Permutation estimation and minimax rates of identifiability

Olivier Collier, Arnak S. Dalalyan

Permutation estimation and minimax rates of identifiability

Olivier Collier
IMAGINE, Université Paris-Est

Arnak Dalalyan
CREST, ENSAE

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 notion of the minimax rate of identifiability is introduced and its expression is obtained as a function of the sample size, noise level and dimensionality. We consider the cases of homoscedastic and heteroscedastic noise and carry out, in each case, upper bounds on the identifiability threshold of several estimators. This upper bounds are shown to be unimprovable in the homoscedastic setting. We also discuss the computational aspects of the estimators and provide empirical evidence of their consistency on synthetic data.

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 $\mathbb{R}^d$ containing many matching elements. That is, for many X_is 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 ensuring the identifiability of π^*.

In order to define a statistical framework making it possible 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 identify 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 *identifiability threshold* and will be the main concept of interest in the present study.

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. [11, 5]. The features are usually local descriptors that serve to summarize the images. The most famous examples of such features are perhaps SIFT [9] and SURF [1]. 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.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume XX of JMLR: W&CP XX.

[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.

1.2 Main contributions

We consider four procedures of permutation estimation: 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 and three estimators defined as minimizers of the 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.

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 dimension d, the noise level σ and the sample size n. This function is the same for the four aforementioned procedures and is given, up to a multiplicative factor, by

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

We then prove that this expression provides the optimal rate of the identifiability threshold in the sense that for some absolute constant c if $\kappa \leq c\kappa^*(d, \sigma, n)$ 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. This result shows that 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 show that the three estimators stemming from the maximum likelihood methodology are efficiently computable by linear programming.

2 Notation and problem formulation

We consider that $n = m \geq 2$ and the two sets of features $\{X_1, \ldots, X_n\}$ and $\{X_1^\#, \ldots, X_n^\#\}$ are randomly generated from the model

$$\begin{cases} X_i = \theta_i + \sigma_i \xi_i \ , \\ X_i^\# = \theta_{\pi^*(i)} + \sigma_i^\# \xi_i^\#, \end{cases} \qquad i = 1, \ldots, n \qquad (1)$$

where

- $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ is a collection of distinct vectors from $\mathbb{R}^d$, corresponding to the original features, which are unavailable,
- $\sigma_1, \ldots, \sigma_n, \sigma_1^\#, \ldots, \sigma_n^\#$ are positive real numbers corresponding to the levels of noise contaminating each feature,
- $\xi_1, \ldots, \xi_n$ and $\xi_1^\#, \ldots, \xi_n^\#$ are two independent sets of i.i.d. random vectors drawn from the Gaussian distribution with zero mean and identity covariance matrix,
- π^* is a permutation of $\{1, \ldots, n\}$.

In this formulation, there are three (sets of) unknown parameters: $\boldsymbol{\theta}$, $\boldsymbol{\sigma} = \{\sigma_i, \sigma_i^\#\}_{i=1,\ldots,n}$ and π^*. However, we will focus our attention on the problem 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 $(X_1, \ldots, X_n, X_1^\#, \ldots, X_n^\#)$ defined by (1). The set of all permutations of $\{1, \ldots, n\}$ will be denoted by $\mathfrak{S}_n$.

Let us denote by $\kappa(\boldsymbol{\theta})$ the smallest Euclidean distance between two distinct features:

$$\kappa(\boldsymbol{\theta}) \triangleq \min_{i \neq j} \|\theta_i - \theta_j\|. \qquad (2)$$

It is clear that if $\kappa(\boldsymbol{\theta}) = 0$, 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 $\kappa(\boldsymbol{\theta}) > 0$ is necessary for the existence of consistent estimators of π^*. Furthermore, good estimators are those consistently estimating π^* even if $\kappa(\boldsymbol{\theta})$ is small. To give a precise sense to these considerations, let $\alpha \in (0, 1)$ be a prescribed tolerance level and let us call *identifiability threshold* of a given estimation procedure $\hat{\pi}$ the quantity

$$\kappa_{\alpha,\boldsymbol{\sigma},n,d}(\hat{\pi}) = \inf\Big\{\kappa > 0 : \sup_{\substack{\boldsymbol{\theta}:\kappa(\boldsymbol{\theta})>\kappa \\ \pi \in \mathfrak{S}_n}} \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\sigma},\pi}(\hat{\pi} \neq \pi) \leq \alpha\Big\}.$$

In the next section, we establish nonasymptotic upper bounds on the identifiability threshold of several natural estimators. We then define the minimax rate of identifiability as

$$\kappa_{\alpha,\boldsymbol{\sigma},n,d} = \inf_{\hat{\pi}} \kappa_{\alpha,\boldsymbol{\sigma},n,d}(\hat{\pi}),$$

where the inf is taken over all possible estimators of π^*. In the homoscedastic case, we will show that the aforementioned upper bounds coincide with the minimax rate of identifiability up to a multiplicative constant.

