Optimal detection of the feature matching map in presence of noise and outliers

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the problem of finding the matching map between two sets of d-dimensional vectors from noisy observations, where the second set contains outliers. The matching map is then an injection, which can be consistently estimated only if the vectors of the second set are well separated. The main result shows that, in the high-dimensional setting, a detection region of unknown injection may be characterized by the sets of vectors for which the inlier-inlier distance is of order at least $d^{1/4}$ and the inlier-outlier distance is of order at least $d^{1/2}$. These rates are achieved using the estimated matching minimizing the sum of logarithms of distances between matched pairs of points. We also prove lower bounds establishing optimality of these rates. 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 estimators studied in this work.

1 Introduction

Finding the best match between two clouds of points is a problem encountered in many real problems. In computer vision, one can look for correspondences between two sets of local descriptors extracted from two images. In text analysis, one can be interested in matching vector representations of the words of two similar texts, potentially in two different languages. The goal of the present work is to gain theoretical understanding of the statistical limits of the matching problem.

In the sequel, we use the notation $[n] = \{1, \dots, n\}$ for any integer n, and define $\|\cdot\|$ as the Euclidean norm in $\mathbb{R}^d$. Assume that two independent sequences $\boldsymbol{X} = (X_i; i \in [n])$ and $\boldsymbol{Y} = (Y_i; i \in [n])$ of independent vectors are generated such that X_i and Y_i are drawn from the same distribution P_i on $\mathbb{R}^d$, for every $i \in [n]$. The statistician observes the sequence $\boldsymbol{X}$ and a shuffled version $\boldsymbol{X}^\#$ of the sequence $\boldsymbol{Y}$. More precisely, $\boldsymbol{X}^\#$ is such that $X_i^\# = Y_{\pi^*(i)}$ for some unobserved permutation π^*. The goal of matching is to estimate the permutation π^* from data $(\boldsymbol{X}, \boldsymbol{X}^\#)$. In the case of Gaussian distributions P_i, this problem has been studied in (Collier and Dalalyan, 2013, 2016). Clearly, consistent estimation of the matching map π^* is impossible if there are two data generating distributions P_i and P_j that are very close. In (Collier and Dalalyan, 2013, 2016), a precise quantification of the separation between these distributions is given that enables consistent estimation of π^*. Furthermore, it is shown that the permutation estimator minimizing the sum of logarithms of pairwise distances between the elements of $\boldsymbol{X}$ and the elements of the shuffled version $\boldsymbol{X}^\#$ is an optimal estimator of π^*.

In this paper, we extend the model studied in (Collier and Dalalyan, 2016) to the case when the set $\boldsymbol{X}^\#$ is contaminated by outliers. The number of outliers is supposed to be known and is equal to $m - n$, where $n = |\boldsymbol{X}|$ and $m = |\boldsymbol{X}^\#|$ are the sizes of considered two sequences, however the indices of the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

	Without outliers (Collier and Dalalyan, 2016)	With outliers this paper
known $\boldsymbol{\sigma}^{\#}$ or all equal $\boldsymbol{\sigma}^{\#}$s	LSNS is optimal $\bar{\kappa} \gtrsim (d \log n)^{1/4}$	LSNS is optimal [Thm. 1] $\bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}} \gtrsim (d \log(nm))^{1/4}$
unknown $\boldsymbol{\sigma}^{\#}$ $\sigma_{\max}/\sigma_{\min} \leq C$	LSL is optimal $\bar{\kappa} \gtrsim (d \log n)^{1/4}$	LSL is optimal [Thm. 4] $\bar{\kappa}_{\text{in-in}} \gtrsim (d \log(nm))^{1/4}$ & $\bar{\kappa}_{\text{in-out}} \gtrsim d^{1/2}$
unknown $\boldsymbol{\sigma}^{\#}$ arbitrary		LSL is optimal [Thm. 2, 3] $\bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}} \gtrsim d^{1/2}$

Table 1: A brief overview of the contributions in the high-dimensional regime $d \geq c \log n$. The table provides the condition on the normalized inlier-inlier distance $\bar{\kappa}_{\text{in-in}}$ and inlier-outlier distance $\bar{\kappa}_{\text{in-out}}$, making it possible to consistently detect the matching map between two sets of d-dimensional vectors. LSL and LSNS refer to least sum of logarithms and least sum of normalized squares, respectively.

outliers are unknown. All the distributions are assumed in this paper to be spherical Gaussian, although all the probabilistic tools used in the proofs have their sub-Gaussian counterparts. Thus, we consider that two spherical Gaussian distributions [1] $P_1 = \mathcal{N}_d(\mu_1, \sigma_1^2 \mathbf{I}_d)$ and $P_2 = \mathcal{N}_d(\mu_2, \sigma_2^2 \mathbf{I}_d)$ are well separated if the "distance to noise ratio" $\kappa(P_1, P_2) = \|\mu_1 - \mu_2\| / \sqrt{\sigma_1^2 + \sigma_2^2}$ is large. Main findings of (Collier and Dalalyan, 2016), in terms of smallest separation distance $\bar{\kappa} = \min_{i \neq j} \kappa(P_i, P_j)$ are summarized in the second columns of Table 1. Likewise, the last column of the table provides a summary of the contributions of the present paper in terms of $\bar{\kappa}_{\text{in-in}} = \min_{i \neq j} \kappa(P_i, P_j)$ and $\bar{\kappa}_{\text{in-out}} = \min_{i,j} \kappa(P_i, Q_j)$, where $Q_1, \ldots, Q_{m-n}$ are the distributions of the outliers.

