# Matching Map Recovery with an Unknown Number of Outliers

**Arshak Minasyan***
CREST, ENSAE, IP Paris
5 av. Henry Le Chatelier
91764 Palaiseau
arshak.minasyan@ensae.fr

**Tigran Galstyan**
RAU, YerevaNN
20 Charents street
0025 Yerevan
tigran@yerevann.com

**Sona Hunanyan**
Yerevan State University
1 Alek Manukyan
0025 Yerevan
hunan.sona@gmail.com

**Arnak Dalalyan**
CREST, ENSAE, IP Paris
5 av. Henry Le Chatelier
91764 Palaiseau
arnak.dalalyan@ensae.fr

## 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 $5(d \log(4nm/\alpha))^{1/4}$, 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 matching 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 a theoretical understanding of 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.*,

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

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. isotropic Gaussian random vectors. Formally,

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

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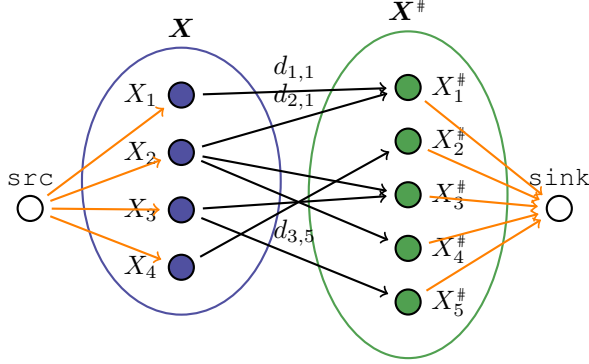
Figure 1: Matching as a Minimum Cost Flow (MCF) problem. The idea is to augment the graph with two nodes, *source* and *sink*, and $n + m$ edges. The capacities of orange edges should be set to 1, while the cost should be set to 0. Setting the total flow sent through the graph to $k$, the solution of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

outlier-free case with equal sizes of sequences $\boldsymbol{X}$ and $\boldsymbol{X}^\sharp$ (*i.e.*, $m = n$ and $S^* = [n]$), whereas Galstyan et al. [2022] 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 Wang et al. [2022], Chen et al. [2022b], Kunisky and Niles-Weed [2022]. These papers obtain minimax-optimal separation rates and, in most cases, despite the discrete nature of the matching problem, provide computationally tractable procedures to achieve 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 the quantity

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

which is the signal-to-noise ratio of the difference $X_i - X_j^\sharp$ 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

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

is bounded away from zero. A recovery procedure $\widehat{\pi}$ is considered to be good, if the threshold $\lambda$ such that $\widehat{\pi}$ recovers $\pi^*$ 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 $\pi^*$ with probability $1 - \alpha$ for $\lambda = 4\{(d\log(4n^2/\alpha))^{1/4} \vee (8\log(4n^2/\alpha))^{1/2}\}$. 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 $d \gtrsim \log n$.

Let us set

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

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, based on maximizing profile likelihood among matching maps between two sets of size $k$, can be efficiently computed using the minimum cost flow problem. We denote the matching obtained using $k$-LSS by $\widehat{\pi}_k^{\text{LSS}}$.

- 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 $\widehat{\pi}_k^{\text{LSS}}$ makes no mistake with probability $1 - \alpha$.

- We design a data-driven model selection algorithm that adaptively chooses $\widehat{k}$ such that with probability $1 - \alpha$, we have $\widehat{k} = k^*$ and $\widehat{\pi}_{\widehat{k}}^{\text{LSS}} = \pi^*$ as soon as $\bar{\kappa}_{\text{all}} \geq (5/4)\lambda_{n,m,d,\alpha}$.

The last item above implies that our data-driven algorithm $\widehat{\pi}_{\widehat{k}}^{\text{LSS}}$ 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    RELATED WORK

In statistical hypothesis 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 [Xing et al., 2020, Wolfer and Kontorovich, 2020, Blanchard et al., 2018, Ramdas et al., 2016, Wei et al., 2019, Collier, 2012]. Beyond the classical setting of two hypotheses, it can also be applied to multiple testing frameworks, for instance, variable selection [Ndaoud and Tsybakov, 2020, Azaïs and de Castro, 2020, Comminges and Dalalyan, 2012, 2013] or the matching problem considered here.

In computer vision, 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. Malkov and Yashunin [2020], Wang et al. [2018], Harwood and Drummond [2016], Jiang et al. [2016]). Another direction is to improve the matching quality by considering alternative local descriptors [Rublee et al., 2011, Chen et al., 2010, Calonder et al., 2010] for given keypoints. The choice of keypoints is considered in Tian et al. [2020], Bai 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 [Fulkerson, 1961, Ahuja et al., 1992]. Since then many algorithms have been proposed for solving minimum cost flow problems in strongly polynomial time [Orlin et al., 1993, Orlin, 1993, 1996, Goldberg and Tarjan, 1989, Galil and Tardos, 1988] with the fastest runtime of around $\mathcal{O}(nm)$. Recent advances for solving MCF 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 [Flammarion et al., 2019, Giraud et al., 2021, Cai and Ma, 2022], noisy sorting [Mao et al., 2018], regression with shuffled data [Pananjady et al., 2017, Slawski and Ben-David, 2019], isotonic regression and matrices [Mao et al., 2020, Pananjady and Samworth, 2020, Ma et al., 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 $\pi^*$, 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

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

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

$$\widehat{\pi}_k^{\mathrm{LSS}} \in \arg\min_{\pi \in \mathcal{P}_k} \sum_{i \in S_\pi} \|X_i - X_{\pi(i)}^\sharp\|_2^2, \tag{3}$$

where $S_\pi$ denotes the support of function $\pi$. In the particular case of $k^* = n$, the optimization above is conducted over all the injective mappings from $[n]$ to $[m]$. This coincides with the LSS method from [Galstyan et al., 2022].

