Journal of Machine Learning Research 17 (2016) 1-31 Submitted 10/13; Revised 3/15; Published 3/16

Minimax Rates in Permutation Estimation for Feature Matching

Olivier Collier OLIVIER.COLLIER@ENPC.FR
Imagine- LIGM
Université Paris EST
Marne-la-Vallée, FRANCE

Arnak S. Dalalyan ARNAK.DALALYAN@ENSAE.FR
Laboratoire de Statistique
ENSAE - CREST
Malakoff, FRANCE

Editor: Gabor Lugosi

Abstract

The problem of matching two sets of features appears in various tasks of computer vision and can be often formalized as a problem of permutation estimation. We address this problem from a statistical point of view and provide a theoretical analysis of the accuracy of several natural estimators. To this end, the minimax rate of separation is investigated and its expression is obtained as a function of the sample size, noise level and dimension of the features. We consider the cases of homoscedastic and heteroscedastic noise and establish, in each case, tight upper bounds on the separation distance of several estimators. These upper bounds are shown to be unimprovable both in the homoscedastic and heteroscedastic settings. Interestingly, these bounds demonstrate that a phase transition occurs when the dimension d of the features is of the order of the logarithm of the number of features n. For $d = O(\log n)$, the rate is dimension free and equals $\sigma(\log n)^{1/2}$, where σ is the noise level. In contrast, when d is larger than $c \log n$ for some constant $c > 0$, the minimax rate increases with d and is of the order of $\sigma(d \log n)^{1/4}$. We also discuss the computational aspects of the estimators and provide empirical evidence of their consistency on synthetic data. Finally, we show that our results extend to more general matching criteria.

Keywords: permutation estimation, minimax rate of separation, feature matching

1. Introduction

In this paper, we present a rigorous statistical analysis of the problem of permutation estimation and multiple feature matching from noisy observations. More precisely, let $\{X_1, \dots, X_n\}$ and $\{X_1^\#, \dots, X_m^\#\}$ be two sets of vectors from $\mathbb{R}^d$, hereafter referred to as noisy features, containing many matching elements. That is, for many X_i's there is a $X_j^\#$ such that X_i and $X_j^\#$ coincide up to an observation noise (or measurement error). Our goal is to estimate an application $\pi^* : \{1, \dots, n\} \to \{1, \dots, m\}$ for which each X_i matches with $X^\#_{\pi^*(i)}$ and to provide tight conditions which make it possible to accurately recover π^* from data.

In order to define a statistical framework allowing us to compare different estimators of π^*, we confine[1] our attention to the case $n = m$, that is when the two sets of noisy features have equal sizes. Furthermore, we assume that there exists a unique permutation of $\{1, \dots, n\}$, denoted π^*, leading to pairs of features $(X_i, X^{\#}_{\pi^*(i)})$ that match up to a measurement error. In such a situation, it is clearly impossible to recover the true permutation π^* if some features within the set $\{X_1, \dots, X_n\}$ are too close. Based on this observation, we propose to measure the quality of a procedure of permutation estimation by the minimal distance between pairs of different features for which the given procedure is still consistent. This quantity will be called *separation distance* and will be the main concept of interest in the present study. In this respect, the approach we adopted is close in spirit to the minimax theory of hypotheses testing (see, for instance, Spokoiny (1996); Ingster and Suslina (2003)).

1.1 A Motivating Example: Feature Matching in Computer Vision

Many tasks of computer vision, such as object recognition, motion tracking or structure from motion, are currently carried out using algorithms that contain a step of feature matching, cf. Szeliski (2010); Hartley and Zisserman (2003). The features are usually local descriptors that serve to summarize the images. The most famous examples of such features are perhaps SIFT (Lowe, 2004) and SURF (Bay et al., 2008). Once the features have been computed for each image, an algorithm is applied to match features of one image to those of another one. The matching pairs are then used for estimating the deformation of the object, for detecting the new position of the followed object, for creating a panorama, etc. In this paper, we are interested in simultaneous matching of a large number of features. The main focus is on the case when the two sets of features are extracted from the images that represent the same scene with a large overlap, and therefore the sets of features are (nearly) of the same size and every feature in the first image is also present in the second one. This problem is made more difficult by the presence of noise in the images, and thus in the features as well. Typically, due to the high resolution of most images, the number of features is large and their dimension is relatively large as well (128 for SIFT and 64 for SURF). It is therefore important to characterize the behavior of various matching procedures as a function of the number of features, the dimension and the noise level.

1.2 Main Contributions

We consider four procedures of permutation estimation that naturally arise in this context. The first one is a greedy procedure that sequentially assigns to each feature X_i the closest feature $X^{\#}_j$ among those features that have not been assigned at an earlier step. The three other estimators are defined as minimizers of the (profiled-)log-likelihood under three different modeling assumptions. These three modeling assumptions are that the noise level is constant across all the features (homoscedastic noise), that the noise level is variable (heteroscedastic noise) but known and that the noise level is variable and unknown. The corresponding estimators are respectively called least sum of squares (LSS) estimator, least sum of normalized squares (LSNS) estimator and least sum of logarithms (LSL) estimator.

1. These assumptions are imposed for the purpose of getting transparent theoretical results and are in no way necessary for the validity of the considered estimation procedures, as discussed later in the paper.

We first consider the homoscedastic setting and show that all the considered estimators are consistent under similar conditions on the minimal distance between distinct features κ. These conditions state that κ is larger than some function of the noise level σ, the sample size n and the dimension d. This function is the same for the four aforementioned procedures and is given, up to a multiplicative factor, by

$$\kappa^*(\sigma, n, d) = \sigma \max((\log n)^{1/2}, (d \log n)^{1/4}). \tag{1}$$

Then, we prove that this expression provides the optimal rate of the separation distance in the sense that for some absolute constant c if $\kappa \leq c\kappa^*(\sigma, n, d)$ then there is no procedure capable of consistently estimating π^*.

In the heteroscedastic case, we provide an upper bound on the identifiability threshold ensuring the consistency of the LSNS and LSL estimators. Up to a proper normalization by the noise level, this bound is of the same form as (1) and, therefore, the ignorance of the noise level does not seriously affect the quality of estimation. Furthermore, the LSL estimator is easy to adapt to the case $n \neq m$ and is robust to the presence of outliers in the features. We carried out a small experimental evaluation that confirms that in the heteroscedastic setting the LSL estimator is as good as the LSNS (pseudo-) estimator and that they outperform the two other estimators: the greedy estimator and the least sum of squares. We also argue that the three estimators stemming from the maximum likelihood methodology are efficiently computable either by linear programming or by the Hungarian algorithm.

Note that different loss functions may be used for measuring the distance between an estimated permutation and the true one. Most results of this paper are established for the 0-1 loss, which equals one if the estimator and the true permutation differ at least at one location and equals 0 otherwise. However, it is of interest to analyze the situation with the Hamming distance as well, since it amounts to controlling the proportion of the mismatched features and, hence, offers a more graduated evaluation of the quality of estimation. We show that in the case of the Hamming distance, in the regime of moderately large dimension (*i.e.,* $d \geq c \log n$ for some constant $c > 0$) the rate of separation is exactly the same as in the case of the 0-1 distance. The picture is more complex in the regime of small dimensionality $d = o(\log(n))$, in which we get the same upper bound as for the 0-1 loss but the lower bound is expressed in terms of the logarithm of the packing number of an ℓ_2 ball of the symmetric group. We conjecture that this quantity is of the order of $n \log(n)$ and check this conjecture for relatively small values of n. If this conjecture is correct, our lower bound coincides up to a multiplicative factor with the upper bound.

Finally, let us mention that some of the results of the present work have been presented in the AI-STATS 2013 conference and published in the proceedings (Collier and Dalalyan, 2013).

1.3 Plan of the Paper

We introduce in Section 2 a model for the problem of matching two sets of features and of estimation of a permutation. The estimating procedures analyzed in this work are presented in Section 3, whereas their performances in terms of rates of separation distance are described in Section 4. In Section 5, computational aspects of the estimating procedures are discussed

while Section 6 is devoted to the statement of some extensions of our results. We report in Section 7 the results of some numerical experiments. The proofs of the theorems and of the lemmas are postponed to Sections 9 and 10, respectively.

2. Notation and Problem Formulation

We begin with formalizing the problem of matching two sets of features $\{X_1, \dots, X_n\}$ and $\{X_1^\#, \dots, X_m^\#\}$ with $n, m \geq 2$. In what follows we assume that the observed features are randomly generated from the model

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

where

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

The task of feature matching consists in finding a bijection π^* between the largest possible subsets S_1 and S_2 of $\{1, \dots, n\}$ and $\{1, \dots, m\}$ respectively, such that

$$\forall\ i \in S_2, \quad \theta_i^\# \equiv \theta_{\pi^*(i)}, \tag{3}$$

where $\equiv$ is an equivalence relation that we call *matching criterion*. The features that do not belong to S_1 or S_2 are called *outliers*. To ease presentation, we mainly focus on the case where the matching criterion is the equality of two vectors. However, as discussed in Section 6.2, most results carry over the equivalence corresponding to equality of two vectors transformed by a given linear transformation. Furthermore, it turns out that statistical inference for the matching problem is already quite involved when no outlier is present in the data. Therefore, we make the following assumption

$$m = n \qquad \text{and} \qquad S_1 = S_2 = \{1, \dots, n\}. \tag{4}$$

Note however that the procedures we consider below admit natural counterparts in the setting with outliers. We will also restrict ourselves to noise levels satisfying some constraints. The two types of constraints we consider, referred to as homoscedastic and heteroscedastic setting, correspond to the relations $\sigma_i = \sigma_i^\# = \sigma$, $\forall i = 1, \dots, n$, and $\sigma_{\pi^*(i)} = \sigma_i^\#$, $\forall i = 1, \dots, n$.

In this formulation, the data generating distribution is defined by the (unknown) parameters $\boldsymbol{\theta}$, $\boldsymbol{\sigma} = (\sigma_i, \dots, \sigma_n)$ and π^*. In the problem of matching, 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 vector $(X_1, \dots, X_n, X_1^\#, \dots, X_n^\#)$ defined by (2) under the conditions (3) and (4). We write $\mathbf{E}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}$

for the expectation with respect to $\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}$. The symmetric group, *i.e.,* the set of all permutations of $\{1,\dots,n\}$, will be denoted by $\mathfrak{S}_n$.

We will use two measures of quality for quantifying the error of an estimator $\hat{\pi}$ of the permutation π^*. These errors are defined as the 0-1 distance and the normalized Hamming distance between $\hat{\pi}$ and π^*, given by

$$\delta_{0\text{-}1}(\hat{\pi},\pi^*) \triangleq \mathbb{1}_{\{\hat{\pi}\neq\pi^*\}}, \qquad \delta_H(\hat{\pi},\pi^*) \triangleq \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{\hat{\pi}(k)\neq\pi^*(k)\}}. \tag{5}$$

Our ultimate goal is to design estimators that have an expected error smaller than a prescribed level α under the weakest possible conditions on the nuisance parameter $\boldsymbol{\theta}$. The estimation of the permutation or, equivalently, the problem of matching is more difficult when the features are hardly distinguishable. To quantify this phenomenon, we introduce the separation distance $\kappa(\boldsymbol{\theta})$ and the relative separation distance $\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})$, which measure the minimal distance between distinct features and the minimal distance-to-noise ratio, respectively. The precise definitions are

$$\kappa(\boldsymbol{\theta}) \triangleq \min_{i\neq j} \|\theta_i-\theta_j\|, \qquad \bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma}) \triangleq \min_{i\neq j} \frac{\|\theta_i-\theta_j\|}{(\sigma_i^2+\sigma_j^2)^{1/2}}. \tag{6}$$

We will see that in the heteroscedastic case the last quantity is more suitable for characterizing the behavior of the estimators than the first one.

Clearly, if $\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})=0$ and σ_i's are all equal, then 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, the condition $\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})>0$ is necessary for the existence of consistent estimators of π^*. Furthermore, good estimators are those consistently estimating π^* even if $\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})$ is small. To give a precise sense to these considerations, let $\alpha\in(0,1)$ be a prescribed tolerance level and let us call *perceivable separation distance* of a given estimation procedure $\hat{\pi}$ the quantity

$$\bar{\kappa}_\alpha(\hat{\pi}) \triangleq \inf\Big\{\kappa>0 \,\Big|\, \max_{\pi\in\mathfrak{S}_n} \sup_{\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})>\kappa} \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi}(\hat{\pi}\neq\pi)\le\alpha\Big\},$$

where we skip the dependence on n, d and $\boldsymbol{\sigma}$. Here, the perceivable separation distance is defined with respect to the 0-1 distance, the corresponding definition for the Hamming distance is obtained by replacing $\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi}(\hat{\pi}\neq\pi)$ by $\mathbf{E}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi}[\delta_H(\hat{\pi},\pi)]$. Finally, we call *minimax separation distance* the smallest possible perceivable separation distance achieved by an estimator $\hat{\pi}$, *i.e.,*

$$\bar{\kappa}_\alpha \triangleq \inf_{\hat{\pi}} \ \bar{\kappa}_\alpha(\hat{\pi}), \tag{7}$$

where the infimum is taken over all possible estimators of π^*. In the following sections, we establish nonasymptotic upper and lower bounds on the minimax separation distance which coincide up to a multiplicative constant independent of n, d and $\boldsymbol{\sigma}$. We also show that a suitable version of the maximum profiled likelihood estimator is minimax-separation-rate-optimal both in the homoscedastic and heteroscedastic settings.