3 Theoretical results

3.1 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

$$\pi^{\mathrm{gr}}(i) = \arg \min_{j\notin\{\pi^{\mathrm{gr}}(1),\dots,\pi^{\mathrm{gr}}(i-1)\}} \|X_j - X_i^{\#}\|. \quad (3)$$

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 in the homoscedastic setting this estimator possesses nice optimality properties.

The three other estimators, termed least sum of squares, least sum of normalized squares and least sum of logarithms, are defined as

$$\pi^{\mathrm{LSS}} = \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^{n} \|X_{\pi(i)} - X_i^{\#}\|^2, \quad (4)$$

$$\pi^{\mathrm{LSNS}} = \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^{n} \frac{\|X_{\pi(i)} - X_i^{\#}\|^2}{\sigma_{\pi(i)}^2 + {\sigma_i^{\#}}^2}, \quad (5)$$

$$\pi^{\mathrm{LSL}} = \arg\min_{\pi\in\mathfrak{S}_n} \sum_{i=1}^{n} \log \|X_{\pi(i)} - X_i^{\#}\|^2. \quad (6)$$

A first remark concerning these three estimators is that the LSS and the LSL are adaptive with respect to the noise level $\boldsymbol{\sigma}$, while the computation of the LSNS requires the knowledge of $\boldsymbol{\sigma}$. A second remark is that these estimators can be seen as maximum likelihood (ML) estimators under different settings. The LSS corresponds to the ML-estimator when $\sigma_i = \sigma_i^{\#} = \sigma$ for every i (homoscedastic noise). The LSNS corresponds to the ML-estimator when the noise is not necessarily homoscedastic, but the noise level $\boldsymbol{\sigma}$ is known. More interestingly, and this is less obvious, the LSL corresponds to the ML-estimator when the noise is heteroscedastic with unknown noise level $\boldsymbol{\sigma}$ satisfying the relation $\sigma_i^{\#} = \sigma_{\pi^*(i)}$ for every i. Finally, a third observation deserving to be mentioned is that the computation of these three estimators can be done efficiently by linear programming, and the complexity of this computation is comparable to that of the greedy algorithm.

3.2 Minimax rates of identifiability

The purpose of this section is to present our main theoretical results providing guarantees for the consistency of the aforementioned estimators in terms of upper bounds on their identifiability threshold. This makes it possible to compare different procedures and to evaluate how far they are from the optimal ones. We start by considering the homoscedastic case, in which upper and lower bounds matching up to a constant are obtained. Note that in this setting the LSNS estimator coincides with the LSS estimator.

Theorem 1. *Let $\alpha \in (0,1)$ be a tolerance level and let the noise level $\boldsymbol{\sigma}$ be a constant vector: $\sigma_i = \sigma_i^{\#} = \sigma$ for all $i \in \{1,\dots,n\}$. Then, if $\hat\pi$ denotes either one of the estimators (3)-(6), we have*

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

An equivalent way of stating this result is that if

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

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

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

for all the estimators defined in the previous subsection. The strength of this result is that it is nonasymptotic and holds for any vector $\boldsymbol{\theta}$ and any tolerance level α. Roughly speaking, it tells us that the identifiability threshold of the procedures under consideration is at most of the order of

$$\sigma \max\Big\{(\log n)^{1/2}, (d\log n)^{1/4}\Big\}. \quad (7)$$

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. Furthermore, as stated in the next theorem, this bound is optimal up to a multiplicative constant.

Theorem 2. *Assume that $n \ge 4$. There exist two absolute constants $c > 0$ and $C > 0$ such that if*

$$\kappa \le \frac{\sigma}{4} \max\Big\{(\log n)^{1/2}, c(d\log n)^{1/4}\Big\},$$

then,

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

where the infimum is taken over all permutation estimators.