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

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

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

In the sequel, we denote by $(\widehat{k}, \bar{\sigma}_{\widehat{k}}, \widehat{\pi}_{\widehat{k}}^{\mathrm{LSS}})$ the output of this procedure. Notice that we start with the value of $k = k_{\min}$, 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 $\gamma$ and $\lambda$, as stated in the next theorem, the described procedure outputs the correct values of $k^*$ and $\pi^*$ with high probability.

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

Since the condition on the separation distance $\bar{\kappa}_{\mathrm{all}}$ 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 challenging than that of the known $k^*$.

In the sequel, without much loss of generality, we assume that the sizes of $\boldsymbol{X}$ and $\boldsymbol{X}^\sharp$ 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 $\boldsymbol{X}$ obtaining equal size sets $\boldsymbol{X}^+$ and $\boldsymbol{X}^\sharp$.

Notice that in the optimization problem (3) the domain of $\pi$ is a finite set of injective functions. For a given value of $k$, the number of such functions is $k!\binom{n}{k}^2$ making thus an exhaustive search computationally infeasible. Instead, we show in Section 5 that the optimization problem formulated in (3) can indeed be solved efficiently with complexity $\widetilde{\mathcal{O}}(\sqrt{k}\,n^2)$, where the notation $\widetilde{\mathcal{O}}$ hides polylogarithmic factors, *i.e.,* up to polylogarithmic factors, the computational cost is of order $\sqrt{k}n^2$.

# 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 Appendix A.

## 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 $k$-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* $\widehat{S} = \mathrm{supp}(\widehat{\pi})$ *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 $\bar{\kappa}_{\mathrm{all}} \geq \lambda_{n,d,\alpha}$ then, with probability at least $1-\alpha$, the support of the estimator $\widehat{\pi}$ is included in $S^*$ and $\widehat{\pi}$ coincides with $\pi^*$ 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

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

are standard Gaussian and define the following quantities

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

For the ease of notation, for any matching map $\pi$ 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 $\pi$ be any matching map that can not be obtained as a restriction of $\pi^*$ 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 $\pi_0$ be the restriction of $\pi^*$ to $S_0$. On the event $\Omega_0 = \{8\zeta_1 \leq \bar{\kappa}_{\mathrm{all}}; 4\sqrt{d}\,\zeta_2 \leq \bar{\kappa}_{\mathrm{all}}^2\}$, we have*

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

Let $\pi$ be any matching map from $\mathcal{P}_k$ that is not a restriction of $\pi^*$. Since $|S_\pi| = k \leq k^*$, there exists necessarily a $\pi_0$ as in Lemma 1 such that $|S_0| = |S_\pi|$. For this $\pi_0$, we have $L(\pi) - L(\pi_0) \geq (1/4)\bar{\kappa}_{\mathrm{all}}^2 > 0$. This implies that $\pi$ cannot be a minimizer of $L(\cdot)$ over $\mathcal{P}_k$. As a consequence, on $\Omega_0$, any minimizer of $L(\cdot)$ over $\mathcal{P}_k$ is a restriction of $\pi^*$. Therefore, on $\Omega_0$, 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 $\Omega_{0,x} = \{8\zeta_1 \leq x\} \cap \{4\sqrt{d}\zeta_2 \leq x^2\}$ with $\zeta_1, \zeta_2$ defined as in* (5). *Then, for every $x > 0$, $\mathbf{P}(\Omega_{0,x}^{\complement})$ is upper bounded by*

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

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

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

This system is equivalent to

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

Therefore, if the signal-to-noise ratio satisfies

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

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

## 4.2 Matching Map Recovery for Unknown $k^*$

If no information on $k^*$ is available, and the goal is to recover the entire mapping $\pi^*$, 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 those using a suitable criterion. To define the selection criterion proposed in this paper, let us remark that

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

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 $\chi^2$ 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 order $\bar{\kappa}_{\text{all}}^2$. Therefore, if $\bar{\kappa}_{\text{all}}^2$ is larger than the deviations of the $\chi_d^2$ 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]

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

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 $\mathbf{P}\big(\widehat{k} = k^* \text{ and } \widehat{\pi}_{\widehat{k}} = \pi^*\big) \geq 1 - \alpha$. 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 $\pi^*$, 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 levels are known. It is also worth mentioning that our procedure requires only $\sigma_0^2 = \sigma^2 + \sigma^{\#2}$, not $\sigma$ and $\sigma^{\#}$ separately.

*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 $\sigma_0^2$ it is more convenient to work with the normalized version of test statistics $\widehat{\Phi}(\cdot)$, 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, $\Omega_0 = \{8\zeta_1 \leq \bar{\kappa}_{\text{all}}; 4\sqrt{d}\,\zeta_2 \leq \bar{\kappa}_{\text{all}}^2\}$, we have $\widehat{L}(k^*+1) - \widehat{L}(k^*) \geq d + \bar{\kappa}_{\text{all}}^2/4$.*

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

Lemma 2 implies that the probability of the event $\Omega_1 = \{8\zeta_1 \leq \lambda_{n,d,\alpha}; 4\sqrt{d}\,\zeta_2 \leq \lambda_{n,d,\alpha}^2\}$ is at least $1 - \alpha$. Since $\Omega_1$ is included in $\Omega_0$, in view of Lemma 4, on $\Omega_1$ we have $\widehat{L}(k+1) - \widehat{L}(k) \leq d + \lambda_{n,d,\alpha}^2/4$ for any $k < k^*$. On the other hand, in view of Lemma 3, on the same event we have $\widehat{L}(k^*+1) - \widehat{L}(k^*) \geq d + \bar{\kappa}_{\text{all}}^2/4 > d + \lambda_{n,d,\alpha}^2/4$. This implies that $\widehat{k} = k^*$ and, therefore, $\widehat{\pi}_{\widehat{k}} = \widehat{\pi}_{k^*}$. Thanks to Theorem 2, on the same event $\Omega_1$, we have $\widehat{\pi}_{k^*} = \pi^*$. $\square$

---

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

## 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 $\sigma$ and $\sigma^{\#}$. Notice that we do not need to estimate parameters $\sigma, \sigma^{\#}$ separately, it is sufficient to estimate only their squared sum, which is denoted by $\sigma_0^2$. In the definition of $\widehat{k}$, we use the value of $\sigma_0^2$ in the threshold for $\widehat{\Phi}(k+1) - \widehat{\Phi}(k)$. When both $k^*$ and $\sigma_0^2$ are unknown, we first estimate $\sigma_0^2$ and then plug it in the selection criterion of $k$.