An unexpected finding of this work is that the "degree" of heteroscedasticity of the model has a strong impact on the separation distances and the detection regions (sets of values of $(\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}})$ for which it is possible to detect the feature map π^*). This is in sharp contrast with the outlier-free case, where consistent estimation requires $\bar{\kappa}$ to be at least of order $(d \log n)^{1/4}$ irrespective from the behaviour of variances of P_i. We prove in this work that in the high dimensional regime $d \geq c \log n$, which is arguably more appealing than the low dimensional regime $d \leq c \log n$, the following statements are true:

- If there is no heteroscedasticity, *i.e.*, when all the variances are equal, consistent estimation of π^* is possible if and only if $\bar{\kappa} = \bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}}$ is at least of order $(d \log(nm))^{1/4}$. This is the same rate as in the outlier-free case.
- If the heteroscedasticity is mild, *i.e.*, all the variances are of the same order, the condition $\bar{\kappa}_{\text{in-in}} \gtrsim (d \log(nm))^{1/4}$ is the same as in the previous item, but the stronger condition $\bar{\kappa}_{\text{in-out}} \gtrsim d^{1/2}$ is needed for the inlier-outlier separation distance.
- Finally, in the general heteroscedastic setting both $\bar{\kappa}_{\text{in-in}}$ and $\bar{\kappa}_{\text{in-out}}$ should be at least of order $d^{1/2}$. Furthermore, in all these cases consistent estimation is performed by the same estimator: the Least Sum of Logarithms (LSL).

Note also that the empirical evaluation done in this paper shows that LSL is interesting not only from the theoretical but also from the practical point of view.

Agenda Section 2 describes the framework of the vector matching problem and introduces the terminology used throughout this paper. Precise statements of the main theoretical results are gathered in Section 3. The prior work is briefly discussed in Section 4. Section 5 contains numerical experiments carried out both for synthetic and real data. A brief summary and some concluding remarks are presented in Section 6. Proofs of all theoretical claims are deferred to the supplemental material.

2 Problem Formulation

We begin with formalizing the problem of matching two sequences of feature vectors $(X_1, \ldots, X_n)$ and $(X_1^{\#}, \ldots, X_m^{\#})$ with different sizes n and m such that $m \geq n \geq 2$. In what follows, we assume

[1] We use the notation $\mathbf{I}_d$ for the $d \times d$ identity matrix

that the observed feature vectors are randomly generated from the model

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

In this model, illustrated in Figure 1, it is assumed that

- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ and $\boldsymbol{\theta}^\# = (\theta_1^\#, \dots, \theta_m^\#)$ are two sequences of vectors from $\mathbb{R}^d$, corresponding to the original features, which are unavailable,
- $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)^\top, \boldsymbol{\sigma}^\# = (\sigma_1^\#, \dots, \sigma_m^\#)^\top$ are positive real numbers corresponding to the magnitudes of the noise contaminating each feature,
- $\xi_1, \dots, \xi_n$ and $\xi_1^\#, \dots, \xi_m^\#$ are two independent sequences of i.i.d. random vectors drawn from the Gaussian distribution with zero mean and identity covariance matrix.

The simplest special case of (1), considered in (Collier and Dalalyan, 2016), corresponds to the situation where a perfect matching exists between the two sequences $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\#$. This means that $m = n$ and, for some bijective mapping $\pi^* : [n] \to [n]$, $\theta_i = \theta^\#_{\pi(i)}$ for all $i \in [n]$. In the general case, both $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{X}^\# = (X_1^\#, \dots, X_m^\#)$ may contain outliers, *i.e.* feature vectors that have no corresponding pair. In such a situation, it is merely assumed that there exists a set $S \subset [n]$ and an injective mapping $\pi^* : S \to [m]$ such that

$$\theta_i = \theta^\#_{\pi^*(i)} \quad \text{and} \quad \sigma_i = \sigma^\#_{\pi^*(i)}, \qquad \forall\, i \in S. \tag{2}$$

In this case we say that the vectors $\{X_i : i \in [n] \setminus S\}$ and $\{X_j^\# : j \in [m] \setminus \pi^*(S)\}$ are outliers. The ultimate goal is to estimate the feature matching map π^*.

In this work we consider the case when $S = [n]$ and $m > n$. This means that only the larger set of feature vectors, namely $\boldsymbol{X}^\#$, contains outliers. Let us also define the set $O_{\pi^*} \triangleq [m] \setminus \mathrm{Im}(\pi^*)$, which contains the indices of outliers and satisfies $|O_{\pi^*}| = m - n$. Naturally, the feature vectors contained in $\boldsymbol{X}$, as well as those vectors from $\boldsymbol{X}^\#$ that are not outliers, are called *inliers*.

In this formulation, the data generating distribution is defined by the parameters $\boldsymbol{\theta}^\#$, $\boldsymbol{\sigma}^\#$ and π^*. We omit the set of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$, since they are automatically identified using π^*, $\boldsymbol{\theta}^\#$ and $\boldsymbol{\sigma}^\#$ by the formula $(\theta_i, \sigma_i) = (\theta^\#_{\pi^*(i)}, \sigma^\#_{\pi^*(i)})$ for $i \in [n]$. Since our goal is to match the feature vectors, we focus our attention on the problem of estimating the parameter π^* only, considering $\boldsymbol{\theta}^\#$ and $\boldsymbol{\sigma}^\#$ as nuisance parameters. In what follows, we denote by $\mathbf{P}_{\boldsymbol{\theta}^\#, \boldsymbol{\sigma}^\#, \pi^*}$ the probability distribution of the sequence $(X_1, \dots, X_n, X_1^\#, \dots, X_m^\#)$ defined by (1) under condition (2) with $S = [n]$.

We are interested in designing estimators that have an expected error smaller than a prescribed level α under the weakest possible conditions on the nuisance parameter $\boldsymbol{\theta}^\#$ and noise level $\boldsymbol{\sigma}^\#$. Clearly, the problem of matching becomes more difficult with hardly distinguishable features. To quantify this phenomenon, we introduce the normalized separation distance $\bar\kappa_{\text{in-in}} = \bar\kappa_{\text{in-in}}(\boldsymbol{\theta}^\#, \boldsymbol{\sigma}^\#, \pi^*)$ and the normalized outlier separation distance $\bar\kappa_{\text{in-out}} = \bar\kappa_{\text{in-out}}(\boldsymbol{\theta}^\#, \boldsymbol{\sigma}^\#, \pi^*)$, which measure the minimal distance-to-noise ratio between inliers and the minimal distance-to-noise ratio between inliers and outliers, respectively. The precise definitions read as

$$\bar\kappa_{\text{in-in}} \triangleq \min_{\substack{i,j \notin O_{\pi^*}, \\ j \neq i}} \frac{\|\theta_i^\# - \theta_j^\#\|}{(\sigma_i^{\#\,2} + \sigma_j^{\#\,2})^{1/2}}, \qquad \bar\kappa_{\text{in-out}} \triangleq \min_{\substack{i \notin O_{\pi^*}, \\ j \in O_{\pi^*}}} \frac{\|\theta_i^\# - \theta_j^\#\|}{(\sigma_i^{\#\,2} + \sigma_j^{\#\,2})^{1/2}}. \tag{3}$$