We switch now to the heteroscedastic setting, which allows us to discriminate between the four procedures. One remarks immediately that the greedy algorithm, the LSS and the LSL have a serious advantage with respect to the LSNS in that they can be computed without knowing the noise level $\boldsymbol{\sigma}$.

The identifiability threshold here is slightly redefined in order to better reflect the variability of the noise level. We assume that for every $i = 1, \dots, n$, $\sigma_i^{\#} = \sigma_{\pi^*(i)}$ and define

$$\tilde{\kappa}(\boldsymbol{\theta}) \triangleq \min_{i \neq j} \frac{\|\theta_i - \theta_j\|}{\sqrt{\sigma_i^2 + \sigma_j^2}}.$$

Theorem 3. *Consider a real α in $(0,1)$. Set*

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

and denote by $\Theta_{\tilde{\kappa}}$ the set of all $\boldsymbol{\theta} \in \mathbb{R}^{n \times d}$ such that $\tilde{\kappa}(\boldsymbol{\theta}) \geq \tilde{\kappa}$. Then, if $\hat{\pi}$ is either π^{LSNS} (if the noise levels $\sigma_i, \sigma_i^{\#}$ are known) or π^{LSL} (when the noise levels are unknown), we have

$$\sup_{\boldsymbol{\theta} \in \Theta_{\tilde{\kappa}}} \max_{\pi^* \in \mathfrak{S}_n} \mathbf{P}_{\boldsymbol{\theta}, \pi^*}(\hat{\pi} \neq \pi^*) \leq \alpha.$$

It follows from this theorem that the identifiability thresholds of the LSNS and LSL estimators are upper bounded by

$$4 \max_{i \neq j} (\sigma_i^2 + \sigma_j^2)^{1/2} \Big\{ \Big(2 \log \frac{8n^2}{\alpha} \Big)^{1/2} \bigvee \Big(d \log \frac{4n^2}{\alpha} \Big)^{1/4} \Big\},$$

which coincides with the upper bound stated in Theorem 1 in the case of constant noise level. We believe that this expression provides the optimal rate of identifiability in the heteroscedastic setting, but we do not have the rigorous proof of this claim.

Note also that Theorem 3 does not tell anything about the theoretical properties of the greedy algorithm and the LSS under heteroscedasticity. In fact, the identifiability thresholds 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 numerical simulations presented in Section 6 below.

Theorem 4. *Let $d \geq 225 \log 6$ and $n = 2$. Consider the heteroscedastic setting described above with $\sigma_1^2 = 3$ and $\sigma_2^2 = 1$. Then, if $\tilde{\kappa} < 0.1(2d)^{1/2}$, we have*

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

This theorem shows that if d is large, the condition necessary for π^{gr} to be consistent is much stronger than the one obtained for π^{LSL} in Theorem 3. Indeed, for the consistency of π^{gr}, $\tilde{\kappa}$ needs to be at least of the order of $d^{1/2}$, whereas $d^{1/4}$ is sufficient for the consistency of π^{LSL}.

4 Computational aspects

At first sight, the computation of the estimators (4)-(6) 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 as soon as $n \geq 20$. In this section, we show how to compute these (maximum likelihood) estimators in polynomial time using linear programming[2].

For instance, let us consider the LSS estimator

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

For every permutation π, we denote by P^π the $n \times n$ (permutation) matrix with coefficients $P_{ij}^\pi = \mathbb{1}_{\{j = \pi(i)\}}$. Then it is equivalent to compute the estimator

$$\pi^{\mathrm{LSS}} = \arg \min_{\pi \in \mathfrak{S}_n} \mathbf{tr}\big(MP^\pi\big), \tag{8}$$

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.