Thus, we define "candidate" estimators of $\sigma_0^2$

$$\big\{\bar{\sigma}_k^2 = \frac{\widehat{\Phi}(k)}{kd}, \ k \in [n]\big\}. \tag{6}$$

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 $\chi^2$ distribution with $d$ degrees of freedom. Hence, after division by $kd$, we obtain an estimator of $\sigma_0^2$. However, from the perspective of testing the values of $k$, we need to slightly overestimate the noise variance. This is done through the multiplication by the inflation factor $1/(1-\gamma)$.

We are now ready to proceed with the proof of our main result stated in Theorem 1.

*Proof of Theorem 1.* We will provide the proof only in the high dimensional setting, that is we assume throughout the proof that $d \geq 800 \log(2n/\sqrt{\alpha})$. First, we show that for every $k < k^*$ the condition from $\widehat{\Phi}(k+1) - \widehat{\Phi}(k) \leq \frac{d+\lambda}{1-\gamma}\bar{\sigma}_k^2$ 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 $\Omega_0 = \{8\zeta_1 \leq \lambda; 4\sqrt{d}\zeta_2 \leq \lambda^2\}$, we have

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

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

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

Now we show that for $k = k^*$ the relative difference of function $\widehat{\Phi}(\cdot)$ at points $k^*+1$ and $k^*$ is large enough. Indeed,

we have

$$\frac{\widehat{\Phi}(k^*+1) - \widehat{\Phi}(k^*)}{\widehat{\Phi}(k^*)} \geq \frac{\widehat{\Phi}(k^*+1) - \widehat{\Phi}(k^*)}{\sum_{i \in S^*} \|X_i - X_{\pi^*(i)}^{\#}\|_2^2}$$

$$\geq \frac{\min_{i \neq \pi^*(j)} \|X_i - X_j^{\#}\|_2^2}{\sigma_0^2 \sum_{i \in S^*} \|\eta_{i,\pi^*(i)}\|_2^2}$$

$$\geq \frac{\bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} + d - \sqrt{d}\zeta_2}{k(d + \sqrt{d}\zeta_2)},$$

where the first inequality follows from the definition of function $\widehat{\Phi}(\cdot)$, while second and third inequalities are consequences of the definitions introduced earlier in this section. Then, on the event $\Omega_0$ we bound the quantities $\zeta_1$ and $\zeta_2$ 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

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

$$\geq \frac{d + \lambda^2/4}{kd(1 - \lambda^2/4d)}.$$

Thus, we have shown that on the event $\Omega_0$ 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 $\Omega_0$ we have $\widehat{\pi}_{k^*} = \pi^*$. Finally, using Lemma 2, we get that the event $\Omega_0$ occurs with probability at least $1 - \alpha$. Therefore, the desired result follows. $\qquad\square$

### 4.4 Lower Bounds

Theorem 2 and Theorem 3 imply that the minimax rate of separation in the problem of recovering $\pi^*$ is at most of the order of $\lambda_{n,m,d,\alpha}$ defined in (2). 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., 2022, Theorem 5] for the general rectangular case $m \geq n$. Our model is more general[2] than those of these two references, the same lower bound applies to our model. 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 separation rate.

## 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 an imperfect matching problem and, to the best

---

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

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 the results of numerical experiments conducted on both synthetic and real data and highlight their relation to the aforementioned theorems stated and proved in Sections 3 and 4, respectively. Our reproducible codes are provided in the supplementary material.

### 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

$$\text{minimize} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} w_{ij} \qquad (7)$$

subject to $\boldsymbol{w} = (w_{ij})_{(i,j) \in [n] \times [m]} \in [0,1]^{n \times m}$ satisfying

$$\sum_{i=1}^{n} w_{i\cdot} \leq 1, \quad \sum_{j=1}^{m} w_{\cdot j} \leq 1, \quad \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} = k, \qquad (8)$$

known as the 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 auxiliary nodes to the graph, called *source* and *sink* (see Fig. 1). We are interested in the flow of the minimal cost, where the cost of each edge except those adjacent to *source* and *sink* is assigned from the distance matrix $\{d_{ij}\}_{i,j=1}^{n,m}$. 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) given the constraints (8) 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. The last constraint in (8) implies that the matching size (number of $w_{ij}$ that are equal to 1) will be $k$. Though the algorithm provided in [Goldberg et al., 2015] has the fastest known asymptotic complexity, the implementation of their algorithm is out of the scope of this paper. Therefore, in our experiments, we used `SimpleMinCostFlow` solver from OR-tools library [Perron and Furnon, 2022].

### 5.2 Numerical Experiments on Synthetic Data

In this part, we conducted several experiments on synthetic data to support our theoretical findings. In these experiments, we constructed two sets of sizes $n = m = 100$ consisting of $d = 100$-dimensional data points. The underlying matching size $k^*$ was set to 60. In other words, in each point cloud, we had 60 *inliers* and 40 *outliers*. We also fixed the confidence level at $1\%$, *i.e.*, $\alpha = 0.01$. The procedure for generating synthetic data was as follows. We set $S^* = [k^*]$ and chose an additional parameter $\tau$ used to control $\bar{\kappa}_{\text{all}}$ throughout the experiments. Then, each coordinate

of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{\#}$ was independently sampled from a Gaussian distribution with $0$ mean and standard deviation $\tau$. Additionally, for every $i \notin S^*$, we incremented every coordinate of $\theta_i$ by $\tau$, *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 $2\tau$. This allows the outliers of each point cloud to be sufficiently far from each other, hence a pair of outliers is less likely to be confused as a pair of inliers. Notice also that such a choice of generating outliers fits within the conditions of Theorem 1. Finally, the sequences $\boldsymbol{X}$ and $\boldsymbol{X}^{\#}$ were generated according to (1) with $\pi^*(i) = i$, for all $i \in S^*$. We measured the performance of our estimator $\widehat{\pi}_{k^*}^{\text{LSS}}$ by its precision, which is the number of correctly matched inliers divided by $k = k^*$.
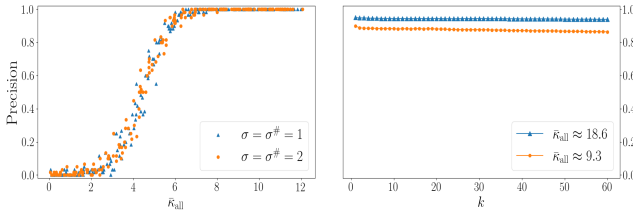