3. Estimation Procedures

As already mentioned, we will consider four estimators. The simplest one, called greedy algorithm and denoted by π^{gr} is defined as follows: $\pi^{\mathrm{gr}}(1) = \arg\min_{j\in\{1,\dots,n\}} \|X_j - X_1^\#\|$ and, for every $i \in \{2,\dots,n\}$, recursively define

$$\pi^{\mathrm{gr}}(i) \triangleq \arg \min_{j\notin\{\pi^{\mathrm{gr}}(1),\dots,\pi^{\mathrm{gr}}(i-1)\}} \|X_j - X_i^\#\|. \tag{8}$$

A drawback of this estimator is that it is not symmetric: the resulting permutation depends on the initial numbering of the features. However, we will show that this estimator is minimax-separation-rate-optimal in the homoscedastic setting.

A common approach for avoiding incremental estimation and taking into consideration all the observations at the same time consists in defining the estimator $\hat\pi$ as a maximizer of the profiled likelihood. In the homoscedastic case, which is the first setting we studied, the computations lead to the estimator

$$\pi^{\mathrm{LSS}} \triangleq \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^n \|X_{\pi(i)} - X_i^\#\|^2. \tag{9}$$

which will be referred to as the *Least Sum of Squares* (LSS) estimator.

The LSS estimator takes into account the distance between the observations irrespectively of the noise levels. The fact of neglecting the noise levels, while harmless in the homoscedastic setting, turns out to cause serious loss of efficiency in terms of the perceivable distance of separation in the setting of heteroscedastic noise. Yet, in the latter setting, the distance between the observations is not as relevant as the signal-to-noise ratio $\|\theta_i - \theta_j\|^2/(\sigma_i^2 + \sigma_j^2)$. Indeed, when the noise levels are small, two noisy but distinct vectors are easier to distinguish than when the noise levels are large.

The computation of the maximum likelihood estimator in the heteroscedastic case with known noise levels also suggests that the signal-to-noise ratio should be taken into account. In this setting, the likelihood maximization leads to the *Least Sum of Normalized Squares* (LSNS) estimator

$$\pi^{\mathrm{LSNS}} \triangleq \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^n \frac{\|X_{\pi(i)} - X_i^\#\|^2}{\sigma_{\pi(i)}^2 + {\sigma_i^\#}^2}. \tag{10}$$

We will often call the LSNS a pseudo-estimator, to underline the fact that it requires the knowledge of the noise levels σ_i which are generally unavailable.

In the general setting, when no information on the noise levels is available, the likelihood is maximized over all nuisance parameters (features and noise levels). But this problem is underconstrained, and the result of this maximization is $+\infty$. This can be circumvented by assuming a proper relation between the noise levels. As mentioned earlier, we chose the assumption

$$\forall\ i \in \{1,\dots,n\}, \quad \sigma_i^\# = \sigma_{\pi^*(i)}. \tag{11}$$

The maximum likelihood estimator under this constraint is the *Least Sum of Logarithms* (LSL) defined as

$$\pi^{\mathrm{LSL}} \triangleq \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^n \log \|X_{\pi(i)} - X_i^\#\|^2. \tag{12}$$

We will prove that this estimator is minimax-rate-optimal both in the homoscedastic and the heteroscedastic cases.

4. Performance of the Estimators

The purpose of this section is to assess the quality of the aforementioned procedures. To this end, we present conditions for the consistency of these estimators in the form of upper bounds on their perceivable separation distance. Furthermore, to compare this bounds with the minimax separation distance, we establish lower bounds on the latter and prove that it coincides up to a constant factor with the perceivable separation distance of the LSL estimator.

4.1 Homoscedastic Setup

We start by considering the homoscedastic case, in which upper and lower bounds matching up to a constant are obtained for all the estimators introduced in the previous section.

Theorem 1 *Let $\alpha \in (0,1)$ be a tolerance level and $\sigma_j = \sigma_j^\# = \sigma$ for all $j \in \{1,\dots,n\}$. If $\hat{\pi}$ denotes any one of the estimators (8)-(12), then*

$$\bar{\kappa}_\alpha(\hat{\pi}) \le 4 \max\left\{ \left(2 \log \frac{8n^2}{\alpha}\right)^{1/2}, \left(d \log \frac{4n^2}{\alpha}\right)^{1/4} \right\}.$$

An equivalent way of stating this result is that if

$$\kappa = 4\sigma \max\left\{ \left(4 \log \frac{8n^2}{\alpha}\right)^{1/2}, \left(4d \log \frac{4n^2}{\alpha}\right)^{1/4} \right\}$$

and Θ_κ is the set of all $\boldsymbol{\theta} \in \mathbb{R}^{n\times d}$ such that $\kappa(\boldsymbol{\theta}) \ge \kappa$, then

$$\max_{\pi^* \in \mathfrak{S}_n} \sup_{\boldsymbol{\theta} \in \Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}(\hat{\pi} \ne \pi^*) \le \alpha$$

for all the estimators defined in Section 3. Note that this result is nonasymptotic. Furthermore, it tells us that the perceivable separation distance of the procedures under consideration is at most of the order of

$$\max\left\{ (\log n)^{1/2}, (d \log n)^{1/4} \right\}. \tag{13}$$

It is interesting to observe that there are two regimes in this rate, the boundary of which corresponds to the case where d is of the order of $\log n$. For dimensions that are significantly smaller than $\log n$, the perceivable distance of separation is dimension free. On the other hand, when d is larger than $c \log n$ for some absolute constant $c > 0$, the perceivable distance of separation deteriorates with increasing d at the polynomial rate $d^{1/4}$. However, this result does not allow us to deduce any hierarchy between the four estimators, since it provides the same upper bound for all of them. Moreover, as stated in the next theorem, this bound is optimal up to a multiplicative constant.

Theorem 2 *Assume that $n \geq 6$, Θ_κ is the set of all $\boldsymbol{\theta} \in \mathbb{R}^{n\times d}$ such that $\kappa(\boldsymbol{\theta}) \geq \kappa$. Then there exist two absolute constants $c, C > 0$ such that for*

$$\kappa = 2^{-5/2}\sigma \max\Big\{(\log n)^{1/2}, (cd\log n)^{1/4}\Big\},$$

the following lower bound holds

$$\inf_{\hat\pi} \max_{\pi^*\in\mathfrak{S}_n} \sup_{\boldsymbol{\theta}\in\Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}\big(\hat\pi \neq \pi^*\big) > C,$$

where the infimum is taken over all permutation estimators.

An equivalent way of stating this result is to say that, for any $\alpha \leq C$, the minimax distance of separation satisfies the inequality

$$\bar\kappa_\alpha \geq \frac{1}{8}\max\big\{(\log n)^{1/2}, (cd\log n)^{1/4}\big\}.$$

Combined with Theorem 1, this implies that the minimax rate of separation is given by the expression $\max\big\{(\log n)^{1/2}, (d\log n)^{1/4}\big\}$.

In order to avoid any possible confusion, we emphasize that the rate obtained in this and subsequent sections concerns the speed of decay of the separation distance and not the estimation risk measured by $\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}\big(\hat\pi \neq \pi^*\big)$. For the latter, considering κ as fixed, one readily derives from Theorem 1 that

$$\max_{\pi^*\in\mathfrak{S}_n} \sup_{\boldsymbol{\theta}\in\Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}\big(\hat\pi \neq \pi^*\big) \leq \max\Big\{8n^2\exp\Big(-\frac{\kappa^2}{2^6\sigma^2}\Big), 4n^2\exp\Big(-\frac{\kappa^4}{2^{10}d\sigma^4}\Big)\Big\} \qquad (14)$$

for the four estimators $\hat\pi$ defined in the previous section by equations (8)-(12). We do not know whether the right-hand side of this inequality is the correct rate for the minimax risk $R^{\mathrm{mmx}} = \inf_{\hat\pi}\sup_{(\boldsymbol{\theta},\pi^*)\in\Theta_\kappa\times\mathfrak{S}_n} \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}\big(\hat\pi \neq \pi^*\big)$. In fact, one can adapt the proof of Theorem 2 to get a lower bound on R^{mmx} which is of the same form as the right-hand side in (14), but with constants 2^6 and 2^{10} replaced by smaller ones. The ratio of such a lower bound and the upper bound in (14) tends to 0 and, therefore, does not provide the minimax rate of estimation, in the most common sense of the term. However, one may note that the minimax rate of separation established in this work is the analogue of the minimax rate of estimation of the Bahadur risk, see Bahadur (1960); Korostelev (1996); Korostelev and Spokoiny (1996).

4.2 Heteroscedastic Setup

We switch now to the heteroscedastic setting, which allows us to discriminate between the four procedures. Note that the greedy algorithm, the LSS and the LSL have a serious advantage over the LSNS since they can be computed without knowing the noise levels $\boldsymbol{\sigma}$.

Theorem 3 *Let $\alpha \in (0,1)$ and condition (11) be fulfilled. If $\hat\pi$ is either π^{LSNS} (if the noise levels σ_i are known) or π^{LSL} (when the noise levels are unknown), then*

$$\bar\kappa_\alpha(\hat\pi) \leq 4\max\Big\{\Big(2\log\frac{8n^2}{\alpha}\Big)^{1/2}, \Big(d\log\frac{4n^2}{\alpha}\Big)^{1/4}\Big\}.$$

This result tells us that the performance of the LSNS and LSL estimators, measured in terms of the order of magnitude of the separation distance, is not affected by the heteroscedasticity of the noise levels. Two natural questions arise: 1) is this performance the best possible over all the estimators of π^* and 2) is the performance of the LSS and the greedy estimator as good as that of the LSNS and LSL?

To answer the first question, we start by discarding the degenerate situations where faster rates can be achieved by estimators that are heavily based on the knowledge of the noise levels $\boldsymbol{\sigma}$. In fact, although considering $\boldsymbol{\sigma}$ as known limits considerably the scope of applications of the methods, the definition (7) involves a minimum over all possible estimators $\hat{\pi}$ which are allowed to depend on $\boldsymbol{\sigma}$. For some specific noise levels $\boldsymbol{\sigma}$, from a purely theoretical point of view, knowing $\boldsymbol{\sigma}$ may lead to a substantially better minimax rate and even to a separation equal to zero, which corresponds to an estimator that has no real practical interest. Indeed, under condition (11), it is possible to estimate the permutation π^* by fitting the noise levels. For instance, we can define an estimator $\hat{\pi}$ as follows: $\hat{\pi}(1) = \arg\min_{i=1,\dots,n} \left|\frac{1}{2d}\|X_i - X_1^\#\|^2 - \sigma_i^2\right|$ and, recursively, for every $j \in \{2,\dots,n\}$,

$$\hat{\pi}(j) = \arg \min_{i \notin \{\hat{\pi}(1),\dots,\hat{\pi}(j-1)\}} \left|\frac{1}{2d}\|X_i - X_j^\#\|^2 - \sigma_i^2\right|.$$

This provides an accurate estimator of π^*—for vectors $\{\theta_j\}$ that are very close—as soon as the noise levels are different enough from each other. In particular, the estimated permutation $\hat{\pi}$ coincides with the true permutation π^* on the event

$$\forall i \in \{1,\dots,n\}, \qquad \left|\frac{1}{2d}\|X_{\pi^*(i)} - X_i^\#\|^2 - \sigma_{\pi^*(i)}^2\right| < \min_{j \neq \pi^*(i)} \left|\frac{\|X_j - X_i^\#\|^2}{2d} - \sigma_j^2\right|,$$

which includes the event:

$$\left|\frac{1}{2d}\|X_{\pi^*(i)} - X_i^\#\|^2 - \sigma_{\pi^*(i)}^2\right| + \left|\frac{\|X_j - X_i^\#\|^2}{2d} - \frac{\sigma_j^2 + \sigma_{\pi^*(i)}^2}{2}\right| < \frac{|\sigma_j^2 - \sigma_{\pi^*(i)}^2|}{2} \qquad (15)$$

for all $(i,j) \in \{1,\dots,n\}^2$ such that $j \neq \pi^*(i)$. Using standard bounds on the tails of the χ^2 distribution recalled in Lemma 10 of Section 9, in the case when all the vectors θ_j are equal, one can check that the left-hand side in (15) is of the order of $(\sigma_{\pi^*(i)}^2 + \sigma_j^2)\max\left\{\sqrt{(\log n)/d}, (\log n)/d\right\}$. This implies that we can consistently identify the permutation when

$$\forall (i,j) \in \{1,\dots,n\}^2, \quad i \neq j, \qquad \left|\frac{\sigma_j^2}{\sigma_i^2} - 1\right| \gg \max\left\{\sqrt{\frac{\log n}{d}}, \frac{\log n}{d}\right\},$$

even if the separation distance is equal to zero. In order to discard such kind of estimators from the competition in the procedure of determining the minimax rates, we restrict our attention to the noise levels for which the values $|\frac{\sigma_j^2}{\sigma_i^2} - 1|$ are not larger than $C\max\left\{\sqrt{(\log n)/d}, (\log n)/d\right\}$ for $j \neq i$.

Theorem 4 *Assume that $n \geq 6$, $\bar{\Theta}_\kappa$ is the set of all $\boldsymbol{\theta} \in \mathbb{R}^{n\times d}$ such that $\bar{\kappa}(\boldsymbol{\theta}, \boldsymbol{\sigma}) \geq \kappa$ and*

$$\frac{\max_i \sigma_i^2}{\min_i \sigma_i^2} - 1 \leq \max\left\{\sqrt{\frac{\log n}{16d}}, \frac{\log n}{16d}\right\}.$$

Then there exist two constants $c, C > 0$ such that $\kappa < (1/8)\max\{(\log n)^{1/2}, (cd\log n)^{1/4}\}$, implies that

$$\inf_{\hat\pi}\max_{\pi^*\in\mathfrak{S}_n}\sup_{\boldsymbol{\theta}\in\bar\Theta_\kappa}\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}(\hat\pi\neq\pi^*) > C,$$

where the infimum is taken over all permutation estimators.