Notice that $\bar\kappa_{\text{in-in}}$ can be rewritten as

$$\bar\kappa_{\text{in-in}} = \min_{\substack{i,j \in [n] \\ i \neq j}} \frac{\|\theta_i - \theta_j\|}{(\sigma_i^2 + \sigma_j^2)^{1/2}}.$$

Clearly, if $\bar\kappa_{\text{in-in}} = 0$ or, $\bar\kappa_{\text{in-out}} = 0$, there are two identical feature vectors in $\boldsymbol{X}^\#$. In such a situation, assuming σ_i's are all equal, the parameter π^* is nonidentifiable, in the sense that there exist two different permutations π_1^* and π_2^* such that the distributions $\mathbf{P}_{\boldsymbol{\theta}^\#, \boldsymbol{\sigma}^\#, \pi_1^*}$ and $\mathbf{P}_{\boldsymbol{\theta}^\#, \boldsymbol{\sigma}^\#, \pi_2^*}$ coincide. Therefore, to ensure the existence of consistent estimators of π^* it is necessary to impose the conditions $\bar\kappa_{\text{in-in}} > 0$ and $\bar\kappa_{\text{in-out}} > 0$. Moreover, good estimators are those consistently estimating π^* even if either $\bar\kappa_{\text{in-in}}$ or $\bar\kappa_{\text{in-out}}$ are small. We are interested here in finding the detection boundary in terms of the order of magnitude of $(\bar\kappa_{\text{in-in}}, \bar\kappa_{\text{in-out}})$. More precisely, for any given $\alpha \in (0, 1)$ we wish to find a region $\mathcal{R}^\alpha_{n,m,d}$ in $\mathbb{R}^2$ such that:

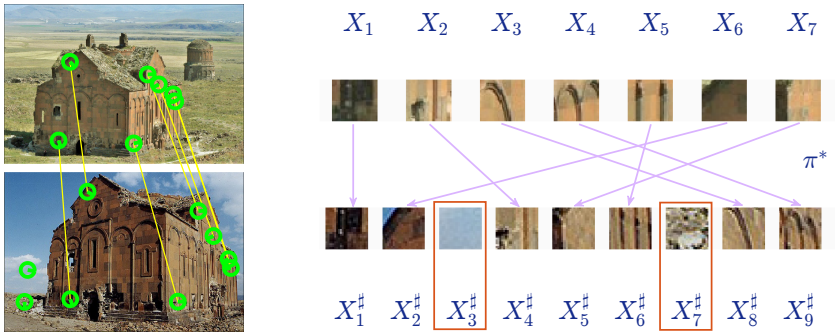

Figure 1: Illustration of the considered framework described in (1). We wish to match a set of 7 patches extracted from the first image to the 9 patches from the second image. The picture on the left shows the locations of patches as well as the true matching map π^* (the yellow lines).

- There is an estimator $\hat{\pi}_{n,m}$ of π^* satisfying $\mathbf{P}_{\boldsymbol{\theta}^\sharp,\boldsymbol{\sigma}^\sharp,\pi^*}(\hat{\pi} \neq \pi^*) \leq \alpha$ for every $(\boldsymbol{\theta}^\sharp, \boldsymbol{\sigma}^\sharp, \pi^*)$ lying in the detection region, *i.e.*, for which $(\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}}) \in \mathcal{R}^\alpha_{n,m,d}$.
- There is a constant $C < 1$ such that for any estimator $\bar{\pi}_{n,m}$ of π^*, we can find a parameter value $(\boldsymbol{\theta}^\sharp, \boldsymbol{\sigma}^\sharp, \pi^*)$ in the region $\{(\boldsymbol{\theta}^\sharp, \boldsymbol{\sigma}^\sharp, \pi^*) : (\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}}) \in C\mathcal{R}^\alpha_{n,m,d}\}$ such that $\bar{\pi}$ fails to detect π^* with a probability larger than α.

Let us make two remarks. First, note that in the outlier-free case considered in (Collier and Dalalyan, 2016), $\bar{\kappa}_{\text{in-out}}$ is meaningless and, therefore, the detection region is one-dimensional $\mathcal{R}^\alpha_{n,m,d}$. Thus, it is necessarily a half-line and is proven to be of the form $\bar{\kappa}_{\text{in-in}} \geq C(\log {}^n\!/_\alpha)^{1/2} \vee (d \log {}^n\!/_\alpha)^{1/4}$ for some universal constant C. Second, the aforementioned definition of the detection region $\mathcal{R}^\alpha_{n,m,d}$ does not guarantee its uniqueness (even up to a scaling by a universal constant). This is in contrast with the outlier-free case. To overcome this difficulty, we look for $\mathcal{R}^\alpha_{n,m,d}$ of the form $[t_{\text{in-in}}, +\infty) \times [t_{\text{in-out}}, +\infty)$ with the smallest possible threshold $t_{\text{in-out}}$ for the normalized inlier-outlier distance $\bar{\kappa}_{\text{in-out}}$.

3 Main theoretical results

In this section, we have collected the main theoretical findings of the paper. When the noise is homoscedastic, *i.e.*, when all $\boldsymbol{\sigma}$'s are equal, the results obtained by Collier and Dalalyan (2016) in the outlier-free setting can be easily extended to the setting with outliers. Therefore, in the present paper, we focus on the heteroscedastic case. For the sake of clarity of exposition, we will present the results in the case of known variances $\boldsymbol{\sigma}, \boldsymbol{\sigma}^\sharp$ prior to investigating the more interesting case of unknown variances.