Theorem 5 (cf., for instance, Budish et al. [2])**.** *Let $\mathcal{P}$ be 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 (8) is equivalent to the following problem of continuous optimization:

$$P^{\mathrm{LSS}} = \arg \min_{P \in \mathcal{P}} \mathbf{tr}\big(MP\big), \tag{9}$$

in the sense that if π is a solution to (8), then P^π is a solution to (9). 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

$$P = \sum_{i=1}^{n!} \alpha_i P^{\pi_i} \text{ and } \sum_{i=1}^{n!} \alpha_i = 1.$$

Therefore,

$$\mathbf{tr}\big(MP\big) = \sum_{i=1}^{n!} \alpha_i \mathbf{tr}\big(MP^{\pi_i}\big) \geq \min_{\pi \in \mathfrak{S}_n} \mathbf{tr}\big(MP^\pi\big)$$

and

$$\mathbf{tr}\big(MP^{\mathrm{LSS}}\big) \geq \mathbf{tr}\big(MP^{\pi^{\mathrm{LSS}}}\big).$$

[2]This idea has been already used in the literature; see, for instance, Jebara [7]

The great advantage of (9) 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. It is clear that the same arguments apply to the estimators π^{LSNS} and π^{LSL} (only the matrix M needs to be changed).

5 Possible extensions

When considering the problem of permutation estimation, it may be relevant to define the risk of an algorithm $\hat{\pi}$ as the average rate of incorrect matches provided by the algorithm:

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

instead of the probability of failing to correctly estimate the whole permutation π^*. All the results presented in previous sections carry over this setting with minor modifications. Because of space limitations, these results are not contained in this paper but will be included in its extended version which is in preparation.

Another interesting extension concerns the case of arrangement estimation, *i.e.,* the case $m \neq n$. In such a situation, without loss of generality, one can assume that $n < m$ and look for an injective function $\pi^* : \{1,\dots,n\} \to \{1,\dots,m\}$. All the estimators presented in Section 3.1 admit natural counterparts in this "rectangular" setting. Furthermore, the computational tricks described in the previous section are valid in this setting as well, and are justified by the extension of the Birkhoff-von Neumann theorem recently proved by Budish et al. [2]. In this case, the minimization should be carried out over the set of all $n\times m$ matrices P such that $P_{i,j} \geq 0$, $\sum_{i=1}^{n} P_{i,j} \leq 1$ and $\sum_{j=1}^{m} P_{i,j} = 1$ for every $(i,j) \in \{1,\dots,n\}\times\{1,\dots,m\}$.

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.

6 Experimental results

We have implemented all the procedures in Matlab and carried out a certain number of numerical experiments on synthetic data. To simplify, we have used the general-purpose solver SeDuMi [10] 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 6 seconds to compute a solution to (9) on a standard PC.

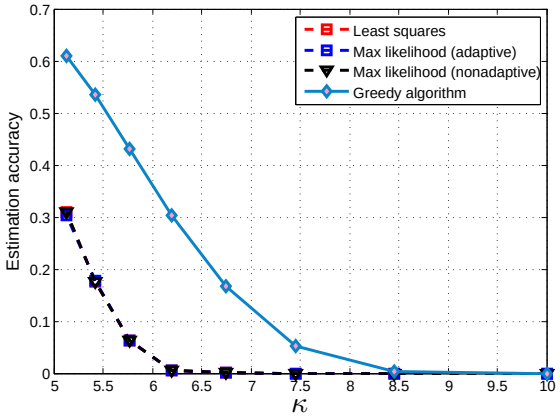

Figure 1: Average error rate of the four estimating procedures in the experiment with homoscedastic noise as a function of the minimal distance κ between distinct features. One can observe that the LSS, LSNS and LSL procedures are indistinguishable and, a bit surprisingly, perform much better than the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

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 (1) 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.

One can clearly observe 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 when the parameter κ is small.

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 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 (1) with $\sigma_i = \sigma^{\#}_{\pi^*(i)} = 1$ for 10 randomly chosen values of i and $\sigma_i = \sigma^{\#}_{\pi^*(i)} = 0.5$ for the others. Using these sets as data, we computed the four estimators of π^* and evaluated the average error rate