It is clear that the constants c and C of the previous theorem are closely related. The inspection of the proof shows that, for instance, if $c \le 1/20$ then C is larger than 17%.

Let us discuss now the second question raised earlier in this section and concerning the theoretical properties of the greedy algorithm and the LSS under heteroscedasticity. In fact, the perceivable distances of separation of these two procedures are significantly worse than those of the LSNS and the LSL especially for large dimensions d. We state the corresponding result for the greedy algorithm, a similar conclusion being true for the LSS as well. The superiority of the LSNS and LSL is also confirmed by the numerical simulations presented in Section 7 below. In the next theorem and in the sequel of the paper, we denote by id the identity permutation defined by $id(i) = i$ for all $i \in \{1, \dots, n\}$.

Theorem 5 *Assume that $d \ge 225\log 6$, $n = 2$, $\sigma_1^2 = 3$ and $\sigma_2^2 = 1$. Then the condition $\kappa < 0.1(2d)^{1/2}$ implies that*

$$\sup_{\boldsymbol{\theta}\in\bar\Theta_\kappa}\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},id}(\pi^{gr}\neq id) \ge 1/2.$$

This theorem shows that if d is large, the necessary condition for $\pi^{\rm gr}$ to be consistent is much stronger than the one obtained for $\pi^{\rm LSL}$ in Theorem 3. Indeed, for the consistency of $\pi^{\rm gr}$, κ needs to be at least of the order of $d^{1/2}$, whereas $d^{1/4}$ is sufficient for the consistency of $\pi^{\rm LSL}$. Hence, the maximum likelihood estimators LSNS and LSL that take into account noise heteroscedasticity are, as expected, more interesting than the simple greedy estimator[2].

5. Computational Aspects

At first sight, the computation of the estimators (9)-(12) requires to perform an exhaustive search over the set of all possible permutations, the number of which, $n!$, is prohibitively large. This is in practice impossible to do on a standard PC as soon as $n \ge 20$. In this section, we show how to compute these (maximum likelihood) estimators in polynomial time using, for instance, algorithms of linear programming[3].

To explain the argument, let us consider the LSS estimator

$$\pi^{\rm LSS} = \arg\min_{\pi\in\mathfrak{S}_n}\sum_{i=1}^n \|X_{\pi(i)} - X_i^\#\|^2.$$

2. It should be noted that the conditions of Theorem 5 are not compatible with those of Theorem 4. Hence, strictly speaking, the former does not imply that the greedy estimator is not minimax under the conditions of the latter.
3. The idea of reducing the problem of permutation estimation to a linear program has been already used in the literature, however, without sufficient theoretical justification: see, for instance, Jebara (2003).

For every permutation π, we denote by P^π the $n \times n$ permutation matrix with coefficients $P^\pi_{ij} = \mathbb{1}_{\{j=\pi(i)\}}$. Then we can give the equivalent formulation

$$\pi^{\mathrm{LSS}} = \arg\min_{\pi \in \mathfrak{S}_n} \mathbf{tr}\big(MP^\pi\big), \tag{16}$$

where M is the matrix with coefficient $\|X_i - X_j^\#\|^2$ at the i^{th} row and j^{th} column. The cornerstone of our next argument is the Birkhoff-von Neumann theorem stated below, which can be found for example in (Budish et al., 2013).

Theorem 6 (Birkhoff-von Neumann Theorem) *Assume that $\mathcal{P}$ is the set of all doubly stochastic matrices of size n, i.e., the matrices whose entries are nonnegative and sum up to 1 in every row and every column. Then every matrix in $\mathcal{P}$ is a convex combination of matrices $\{P^\pi : \pi \in \mathfrak{S}_n\}$. Furthermore, permutation matrices are the vertices of the simplex $\mathcal{P}$.*

In view of this result, the combinatorial optimization problem (16) is equivalent to the following problem of continuous optimization:

$$P^{\mathrm{LSS}} = \arg\min_{P \in \mathcal{P}} \mathbf{tr}\big(MP\big), \tag{17}$$

in the sense that π is a solution to (16) if and only if P^π is a solution to (17). To prove this claim, let us remark that for every $P \in \mathcal{P}$, there exist coefficients $\alpha_1, \dots, \alpha_{n!} \in [0,1]$ such that $P = \sum_{i=1}^{n!} \alpha_i P^{\pi_i}$ and $\sum_{i=1}^{n!} \alpha_i = 1$. Therefore, we have $\mathbf{tr}\big(MP\big) = \sum_{i=1}^{n!} \alpha_i \mathbf{tr}\big(MP^{\pi_i}\big) \geq \min_{\pi \in \mathfrak{S}_n} \mathbf{tr}\big(MP^\pi\big)$ and $\mathbf{tr}\big(MP^{\mathrm{LSS}}\big) \geq \mathbf{tr}\big(MP^{\pi^{\mathrm{LSS}}}\big)$.

The great advantage of (17) is that it concerns the minimization of a linear function under linear constraints and, therefore, is a problem of linear programming that can be efficiently solved even for large values of n. The same arguments apply to the estimators π^{LSNS} and π^{LSL}, only the matrix M needs to be changed.

There is a second way to compute the estimators LSS, LSNS and LSL efficiently. Indeed, the computation of the aforementioned maximum likelihood estimators is a particular case of the assignment problem, which consists in finding a minimum weight matching in a weighted bipartite graph, where the matrix of the costs is the matrix M from above. This means that the cost of assigning the i^{th} feature of the first image to the j^{th} feature of the second image is either

- the squared distance $\|X_i - X_j^\#\|^2$,
- or the normalized squared distance $\|X_i - X_j^\#\|^2/(\sigma_i^2 + (\sigma_j^\#)^2)$,
- or the logarithm of the squared distance $\log \|X_i - X_j^\#\|^2$.

The so-called Hungarian algorithm presented in Kuhn (1955) solves the assignment problem in time $O(n^3)$.

6. Extensions

In this section, we briefly discuss possible extensions of the foregoing results to other distances, more general matching criteria and to the estimation of an arrangement.

$n=$	4	5	6	7	8	9	10	11	12
$M_n \geq$	19	57	179	594	1939	3441	11680	39520	86575
$\frac{\log M_n}{n \log n} \geq$	0.53	0.50	0.48	0.47	0.455	0.412	0.407	0.401	0.381
$\frac{\log M_n}{n \log n} \leq$	0.53	0.50	0.48	0.47	0.455	0.445	0.436	0.427	0.420

Table 1: The values of $M_n = \mathcal{M}(1/4, B_{2,n}(2), \delta_H)$ for $n \in \{4, \dots, 12\}$. The lower bound is just the cardinality of one ϵ-packing, not necessarily the largest one. The upper bound is merely the cardinality of the ℓ_2-ball.

6.1 Minimax Rates for the Hamming Distance

In the previous sections, the minimax rates were obtained for the error of estimation measured by the risk $\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}(\hat{\pi} \neq \pi^*) = \mathbf{E}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}[\delta_{0\text{-}1}(\hat{\pi}, \pi^*)]$, which may be considered as too restrictive. Indeed, one could find acceptable an estimate having a small number of mismatches, if it makes it possible to significantly reduce the perceivable distance of separation. These considerations lead to investigating the behavior of the estimators in terms of the Hamming loss, *i.e.*, to studying the risk

$$\mathbf{E}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}[\delta_H(\hat{\pi}, \pi^*)] = \mathbf{E}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}\Big[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{\pi}(i) \neq \pi^*(i)\}}\Big]$$

corresponding to the expected average number of mismatched features. Another advantage of studying the Hamming loss instead of the 0-1 loss is that the former sharpens the difference between the performances of various estimators. Note, however, that thanks to the inequality $\delta_H(\hat{\pi}, \pi^*) \leq \delta_{0\text{-}1}(\hat{\pi}, \pi^*)$, all the upper bounds established for the minimax rate of separation under the 0-1 loss directly carry over to the case of the Hamming loss. This translates into the following theorem.

Theorem 7 *Let $\alpha \in (0,1)$ and condition (11) be fulfilled. If $\hat{\pi}$ is either π^{LSNS} or π^{LSL}, then $\bar{\kappa}_\alpha(\hat{\pi}) \leq 4 \max\big\{\big(2\log\frac{8n^2}{\alpha}\big)^{1/2}, \big(d\log\frac{4n^2}{\alpha}\big)^{1/4}\big\}$. That is, if*

$$\kappa = 4\max\Big\{\Big(2\log\frac{8n^2}{\alpha}\Big)^{1/2}, \Big(d\log\frac{4n^2}{\alpha}\Big)^{1/4}\Big\}$$

and $\bar{\Theta}_\kappa$ is the set of all $\boldsymbol{\theta} \in \mathbb{R}^{n\times d}$ such that $\bar{\kappa}(\boldsymbol{\theta}, \sigma) \geq \kappa$, then

$$\max_{\pi^* \in \mathfrak{S}_n} \sup_{\boldsymbol{\theta} \in \bar{\Theta}_\kappa} \mathbf{E}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}[\delta_H(\hat{\pi}, \pi^*)] \leq \alpha.$$

While this upper bound is an immediate consequence of Theorem 3, getting lower bounds for the Hamming loss appears to be more difficult. To state the corresponding result, let is consider the case of homoscedastic noise and introduce some notation. We denote by $\delta_2(\cdot,\cdot)$ the normalized ℓ^2-distance on the space of permutations $\mathfrak{S}_n$: $\delta_2(\pi,\pi')^2 = \frac{1}{n}\sum_{k=1}^{n}\big(\pi(k) - \pi'(k)\big)^2$. Let $B_{2,n}(R)$ be the ball of $(\mathfrak{S}_n, \delta_2)$ with radius R centered at *id*. As usual, we denote by $\mathcal{M}(\epsilon, B_{2,n}(R), \delta_H)$ the ϵ-packing number of the ℓ_2-ball $B_{2,n}(R)$ in the metric δ_H. This

means that $\mathcal{M}(\epsilon, B_{2,n}(R), \delta_H)$ is the largest integer M such that there exist permutations $\pi_1, \dots, \pi_M \in B_{2,n}(R)$ satisfying $\delta_H(\pi_i, \pi_j) \geq \epsilon$ for every $i \neq j$. One can easily check that replacing $B_{2,n}(R)$ by any other ball of radius R leaves the packing number $\mathcal{M}(\epsilon, B_{2,n}(R), \delta_H)$ unchanged. We set $M_n = \mathcal{M}(1/4, B_{2,n}(2), \delta_H)$.

Theorem 8 *Let $\sigma_k = \sigma$ for all $k \in \{1, \dots, n\}$ and $\bar{\Theta}_\kappa$ is the set of all $\boldsymbol{\theta} \in \mathbb{R}^{n \times d}$ such that $\bar{\kappa}(\boldsymbol{\theta}, \boldsymbol{\sigma}) \geq \kappa$. Furthermore, assume that one of the following two conditions is fulfilled:*

- *$n \geq 3$ and $\kappa = (1/4)\big(\frac{\log M_n}{n}\big)^{1/2}$,*
- *$n \geq 26$, $d \geq 24 \log n$ and $\kappa \leq (1/8)(d \log n)^{1/4}$.*

Then $\inf_{\hat{\pi}} \max_{\pi^* \in \mathfrak{S}_n} \sup_{\boldsymbol{\theta} \in \bar{\Theta}_\kappa} \mathbf{E}_{\boldsymbol{\theta}, \boldsymbol{\sigma}, \pi^*}[\delta_H(\hat{\pi}, \pi^*)] > 2.15\%$.

This result implies that in the regime of moderately large dimension, $d \geq 24 \log n$, the minimax rate of the separation is the same as the one under the 0-1 loss and it is achieved by the LSL estimator. The picture is less clear in the regime of small dimensions, $d = o(\log n)$. If one proves that for some $c > 0$, the inequality $\log M_n \geq cn \log n$ holds for every $n \geq 3$, then the lower bound of the last theorem matches the upper bound of Theorem 7 up to constant factors and leads to the minimax rate of separation $\max\{(\log n)^{1/2}, (d \log n)^{1/4}\}$. Unfortunately, we were unable to find any result on the order of magnitude of $\log M_n$, therefore, we cannot claim that there is no gap between our lower bound and the upper one. However, we did a small experiment for evaluating M_n for small values of n. The result is reported in Table 1.