Figure 2: Left: the dependence of the matching precision of $\widehat{\pi}_{k^*}^{\text{LSS}}$ on $\bar{\kappa}_{\text{all}}$ for $\sigma = \sigma^{\#} = 1$ (blue triangles) and $\sigma = \sigma^{\#} = 2$ (orange circles). Right: the accuracy of subset recovery for a known $k \leq k^*$. In both plots, for $\bar{\kappa}_{\text{all}}$ large enough we observe that the estimated matching $\widehat{\pi}_{k^*}^{\text{LSS}}$ indeed coincides with $\pi^*$ or yields a subset of it (right plot).

Let us now comment on the results obtained in Fig. 2. On the left plot, given that $\bar{\kappa}_{\text{all}}$ is large enough we observe that $\widehat{\pi}_{k^*}^{\text{LSS}} = \pi^*$ holds. Another interesting feature that can be inferred from Fig. 2 is that the precision is independent of the noise levels $\sigma, \sigma^{\#}$, which is also consistent with Theorem 1 through the definition of $\lambda_{n,m,d,\alpha}$ which is independent of $\sigma, \sigma^{\#}$. The right plot supports the findings of Theorem 2, that is, if one computes $\widehat{\pi}_k^{\text{LSS}}$ for some $k \leq k^*$ it guarantees that the support of $\widehat{\pi}_k^{\text{LSS}}$ will be included in the support of $\pi^*$, and moreover on this support the values of $\widehat{\pi}_k^{\text{LSS}}$ and $\pi^*$ coincide. Moreover, plugging in the values of this experiment into (4) we have $\lambda_{n,d,\alpha} \approx 44$, which means that the results proved in Theorems 1 and 2 hold whenever $\bar{\kappa}_{\text{all}} \geq \frac{5}{4}\lambda_{n,d,\alpha}$ and $\bar{\kappa}_{\text{all}} \geq \lambda_{n,d,\alpha}$, respectively. The results reported in Fig. 2 are the average over 200 independent trials of the same experiment.

In the previous experiment, we focused on the case of known $k^*$ and noise levels. We first consider the case of known noise levels and apply the estimator proposed in Section 4.2. In Fig. 3 we illustrate how the unknown value of $k^*$ is estimated for two different settings of known noise levels. Namely, given that $\bar{\kappa}_{\text{all}}$ is "small" ($\bar{\kappa}_{\text{all}} < 15$) all the pairs are considered as inliers and $\widehat{k} = 100$ as opposed to $k^* = 60$. However, given that the value of $\bar{\kappa}_{\text{all}}$ is large enough we see that the algorithm starts to differentiate inliers from outliers, hence we have $\widehat{k} = 60$ starting from $\bar{\kappa}_{\text{all}} \approx 22$, which
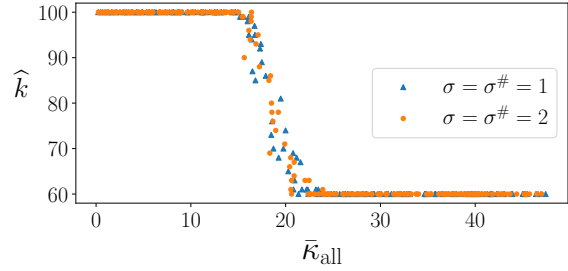


Figure 3: Dependence of the estimate of $k^*$ on $\bar{\kappa}_{\text{all}}$ when the noise levels $\sigma$ and $\sigma^{\#}$ are known. Observe that given that the value of $\bar{\kappa}_{\text{all}}$ is large enough our procedure recovers the true value of $k^* = 60$ while failing to identify the outliers when the signal-to-noise ratio is small and hence estimating $\widehat{k} = 100$.

confirms the result of Theorem 3. Recall that the result of Theorem 3 holds whenever $\bar{\kappa}_{\text{all}} \geq \lambda_{n,d,\alpha} \approx 44$. Moreover, similarly to the left plot of Fig. 2, the threshold after which the estimation becomes exact does not depend on noise levels. The results reported in Fig. 3 are the average over 200 independent trials of the same experiment.

Finally, we consider the case when no additional information is available neither about $k^*$ nor about $\sigma_0^2$, which corresponds to the most generic setting of Theorem 1. Recall the sequential procedure from Section 3 of estimating the triplet $(k^*, \sigma_0^2, \pi^*)$. In Fig. 4, we plot the dependencies of estimates of $k^*$ and $\sigma_0^2$ on $\bar{\kappa}_{\text{all}}$. There are several key features that are worth noticing. First, for small values of $\bar{\kappa}_{\text{all}}$ it is impossible to distinguish inliers from outliers, hence all points are treated as inliers and $\widehat{k} = 100$. As a consequence $\sigma_0^2$ is overestimated. Second, for large enough values of $\bar{\kappa}_{\text{all}}$ both $k^*$ and $\sigma_0^2$ are accurately estimated, therefore the precision of $\widehat{\pi}_{\widehat{k}}^{\text{LSS}}$ is (close to) 1. To make a link with Theorem 1 we also include the theoretical value of threshold $\lambda_{n,m,d,\alpha}$ in the plots.