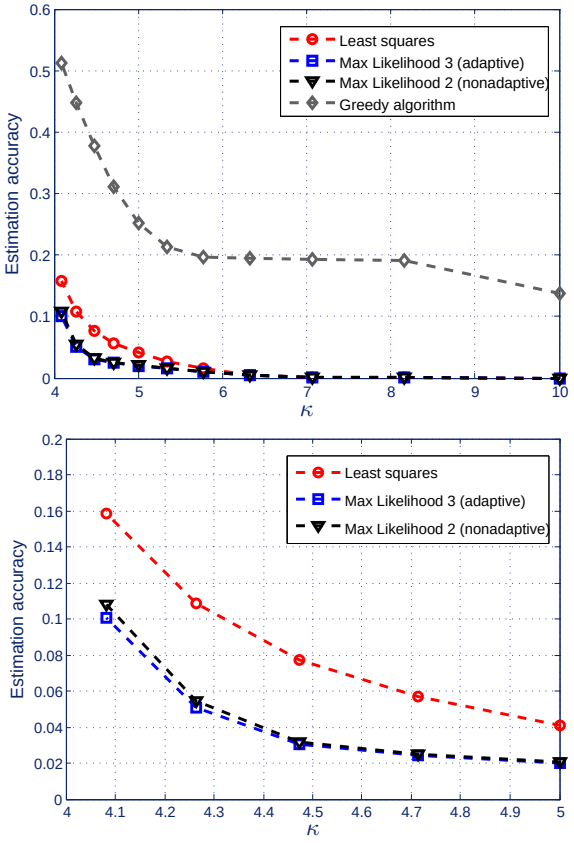

Figure 2: Top: Average error rate of the four estimating procedures in the experiment with heteroscedastic noise as a function of the minimal distance κ between distinct features. Bottom: zoom on the same plots. One can observe that the LSNS and LSL are almost indistinguishable and, as predicted by the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

$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{\pi}(i) \neq \pi^*(i))$. The result, averaged over 500 independent trials, is plotted in Fig. 2.

One can observe 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.

7 Conclusion

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 identification, which plays the same role as the minimax rate of separation in the statistical hypotheses testing theory [6]. 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 case. However, we have shown that the least sum of logarithms estimator outperforms the other procedures either theoretically or empirically.

8 Proofs of Theorems

In this section we collect the proofs of the theorems stated in Section 3.2. We start with the proof of Theorem 3, since it concerns the more general setting and the proof of Theorem 1 follows from that of Theorem 3 by simple arguments. We then prove Theorem 2 and postpone the proofs of some technical lemmas to the appendix.

8.1 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 $\Omega = \{\hat{\pi} \neq \pi^*\}$.

Let us first denote by $\hat{\pi}$ the maximum likelihood estimator π^{LSL} defined by (6). We have

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

where

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

On the one hand, for every permutation π, using the concavity of the logarithm we get

$$\begin{aligned}\sum_{\pi(i)\neq i} \log\big(\tfrac{2\sigma_i^2}{\sigma_i^2+\sigma_{\pi(i)}^2}\big) &= \sum_{i=1}^{n} \big(\log(2\sigma_i^2) - \log(\sigma_i^2+\sigma_{\pi(i)}^2)\big) \\ &= \sum_{i=1}^{n} \frac{\log(2\sigma_i^2) + \log(2\sigma_{\pi(i)}^2)}{2} - \log(\sigma_i^2+\sigma_{\pi(i)}^2) \\ &\leq 0.\end{aligned}$$

Therefore,

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

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

$$\bar{\Omega} = \bigcup_{i=1}^{n} \bigcup_{i\neq j} \Big\{ \frac{\|X_i - X_i^\#\|^2}{2\sigma_i^2} \geq \frac{\|X_j - X_i^\#\|^2}{\sigma_j^2+\sigma_i^2} \Big\}. \quad (10)$$

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

Let us denote

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

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

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

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,

$$\|X_j - X_i^\#\|^2 \ge \|\theta_j-\theta_i\|^2 + (\sigma_i^2+\sigma_j^2)(d - \sqrt{d}\zeta_2) - 2\|\theta_j-\theta_i\|\sqrt{\sigma_i^2+\sigma_j^2}\zeta_1.$$

This implies that on the event $\Omega_1 = \{\tilde\kappa(\boldsymbol\theta) \ge \zeta_1\}$ it holds that

$$\frac{\|X_j - X_i^\#\|^2}{\sigma_i^2+\sigma_j^2} \ge \tilde\kappa(\boldsymbol\theta)^2 - 2\tilde\kappa(\boldsymbol\theta)\zeta_1 + d - \sqrt{d}\zeta_2.$$