6.2 More General Matching Criteria

In the previous sections, we were considering two vectors θ_i and $\theta_j^\#$ as matching if $\theta_i \equiv \theta_j^\#$, and $\equiv$ was the usual equality. In this part, we show that our results can be extended to more general matching criteria, defined as follows. Let A, $A^\#$ be two known $p \times d$ matrices with some $p \in \mathbb{N}$ and b, $b^\#$ be two known vectors from $\mathbb{R}^d$. We write $\theta \equiv_{A,b} \theta^\#$, if

$$A(\theta - b) = A^\#(\theta^\# - b^\#). \tag{18}$$

Note that the case of equality studied in previous sections is obtained for $A = A^\# = \mathbf{I}_d$ and $b = b^\# = 0$, where $\mathbf{I}_d$ is the identity matrix of size d. Let us first note that without loss of generality, by a simple transformation of the features, one can replace (18) by the simpler relation

$$\bar{\theta} = B\bar{\theta}^\#, \tag{19}$$

where $\bar{\theta} \in \mathbb{R}^{d_1}$ for $d_1 = \operatorname{rank}(A)$, $\bar{\theta}^\# \in \mathbb{R}^{d_2}$ for $d_2 = \operatorname{rank}(A^\#)$ and B is a $d_1 \times d_2$ known matrix. Indeed, let $A = U^\top \Lambda V$ (resp. $A^\# = \tilde{U}^\top \tilde{\Lambda} \tilde{V}$) be the singular value decomposition of A (resp. $A^\#$), with orthogonal matrices $U \in \mathbb{R}^{d_1 \times p}$, $V \in \mathbb{R}^{d_1 \times d}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{d_1 \times d_1}$ with positive entries. Then, one can deduce (19) from (18) by setting $\bar{\theta} = V(\theta - b)$, $\bar{\theta}^\# = \tilde{V}(\theta^\# - b^\#)$ and $B = \Lambda^{-1} U \tilde{U}^\top \tilde{\Lambda}$. Of course, the same transformation should be applied to the observed noisy features, which leads to $\bar{X}_i = V(X_i - b)$ and $\bar{X}_i^\# = \tilde{V}(X^\# - b^\#)$. Since V and $\tilde{V}$ are orthogonal matrices, *i.e.*, satisfy the relations $VV^\top = \mathbf{I}_{d_1}$ and $\tilde{V}\tilde{V}^\top = \mathbf{I}_{d_2}$, the noise component in the transformed noisy features is still white Gaussian.

All the four estimators introduced in Section 3 can be adapted to deal with such type of criterion. For example, denoting by M the matrix $B(B^\top B)^+B^\top + BB^\top$ where M^+ is the Moore-Penrose pseudoinverse of the matrix M, the LSL estimator should be modified as follows

$$\pi^{\mathrm{LSL}} \triangleq \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^{n} \log \|M^+(\bar{X}_{\pi(i)} - B\bar{X}_i^\#)\|^2. \tag{20}$$

All the results presented before can be readily extended to this case. In particular, if we assume that all the nonzero singular values of B are bounded and bounded away from 0 and $\mathrm{rank}(B) = q$, then the minimax rate of separation is given by $\max\left\{(\log n)^{1/2}, (q\log n)^{1/4}\right\}$. This rate is achieved, for instance, by the LSL estimator.

Let us briefly mention two situations in which this kind of general affine criterion may be useful. First, if each feature θ (resp. $\theta^\#$) corresponds to a patch in an image I (resp. $I^\#$), then for detecting pairs of patches that match each other it is often useful to neglect the changes in illumination. This may be achieved by means of the criterion $A\theta = A\theta^\#$ with $A = \mathbf{I}_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top$. Indeed, the multiplication of the vector θ by A corresponds to removing from pixel intensities the mean pixel intensity of the patch. This makes the feature invariant by change of illumination. The method described above applies to this case and the corresponding rate is $\max\left\{(\log n)^{1/2}, ((d-1)\log n)^{1/4}\right\}$ since the matrix A is of rank $d-1$. Second, consider the case when each feature combines the local descriptor of an image and the location in the image at which this descriptor is computed. If we have at our disposal an estimator of the transformation that links the two images and if this transformation is linear, then we are in the aforementioned framework. For instance, let each θ be composed of a local descriptor $\mathbf{d} \in \mathbb{R}^{d-2}$ and its location $\mathbf{x} \in \mathbb{R}^2$. Assume that the first image I from which the features $\theta_i = [\mathbf{d}_i, \mathbf{x}_i]$ are extracted is obtained from the image $I^\#$ by a rotation of the plane. Let R be an estimator of this rotation and $\theta_i^\# = [\mathbf{d}_i^\#, \mathbf{x}_i^\#]$ be the features extracted from the image $I^\#$. Then, the aim is to find the permutation π such that $\mathbf{d}_i^\# = \mathbf{d}_{\pi(i)}$ and $\mathbf{x}_i^\# = R\mathbf{x}_{\pi(i)}$ for every $i = 1, \dots, n$. This corresponds to taking in (19) the matrix B given by

$$B = \begin{pmatrix} \mathbf{I}_{d-2} & 0 \\ 0 & R \end{pmatrix}.$$

This matrix B being orthogonal, the resulting minimax rate of separation is exactly the same as when $B = \mathbf{I}_d$.

Remark 9 *An interesting avenue for future research concerns the determination of the minimax rates in the case when the equivalence of two features is understood under some transformation A which is not completely determined. For instance, one may consider that the features θ and θ' match if there is a matrix A in a given parametric family $\{A_\tau : \tau \in \mathbb{R}\} \subset \mathbb{R}^{d\times d}$ for which $\theta = A\theta'$. In other terms, $\theta \equiv \theta'$ is understood as $\inf_\tau \|\theta - A_\tau\theta'\| = 0$. Such a criterion of matching may be useful for handling various types of invariance (see Collier and Dalalyan (2015); Collier (2012) for invariance by translation).*

6.3 Estimation of an Arrangement

An interesting extension concerns the case of the estimation of a general arrangement, *i.e.,* the case when m and n are not necessarily identical. In such a situation, without loss of

generality, one can assume that $n \leq m$ and look for an injective function

$$\pi^* : \{1, \ldots, n\} \to \{1, \ldots, m\}.$$

All the estimators presented in Section 3 admit natural counterparts in this rectangular setting. Furthermore, the computational method using the Birkhoff-von Neumann theorem is still valid in this setting, and is justified by the extension of the Birkhoff-von Neumann theorem recently proved by Budish et al. (2013). In this case, the minimization should be carried out over the set of all matrices P of size (n, m) such that $P_{i,j} \geq 0$, and

$$\begin{cases} \sum_{i=1}^{n} P_{i,j} \leq 1 \\ \sum_{j=1}^{m} P_{i,j} = 1 \end{cases}, \quad (i,j) \in \{1, \ldots, n\} \times \{1, \ldots, m\}.$$

From a practical point of view, it is also important to consider the issue of robustness with respect to the presence of outliers, *i.e.,* when for some i there is no $X_j^\#$ matching with X_i. The detailed exploration of this problem being out of scope of the present paper, let us just underline that the LSL-estimator seems to be well suited for such a situation because of the robustness of the logarithmic function. Indeed, the correct matches are strongly rewarded because $\log(0) = -\infty$ and the outliers do not interfere too much with the estimation of the arrangement thanks to the slow growth of log in $+\infty$.

7. Experimental Results

We have implemented all the procedures in Matlab and carried out numerical experiments on synthetic data. To simplify, we have used the general-purpose solver SeDuMi (Sturm, 1999) for solving linear programs. We believe that it is possible to speed-up the computations by using more adapted first-order optimization algorithms, such as coordinate gradient descent. However, even with this simple implementation, the running times are reasonable: for a problem with $n = 500$ features, it takes about six seconds to compute a solution to (17) on a standard PC.

7.1 Homoscedastic noise

We chose $n = d = 200$ and randomly generated a $n \times d$ matrix $\boldsymbol{\theta}$ with i.i.d. entries uniformly distributed on $[0, \tau]$, with several values of τ varying between 1.4 and 3.5. Then, we randomly chose a permutation π^* (uniformly from $\mathfrak{S}_n$) and generated the sets $\{X_i\}$ and $\{X_i^\#\}$ according to (2) with $\sigma_i = \sigma_i^\# = 1$. Using these sets as data, we computed the four estimators of π^* and evaluated the average error rate $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{\pi}(i) \neq \pi^*(i)\}}$. The result, averaged over 500 independent trials, is plotted in Fig. 1.

Note that the three estimators originating from the maximum likelihood methodology lead to the same estimators, while the greedy algorithm provides an estimator which is much worse than the others when the parameter κ is small.

7.2 Heteroscedastic noise

This experiment is similar to the previous one, but the noise level is not constant. We still chose $n = d = 200$ and defined $\boldsymbol{\theta} = \tau I_d$, where I_d is the identity matrix and τ varies between

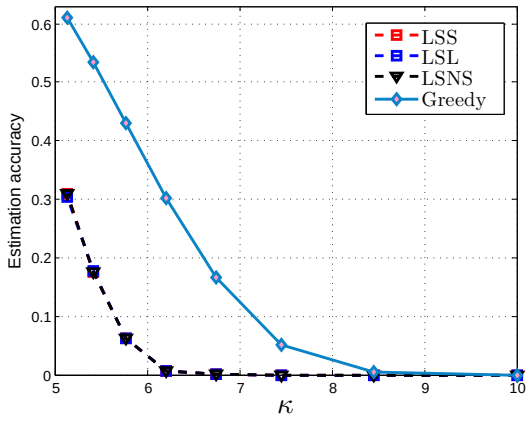

Figure 1: Average error rate of the four estimating procedures in the experiment with homoscedastic noise as a function of the minimal distance κ between distinct features. One can observe that the LSS, LSNS and LSL procedures are indistinguishable and perform much better than the greedy algorithm.

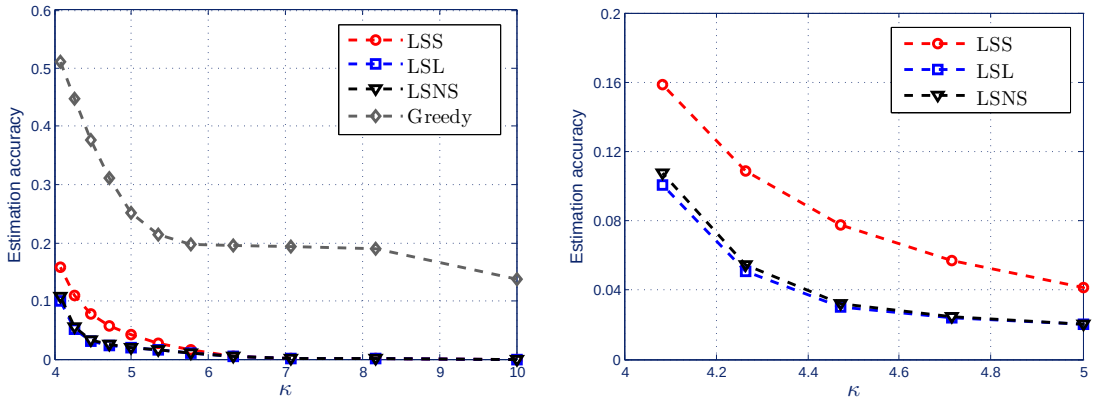

Figure 2: Left: Average error rate of the four estimating procedures in the experiment with heteroscedastic noise as a function of the minimal distance κ between distinct features. Right: zoom on the same plots. One can observe that the LSNS and LSL are almost indistinguishable and, as predicted by the theory, perform better than the LSS and the greedy algorithm.

4 and 10. Then, we randomly chose a permutation π^* (uniformly from $\mathfrak{S}_n$) and generated the sets $\{X_i\}$ and $\{X_i^\#\}$ according to (2) with $\sigma_{\pi^*(i)} = \sigma_i^\# = 1$ for 10 randomly chosen values of i and $\sigma_{\pi^*(i)} = \sigma_i^\# = 0.5$ for the others. Using these sets as data, we computed the four estimators of π^* and evaluated the average error rate $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\hat\pi(i)\neq\pi^*(i)\}}$. The result, averaged over 500 independent trials, is plotted in Fig. 2.

Note that among the noise-level-adaptive estimators, LSL outperforms the two others and is as accurate as, and even slightly better than the LSNS pseudo-estimator. This confirms the theoretical findings presented in foregoing sections.

8. Conclusion and Future Work

Motivated by the problem of feature matching, we proposed a rigorous framework for studying the problem of permutation estimation from a minimax point of view. The key notion in our framework is the minimax rate of separation, which plays the same role as in the statistical hypotheses testing theory (Ingster and Suslina, 2003). We established theoretical guarantees for several natural estimators and proved the optimality of some of them. The results appeared to be quite different in the homoscedastic and in the heteroscedastic cases. However, we have shown that the least sum of logarithms estimator outperforms the other procedures both theoretically and empirically.

Several avenues of future work have been already mentioned in previous sections. In particular, investigating the statistical properties of the arrangement estimation problem described in Section 6.3 and considering the case of unspecified transformation relating the features may have a significant impact on the practice of feature matching.

Another interesting question is to extend the statistical inference developed here for the problem of feature matching to the more general assignment problem. The latter aims at assigning m tasks to n agents such that the cost of assignment is as small as possible. Various settings of this problem have been considered in the literature (Pentico, 2007) and many algorithms for solving the problem have been proposed (Romeijn and Morales, 2000). However, to the best of our knowledge, the statistical aspects of the problem in the case where the cost matrix is corrupted by noise have not been studied so far.

9. Proofs of the Theorems

In this section we collect the proofs of the theorems. We start with the proof of Theorem 3, since it concerns the more general setting and the proof of Theorem 1 can be deduced from that of Theorem 3 by simple arguments. We then prove the other theorems in the usual order and postpone the proofs of some technical lemmas to the next section.

Proof of Theorem 3 To ease notation and without loss of generality, we assume that π^* is the identity permutation denoted by id. Furthermore, since there is no risk of confusion, we write $\mathbf{P}$ instead of $\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi^*}$. We wish to bound the probability of the event $\Omega = \{\hat{\pi} \neq id\}$.

