B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '96, page 474–481, USA, 1996. Society for Industrial and Applied Mathematics. ISBN 0898713668.

James B. Orlin, Serge A. Plotkin, and Éva Tardos. Polynomial dual network simplex algorithms. *Math. Program.*, 60(1–3):255–276, 1993.

Ashwin Pananjady and Richard J Samworth. Isotonic regression with unknown permutations: Statistics, computation, and adaptation. *arXiv preprint arXiv:2009.02609*, 2020.

Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64(5):3286–3300, 2017.

Laurent Perron and Vincent Furnon. Or-tools, 2019. URL https://developers.google.com/optimization/.

Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

Aaditya Ramdas, David Isenberg, Aarti Singh, and Larry A. Wasserman. Minimax lower bounds for linear independence testing. In *IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016*, pages 965–969. IEEE, 2016.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf.

In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. doi: 10.1109/ICCV.2011.6126544.

Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 67(6):4162–4184, 2021. doi: 10.1109/TIT.2020.3045613.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, 13(1):1–36, 2019.

Yurun Tian, Vassileios Balntas, Tony Ng, Axel Barroso-Laguna, Yiannis Demiris, and Krystian Mikolajczyk. D2d: Keypoint extraction with describe to detect approach. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yoloh. Random graph matching in geometric models: the case of complete graphs. *arXiv preprint arXiv:2202.10662*, 2022.

Ke Wang, Ningyu Zhu, Yao Cheng, Ruifeng Li, Tianxiang Zhou, and Xuexiong Long. Fast feature matching based on r -nearest k -means searching. *CAAI Transactions on Intelligence Technology*, 3(4):198–207, 2018. doi: https://doi.org/10.1049/trit.2018.1041. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/trit.2018.1041.

Yuting Wei, Martin J. Wainwright, and Adityanand Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *The Annals of Statistics*, 47(2):994 – 1024, 2019.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax testing of identity to a reference ergodic markov chain. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 191–201. PMLR, 2020.

Xin Xing, Meimei Liu, Ping Ma, and Wenxuan Zhong. Minimax nonparametric parallelism test. *Journal of Machine Learning Research*, 21(94):1–47, 2020. URL http://jmlr.org/papers/v21/19-800.html.