Combining these bounds, we get that $\Omega\cap\Omega_1 \subset \Big\{ d + \sqrt{d}\zeta_2 > \tilde\kappa(\boldsymbol\theta)^2 - 2\tilde\kappa(\boldsymbol\theta)\zeta_1 + d - \sqrt{d}\zeta_2 \Big\}$, which implies that

$$\begin{aligned}\mathbf{P}(\Omega) &\le \mathbf{P}(\Omega_1^\complement) + \mathbf{P}\big(\Omega\cap\Omega_1\big)\\ &\le \mathbf{P}\big(\zeta_1 \ge \tilde\kappa(\boldsymbol\theta)\big) + \mathbf{P}(2\sqrt{d}\zeta_2 + 2\tilde\kappa(\boldsymbol\theta)\zeta_1 > \tilde\kappa(\boldsymbol\theta)^2)\\ &\le 2\mathbf{P}\big(\zeta_1 \ge \tfrac{\tilde\kappa(\boldsymbol\theta)}{4}\big) + \mathbf{P}\big(\zeta_2 > \tfrac{\tilde\kappa(\boldsymbol\theta)^2}{4\sqrt{d}}\big). \qquad (11)\end{aligned}$$

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

$$\begin{aligned}\mathbf{P}\big(\zeta_1 \ge \tfrac14\tilde\kappa(\boldsymbol\theta)\big) &\le \sum_{i\neq j}\mathbf{P}\big(|\zeta_{i,j}| \ge \tfrac14\tilde\kappa(\boldsymbol\theta)\big)\\ &\le 2n^2 e^{-\frac{1}{32}\tilde\kappa(\boldsymbol\theta)^2}. \qquad (12)\end{aligned}$$

Similarly, using the tail bound of Lemma 4 stated in the supplementary material and borrowed from Laurent and Massart [8], we get

$$\mathbf{P}\big(\zeta_2 > \tfrac{\tilde\kappa(\boldsymbol\theta)^2}{4\sqrt{d}}\big) \le 2n^2 e^{-\frac{(\tilde\kappa(\boldsymbol\theta)/16)^2}{d}(\tilde\kappa^2(\boldsymbol\theta)\wedge 8d)}. \qquad (13)$$

Combining inequalities (11)-(13), we obtain that as soon as

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

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

8.2 Proof of Theorem 1

It is evident that on the event

$$\mathcal{A} = \bigcap_{i=1}^n \bigcap_{j\neq i} \Big\{ \|X_{\pi^*(i)} - X_i^\#\| < \|X_{\pi^*(i)} - X_j^\#\| \Big\}$$

all the four estimators coincide with the true permutation π^*. Therefore, we have

$$\{\hat\pi \neq \pi^*\} \subseteq \bigcup_{i=1}^n \bigcup_{j\neq i} \Big\{ \|X_{\pi^*(i)} - X_i^\#\| > \|X_{\pi^*(i)} - X_j^\#\| \Big\}.$$

The latter event coincides with $\bar\Omega$ at the right-hand side of (10), the probability of which has been already evaluated in the previous proof. This completes the proof.

8.3 Proof of Theorem 2

For two probability measures $\mathbf{P}$ and $\mathbf{Q}$ such that $\mathbf{P}$ is absolutely continuous with respect to $\mathbf{Q}$, we denote by $K(\mathbf{P},\mathbf{Q})$ the Kullback-Leibler divergence between $\mathbf{P}$ and $\mathbf{Q}$ defined by

$$K(\mathbf{P},\mathbf{Q}) = \int \log\frac{d\mathbf{P}}{d\mathbf{Q}}\, d\mathbf{P}.$$

In our proof, we decided to separate the cases when

$$\max\Big\{(\log n)^{1/2}, c(d\log n)^{1/4}\Big\} = (\log n)^{1/2}$$

and when

$$\max\Big\{(\log n)^{1/2}, c(d\log n)^{1/4}\Big\} = c(d\log n)^{1/4}.$$

Besides, we will repeatedly use the fact that for $n \ge 4$, $\log(n/2) \ge \frac12\log n$ and, to ease notation, we will omit the subscript $\boldsymbol\sigma$ in $\mathbf{P}_{\boldsymbol\theta,\boldsymbol\sigma,\pi}$.

First part: $\kappa \le (\sigma/4)\sqrt{\log n}$
Notice that in this first case $\kappa \le (\sigma/\sqrt{8})\sqrt{\log(n/2)}$, which is the bound we will use from now on.