Let us first denote by $\hat{\pi}$ the maximum likelihood estimator π^{LSL} defined by (12). We have

$$\Omega \subset \bigcup_{\pi \neq id} \Omega_\pi,$$

where

$$\Omega_\pi = \Big\{ \sum_{i=1}^n \log \frac{\|X_i - X_i^\#\|^2}{\|X_{\pi(i)} - X_i^\#\|^2} \geq 0 \Big\} = \Big\{ \sum_{i:\pi(i)\neq i} \log \frac{\|X_i - X_i^\#\|^2}{\|X_{\pi(i)} - X_i^\#\|^2} \geq 0 \Big\}.$$

On the one hand, for every permutation π,

$$\sum_{\pi(i)\neq i} \log\Big(\tfrac{2\sigma_i^2}{\sigma_i^2+\sigma_{\pi(i)}^2}\Big) = \sum_{i=1}^n \big(\log(2\sigma_i^2) - \log(\sigma_i^2+\sigma_{\pi(i)}^2)\big)$$
$$= \sum_{i=1}^n \frac{\log(2\sigma_i^2)+\log(2\sigma_{\pi(i)}^2)}{2} - \log(\sigma_i^2+\sigma_{\pi(i)}^2)$$

so, using the concavity of the logarithm, this quantity is nonpositive. Therefore,

$$\Omega_\pi \subset \Big\{ \sum_{i:\pi(i)\neq i} \log \tfrac{\|X_i - X_i^\#\|^2/(2\sigma_i^2)}{\|X_{\pi(i)} - X_i^\#\|^2/(\sigma_i^2+\sigma_{\pi(i)}^2)} \geq 0\Big\}$$
$$\subset \bigcup_{i=1}^n \bigcup_{j\neq i} \Big\{ \frac{\|X_i - X_i^\#\|^2}{2\sigma_i^2} \geq \frac{\|X_j - X_i^\#\|^2}{\sigma_j^2+\sigma_i^2}\Big\}.$$

This readily yields $\Omega \subset \bar\Omega$, where

$$\bar\Omega = \bigcup_{i=1}^n \bigcup_{j\neq i} \Big\{ \frac{\|X_i - X_i^\#\|^2}{2\sigma_i^2} \geq \frac{\|X_j - X_i^\#\|^2}{\sigma_j^2+\sigma_i^2}\Big\}. \tag{21}$$

Furthermore, the same inclusion is true for the LSNS estimator as well. Therefore, the rest of the proof is common for the estimators LSNS and LSL.

We set $\sigma_{i,j} = (\sigma_i^2+\sigma_j^2)^{1/2}$ and

$$\zeta_1 = \max_{i\neq j} \Big| \frac{(\theta_i-\theta_j)^\top(\sigma_i\xi_i - \sigma_j\xi_j^\#)}{\|\theta_i-\theta_j\|\sigma_{i,j}}\Big|, \quad \zeta_2 = d^{-1/2}\max_{i,j}\Big|\Big\|\frac{\sigma_i\xi_i-\sigma_j\xi_j^\#}{\sigma_{i,j}}\Big\|^2 - d\Big|.$$

Since $\pi^* = id$, it holds that for every $i\in\{1,\dots,n\}$,

$$\|X_i - X_i^\#\|^2 = \sigma_i^2\|\xi_i - \xi_i^\#\|^2 \leq 2\sigma_i^2(d+\sqrt{d}\zeta_2).$$

Similarly, for every $j\neq i$,

$$\|X_j - X_i^\#\|^2 = \|\theta_j-\theta_i\|^2 + \|\sigma_j\xi_j - \sigma_i\xi_i^\#\|^2 + 2(\theta_j-\theta_i)^\top(\sigma_j\xi_j - \sigma_i\xi_i^\#).$$

Therefore,

$$\|X_j - X_i^\#\|^2 \geq \|\theta_j-\theta_i\|^2 + \sigma_{i,j}^2(d-\sqrt{d}\zeta_2) - 2\|\theta_j-\theta_i\|\sigma_{i,j}\zeta_1.$$

This implies that on the event $\Omega_1 = \{\bar\kappa(\boldsymbol{\theta},\boldsymbol{\sigma}) \geq \zeta_1\}$ it holds that

$$\frac{\|X_j - X_i^\#\|^2}{\sigma_{i,j}^2} \geq \bar\kappa(\boldsymbol{\theta},\boldsymbol{\sigma})^2 - 2\bar\kappa(\boldsymbol{\theta},\boldsymbol{\sigma})\zeta_1 + d - \sqrt{d}\zeta_2.$$

Combining these bounds, we get that

$$\Omega\cap\Omega_1 \subset \Big\{ d+\sqrt{d}\zeta_2 \geq \bar\kappa(\boldsymbol{\theta},\boldsymbol{\sigma})^2 - 2\bar\kappa(\boldsymbol{\theta},\boldsymbol{\sigma})\zeta_1 + d - \sqrt{d}\zeta_2\Big\},$$

which implies that

$$\begin{aligned}\mathbf{P}(\Omega) &\leq \mathbf{P}(\Omega_1^{\complement}) + \mathbf{P}\big(\Omega \cap \Omega_1\big) \\ &\leq \mathbf{P}\big(\zeta_1 \geq \bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})\big) + \mathbf{P}(2\sqrt{d}\zeta_2 + 2\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})\zeta_1 \geq \bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})^2) \\ &\leq 2\mathbf{P}\big(\zeta_1 \geq \tfrac{\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})}{4}\big) + \mathbf{P}\big(\zeta_2 \geq \tfrac{\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})^2}{4\sqrt{d}}\big). \end{aligned} \tag{22}$$

Finally, one easily checks that for suitably chosen random variables $\zeta_{i,j}$ drawn from the standard Gaussian distribution, it holds that $\zeta_1 = \max_{i\neq j} |\zeta_{i,j}|$. Therefore, using the well-known tail bound for the standard Gaussian distribution in conjunction with the union bound, we get

$$\mathbf{P}\big(\zeta_1 \geq \tfrac{1}{4}\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})\big) \leq \sum\nolimits_{i\neq j} \mathbf{P}\big(|\zeta_{i,j}| \geq \tfrac{1}{4}\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})\big) \leq 2n^2 e^{-\frac{1}{32}\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})^2}. \tag{23}$$

To bound the large deviations of the random variable ζ_2, we rely on the following result.

Lemma 10 (Laurent and Massart (2000), Eq. (4.3) and (4.4)) *If Y is drawn from the chi-squared distribution $\chi^2(D)$, where $D \in \mathbb{N}^*$, then, for every $x > 0$,*

$$\begin{cases} \mathbf{P}\big(Y - D \leq -2\sqrt{Dx}\big) \leq e^{-x}, \\ \mathbf{P}\big(Y - D \geq 2\sqrt{Dx} + 2x\big) \leq e^{-x}. \end{cases}$$

As a consequence, $\forall y > 0$, $\mathbf{P}\big(D^{-1/2}|Y - D| \geq y\big) \leq 2\exp\big\{-\frac{1}{8}y(y \wedge \sqrt{D})\big\}$.

This inequality, combined with the union bound, yields

$$\mathbf{P}\Big(\zeta_2 \geq \tfrac{\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})^2}{4\sqrt{d}}\Big) \leq 2n^2 \exp\Big\{-\frac{(\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma})/16)^2}{d}(\bar{\kappa}^2(\boldsymbol{\theta},\boldsymbol{\sigma}) \wedge 8d)\Big\}. \tag{24}$$

Combining inequalities (22)-(24), we obtain that as soon as

$$\bar{\kappa}(\boldsymbol{\theta},\boldsymbol{\sigma}) \geq 4\Big(\sqrt{2\log(8n^2/\alpha)} \vee \big(d\log(4n^2/\alpha)\big)^{1/4}\Big),$$

we have $\mathbf{P}(\hat{\pi} \neq \pi^*) = \mathbf{P}(\Omega) \leq \alpha$.

Proof of Theorem 1 Without loss of generality, we assume $\pi^* = id$. It holds that, on the event

$$\mathcal{A} = \bigcap_{i=1}^{n} \bigcap_{j\neq i} \Big\{\|X_i - X_i^{\#}\| < \|X_j - X_i^{\#}\|\Big\},$$

all the four estimators coincide with the true permutation id. Therefore, we have

$$\{\hat{\pi} \neq id\} \subseteq \bigcup_{i=1}^{n} \bigcup_{j\neq i} \Big\{\|X_i - X_i^{\#}\| \geq \|X_j - X_i^{\#}\|\Big\}.$$

The latter event is included in $\bar{\Omega}$ at the right-hand side of (21), the probability of which has been already shown to be small in the previous proof.

Proof of Theorem 2 We refer the reader to the proof of Theorem 4 below, which concerns the more general situation. Indeed, when all the variances σ_j are equal, Theorem 4 boils down to Theorem 2.

Proof of Theorem 4 To establish lower bounds for various types of risks we will use the following lemma:

Lemma 11 (Tsybakov (2009), Theorem 2.5) *Assume that for some integer $M \geq 2$ there exist distinct permutations $\pi_0, \dots, \pi_M \in \mathfrak{S}_n$ and mutually absolutely continuous probability measures $\mathbf{Q}_0, \dots, \mathbf{Q}_M$ defined on a common probability space $(\mathcal{Z}, \mathscr{Z})$ such that*

$$\frac{1}{M}\sum_{j=1}^{M} K(\mathbf{Q}_j, \mathbf{Q}_0) \leq \frac{1}{8}\log M.$$

Then, for every measurable mapping $\tilde{\pi} : \mathcal{Z} \to \mathfrak{S}_n$,

$$\max_{j=0,\dots,M} \mathbf{Q}_j(\tilde{\pi} \neq \pi_j) \geq \frac{\sqrt{M}}{\sqrt{M}+1}\Big(\frac{3}{4} - \frac{1}{2\sqrt{\log M}}\Big).$$

To prove Theorem 4, we split the inequality $8\kappa \leq \max\{(\log n)^{1/2}, (cd\log n)^{1/4}\}$ into two cases

$$\begin{aligned} &\textbf{Case 1:} \qquad 8\kappa \leq (\log n)^{1/2}, \\ &\textbf{Case 2:} \qquad (\log n)^{1/2} \leq 8\kappa \leq (cd\log n)^{1/4}. \end{aligned}$$

Case 1: We assume that $8\kappa \leq (\log n)^{1/2}$.
Denote by m the largest integer such that $2m \leq n$. We assume without loss of generality that the noise levels are ranked in increasing order: $\sigma_1 \leq \ldots \leq \sigma_n$. Then, we construct a least favorable set of vectors for the estimation of the permutation. To ease notation, we set $\sigma_{i,j} = (\sigma_i^2 + \sigma_j^2)^{1/2}$.

Lemma 12 *Assume that m is the largest integer such that $2m \leq n$. Then there is a set of vectors $\boldsymbol{\theta}$ such that*

$$\frac{\|\theta_1 - \theta_2\|}{\sigma_{1,2}} = \ldots = \frac{\|\theta_{2m-1} - \theta_{2m}\|}{\sigma_{2m-1,2m}} = \kappa,$$

and for every pair $\{i, j\}$ different from the pairs $\{1,2\}, \dots, \{2m-1, 2m\}$ we have

$$\frac{\|\theta_i - \theta_j\|}{\sigma_{i,j}} > \kappa\bigg(1 + \frac{\max_{1\leq\ell\leq n}\sigma_\ell}{\min_{1\leq\ell\leq n}\sigma_\ell}\bigg).$$

Let $\boldsymbol{\theta}^0$ be the set constructed in Lemma 12. The latter implies that $\boldsymbol{\theta}^0$ belongs to $\bar{\Theta}_\kappa$, so that, denoting for every $k \in \{1, \dots, m\}, \pi_k = (2k-1\ 2k)$ the transposition of $\mathfrak{S}_n$ that only permutes $2k-1$ and $2k$, and $\pi_0 = id$, we get the following lower bound for the risk:

$$\inf_{\hat{\pi}} \sup_{(\pi,\boldsymbol{\theta})\in\mathfrak{S}_n\times\bar{\Theta}_\kappa} \mathbf{P}_{\boldsymbol{\theta},\pi}(\hat{\pi} \neq \pi) \geq \inf_{\hat{\pi}} \max_{k=0,\dots,m} \mathbf{P}_{\boldsymbol{\theta}^0,\pi_k}(\hat{\pi} \neq \pi_k).$$

In order to use Lemma 11 with $\mathbf{Q}_j = \mathbf{P}_{\boldsymbol{\theta}^0,\pi_j}$, we compute for every $k \in \{1,\dots,m\}$

$$K(\mathbf{P}_{\boldsymbol{\theta}^0,\pi_k}, \mathbf{P}_{\boldsymbol{\theta}^0,\pi_0}) = \|\theta^0_{2k-1} - \theta^0_{2k}\|^2 \Big(\frac{1}{2\sigma^2_{2k}} + \frac{1}{2\sigma^2_{2k-1}}\Big) + \frac{d}{2}\Big(\frac{\sigma^2_{2k-1}}{\sigma^2_{2k}} + \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 2\Big).$$

Using the fact that $\|\theta^0_{2k-1} - \theta^0_{2k}\|^2 = \kappa^2(\sigma^2_{2k-1} + \sigma^2_{2k})$, we get

$$K(\mathbf{P}_{\boldsymbol{\theta}^k,\pi_k}, \mathbf{P}_{\boldsymbol{\theta}^0,\pi_0}) = \frac{\kappa^2}{2}\big(2 + \sigma^2_{2k-1}\sigma^{-2}_{2k} + \sigma^2_{2k}\sigma^{-2}_{2k-1}\big) + \frac{d}{2}\Big(\frac{\sigma^2_{2k-1}}{\sigma^2_{2k}} + \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 2\Big).$$

The inequality $\sigma^2_{2k-1} \le \sigma^2_{2k}$ implies the following two relations:

$$\begin{aligned}
&\frac{\sigma^2_{2k-1}}{\sigma^2_{2k}} + \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 2 \le \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 1\\
&\frac{\sigma^2_{2k-1}}{\sigma^2_{2k}} + \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 2 = \frac{\sigma^2_{2k-1}}{\sigma^2_{2k}}\Big(\frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 1\Big)^2 \le \Big(\frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 1\Big)^2.
\end{aligned}$$

This readily yields $\frac{\sigma^2_{2k-1}}{\sigma^2_{2k}} + \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}} - 2 \le \frac{\log n}{16d}$ and, therefore,

$$K(\mathbf{P}_{\boldsymbol{\theta}^k,\pi_k}, \mathbf{P}_{\boldsymbol{\theta}^0,\pi_0}) \le \kappa^2\Big(\frac{3}{2} + \frac{\sigma^2_{2k}}{2\sigma^2_{2k-1}}\Big) + \frac{\log n}{32}.$$

Next, we apply the following result.