The detection regions we study below are based on the profile maximum likelihood estimator. The model presented in (1) has the parameter $\boldsymbol{\Xi} = (\boldsymbol{\theta}^\sharp, \boldsymbol{\sigma}^\sharp, \pi)$, while the observations are the sequences of feature vectors $\boldsymbol{X}$ and $\boldsymbol{X}^\sharp$. The negative log-likelihood of this model is given by

$$\ell_n(\boldsymbol{\Xi}; \{\boldsymbol{X}, \boldsymbol{X}^\sharp\}) = \sum_{i=1}^n \left(\frac{\|X_i - \theta^\sharp_{\pi(i)}\|_2^2}{2\sigma^{\sharp 2}_{\pi(i)}} + \frac{1}{2}\log(\sigma^{\sharp 2}_{\pi(i)}) \right) + \sum_{j=1}^m \left(\frac{\|X^\sharp_j - \theta^\sharp_j\|_2^2}{2\sigma^{\sharp 2}_j} + \frac{1}{2}\log(\sigma^{\sharp 2}_j) \right).$$

The profile negative log-likelihood is then defined as the minimum with respect to $(\boldsymbol{\theta}^\sharp, \boldsymbol{\sigma}^\sharp)$ of the log-likelihood $\ell_n(\boldsymbol{\Xi}; \{\boldsymbol{X}, \boldsymbol{X}^\sharp\})$.

3.1 Warming up: known variances $\boldsymbol{\sigma}, \boldsymbol{\sigma}^\sharp$

One can check that the minimization with respect to $\boldsymbol{\theta}^\sharp$ leads to the variance-dependent cost function

$$\ell_n(\pi, \boldsymbol{\sigma}^\sharp; \{\boldsymbol{X}, \boldsymbol{X}^\sharp\}) = \sum_{i=1}^n \frac{\|X_i - X^\sharp_{\pi(i)}\|^2}{\sigma_i^2 + {\sigma^\sharp_{\pi(i)}}^2} + \sum_{i=1}^n \frac{1}{2}\log(\sigma^{\sharp 2}_{\pi(i)}) + \sum_{j=1}^m \frac{1}{2}\log(\sigma^{\sharp 2}_j). \tag{4}$$

When $m = n$ and there is no outlier, the last two sums of the last display do not depend on π and, therefore, the maximum profile likelihood estimator of π^* is obtained by the Least Sum of

Normalized Squares (LSNS) criterion

$$\hat{\pi}^{\text{LSNS}}_{n,m} \in \arg\min_{\pi} \sum_{i=1}^{n} \frac{\|X_i - X^{\#}_{\pi(i)}\|^2}{\sigma_i^2 + {\sigma^{\#}_{\pi(i)}}^2}, \tag{5}$$

where the minimum is over all injective mappings $\pi : [n] \to [m]$. This, and the other estimators defined in this work, can be efficiently computed using suitable versions of the Hungarian algorithm (Kuhn, 1955, 2010; Munkres, 1957). As shows the next theorem, it turns out that even when $m > n$, the estimator $\hat{\pi}^{\text{LSNS}}_{n,m}$ defined above leads to an optimal detection region.

Theorem 1 (Upper bound for LSNS) *Let $\alpha \in (0,1)$ and condition (2) be fulfilled. If the separation distances $\bar{\kappa}_{\text{in-in}}$ and $\bar{\kappa}_{\text{in-out}}$ corresponding to $(\boldsymbol{\theta}^{\#}, \boldsymbol{\sigma}^{\#}, \pi^*)$ and defined in (3) satisfy the condition*

$$\min\{\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}}\} \geq 4\Big\{ \big(d \log(4nm/\alpha)\big)^{1/4} \vee \big(2 \log(8nm/\alpha)\big)^{1/2} \Big\} \tag{6}$$

then the LSNS estimator defined in (5) detects the true matching map π^* with probability at least $1-\alpha$, that is

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

The similarity—both its statement and its proof— of this result is to its counterpart in the outlier-free setting might suggest that the presence of outliers does not make the problem any harder from a statistical point of view. However, this is not true in the more appealing setting of unknown variances.

3.2 Detection of π^* for unknown and arbitrary variances $\boldsymbol{\sigma}, \boldsymbol{\sigma}^{\#}$

The LSNS procedure analyzed in Theorem 1 exploits the values of known noise variances to normalize the squares of distances between vectors X_i and $X^{\#}_{\pi(i)}$. Therefore, LSNS is inapplicable in the case of unknown noise variances. Instead, we consider the Least Sum of Logarithms (LSL) estimator

$$\hat{\pi}^{\text{LSL}}_{n,m} \triangleq \arg\min_{\pi : [n] \to [m]} \sum_{i=1}^{n} \log \|X_i - X^{\#}_{\pi(i)}\|^2, \tag{7}$$

where the minimum is over all injective maps $\pi : [n] \to [m]$. This estimator can be seen as the minimizer of a criterion defined as the minimum of the cost function from (4) with respect to $\boldsymbol{\sigma}^{\#}$ under the constraint $\min_{j \notin \text{Im}(\pi)} \sigma^{\#}_j \geq \sigma_{\min}$, for some fixed (but unknown) constant $\sigma_{\min} > 0$.

To provide a quick overview of what follows, let us stick in the remaining of this paragraph to the case $\log(nm) = O(d)$ so that the right hand side of (6) is of order $\big(d \log(nm)\big)^{1/4}$. Recall that in the outlier-free case, the LSL estimator has been shown to perform as well as the LSNS while having the advantage of not requiring the knowledge of variances $\boldsymbol{\sigma}^{\#}$ (Collier and Dalalyan, 2016). Somewhat unexpectedly, the situation is significantly different in the presence of outliers. Indeed, the best we manage to prove in the presence of outliers is that the detection of the matching map by LSL is possible whenever $\min\{\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}}\} \geq C\sqrt{d}$ for some sufficiently large constant C. The precise statement being given in the next theorem, let us mention right away that the discrepancy between this rate $\sqrt{d}$ and the rate $\big(d \log(nm)\big)^{1/4}$ in (6) is due to the inherent hardness of the setting and not merely an artefact of the proof. This will be made clear below.