There are several takeaways that we would like to highlight. In the theoretical part, we proved that provided $\bar{\kappa}_{\text{all}}$ is large than the given threshold, it is possible to recover $\pi^*$ and to estimate $k^*$. The threshold provided in theorems is rate optimal, however, it is not sharp. We observe that for different tasks the sharp threshold might differ up to a constant factor. For instance, in the case of known $k^*$ and $\sigma_0^2$ we see that the threshold seems close to 7 (see Fig. 2), while it is around 32 in the more difficult setting of unknown $k^*$ and $\sigma_0^2$, as it is shown in Fig. 4. The setting of Fig. 3 is somewhat in between, since $\sigma_0^2$ is assumed to be known but $k^*$ is unknown, hence the "change point" occurs around 22, which is between 7 and 32. In practice, one could replace the theoretical quantiles of $\chi_d^2$ distribution with the empirical counterparts of the squared distances between matching pairs $X_i$ and $X_{\widehat{\pi}(i)}$ for a given estimator $\widehat{\pi}$.
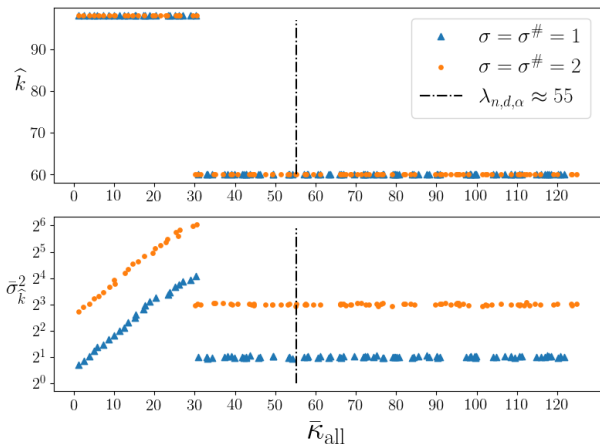
Figure 4: Two settings were considered: $\sigma, \sigma^\# = 1$ (blue triangles) and $\sigma, \sigma^\# = 2$ (orange circles). The top plot shows the dependence of the estimate of $k^*$ on $\bar{\kappa}_{\text{all}}$, while the bottom one—the dependence of the estimate (6) of $\sigma_0^2$ on $\bar{\kappa}_{\text{all}}$.

## 5.3 Estimation of $k^*$ for Real Data

In what follows we carry out experiments on real data. In this experiment, we only focus on the estimation of the matching size in the setting of matching the keypoints (SIFT descriptors [Lowe, 2004]) of two different images of the same scene. The goal is to be able to match the same keypoints considering the noise present in the images, and not match the keypoints that are present in only one of the two images. We refer to Appendix B for more description and additional experiments in the keypoint matching problem.

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. We used images from the "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 chose 1000 distinct image pairs of the same scene. Then, a scene point cloud is used to obtain pseudo-ground-truth matching between keypoints on different images of the same scene. Using these pseudo-ground-truth matching keypoints we then choose $k^* = 60$ keypoints that are present in both images and add to each of them 40 outlier keypoints. Hence, we have $n = m = 100$ with $k^* = 60$ and $d = 128$-dimensional SIFT descriptors. Notice that in the case of images, no information on $\sigma$ and $\sigma^\#$ is available. Moreover, noise levels might not be homoscedastic. In Fig. 5, we present the histogram of the estimated value $\hat{k}$ and observe that the procedure described in Section 4.3 provides a reasonably good estimate of $k^* = 60$.

Recall that at each iteration of estimating $k^*$ we solve an MCF problem, which can be computationally costly given the sample sizes are large. However, there are several aspects where we can speed up this procedure. First, using
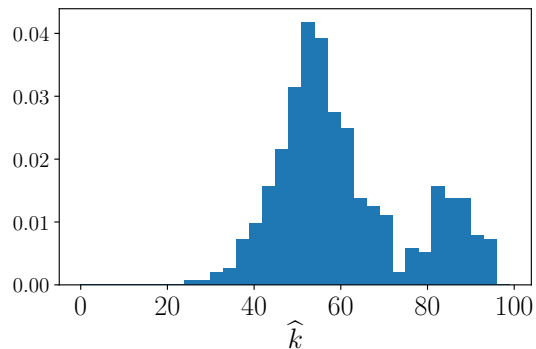


Figure 5: The histogram of estimates values $\hat{k}$ computed on 1000 distinct image pairs. In each image, a set of keypoints of sizes $n = m = 100$ were chosen, with only $k^* = 60$ inlier keypoints. Each keypoint is represented by its 128-dimensional SIFT descriptor.

the Greedy algorithm to match the feature vectors (as done in OpenCV) will allow us to compute only one distance per iteration instead of solving MCF from scratch. Second, the stepsize from the second step of the procedure from Section 4.3 can be increased by considering the difference $\hat{\Phi}(k + 10) - \hat{\Phi}(k)$, then with a proper adjustment to the threshold, we will obtain a 10 times speedup in the number of iterations.

## 5.4 Experiments on Biomedical Data

First, we tested our estimator $\hat{\pi}_k^{\text{LSS}}$ from (3) in the setting of [Chen et al., 2022b][Sec. 5.1]. The setting considered there is to recover the matching between two datasets[3] collected from human pancreatic islets using technologies (CEL-seq2 [Hashimshony et al., 2016] and Smart-seq2 [Picelli et al., 2013]). CEL-seq2 data contain measurements on 34363 RNAs in 2285 cells, and Smart-seq2 data contain 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 $2808 \times 2285$ and $2808 \times 2394$. Each cell in both datasets has a human-annotated type (out of 13 cell types). We randomly downsample cells to get an equal number of cells-per-type in both datasets, eventually getting two datasets of size $2808 \times 1935$. We match cells in two datasets using $\hat{\pi}_k^{\text{LSS}}$ from (3). However, we have no information on the ground truth matching, therefore as suggested in Chen et al. [2022b] the accuracy is calculated on the cell-type level. This means that a single match is considered correct if it matches two cells of the same type. Our matching estimator achieves $97.88\%$ cell-type level ac-
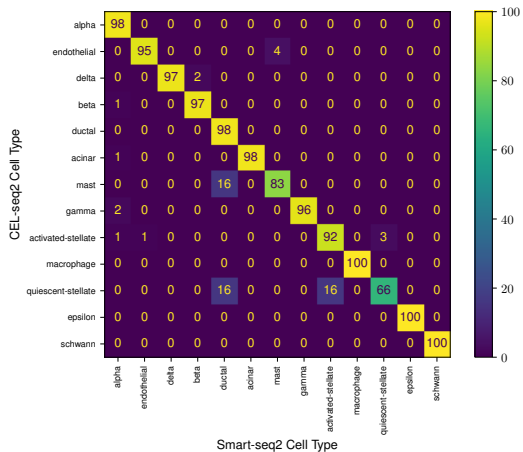
---

[3] `SeuratData` R package [Hao et al., 2021].