Lemma 13 *Let $a_1, a_2, \dots, a_m$ be real numbers larger than one such that $\prod_{k=1}^m a_k \le A$. Then, $\sum_{k=1}^m a_k \le m + \log A \max_k a_k$.*

Proof We use the simple inequality $e^x \le 1 + xe^x$ for all $x \ge 0$. Replacing x by $\log a_k$ and summing over $k = 1,\dots,m$ we get

$$\sum_{k=1}^m a_k \le \sum_{k=1}^m 1 + a_k \log a_k \le m + \max_{k=1,\dots,m} a_k \log \prod_{k=1}^m a_k \le m + \max_{k=1,\dots,m} a_k \log A.$$

This completes the proof of the lemma. ∎

We apply this lemma to $a_k = \sigma^2_{2k}/\sigma^2_{2k-1}$. Since the variances are sorted in increasing order, we have $\prod_{k=1}^m a_k \le \prod_{i=1}^{n-1} \sigma^2_{i+1}/\sigma^2_i = \sigma^2_n/\sigma^2_1 \le 1 + \gamma_{n,d}$ with $\gamma_{n,d} = \max\big((\frac{\log n}{16d})^{1/2}, \frac{\log n}{16d}\big)$. In conjunction with the inequality $\log(1+x) \le x$, this entails that $\sum_{k=1}^m \sigma^2_{2k}/\sigma^2_{2k-1} \le m + \gamma_{n,d}(1+\gamma_{n,d})$. Then, since $\log n \le 1.8 \log m$ for $n \ge 6$ and $\gamma_{n,d} \le 0.2 \log n \le 0.36 \log m$, we get

$$\begin{aligned}
\frac{1}{m}\sum_{k=1}^m K(\mathbf{P}_{\boldsymbol{\theta}^k,\pi_k}, \mathbf{P}_{\boldsymbol{\theta}^0,\pi_0}) &\le \kappa^2\left(\frac{3}{2} + \frac{1}{2m}\sum_{k=1}^m \frac{\sigma^2_{2k}}{\sigma^2_{2k-1}}\right) + \frac{9\log m}{160}\\
&\le \kappa^2\Big(2 + \frac{\gamma_{n,d}}{2m}\big\{1 + \gamma_{n,d}\big\}\Big) + \frac{9\log m}{160}\\
&\le \kappa^2\Big(2 + \frac{0.18\log m}{m}\big\{1 + 0.36\log m\big\}\Big) + \frac{9\log m}{160}.
\end{aligned}$$

Finally, using the fact that $m \geq 3$, we get $\frac{1}{m}\sum_{k=1}^m K(\mathbf{P}_{\boldsymbol{\theta}^k,\pi_k}, \mathbf{P}_{\boldsymbol{\theta}^0,\pi_0}) \leq 2.1\kappa^2 + \frac{9\log m}{160} \leq \frac{\log m}{8}$ since $\kappa^2 \leq \frac{1}{64}\log n \leq \frac{9}{320}\log m$.

We conclude by Lemma 11 and by the monotonicity of the function $m \mapsto \frac{\sqrt{m}}{1+\sqrt{m}}(\frac{3}{4} - \frac{1}{2\sqrt{\log m}})$ that

$$\inf_{\hat\pi} \sup_{(\pi,\boldsymbol{\theta})\in\mathfrak{S}_n\times\bar\Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\pi}(\hat\pi \neq \pi) \geq \frac{\sqrt{3}}{\sqrt{3}+1}\Big(\frac{3}{4} - \frac{1}{2\sqrt{\log 3}}\Big) \geq 0.17.$$

Case 2: We assume that $(\log n)^{1/2} \leq 8\kappa \leq (cd\log n)^{1/4}$ with $c \leq 1/20$.
In this case, we have $d \geq \frac{1}{c}\log n$. To get the desired result, we use Lemma 11 for a properly chosen family of probability measures described below.

Lemma 14 *Let $\epsilon_1, \ldots, \epsilon_n$ be real numbers defined by*

$$\epsilon_k = \sqrt{2/d}\,\kappa\sigma_k, \qquad \forall k \in \{1,\ldots,n\},$$

and let μ be the uniform distribution on $\mathcal{E} = \{\pm\epsilon_1\}^d \times \ldots \times \{\pm\epsilon_n\}^d$. We denote by $\mathbf{P}_{\mu,\pi}$ the probability measure on $\mathbb{R}^{d\times n}$ defined by $\mathbf{P}_{\mu,\pi}(A) = \int_{\mathcal{E}} \mathbf{P}_{\boldsymbol{\theta},\pi}(A)\,\mu(d\boldsymbol{\theta})$. Assume that $\sigma_1 \leq \ldots \leq \sigma_n$. For two positive integers $k < k' \leq n$, set $\gamma = \frac{\sigma_{k'}^2}{\sigma_k^2}$ and let $\pi = (k\ k')$ be the transposition that only permutes k and k'. Then

$$K(\mathbf{P}_{\mu,\pi}, \mathbf{P}_{\mu,id}) \leq 4\kappa^2\big(1-\gamma^{-1}\big) + \frac{8\kappa^4}{d}\big(2 + \big(1+(2/d)\kappa^2\big)^2\gamma^2\big) + \frac{1}{2}\big(d+2\kappa^2\big)(\gamma-1)^2$$

and $\mu(\mathcal{E}\backslash\bar\Theta_\kappa) \leq (n(n-1)/2)\ e^{-d/8}$.

The assumption on the noise levels entails that, for any integer $k \in \{1,\ldots,k'\}$, $1 \leq \frac{\sigma_{k'}^2}{\sigma_k^2} \leq 1 + \frac{1}{4}(\frac{\log n}{d})^{1/2}$, and consequently, $(\gamma-1)^2 = \big(\frac{\sigma_{k'}^2}{\sigma_k^2} - 1\big)^2 \leq 4^{-2}\big(\frac{\log n}{d}\big)$. Furthermore, $\frac{\kappa^2}{d} \leq \frac{c}{64} \leq \frac{1}{64}$ provided that $c \leq 1$. Finally, for the Kullback-Leibler divergence between $\mathbf{P}_{\mu,\pi_{k,k'}}$ and $\mathbf{P}_{\mu,id}$, where $\pi_{k,k'} = (k\ k')$ is the transposition from $\mathfrak{S}_n$ permuting only k and k', it holds

$$\begin{aligned} K(\mathbf{P}_{\mu,\pi_{k,k'}}, \mathbf{P}_{\mu,id}) &\leq \kappa^2\sqrt{\frac{\log n}{d}} + \frac{8\kappa^4}{d}\Big(2 + \frac{33^2}{32^2}\big(1+0.25\big)^2\Big) + \frac{33d}{64} \times \frac{\log n}{16d} \\ &\leq \frac{\log n}{8} \leq \frac{\log n(n-1)/2}{8}, \end{aligned}$$

where we have used once again the facts that $c \leq 1$ and $n \geq 3$. Applying Lemma 11 with $M = n(n-1)/2$, $\mathbf{Q}_0 = \mathbf{P}_{\mu,id}$ and $\{\mathbf{Q}_j\}_{j=1,\ldots,M} = \{\mathbf{P}_{\mu,\pi_{k,k'}}\}_{k\neq k'}$, we obtain

$$\begin{aligned} \max_{\pi^*\in\mathfrak{S}_n} \sup_{\boldsymbol{\theta}\in\bar\Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\pi^*}\big(\hat\pi \neq \pi^*\big) &\geq \max_{\pi^*\in\{id\}\cup\{\pi_{k,k'}\}} \int_{\bar\Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\pi^*}\big(\hat\pi \neq \pi^*\big)\frac{\mu(d\boldsymbol{\theta})}{\mu(\bar\Theta_\kappa)} \\ &\geq \max_{\pi^*\in\{id\}\cup\{\pi_{k,k'}\}} \mathbf{P}_{\mu,\pi^*}\big(\hat\pi \neq \pi^*\big) - \mu(\mathcal{E}\backslash\bar\Theta_\kappa) \\ &\geq \frac{\sqrt{15}}{\sqrt{15}+1}\Big(\frac{3}{4} - \frac{1}{2\sqrt{\log 15}}\Big) - \frac{n(n-1)}{2}e^{-d/8}. \end{aligned}$$

In view of the inequalities $d \geq (1/c)\log n$, $c \leq 1/20$ and $n \geq 6$, we get the inequality $\max_{\pi^*\in\mathfrak{S}_n} \sup_{\boldsymbol{\theta}\in\bar\Theta_\kappa} \mathbf{P}_{\boldsymbol{\theta},\pi^*}\big(\hat\pi \neq \pi^*\big) \geq 22.4\%$.

Proof of Theorem 5 Since the event $\{\pi^{\mathrm{gr}} \neq id\}$ includes the event

$$\Omega_2 = \{\|X_1 - X_1^{\#}\|^2 > \|X_2 - X_1^{\#}\|^2\},$$

it is sufficient to bound from below the probability of Ω_2. To this end, we choose any $\boldsymbol{\theta} \in \mathbb{R}^{n\times d}$ satisfying $\|\theta_1 - \theta_2\| = 2\kappa$. This readily implies that $\boldsymbol{\theta}$ belongs to $\bar{\Theta}_\kappa$. Furthermore, for suitably chosen random variables $\eta_1 \sim \chi^2_d$, $\eta_2 \sim \chi^2_d$ and $\eta_3 \sim \mathcal{N}(0,1)$, it holds that $\|X_1 - X_1^{\#}\|^2 - \|X_2 - X_1^{\#}\|^2 = 6\eta_1 - 4\kappa^2 - 8\kappa\eta_3 - 4\eta_2$. The random terms in the last sum can be controlled using Lemma 10. More precisely, for every $x > 0$, each one of the following three inequalities holds true with probability at least $1 - e^{-x^2}$:

$$\eta_1 \geq d - 2\sqrt{d}x, \quad \eta_2 \leq d + 2\sqrt{d}x + 2x^2, \quad \eta_3 \leq \sqrt{2}x.$$

This implies that with probability at least $1 - 3e^{-x^2}$, we have

$$\|X_1 - X_1^{\#}\|^2 - \|X_2 - X_1^{\#}\|^2 \geq 2d - 20\sqrt{d}x - 4(\kappa + \sqrt{2}x)^2.$$

If $x = \sqrt{\log 6}$, then the conditions imposed on κ and d ensure that the right-hand side of the last inequality is positive. Therefore, $\mathbf{P}(\Omega_2) \geq 1 - 3e^{-x^2} = 1/2$.

Proof of Theorem 8 The proof is split into two parts. In the first part, we consider the case $\kappa \leq \frac{1}{4}\sqrt{\frac{\log M_n}{n}}$, while in the second part the case $\kappa \leq \frac{1}{8}\big(\frac{\log n}{d}\big)^{1/4}$ with $d \geq 24 \log n$ and is analyzed. In both cases, the main tool we use is the following result.

Lemma 15 (Tsybakov (2009), Theorem 2.5) *Assume that for some integer $M \geq 2$ there exist distinct permutations $\pi_0, \dots, \pi_M \in \mathfrak{S}_n$ and mutually absolutely continuous probability measures $\mathbf{Q}_0, \dots, \mathbf{Q}_M$ defined on a common probability space $(\mathcal{Z}, \mathscr{Z})$ such that*

$$\begin{cases} \exists\ s > 0,\ \forall\ i \neq j,\ \delta(\pi_i, \pi_j) \geq 2s, \\ \frac{1}{M}\sum_{j=1}^M K(\mathbf{Q}_j, \mathbf{Q}_0) \leq \frac{1}{8}\log M. \end{cases}$$

Then, for every measurable mapping $\tilde{\pi} : \mathcal{Z} \to \mathfrak{S}_n$,

$$\max_{j=0,\dots,M} \mathbf{Q}_j\big(\delta(\hat{\pi}, \pi_j) \geq s\big) \geq \frac{\sqrt{M}}{\sqrt{M}+1}\Big(\frac{3}{4} - \frac{1}{2\sqrt{\log M}}\Big).$$

We now have to choose M and $\pi_0, \dots, \pi_M$ in a suitable manner, which will be done differently according to the relationship between n and d.