Theorem 2 (Upper bound for LSL) *Let $\alpha \in (0, 1/2)$ and condition (2) be fulfilled. If the separation distances $\bar{\kappa}_{\text{in-in}}$ and $\bar{\kappa}_{\text{in-out}}$ corresponding to $(\boldsymbol{\theta}^{\#}, \boldsymbol{\sigma}^{\#}, \pi^*)$ and defined by (3) satisfy*

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

then the LSL estimator (7) detects the matching map π^* with probability at least $1-\alpha$, that is

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

This result is disappointing since it requires the distance between different feature vectors to be larger than $\sqrt{2d}$ in order to be able to consistently estimate the matching map π^*. As we show below, without any further condition (for instance, on the noise variances), this rate cannot be improved.

Furthermore, the rate $\sqrt{d}$ is optimal not only for LSL but also for the larger class of so called *distance based M-estimators.*

We will say that an estimator $\hat{\pi}_n$ of π^* is a distance based M-estimator, if for a sequence of non-decreasing functions $\rho_i : \mathbb{R}_+ \to \mathbb{R}$, $i = 1, \dots, n$, the following is correct

$$\hat{\pi}_n \in \arg \min_{\pi:[n]\to[m]} \sum_{i=1}^{n} \rho_i\big(\|X_i - X^{\#}_{\pi(i)}\|\big),$$

where the minimum is over all injective mappings $\pi : [n] \to [m]$. We denote by $\mathcal{M}$ the set of all distance based M-estimators. We show that there is indeed a setup where $\bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}}$ is as large as $0.2\sqrt{d}$ but any estimator from $\mathcal{M}$ fails to detect π^* with probability at least $1/4$. The next theorem formalizes the described result.

Theorem 3 (Lower bound) *Assume that $m > n \geq 4$ and $d \geq 422 \log(4n)$. There exists a triplet $(\boldsymbol{\sigma}^{\#}, \boldsymbol{\theta}^{\#}, \pi^*)$ such that $\bar{\kappa}_{\text{in-in}} = \bar{\kappa}_{\text{in-out}} = \sqrt{d/20}$ and*

$$\inf_{\hat{\pi}\in\mathcal{M}} \mathbf{P}_{\boldsymbol{\theta}^{\#},\boldsymbol{\sigma}^{\#},\pi^*}(\hat{\pi} \neq \pi^*) > 1/4. \tag{9}$$

The proof of this theorem, postponed to the appendix, is constructive. This means that we exhibit a triplet $(\boldsymbol{\sigma}^{\#}, \boldsymbol{\theta}^{\#}, \pi^*)$ satisfying (9). Careful inspection shows that in the case $d = O(\log(nm))$ the same triplet satisfies $\bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}} \asymp \sqrt{\log(nm)}$ and (9) is still true. This implies that the order of magnitude of the right hand side of (8) is optimal both in the high-dimensional regime $d \geq 422 \log(4n)$ and in the low-dimensional regime $d < 422 \log(4n)$. This shows the optimality of LSL among all estimators from $\mathcal{M}$. Note that the estimator $\hat{\pi}^{\text{LSNS}}_{n,m}$ does not belong to the family of *distance based M-estimators*. Furthermore, in the low dimensional regime $d = O(\log(nm))$, the separation rate of the LSL, $\sqrt{\log(nm)}$, is the same as that of the LSNS.

In the next section we show that under some mild conditions on $\boldsymbol{\sigma}^{\#}$ it is indeed possible to obtain different rates for $\bar{\kappa}_{\text{in-in}}$ and $\bar{\kappa}_{\text{in-out}}$, namely we show that if $\bar{\kappa}_{\text{in-in}} \gtrsim d^{1/4}$ and $\bar{\kappa}_{\text{in-out}} \gtrsim d^{1/2}$ then the LSL estimator detects correct matching with high probability.

3.3 Detection of π^* for unknown and mildly varying variances $\boldsymbol{\sigma}, \boldsymbol{\sigma}^{\#}$

The results of the last two theorems are disappointing, since they indicate that the features should be very different from one another for detection of the matching map to be possible. An interesting finding, presented below, is that strong constraint can be significantly alleviated in the context of mild heteroscedasticity. By mild heteroscedasticity we understand here the situation in which all variances $\sigma_i^{\#}$ are of the same order of magnitude.

Theorem 4 (Upper bound under mild heteroscedasticity) *Let $r_\sigma = \max_{i,j}(\sigma_i^{\#}/\sigma_j^{\#}) < \infty$. If the separation distances $\bar{\kappa}_{\text{in-in}}$ and $\bar{\kappa}_{\text{in-out}}$ defined in (3) satisfy*

$$\begin{aligned}
\bar{\kappa}_{\text{in-in}} &\geq 2\big(4d\log(4nm/\alpha)\big)^{1/4} + 2\big(2\log(4nm/\alpha)\big)^{1/2} \\
\bar{\kappa}_{\text{in-out}} &\geq \sqrt{2(r_\sigma - 1)d} + 2\big(4r_\sigma^2 d\log(4nm/\alpha)\big)^{1/4} + 2\big(2r_\sigma\log(4nm/\alpha)\big)^{1/2},
\end{aligned}$$

then the LSL estimator (7) detects the matching map π^* with probability at least $1-\alpha$, that is

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

Note that a lower bound similar to that of Theorem 3 can be proved in the case of mild heterescodestacity as well, showing that there is an example for which $\bar{\kappa}_{\text{in-in}}$ is of order $d^{1/4}$, $\bar{\kappa}_{\text{in-out}}$ is of order $d^{1/2}$ and any estimator from $\mathcal{M}$ fails to detect π^* with probability at least $1/4$.