Figure 6: Cell-type level confusion matrix of $\widehat{\pi}_n^{\mathrm{LSS}}$ on Cel-seq2 and Smart-seq2 datasets of the same size $n = 1935$.



Figure 7: Cell-type frequencies after type-balancing Cel-seq2 and Smart-seq2 datasets of size $n = 1935$ with 13 cell-types.

curacy, which is almost the same as the result of $97.93\%$ of [Chen et al., 2022b] without any dimensionality reduction and using a simpler approach. The resulting confusion matrix is shown in Fig. 6.

To demonstrate the result of Theorem 1, we design the following experiment. After performing the same pre-processing and balancing steps from the previous experiment, we proceed to remove cells of one fixed type from the CEL-seq2 dataset and cells of another fixed type from the Smart-seq2 dataset. This way, we can ensure the presence of outliers in both sets. Number of cells-per-type are shown in Fig. 7. We only experiment with types *beta, gamma, delta, ductal*, and *acinar* because cells of type *alpha* constitute almost half of the whole dataset and other types have very few cells for the results to be significant. Then we fix a pair of different cell types, remove each from the corresponding dataset, and estimate the real number of inliers (here meaning the number of cells of a type that appears in both datasets) using the algorithm described in Section 3. The results are reported in Table 1. We observe that in most cases the estimated value $\widehat{k}$ is slightly underestimated but is an extremely accurate estimate of the true matching size $k^*$.

## 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 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 $\pi^*$ with high probability. Once again, this holds under the minimal assumption that the signal-to-noise ratio exceeds the minimax separation rate.

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.

| | $\beta$ | $\gamma$ | $\delta$ | duct. | aci. |
|---|---|---|---|---|---|
| $\beta$ | | 1517 / 1593 | 1500 / 1592 | 1370 / 1362 | 1439 / 1421 |
| $\gamma$ | 1517 / 1594 | | 1698 / 1777 | 1568 / 1557 | 1637 / 1618 |
| $\delta$ | 1500 / 1594 | 1698 / 1772 | | 1551 / 1540 | 1620 / 1602 |
| duct. | 1370 / 1362 | 1568 / 1559 | 1551 / 1544 | | 1490 / 1557 |
| aci. | 1439 / 1434 | 1637 / 1631 | 1620 / 1614 | 1490 / 1592 | |

Table 1: Each row corresponds to the experiment where the particular type has been removed from the CEL-seq2 dataset and each column—to the type removed from Smart-seq2. Each cell has two numbers; the true number of inliers / its estimate by the algorithm described in Section 3.
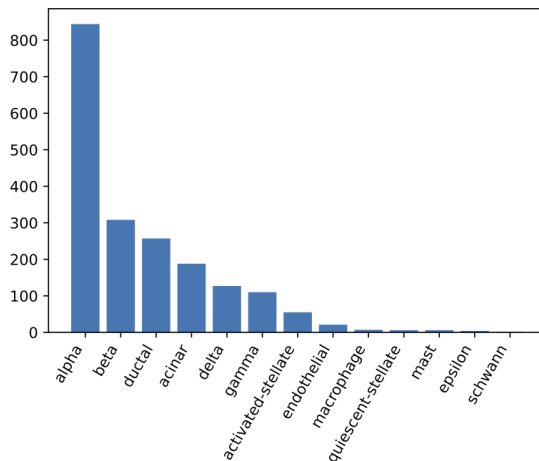
## References

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 S. Dalalyan. Optimal detection of the feature matching map in presence of noise and outliers. *Electronic Journal of Statistics*, 16(2):5720 – 5750, 2022. doi: 10.1214/22-EJS2076. URL https://doi.org/10.1214/22-EJS2076.

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.

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, 2022. 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.

# Appendix

The purpose of this appendix is twofold: to present the proofs of the lemmas used in the main paper and to provide additional experimental evidence showing that it is indeed possible to obtain an accurate estimator for unknown $k^*$ even when the noise levels are unknown and potentially heterogeneous. The reproducible code of all the experiments can be found in the supplementary material.

## A  Proofs of Lemmas from Section 4

We start by presenting the definitions that we use in this supplementary material. Recall the definitions of the test statistics $\widehat{\Phi}(\cdot)$ and its normalized version $\widehat{L}(\cdot)$, which depends on the quantity $\sigma_0^2$

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

For completeness, we also recall the definition of the standard Gaussian random vectors $\eta_{ij}$

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

The quantities associated with $\eta_{ij}$ which will be used in the proofs are $\zeta_1$ and $\zeta_2$, which are defined as follows

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

Recall also that for any matching map $\pi$ we define $L(\pi)$ as follows

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

In this section, we present the proofs of lemmas used in Section 3 for proving Theorem 2 and Theorem 3. For the reader's convenience, we include the statements of the lemmas as well.