Case 1: We assume that $\kappa \leq \frac{1}{4}\sqrt{\frac{\log M_n}{n}}$.
Let $M = M_n$ with $M_n = \mathcal{M}(1/4, B_{2,n}(2), \delta_H)$ and let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ be the set of vectors $\theta_k = k\kappa\sigma(1, 0, \dots, 0) \in \mathbb{R}^d$. Clearly, $\boldsymbol{\theta}$ belongs to $\bar{\Theta}_\kappa$. By definition of the packing number, there exist $\pi_1, \dots, \pi_{M_n}$, permutations from $\mathfrak{S}_n$, such that

$$\delta_2(\pi_j, id) \leq 2, \qquad \delta_H(\pi_i, \pi_j) \geq \frac{1}{4}; \qquad \forall i, j \in \{1, \dots, M_n\},\ i \neq j.$$

Defining $\mathbf{Q}_j = \mathbf{P}_{\boldsymbol{\theta},\pi_j}$ for $j = 1, \ldots, M_n$ and $\mathbf{Q}_0 = \mathbf{P}_{\boldsymbol{\theta},id}$ we get

$$\begin{aligned} K(\mathbf{Q}_j, \mathbf{Q}_0) &= \frac{1}{2\sigma^2} \sum_{k=1}^{n} \|\theta_{\pi_j(k)} - \theta_k\|^2 = \frac{\kappa^2}{2} \sum_{k=1}^{n} (\pi_j(k) - k)^2 \\ &= \frac{n\kappa^2}{2}\, \delta_2(\pi_j, id)^2 \le 2n\kappa^2. \end{aligned}$$

Therefore, using Lemma 15 with $s = 1/8$ we infer from $\kappa \le \frac{1}{4}\big(\frac{\log M_n}{n}\big)^{1/2}$ that

$$\min_{\hat\pi} \max_{j=0,\ldots,M_n} \mathbf{P}_{\boldsymbol{\theta},\pi_j}(\delta_H(\hat\pi, \pi_j) \ge 1/8) \ge \frac{\sqrt{3}}{\sqrt{3}+1}\Big(\frac{3}{4} - \frac{1}{2\sqrt{\log 3}}\Big) \approx 17.31\%.$$

As a consequence, we obtain that

$$\begin{aligned} \min_{\hat\pi} \max_{(\pi,\boldsymbol{\theta})\in\mathfrak{S}_n\times\bar\Theta_\kappa} \mathbf{E}_{\boldsymbol{\theta},\pi}[\delta_H(\hat\pi,\pi)] &\ge \min_{\hat\pi} \max_{j=0,\ldots,M} \mathbf{E}_{\boldsymbol{\theta},\pi}[\delta_H(\hat\pi,\pi)\mathbb{1}_{\{\delta_H(\hat\pi,\pi)\ge 1/8\}}] \\ &\ge \frac{1}{8} \min_{\hat\pi} \max_{j=0,\ldots,M} \mathbf{E}_{\boldsymbol{\theta},\pi}[\mathbb{1}_{\{\delta_H(\hat\pi,\pi)\ge 1/8\}}] \\ &\ge 2.15\%. \end{aligned}$$

This completes the proof of the first case.

Case 2: We assume that $d \ge 24 \log n$ and $\kappa \le \frac{1}{8}(d \log n)^{1/4}$. Let μ be the uniform distribution on $\{\pm\epsilon\}^{m\times d}$ with $\epsilon = \sqrt{2/d}\,\sigma\kappa$, as in Lemma 14. For any set of permutations $\{\pi_0, \ldots, \pi_M\} \subset \mathfrak{S}_n$, in view of Markov's inequality,

$$\sup_{(\pi,\boldsymbol{\theta})\in\mathfrak{S}_n\times\bar\Theta_\kappa} \mathbf{E}_{\boldsymbol{\theta},\pi}\big[\delta_H(\hat\pi,\pi)\big] \ge \frac{3}{16}\Big(\max_{i=0,\ldots,M} \mathbf{P}_{\mu,\pi_i}\Big(\delta_H(\hat\pi,\pi_i) \ge \frac{3}{16}\Big) - \mu(\bar\Theta_\kappa^{\complement})\Big).$$

We choose M and $\pi_0, \ldots, \pi_M$ as in the following lemma.

Lemma 16 *For any integer $n \ge 4$ there exist permutations $\pi_0, \ldots, \pi_M \in \mathfrak{S}_n$ such that*

$$\pi_0 = id, \qquad M \ge (n/24)^{n/6},$$

each π_i is a composition of at most $n/2$ transpositions with disjoint supports, and for every distinct pair of indices $i, j \in \{0, \ldots, M\}$ we have

$$\delta_H(\pi_i, \pi_j) \ge 3/8.$$

As π_i is a product of transpositions, the Kullback-Leibler divergence between $\mathbf{P}_{\mu,\pi_i}$ and $\mathbf{P}_{\mu,\pi_0}$ can be computed by independence thanks to Lemma 14:

$$\frac{1}{M}\sum_{i=1}^{M} K(\mathbf{P}_{\mu,\pi_i}, \mathbf{P}_{\mu,\pi_0}) \le \frac{n}{2} \times \frac{8\kappa^4}{d}\Big(2 + \Big[1 + \frac{2\kappa^2}{d}\Big]^2\Big) = \frac{16n\kappa^4}{d},$$

where the last inequality follows from the bound $\kappa \le 0.45 d^{1/2}$. For $n \ge 26$, it holds that $M \ge 2$ and

$$\log M \ge \frac{n(\log n - \log 24)}{6} \ge \frac{\log n}{512}.$$

Consequently, $\frac{16n\kappa^4}{d} \le \frac{1}{8}\log M$ which allows us to apply Lemma 15. This yields

$$\begin{aligned}\inf_{\hat\pi}\sup_{(\pi,\boldsymbol{\theta})\in\mathfrak{S}_n\times\bar\Theta_\kappa}\mathbf{E}_{\boldsymbol{\theta},\pi}[\delta_H(\hat\pi,\pi)] &\ge \frac{3}{16}\Big[\frac{\sqrt{2}}{\sqrt{2}+1}\Big(\frac{3}{4}-\frac{1}{2\sqrt{\log 2}}\Big)-\frac{n^2}{2}e^{-d/8}\Big]\\ &\ge \frac{3}{16}\Big[0.077-\frac{n^2}{2}e^{-24\log(n)/8}\Big]\\ &\ge \frac{3}{16}\Big[0.077-\frac{1}{2n}\Big]\ge 5.81\%.\end{aligned}$$

10. Proofs of the Lemmas

Proof of Lemma 12 Let us denote $r_\sigma = \max_{1\le\ell\le n}\sigma_\ell/\min_{1\le\ell\le n}\sigma_\ell$. It suffices to set $\theta_1 = 0\in\mathbb{R}^d$,

$$\begin{aligned}\theta_{2k+1} &= \kappa\Big(\sigma_{1,2}+\ldots+\sigma_{2k-1,2k}+k\big(1+r_\sigma\big),0,\ldots,0\Big)\in\mathbb{R}^d,\\ \theta_{2k} &= \kappa\Big(\sigma_{1,2}+\ldots+\sigma_{2k-1,2k}+(k-1)\big(1+r_\sigma\big),0,\ldots,0\Big)\in\mathbb{R}^d\end{aligned}$$

for all $k=1,\ldots,m-1$. If n is impair, one can set $\theta_n=\theta_{n-1}+\kappa\big(1+r_\sigma\big)(1,0,\ldots,0)$. One readily checks that these vectors satisfy the desired conditions.

Proof of Lemma 14 Without loss of generality, we assume hereafter that $\pi\in\mathfrak{S}_n$ is the transposition permuting 1 and 2. Recall that the uniform distribution on $\{\pm\epsilon_1\}^d\times\ldots\times\{\pm\epsilon_n\}^d$ can also be written as the product $\mu=\bigotimes_{\ell=1}^n\mu_\ell$, where μ_ℓ is the uniform distribution on $\{\pm\epsilon_\ell\}^d$. Let us introduce an auxiliary probability distribution $\tilde\mu$ on $\mathbb{R}^{n\times d}$ defined as $\tilde\mu=\delta_{\mathbf{0}}\otimes\delta_{\mathbf{0}}\otimes\mu_2\otimes\ldots\otimes\mu_m$ with $\delta_{\mathbf{0}}$ being the Dirac delta measure at $\mathbf{0}\in\mathbb{R}^d$. We set $\mathbf{P}_{\tilde\mu,id}(\cdot)=\int_\Theta\mathbf{P}_{\boldsymbol{\theta},id}(\cdot)\tilde\mu(d\boldsymbol{\theta})$.

We first compute the density of $\mathbf{P}_{\mu,\pi}$ w.r.t. $\mathbf{P}_{\mu,id}$, which can be written as

$$\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\mu,id}}(\boldsymbol{X},\boldsymbol{X}^\#)=\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\tilde\mu,id}}(\boldsymbol{X},\boldsymbol{X}^\#)\bigg/\bigg(\frac{d\mathbf{P}_{\mu,id}}{d\mathbf{P}_{\tilde\mu,id}}(\boldsymbol{X},\boldsymbol{X}^\#)\bigg),\quad \boldsymbol{X},\boldsymbol{X}^\#\in\mathbb{R}^{n\times d}.$$

For every $\theta_i\in\mathbb{R}^d$ we denote by $\mathbf{P}_{\theta_i,\sigma_i}$ the probability distribution of X_i from (2), given by

$$\frac{d\mathbf{P}_{\theta_i,\sigma_i}}{d\mathbf{P}_{0,\sigma_j}}(x)=\exp\Big\{-\frac{\|\theta_i\|^2}{2\sigma_i^2}+\frac{1}{\sigma_i^2}(x,\theta_i)-\frac{\|x\|^2}{2}(\sigma_i^{-2}-\sigma_j^{-2})\Big\},\quad\forall x\in\mathbb{R}^d.$$

With this notation, we have

$$\begin{aligned}\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\tilde\mu,id}}(\boldsymbol{X},\boldsymbol{X}^\#) &= \mathbf{E}_\mu\bigg[\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_1}}(X_1)\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_2}}(X_2)\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_1}}(X_1^\#)\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_2}}(X_2^\#)\bigg]\\ &= \mathbf{E}_\mu\bigg[\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_1}}(X_1)\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_2}}(X_2^\#)\bigg]\times\mathbf{E}_\mu\bigg[\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_2}}(X_2)\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_1}}(X_1^\#)\bigg]\\ &= \prod_{k=1}^d\cosh\Big(\frac{\epsilon_1}{\sigma_1^2}(X_{1,k}+X_{2,k}^\#)\Big)\cosh\Big(\frac{\epsilon_2}{\sigma_2^2}(X_{2,k}+X_{1,k}^\#)\Big)\\ &\quad\times\exp\Big\{-\frac{1}{2}(\|X_1^\#\|^2-\|X_2^\#\|^2)(\sigma_2^{-2}-\sigma_1^{-2})\Big\}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{d\mathbf{P}_{\mu,id}}{d\mathbf{P}_{\tilde{\mu},id}}(\boldsymbol{X},\boldsymbol{X}^\#) &= \mathbf{E}_\mu\left[\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_1}}(X_1)\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_2}}(X_2)\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_1}}(X_1^\#)\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_2}}(X_2^\#)\right]\\ &= \mathbf{E}_\mu\left[\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_1}}(X_1)\frac{d\mathbf{P}_{\theta_1,\sigma_1}}{d\mathbf{P}_{0,\sigma_1}}(X_1^\#)\right]\times\mathbf{E}_\mu\left[\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_2}}(X_2)\frac{d\mathbf{P}_{\theta_2,\sigma_2}}{d\mathbf{P}_{0,\sigma_2}}(X_2^\#)\right]\\ &= \prod_{k=1}^d \cosh\Big(\frac{\epsilon_1}{\sigma_1^2}(X_{1,k}+X_{1,k}^\#)\Big)\cosh\Big(\frac{\epsilon_2}{\sigma_2^2}(X_{2,k}+X_{2,k}^\#)\Big).\end{aligned}$$

Thus, we get that

$$\begin{aligned}\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\mu,id}}(\boldsymbol{X},\boldsymbol{X}^\#) = \prod_{k=1}^d \frac{\cosh\big(\frac{\epsilon_1}{\sigma_1^2}(X_{1,k}+X_{2,k}^\#)\big)}{\cosh\big(\frac{\epsilon_1}{\sigma_1^2}(X_{1,k}+X_{1,k}^\#)\big)}\times\prod_{k=1}^d \frac{\cosh\big(\frac{\epsilon_2}{\sigma_2^2}(X_{2,k}+X_{1,k}^\#)\big)}{\cosh\big(\frac{\epsilon_2}{\sigma_2^2}(X_{2,k}+X_{2,k}^\#)\big)}\\ \times\exp\Big\{-\frac{1}{2}(\|X_1^\#\|^2-\|X_2^\#\|^2)(\sigma_2^{-2}-\sigma_1^{-2})\Big\}.\end{aligned}$$

Then, we compute the Kullback-Leibler divergence,

$$\begin{aligned}K(\mathbf{P}_{\mu,\pi},\mathbf{P}_{\mu,id}) &= \int\log\Big(\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\mu,id}}(\boldsymbol{X},\boldsymbol{X}^\#)\Big)d\mathbf{P}_{\mu,\pi}(\boldsymbol{X},\boldsymbol{X}^\#)\\ &= \sum_{k=1}^d\sum_{j=1}^2\Big\{\mathbf{E}_\mu\Big[\int\log\cosh\Big[\frac{\epsilon_j}{\sigma_j^2}(2\theta_{j,k}+\sigma_j\sqrt{2}x)\Big]\varphi(x)dx\Big]\\ &\quad-\mathbf{E}_\mu\Big[\int\log\cosh\Big[\frac{\epsilon_j}{\sigma_j^2}(\theta_{1,k}+\theta_{2,k}+\sigma_{12}x)\Big]\varphi(x)dx\Big]\Big\}\\ &\quad+\frac{d}{2}\mathbf{E}_\mu\Big[\int_{\mathbb{R}}((\theta_{1,1}+\sigma_1x)^2-(\theta_{2,1}+\sigma_2x)^2)\varphi(x)\,dx\Big](\sigma_2^{-2}-\sigma_1^{-2}),\end{aligned}$$

where φ is the density function of the standard Gaussian distribution. We evaluate the first two terms of the last display using the following inequalities:

$$\forall u\in\mathbb{R},\qquad \frac{u^2}{2}-\frac{u^4}{12}\le\log\cosh(u)\le\frac{u^2}{2},\tag{25}$$

while for the third term the exact computation yields:

$$\begin{aligned}\mathbf{E}_\mu\Big[\int_{\mathbb{R}}((\theta_{1,1}+\sigma_1x)^2-(\theta_{2,1}+\sigma_2x)^2)\varphi(x)\,dx\Big] &= \epsilon_1^2+\sigma_1^2-\epsilon_2^2-\sigma_2^2\\ &= (\sigma_1^2-\sigma_2^2)(1+(2/d)\kappa^2).\end{aligned}$$