We complete this section by summarizing the joint contribution of Theorems 1 to 4. To simplify this discussion, we consider two cases: high-dimensional case refers to $d \geq \log(4nm/\alpha)$ (presented in Table 1) and low-dimensional case refers to the condition $d < \log(4nm/\alpha)$. In the high dimensional setting with arbitrary noise variances, the detection region for the LSL estimator is given by $\{\bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}} \geq 15\sqrt{d}\}$, which is much worse than the detection region for LSNS, $\{\bar{\kappa}_{\text{in-in}} \wedge \bar{\kappa}_{\text{in-out}} \geq 8(d\log(4nm/\alpha))^{1/4}\}$, obtained in the known-variance scenario. Somewhat surprisingly, in such a

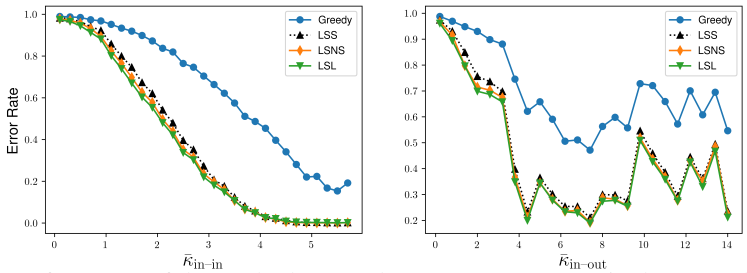

Figure 2: The performance of the methods (Greedy, LSS, LSL, LSNS) in the setup described in Exp. 1. Plots show the error rate (percentage of repetitions in which the estimated matching map differs from the true one) as a function of separation distances. The left plot illustrates that LSS, LSL and LSNS require much lower value of $\bar{\kappa}_{\text{in-in}}$ in order to find the correct mapping. The left plot shows that there is a clear improvement of the error when the inlier-inlier separation distance increases, while the right plot shows that the error rate might not be made small by augmenting $\bar{\kappa}_{\text{in-out}}$ only.

setting, even a strong assumption on the outliers, such as requiring them to be at least at a distance $0.2\sqrt{d}$ of the inliers, is not enough for relaxing the assumption on the inlier-inlier separation distance. Finally, on a positive note, in the intermediate case of mildly varying variances, the detection region for the LSL estimator is of the form $\{\bar{\kappa}_{\text{in-in}} \geq 7(d\log(4nm/\alpha))^{1/4};\ \bar{\kappa}_{\text{in-out}} \geq 10\sqrt{d}\}$. This means that if the outliers are at a distance $\Omega(\sqrt{d})$ of the inliers, then the LSL recovers the true matching under the same condition on $\bar{\kappa}_{\text{in-in}}$ as in the outlier-free setting.

4 Other related work

Measuring the quality of the various statistical procedures of decision making by their minimal separation rates became the standard in hypotheses testing, see the seminal papers (Burnashev, 1979; Ingster, 1982) and 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 hypotheses testing framework, for instance, variable selection (Ndaoud and Tsybakov, 2020; Azaïs and de Castro, 2020; Comminges and Dalalyan, 2012) or the matching problem considered here.

On the other hand, feature matching is a well studied problem in computer vision. In recent years, a great deal of attention was devoted to the acceleration of greedy matching algorithms, based on approximate and fast methods of finding nearest neighbors (e.g. Jiang et al. (2016); Wang et al. (2018); Wang (2011); Harwood and Drummond (2016); Malkov and Yashunin (2020)). Another direction that helps to improve feature matching problem is using alternative local descriptors (Rublee et al., 2011; Chen et al., 2010; Calonder et al., 2010) for given keypoints. Naturally, the question of how to chose keypoints arises, which is addressed, for instance, in (Bai et al., 2020; Tian et al., 2020). For more complete overview of the field we refer to (Ma et al., 2021) and references therein.

Finally, permutation estimation and related problems have been recently investigated in different contexts such as statistical seriation (Flammarion et al., 2019), 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), and recovery of general discrete structure (Gao and Zhang, 2019).

5 Numerical results

In this section, we report the results of some numerical experiments carried out on simulated and real data. We applied aforementioned methods LSNS and LSL and computed different measures of their performance. To get a more complete picture, we included in this study the Least Sum of Squeres (LSS) estimator and the greedy estimator. LSS is an unnormalized version of LSNS, given by

$$\hat{\pi}^{\text{LSS}}_{n,m} \in \arg\min_{\pi} \sum_{i=1}^{n} \|X_i - X^{\#}_{\pi(i)}\|^2. \tag{10}$$

It coincides with LSNS in the case of homoscedastaic noise. The greedy estimator is obtained by sequentially matching each vector from $\boldsymbol{X}$ to the nearest vector from $\boldsymbol{X}^{\#}$. Experiments were implemented using python or matlab. For solving linear sum assignment problems such as (7) or (10), the generalized and improved versions of the Hungarian algorithm were used (Kuhn, 1955, 2010; Munkres, 1957; Duff and Koster, 2001), implemented in SciPy library (Virtanen et al., 2020).

Experiment 1: Synthetic data with random features We first randomly generated π^*, $\boldsymbol{\theta}^{\#}$ and $\boldsymbol{\sigma}^{\#}$ as follows. We randomly sampled from uniform distribution on $[0, 2]$ independent variables τ_{ij}, $i \in [m]$, $j \in [d]$. Then, $(\theta_i^{\#})_j$ are independently sampled from the Gaussian distribution with 0 mean and variance τ_{ij}. Additionally, for every $\theta_i^{\#} \in \boldsymbol{\theta}^{\#}$ such that $i \notin J_{\pi^*}$ ($\theta_i^{\#}$ is an outlier), we incremented every coordinate of $\theta_i^{\#}$ by i. Entries of $\boldsymbol{\sigma}^{\#}$ were sampled from the uniform distribution over $[0.5, 2]$. Sequences $\boldsymbol{X}$ and $\boldsymbol{X}^{\#}$ were generated according to Section 2 with $\pi^*(i) = i$ for $i \in [n]$. We applied to this data the following matching algorithms: Greedy, LSS, LSNS and LSL.

We chose $n = 100$, $m = 130$ and $d = 50$, and generated $N = 50$ datasets according to the foregoing process. For each dataset, we computed the 0-1 error of the considered estimators and the values of $(\bar{\kappa}_{\text{in-in}}, \bar{\kappa}_{\text{in-out}})$. We plotted in the left (resp., the right) panel of Figure 2 the error averaged over all datasets with a given value of $\bar{\kappa}_{\text{in-in}}$ (resp. $\bar{\kappa}_{\text{in-out}}$). In this specific example, we see that the error decreases fast with $\bar{\kappa}_{\text{in-in}}$, while there is no monotonicity of the error rate as a function of $\bar{\kappa}_{\text{in-out}}$.