**Lemma 1.** *Let $\pi$ be any matching map that can not be obtained as a restriction of $\pi^*$ 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 $\pi_0$ be the restriction of $\pi^*$ to $S_0$. On the event $\Omega_0 = \{8\zeta_1 \leq \bar{\kappa}_{\mathrm{all}}; 4\sqrt{d}\,\zeta_2 \leq \bar{\kappa}_{\mathrm{all}}^2\}$, we have*

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

*Proof of Lemma 1.* Let us recall the definition of the individual signal-to-noise ratios $\kappa_{i,j} \triangleq \frac{\|\theta_i - \theta_j^\#\|_2}{\sqrt{\sigma^2 + \sigma^{\#2}}}$. For any matching map $\pi$ and for any $i \in S_\pi$, we have

$$
\begin{aligned}
\frac{\|X_i - X_{\pi(i)}^\#\|_2^2}{\sigma^2 + \sigma^{\#2}} &= \frac{\|\theta_i - \theta_{\pi(i)}^\#\|_2^2}{\sigma^2 + \sigma^{\#2}} + 2\frac{(\theta_i - \theta_{\pi(i)}^\#)^\top \eta_{i,\pi(i)}}{\sqrt{\sigma^2 + \sigma^{\#2}}} + \|\eta_{i,\pi(i)}\|_2^2 \\
&\geq \frac{\|\theta_i - \theta_{\pi(i)}^\#\|_2^2}{\sigma^2 + \sigma^{\#2}} - 2\zeta_1 \frac{\|\theta_i - \theta_{\pi(i)}^\#\|_2}{\sqrt{\sigma^2 + \sigma^{\#2}}} + \|\eta_{i,\pi(i)}\|_2^2 \\
&= \kappa_{i,\pi(i)}^2 - 2\zeta_1 \kappa_{i,\pi(i)} + \|\eta_{i,\pi(i)}\|_2^2. \qquad (9)
\end{aligned}
$$

Note that if $i \in S_\pi$ is such that $\pi^*(i) = \pi(i)$ (correct matching), then $\kappa_{i,\pi(i)} = 0$. For all the other $i \in S_\pi$, we have $\kappa_{i,\pi(i)} \geq \bar{\kappa}_{\mathrm{all}}$. Therefore, denoting $S_\pi^+ = \{i \in S_\pi \cap S_{\pi^*} : \pi(i) = \pi^*(i)\}$ and $S_\pi^- = S_\pi \setminus S_\pi^+$, Eq. (9) implies that on the

event $\{\bar{\kappa}_{\text{all}} \geq \zeta_1\} \supset \Omega_0$, we have

$$
L(\pi) \geq \sum_{i \in S_\pi^-} \left( \kappa_{i,\pi(i)}^2 - 2\zeta_1 \kappa_{i,\pi(i)} \right) + \sum_{i \in S_\pi} \|\eta_{i,\pi(i)}\|_2^2
$$

$$
\geq |S_\pi^-| \left( \bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} \right) + \sum_{i \in S_\pi} \|\eta_{i,\pi(i)}\|_2^2.
$$

Let us choose any $S_0$ such that $S_\pi^+ \subset S_0 \subset S^*$ and $|S_0| \leq |S_\pi|$. We define $\pi_0$ as the restriction of $\pi^*$ on $S_0$. If, in addition, we set $S_0^- = S_0 \setminus S_\pi^+$, we can infer from the last display that

$$
L(\pi) - L(\pi_0) \geq |S_\pi^-| \left( \bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} \right) + \sum_{i \in S_\pi^-} \|\eta_{i,\pi(i)}\|_2^2 - \sum_{i \in S_0^-} \|\eta_{i,\pi^*(i)}\|_2^2
$$

$$
\geq |S_\pi^-| \left( \bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} \right) + d(|S_\pi^-| - |S_0^-|) - \sqrt{d}\, \zeta_2(|S_\pi^-| + |S_0^-|)
$$

$$
\geq |S_\pi^-| \left( \bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} - 2\sqrt{d}\, \zeta_2 \right) + d(|S_\pi^-| - |S_0^-|)
$$

$$
= |S_\pi^-| \left( \bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} - 2\sqrt{d}\, \zeta_2 \right) + d(|S_\pi| - |S_0|). \tag{10}
$$

On the event $\Omega_0$, we have $\bar{\kappa}_{\text{all}}^2 - 2\zeta_1 \bar{\kappa}_{\text{all}} - 2\sqrt{d}\, \zeta_2 \geq \bar{\kappa}_{\text{all}}^2/4$. Moreover, since $\pi \neq \pi_{S_\pi}^*$, we have $|S_\pi^-| \geq 1$. These two inequalities combined with (10) complete the proof of the lemma. $\qquad\square$

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

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

*Proof of Lemma 2.* The union bound implies that

$$
\mathbf{P}(\Omega_{0,x}^{\complement}) \leq \mathbf{P}\left( 8\zeta_1 \geq x \right) + \mathbf{P}\left( 4\sqrt{d}\, \zeta_2 \geq x^2 \right)
$$

$$
= \mathbf{P}\left( \zeta_1 \geq \tfrac{1}{8} x \right) + \mathbf{P}\left( \zeta_2 \geq \tfrac{1}{4\sqrt{d}} x^2 \right). \tag{11}
$$

Notice that $\zeta_1$ can be represented as the maximum of absolute values of standard Gaussian random variables, *i.e.*, $\zeta_1 = \max_{i \neq j} |\zeta_{i,j}|$. Applying the well-known Gaussian tail bounds together with the union bound yields

$$
\mathbf{P}\left( \zeta_1 \geq \tfrac{1}{8} x \right) \leq \sum_{i \neq j} \mathbf{P}\left( |\zeta_{i,j}| \geq \tfrac{1}{8} x \right) \leq 2n^2 \exp \left( - x^2/128 \right). \tag{12}
$$

To bound the second term of (11), we use Lemma 1 from Galstyan et al. [2022] which bounds the tails of a random variable $\zeta_2$. Thus, combining it with a union bound we arrive at the following inequality

$$
\mathbf{P}\left( \zeta_2 \geq \frac{x^2}{4\sqrt{d}} \right) \leq 2n^2 \exp \left\{ - \frac{x^2}{32\sqrt{d}} \left( \frac{x^2}{4\sqrt{d}} \wedge \sqrt{d} \right) \right\}
$$

$$
= 2n^2 \exp \left\{ - \frac{(x/16)^2}{d} \left( 2x^2 \wedge 8d \right) \right\}. \tag{13}
$$