In conjunction with the facts that $\epsilon_1/\sigma_1 = \epsilon_2/\sigma_2$, $\sigma_1 \le \sigma_2$ and $\epsilon_1 \le \epsilon_2$, this leads to

$$\begin{aligned}
&\mathbf{E}_\mu\Big[\int \log\cosh\Big[\frac{\epsilon_j}{\sigma_j^2}(2\theta_{j,k} + \sigma_j\sqrt{2}x)\Big]\varphi(x)dx \le \frac{\epsilon_j^2}{\sigma_j^2} + 2\frac{\epsilon_j^4}{\sigma_j^4} = \frac{\epsilon_2^2}{\sigma_2^2} + 2\frac{\epsilon_2^4}{\sigma_2^4},\\
&\mathbf{E}_\mu\Big[\int \log\cosh\Big[\frac{\epsilon_j}{\sigma_j^2}(\theta_{1,k} + \theta_{2,k} + \sigma_{1,2}x)\Big]\varphi(x)dx\Big]\\
&\quad \ge \frac{\epsilon_j^2}{2\sigma_j^4}\Big(\epsilon_1^2 + \epsilon_2^2 + \sigma_1^2 + \sigma_2^2\Big)\\
&\qquad - \frac{\epsilon_j^4}{12\sigma_j^8}\Big(\epsilon_1^4 + \epsilon_2^4 + 3(\sigma_1^2 + \sigma_2^2)^2 + 6\epsilon_1^2\epsilon_2^2 + 6(\sigma_1^2 + \sigma_2^2)(\epsilon_1^2 + \epsilon_2^2)\Big)\\
&\quad \ge \frac{\epsilon_2^2(\epsilon_1^2 + \sigma_1^2)}{\sigma_2^4} - \frac{\epsilon_1^4(\epsilon_2^2 + \sigma_2^2)^2}{\sigma_1^8}.
\end{aligned}$$

Thus, we get that

$$\begin{aligned}
(1/d)K(\mathbf{P}_{\mu,\pi}, \mathbf{P}_{\mu,id}) &\le \frac{2\epsilon_2^2}{\sigma_2^2} + \frac{4\epsilon_2^4}{\sigma_2^4} - \frac{2\epsilon_2^2(\epsilon_1^2 + \sigma_1^2)}{\sigma_2^4} + \frac{2\epsilon_1^4(\epsilon_2^2 + \sigma_2^2)^2}{\sigma_1^8}\\
&\quad + \frac{1}{2}\big(1 + (2/d)\kappa^2\big)(\sigma_1^2 - \sigma_2^2)(\sigma_2^{-2} - \sigma_1^{-2})\\
&\le \frac{4\kappa^2}{d}\Big(1 - \frac{\sigma_1^2}{\sigma_2^2}\Big) + \frac{16\kappa^4}{d^2} + \frac{8\kappa^4}{d^2}\big(1 + (2/d)\kappa^2\big)^2\frac{\sigma_2^4}{\sigma_1^4}\\
&\quad + \frac{1}{2}\big(1 + (2/d)\kappa^2\big)\Big(\frac{\sigma_2^2}{\sigma_1^2} - 1\Big)^2.
\end{aligned}$$

To complete the proof, we need to evaluate $\mu(\mathcal{E} \setminus \bar{\Theta}_\kappa)$. We note that in view of the union bound,

$$\begin{aligned}
\mu(\mathcal{E} \setminus \bar{\Theta}_\kappa) &= \mu\Big(\bigcup_{k=1}^n \bigcup_{k' \ne k} \{\boldsymbol{\theta} : \|\theta_k - \theta_{k'}\| < \kappa\sigma_{k,k'}\}\Big)\\
&\le \frac{n(n-1)}{2}\max_{k \ne k'} \mu\big(\{\boldsymbol{\theta} : \|\theta_k - \theta_{k'}\|^2 < \kappa^2\sigma_{k,k'}\}\big)\\
&= \frac{n(n-1)}{2}\max_{k \ne k'} \mathbf{P}\big(d\epsilon_k^2 + d\epsilon_{k'}^2 - 2d\epsilon_k\epsilon_{k'}\bar{\zeta} < \kappa^2\sigma_{k,k'}^2\big),
\end{aligned}$$

where $\bar{\zeta} = \frac{1}{d}\sum_{j=1}^d \zeta_j$ with $\zeta_1, \dots, \zeta_d$ being i.i.d. Rademacher random variables (*i.e.*, random variables taking the values $+1$ and -1 with probability $1/2$). One easily checks that

$$\frac{d\epsilon_k^2 + d\epsilon_{k'}^2 - \kappa^2\sigma_{k,k'}^2}{2d\epsilon_k\epsilon_{k'}} = \frac{2\sigma_k^2 + 2\sigma_{k'}^2 - (\sigma_k^2 + \sigma_{k'}^2)}{4\sigma_k\sigma_{k'}} \ge \frac{1}{2}.$$

Therefore, using the Hoeffding inequality, we get $\mu(\mathcal{E} \setminus \bar{\Theta}_\kappa) \le \frac{1}{2}n(n-1)\mathbf{P}\big(\bar{\zeta} > 1/2\big) \le \frac{1}{2}n(n-1)e^{-d/8}$.

Proof of Lemma 16 We first prove an auxiliary result.

Lemma 17 *For any integer $n \geq 2$ there exist permutations $\pi_0, \pi_1, \ldots, \pi_M$ in $\mathfrak{S}_n$ such that*

$$\pi_0 = id, \qquad M \geq (n/8)^{n/2}$$

and for any pair $i, j \in \{0, \ldots, M\}$ of distinct indices we have $\delta_H(\pi_i, \pi_j) \geq \frac{1}{2}$.

Proof When $n \leq 8$, the claim of this lemma is trivial since one can always find at least one permutation that differs from the identity at all the positions and thus $M \geq 1 \geq (n/8)^{n/2}$. Let us consider the case $n > 8$. For every $\pi \in \mathfrak{S}_n$, denote

$$E_\pi \triangleq \Big\{\pi' \in \mathfrak{S}_n \,\Big|\, \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\pi(i)\neq\pi'(i)\}} \geq 1/2\Big\}.$$

We first notice that for every $\pi \in \mathfrak{S}_n$, there is a one-to-one correspondence between E_{id} and E_π through the bijection

$$\phi : \left| \begin{array}{ccc} E_{id} & \longrightarrow & E_\pi \\ \pi' & \longmapsto & \pi \circ \pi' \end{array} \right. ,$$

so that $\#E_\pi = \#E_{id}$. The following lemma, proved later on in this section, gives a bound for this number.

Lemma 18 *Let $n \geq 2$ be an integer and m be the smallest integer such that $2m \geq n$. Then*

$$\#E_{id}^{\complement} \leq \frac{4n!}{m!}.$$

Now we denote $\pi_0 = id$ and choose π_1 in E_{id}. Then, it is sufficient to choose π_2 as any element from $E_{id} \cap E_{\pi_1}$, the latter set being nonempty since

$$\begin{aligned} \#\big(E_{\pi_0} \cap E_{\pi_1}\big) &\geq \#\mathfrak{S}_n - \#E_{\pi_0}^{\complement} - \#E_{\pi_1}^{\complement} \\ &\geq n! \times \Big(1 - \frac{8}{m!}\Big) > 0. \end{aligned}$$

We can continue the construction until π_i if

$$1 - \frac{4i}{m!} > 0 \qquad \Longleftrightarrow \qquad i < \frac{m!}{4}.$$

To conclude, we observe that

$$\frac{m!}{4} > \frac{1}{4}\Big(\frac{n}{2e}\Big)^{n/2} \geq \Big(\frac{n}{8}\Big)^{n/2}.$$

■

Let us denote by m the largest integer such that $2m \leq n$, and choose

$$\tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_M \in \mathfrak{S}_m \quad \text{ with } \quad M \geq (m/8)^{m/2}$$

as in Lemma 17, so that for every $i \neq j \in \{0, \ldots, M\}$, $\delta_H(\tilde{\pi}_i, \tilde{\pi}_j) \geq \frac{1}{2}$. We use each permutation $\tilde{\pi}_i \in \mathfrak{S}_m$ to construct a permutation $\pi_i \in \mathfrak{S}_n$. The idea of the construction

is as follows: the permutation π_i is a product of m transpositions of distinct supports, and each transposition permutes an even integer with an odd one. We set $\pi_0 = id$ and for every i in $\{1, \dots, M\}$,

$$\pi_i = \big(1\ 2\tilde{\pi}_i(1)\big) \circ \big(3\ 2\tilde{\pi}_i(2)\big) \circ \ldots \circ \big(2m-1\ 2\tilde{\pi}_i(m)\big) \in \mathfrak{S}_n.$$

With these choices, the number of differences between π_i and π_j is exactly twice as much as the number of differences between $\tilde{\pi}_i$ and $\tilde{\pi}_j$. To sum up, for every pair of distinct indices $i, j \in \{0, \dots, M\}$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{\pi_i(k)\neq\pi_j(k)\}} = \frac{2m}{n} \times \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}_{\{\tilde{\pi}_i(k)\neq\tilde{\pi}_j(k)\}} \geq \frac{m}{n} \geq \frac{3}{8}, \qquad \forall n \geq 4.$$

To complete the proof, we note that $m \geq n/3$.

Proof of Lemma 18 For every $\ell \in \{m, \dots, n\}$, counting all the permutations π such that $\sum_{k=1}^{n} \mathbb{1}_{\{\pi(k)\neq k\}} = \ell$, we get

$$\#E_{id} = !n + !(n-1)\binom{n}{1} + \ldots + !(n-m)\binom{n}{m},$$

where $!\ell$ is the number of derangements, the permutations such that none of the elements appear in their original position, in $\mathfrak{S}_\ell$ for $\ell \geq 1$. We know that

$$\forall\ \ell \geq 1, \qquad !\ell = \ell! \times \sum_{j=0}^{\ell} \frac{(-1)^j}{j!},$$

which, using the alternating series test, yields

$$\forall\ \ell \geq 1, \qquad !\ell \geq \ell! \times \Big(e^{-1} - \frac{1}{(\ell+1)!}\Big).$$

It follows that

$$\begin{aligned}
\#E_{id} &\geq n! \times \Big(e^{-1} - \frac{1}{(n-m+1)!}\Big) \times \Big(1 + \frac{1}{1!} + \ldots + \frac{1}{m!}\Big) \\
&\geq n! \times \Big(e^{-1} - \frac{1}{(n-m+1)!}\Big) \times \Big(e - \frac{e}{(m+1)!}\Big) \\
&\geq n! \times \Big(1 - \frac{e}{(n-m+1)!} - \frac{1}{(m+1)!}\Big).
\end{aligned}$$

Therefore,

$$\#E_{id}^{\complement} \leq n! \times \Big(\frac{e}{(n-m+1)!} + \frac{1}{(m+1)!}\Big) \leq \frac{4n!}{m!}.$$

Acknowledgments

This work was partially supported by the grants Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and CALLISTO. The authors thank the Reviewers for many valuable suggestions. Special thanks to an anonymous Referee for pointing a mistake in the proof of Theorem 4.

References

R. R. Bahadur. On the asymptotic efficiency of tests and estimates. *Sankhyā*, 22:229–252, 1960.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, jun 2008.

Eric Budish, Yeon-Koo Che, Fuhito Kojima, and Paul Milgrom. Designing random allocation mechanisms: Theory and applications. *American Economic Review*, 103(2):585–623, 2013.

Olivier Collier. Minimax hypothesis testing for curve registration. *Electron. J. Stat.*, 6: 1129–1154, 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. Curve registration by nonparametric goodness-of-fit testing. *J. Statist. Plann. Inference*, 162:20–42, July 2015.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, second edition, 2003.

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.

Tony Jebara. Images as bags of pixels. In *ICCV*, pages 265–272. IEEE Computer Society, 2003.

A. P. Korostelev and V. G. Spokoiny. Exact asymptotics of minimax Bahadur risk in Lipschitz regression. *Statistics*, 28(1):13–24, 1996.

Alexander Korostelev. A minimaxity criterion in nonparametric regression based on large-deviations probabilities. *Ann. Statist.*, 24(3):1075–1083, 1996.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, 2:83–97, 1955.

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

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

David W. Pentico. Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, 176(2):774–793, 2007.

H. Edwin Romeijn and Dolores Romero Morales. A class of greedy algorithms for the generalized assignment problem. *Discrete Applied Mathematics*, 103(1-3):209–235, 2000.

V.G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6): 2477–2498, 1996.

Jos F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optim. Methods Softw.*, 11/12(1-4):625–653, 1999.

Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.