Experiment 2: Synthetic data with deterministic features The second experiment is conducted on data generated by features $\boldsymbol{\theta}^{\#}$ and variances $\boldsymbol{\sigma}^{\#}$ inspired by the example constructed in the proof of Theorem 3. More precisely, for some real numbers a and b representing, respectively, the scale of inlier-inlier distance $\bar{\kappa}_{\text{in-in}}$ and inlier-outlier distance $\bar{\kappa}_{\text{in-out}}$, we set $\theta_k^{\#} = [ka, 0, \dots, 0]^\top$ for $k \in [n]$ and $\theta_{n+k}^{\#} = [na + kb, 0, \dots, 0]^\top$. We also used decreasing variances $\sigma_k^{\#} = 1/k^{3/2}$ for $k \in [m]$ and true identity mapping $\pi^*(k) = k$ for $k \in [n]$. We chose $n = 100$, $m = 120$ and dimension d varying in the set $\{10, 20, 40\}$. For each pair of values (a, b) in a suitably chosen grid, we repeated $n_{\text{rep}} = 400$ times the experiment that consisted in generating data according to (1) and computing estimators $\hat{\pi}_{n,m}^{\text{LSS}}$ and $\hat{\pi}_{n,m}^{\text{LSL}}$ defined respectively by (10) and (7). We then computed, for each pair (a, b) and for each estimator LSS and LSL, the percentage of successful detection among n_{rep} repetitions.

The obtained detection regions are depicted in Figure 3 in the form of heatmaps. This visualisation allows us to grasp the forms of the detection regions for the specific choice of parameters considered in this example. The first observation is that LSL is clearly superior to LSS for all the considered values of the dimension. Second, we clearly see the deterioration of the detection region when the dimension d becomes larger. Third, the values of $\bar{\kappa}_{\text{in-out}}$ used in the plots are at least one order of magnitude larger than those of $\bar{\kappa}_{\text{in-in}}$. This is in line with the claim of Theorem 4. We also observe in these pictures that successful detection occurs when $\bar{\kappa}_{\text{in-out}}$ is larger than some threshold even if $\bar{\kappa}_{\text{in-in}}$ is small. This must be a nice feature of LSL and LSS in this specific example, which unfortunately does not generalize to other examples as shown by our theoretical results.

Experiment 3: Real data example This experiment is conducted on the IMC-PT 2020 dataset from Jin et al. (2020) that consists of images of 16 different scenes with corresponding 3D point-clouds of landmarks, which are used to obtain (pseudo) ground-truth local keypoint matchings. For a given scene, we sampled 1000 pairs of distinct images of the same landmark with different camera locations, angles, weather conditions etc. For each image pair we generated 2D keypoints from original set of 3D points (note that the same 3D point appears in both of the images, so we have ground truth keypoint matching between 2 images). Subsequently, we computed SIFT descriptors (Lowe, 2004) for every keypoint in images using Python OpenCV interface (Itseez, 2015). Some pairs of images being more challenging than others, we split the dataset into two sets of image pairs in order to gain more understanding on the behaviour of the algorithms. The challenging pairs are those for which the OpenCV default matching algorithm has accuracy less than 0.5. Then, for every image pair, we fixed randomly chosen 100 keypoints in the first image (and corresponding keypoints in the second image) and added outliers to the second image from the remaining keypoints. The outlier rate is chosen to be between 0% to 70%. Finally, 3 descriptor matching algorithms were applied (OpenCV default matching algorithm, LSS and LSL). Note that $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{\#}$ from (1) are unknown, hence LSNS is not applicable. One can consider using the estimates $\hat{\boldsymbol{\sigma}}$ and $\hat{\boldsymbol{\sigma}}^{\#}$ instead of $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^{\#}$ in (5), but this is beyond the scope of this paper.. Further details on this experiment along with some additional results are deferred to the supplemental material (Appendix D).

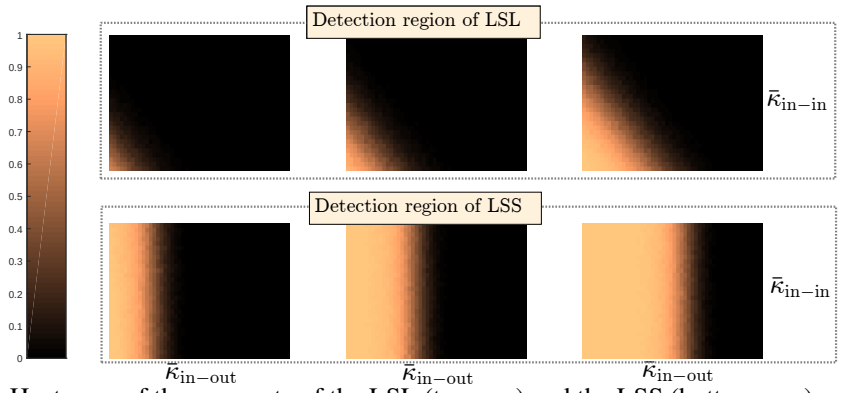

Figure 3: Heatmaps of the error rate of the LSL (top row) and the LSS (bottom row) estimators in Experiment 2. We chose $n = 100$, $m = 120$ and $d \in \{10, 20, 40\}$ from left to right. The parameter a representing the scale of $\bar{\kappa}_{\text{in-in}}$ and corresponding to y-axis varies from 0.02 to 0.08, whereas b representing the scale of $\bar{\kappa}_{\text{in-out}}$ and corresponding to x-axis varies from 0.3 to 10. Dark colour means that probability of successful detection is close to 1 (error rate close to zero).