Then, plugging the bounds obtained in (12) and (13) into (11) concludes the proof of the lemma. $\qquad\square$

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

*Proof of Lemma 3.* This claim is a consequence of Lemma 1. We have already seen in the proof of Theorem 2 that $\widehat{\pi}_{k^*} = \pi^*$ on $\Omega_0$. Therefore,

$$
\widehat{L}(k^*+1) - \widehat{L}(k^*) = L(\widehat{\pi}_{k^*+1}) - L(\pi^*).
$$

If we apply Lemma 1 to $\pi = \widehat{\pi}_{k^*+1}$, it is clear that we can choose as $\pi_0$ the true matching map $\pi^*$. The claim of Lemma 1 then yields

$$
L(\widehat{\pi}_{k^*+1}) - L(\pi^*) \geq \frac{1}{4}\bar{\kappa}_{\text{all}}^2 + d(k^*+1-k^*) = \frac{1}{4}\bar{\kappa}_{\text{all}}^2 + d
$$

and the claim of the lemma follows. $\qquad\square$

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

*Proof of Lemma 4.* Let $\widehat{\pi}_k$ be a matching map from $\mathcal{P}_k$ minimizing $L(\cdot)$, *i.e.*, such that $L(\widehat{\pi}_k) = \widehat{L}(k)$. According to Lemma 1, we have $\widehat{\pi}_k(i) = \pi^*(i)$ for every $i \in \widehat{S}_k \triangleq S_{\widehat{\pi}_k}$. One easily checks that there exists a set $\widehat{S}_{k+1} \subset S^*$ of cardinality $k+1$ such that $\widehat{S}_k \subset \widehat{S}_{k+1}$ and $\widehat{L}(k+1) = L(\widehat{\pi}_{k+1})$ where $\widehat{\pi}_{k+1}$ is the restriction of $\pi^*$ to $\widehat{S}_{k+1}$. Indeed, if $\pi$ is any element of $\mathcal{P}_{k+1}$ minimizing $L(\cdot)$, we know that it is defined as a restriction of $\pi^*$ on some set $S$ of cardinality $k+1$. If we replace arbitrary $k$ elements of $S$ by those of $\widehat{S}_k$, and modify $\pi$ accordingly, then we will get a new mapping from $\mathcal{P}_{k+1}$, for which the value of $L(\cdot)$ is less than or equal to $L(\pi)$. Therefore, we have found a mapping map that minimizes $L(\cdot)$ over $\mathcal{P}_{k+1}$ and has a support that is obtained by adding one point to $\widehat{S}_k$. This implies that

$$\widehat{L}(k+1) - \widehat{L}(k) = L(\widehat{\pi}_{k+1}) - L(\widehat{\pi}_k) = \sum_{i \in \widehat{S}_{k+1}} \|\eta_{i,\pi^*(i)}\|_2^2 - \sum_{i \in \widehat{S}_k} \|\eta_{i,\pi^*(i)}\|_2^2$$

$$= \sum_{i \in \widehat{S}_{k+1} \setminus \widehat{S}_k} \|\eta_{i,\pi^*(i)}\|_2^2 \leq d + \sqrt{d}\,\zeta_2.$$

This completes the proof of the lemma. □

## B  Additional experiments on real data

In this section, we perform experiments on a pair of images and respectively choose keypoints on each of them to showcase the behavior of the proposed procedure. In Fig. 5 of the main manuscript, we show the histogram of the choice of matching size $k^*$ for 1000 distinct image pairs from "Reichstag" scene of IMC-PT 2020 dataset [Jin et al., 2020]. In this dataset, we only have the *pseudo* ground truths, and sometimes these ground truths are incorrect (different points are matched), which makes the results unreliable. Therefore, to have a more controlled experiment we take one image of Sacré Coeur of Paris and crop it in half on each axis. Then, we add noise into the cropped image by interpolating the pixels such that both images have the same resolution. This procedure is in line with the studied model, presented in (1). Afterward, we detect and compute SIFT descriptors of $m = 2n - k^*$ keypoints from the cropped image and translate them into the original image. Then, we fix $k^*$ *inlier* keypoints in both images and add $n - k^*$ distinct points to each image, which will be considered as *outliers*.

We then run our procedure for the estimation of the matching size $k^*$ and the recovery of the matching map $\pi^*$ with the tuning parameters chosen as shown in Theorem 1. The results for different values of $n$ and $k^*$ are summarized in the figure below. The estimated value $\widehat{k}$ is close to $k^*$ and is slightly underestimated in all cases. Slight underestimation is not a problem, whereas slight overestimation would surely cause more incorrect matching pairs.

For all the plots, we see that the value of $k^*$ is estimated accurately and is slightly underestimated (as shown also in Fig. 5). The accuracy of the estimation of $\pi^*$ (number of green lines) is also very high with only a few mistakes. It is worth mentioning that the estimation of $k^*$ is a procedure that can be of interest by itself because after having an accurate estimator for $k^*$ one is free to apply any matching algorithm to circumvent other purposes. For example, one can use fast approximate methods to accelerate matching algorithms (see e.g. Malkov and Yashunin [2020], Harwood and Drummond [2016], Jiang et al. [2016]). Another possible direction is to consider wider or narrower classes of mappings, e.g. 1-to-many matching maps.

Figure 8: We fixed the tolerance level $\alpha = 0.001$. For illustration purposes, we have selected 3 different scenarios to demonstrate the quality of the model selection and the matching recovery. In the uppermost plot $n = 50$ and $k^* = 25$. The procedure from Section 3 outputs $\widehat{k} = 22$ and $\widehat{\pi}_{\widehat{k}}^{\mathsf{LSS}} = \pi^*$ (perfect matching). In the middle plot $n = 100$ and $k^* = 50$, the estimated value of $k^*$ is $\widehat{k} = 44$. In the bottom plot we selected $n = 350$ keypoints from which $k^* = 250$ were inliers ($\widehat{k} = 213$). In last two cases matching map contained only few mistakes.