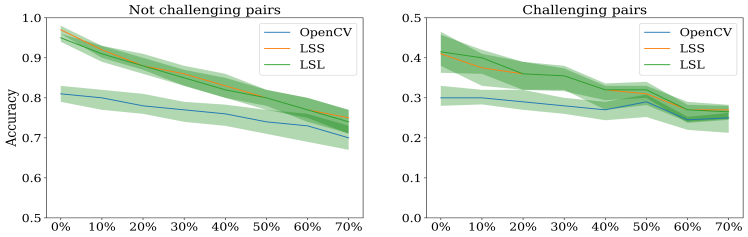

Figure 4: The estimation accuracy measured in the Hamming loss of the estimated matching in Exp. 3 for different values of the outlier rate, $(m-n)/n$, varying from 0% to 70%. The medians of estimation accuracy both for challenging pairs (right plot) and simple pairs (left plot) of images from Temple Nara scene was computed using OpenCV, LSS and LSL matchers. The green region represents the interquartile range (lower and upper bounds being 25% and 75% percentiles, respectively).

The median estimation accuracy measured in the Hamming loss—for the image pairs from Temple Nara Japan scene—is plotted in Figure 4. Similar results for other scenes are presented in the supplemental material. The error bars with borderlines corresponding to 75% and 25% percentiles are also displayed. The first observation is that LSS and LSL outperform the OpenCV matcher in terms of Hamming distance. Second, the task of feature matching becomes harder with the growth of outlier rate, with a deterioration that seems to be linear in the rate of outliers. Notice that in this experiment the outliers can be very similar to the inliers and the separation condition imposed on $\bar{\kappa}_{\text{in-out}}$, of Theorems 2 and 4, is violated. This implies that the larger the outlier rate the harder it is to find the correct match, causing a larger number of mistakes.

6 Conclusion

We have investigated the detection regions in the problem of estimation of the matching map between two sequences of noisy vectors. We have shown that the presence of outliers in one of the two sequences has a strong negative impact on the detection region. Interestingly, this negative impact is mitigated in the regime of mild heteroscedasticity, *i.e.*, when noise variances are of the same order of magnitude. In the extremely favorable case of homoscedastic noise (all variances are equal), the presence of outliers does not make the problem any harder, provided that the outliers are at least as different from inliers as two distinct inliers are different one another. Precise forms of the detection region in these different cases can be found in Table 1. The results of the numerical experiments conducted on both synthetic and real data confirm our findings and, furthermore, show the good behaviour of the LSL estimator in terms of its robustness to noise and to outliers, not only in the problem of detection but also in the problem of estimation. In the future, we plan to investigate the case when both sequences contain outliers and to obtain theoretical guarantees on the estimation error measured by the Hamming distance.

References

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

Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., and Tai, C.-L. (2020). 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)*.

Blanchard, G., Carpentier, A., and Gutzeit, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electron. J. Stat.*, 12(2):3713–3735.

Burnashev, M. V. (1979). On the minimax detection of an inaccurately known signal in a white Gaussian noise background. *Theory Probab. Appl.*, 24:107–119.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). Brief: Binary robust independent elementary features. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *Computer Vision – ECCV 2010*, pages 778–792, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chen, J., Tian, J., Lee, N., Zheng, J., Smith, R. T., and Laine, A. F. (2010). A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Transactions on Biomedical Engineering*, 57(7):1707–1718.

Collier, O. (2012). Minimax hypothesis testing for curve registration. In Lawrence, N. D. and Girolami, M. A., 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.

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

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

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Stat.*, 40(5):2667–2696.

Duff, I. S. and Koster, J. (2001). On algorithms for permuting large entries to the diagonal of a sparse matrix. *SIAM Journal on Matrix Analysis and Applications*, 22(4):973–996.

Flammarion, N., Mao, C., and Rigollet, P. (2019). Optimal rates of statistical seriation. *Bernoulli*, 25(1):623–653.

Gao, C. and Zhang, A. Y. (2019). Iterative algorithm for discrete structure recovery. *arXiv preprint arXiv:1911.01018*.

Harwood, B. and Drummond, T. (2016). Fanng: Fast approximate nearest neighbour graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5713–5722.

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

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

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

Jiang, Z., Xie, L., Deng, X., Xu, W., and Wang, J. (2016). Fast nearest neighbor search in the hamming space. In Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., and Liu, X., editors, *MultiMedia Modeling*, pages 325–336, Cham. Springer International Publishing.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K. M., and Trulls, E. (2020). Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*.

Juditsky, A. and Nemirovski, A. (2020). *Statistical inference via convex optimization*. Princeton, NJ: Princeton University Press.

Kuhn, H. W. (1955). The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Kuhn, H. W. (2010). *The Hungarian Method for the Assignment Problem*, pages 29–47. Springer Berlin Heidelberg, Berlin, Heidelberg.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.

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

Ma, J., Jiang, X., Fan, A., Jiang, J., and Yan, J. (2021). Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129.

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

Malkov, Y. A. and Yashunin, D. A. (2020). 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.

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

Mao, C., Weed, J., and Rigollet, P. (2018). Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory*, pages 821–847. PMLR.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Ndaoud, M. and Tsybakov, A. B. (2020). Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory*, 66(4):2517–2532.

Pananjady, A. and Samworth, R. J. (2020). Isotonic regression with unknown permutations: Statistics, computation, and adaptation. *arXiv preprint arXiv:2009.02609*.

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

Ramdas, A., Isenberg, D., Singh, A., and Wasserman, L. A. (2016). 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.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. (2021). A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 67(6):4162–4184.

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

Tian, Y., Balntas, V., Ng, T., Barroso-Laguna, A., Demiris, Y., and Mikolajczyk, K. (2020). D2d: Keypoint extraction with describe to detect approach. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wang, K., Zhu, N., Cheng, Y., Li, R., Zhou, T., and Long, X. (2018). Fast feature matching based on r -nearest k -means searching. *CAAI Transactions on Intelligence Technology*, 3(4):198–207.

Wang, X. (2011). A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality. In *The 2011 International Joint Conference on Neural Networks*, pages 1293–1299.

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

Wolfer, G. and Kontorovich, A. (2020). Minimax testing of identity to a reference ergodic markov chain. In Chiappa, S. and Calandra, R., 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.

Xing, X., Liu, M., Ma, P., and Zhong, W. (2020). Minimax nonparametric parallelism test. *Journal of Machine Learning Research*, 21(94):1–47.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 1 and Section 2 for informal statements, precise assumptions/limitations are presented in Section 3.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplemental material (Appendix B).
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and supplemental material (Appendix D).
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5 and supplemental material (Appendix D).
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]