We choose $\boldsymbol\theta_0 \in \mathbb{R}^{n\times d}$ such that $\theta_{0,i} = (i\times\kappa, 0, \dots, 0)^\top \in \mathbb{R}^d$ for every $i \in \{1,\dots,n\}$. We reduce our problem from considering Θ_κ to considering only $\boldsymbol\theta_0 \in \Theta_\kappa$:

$$\begin{aligned}&\inf_{\hat\pi}\sup_{\boldsymbol\theta\in\Theta_\kappa}\max_{\pi^*\in\mathfrak{S}_n}\mathbf{P}_{\boldsymbol\theta,\pi^*}\big(\hat\pi\neq\pi^*\big)\\ &\ge \inf_{\hat\pi}\max_{\pi^*\in\mathfrak{S}_n}\mathbf{P}_{\boldsymbol\theta_0,\pi^*}\big(\hat\pi\neq\pi^*\big).\end{aligned}$$

A lower bound for this quantity can be obtained from the following lemma:

Lemma 1 (Tsybakov [12]). *Let M be an integer larger than 2. Assume that there exist distinct permutations $\pi_0, \dots, \pi_M \in \mathfrak{S}_n$ and mutually absolutely continuous probability measures $\mathbf{Q}_0, \dots, \mathbf{Q}_M$ such that*

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

then

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

where the infimum is taken over all permutation estimators.

We will apply this lemma with $M = n-1$ and, for $i = 1, \dots, n$, π_i being the permutation which leaves all $j \in \{1, \dots, n\} \setminus \{i, i+1\}$ invariant and switches i and $i+1$. Then, we have

$$K\big(\mathbf{Q}_i, \mathbf{Q}_0\big) = \frac{1}{2\sigma^2}\sum_{k=1}^{n} \|\theta_{0,\pi_i(k)} - \theta_{0,k}\|^2 \le \frac{\kappa^2}{\sigma^2}.$$

Thus, if $n \ge 3$ and $\kappa \le (\sigma/\sqrt{8})\sqrt{\log(n/2)}$, then

$$\frac{1}{M}\sum_{i=1}^{M} K\big(\mathbf{Q}_i, \mathbf{Q}_0\big) = \frac{1}{8}\ \log(n/2) \le \frac{1}{8}\log M$$

and Lemma 1 yields the desired result.

Second part: $\kappa \le (c\sigma/4)(d\log n)^{1/4}$
In this second part, we suppose that $c \le 1$, so that $\kappa \le (\sigma/4)\big(d\log n\big)^{1/4}$, which is the bound we need to prove the result. Furthermore, the hypothesis made on the maximum implies that

$$d \ge \frac{1}{c^4}\log n \ge \log n.$$

Now, let μ be a (prior) probability measure on $\mathbb{R}^{n\times d}$. Define the (posterior) probability $\mathbf{P}_{\mu,\pi} = \int_{\mathbb{R}^{nd}} \mathbf{P}_{\boldsymbol\theta,\pi}\ \mu(d\boldsymbol\theta)$. It holds that

$$\begin{aligned}
&\sup_{\boldsymbol\theta\in\Theta_\kappa} \max_{\pi^*\in\mathfrak{S}_n} \mathbf{P}_{\boldsymbol\theta,\pi^*}\big(\hat\pi \neq \pi^*\big)\\
&\ge \max_{\pi^*\in\{id\}\cup\Pi} \int_{\Theta_\kappa} \mathbf{P}_{\boldsymbol\theta,\pi^*}\big(\hat\pi \neq \pi\big)\frac{\mu(d\boldsymbol\theta)}{\mu(\Theta_\kappa)}\\
&\ge \max_{\pi^*\in\{id\}\cup\Pi} \mathbf{P}_{\mu,\pi^*}\big(\hat\pi \neq \pi_j\big) - \mu(\mathbb{R}^{nd}\backslash\Theta_\kappa).
\end{aligned}$$

We will use Lemma 1 again with

$$\begin{cases} M = n, \\ \pi_0 = id \text{ and } \pi_1, \dots, \pi_M \ M \text{ distinct transpositions}, \\ \forall i \in \{0, \dots, M\}, \mathbf{Q}_i = \mathbf{P}_{\mu,\pi_i}. \end{cases}$$

To this end, we state the following lemma, that allows us to bound the Kullback-Leibler divergence from above.

Lemma 2. *Let ϵ be a positive real number with $\epsilon \le \sigma/2$ and let μ be the uniform distribution on $\{\pm\epsilon\}^{n\times d}$. Then, for any transposition π, we have*

$$K(\mathbf{P}_{\mu,\pi}, \mathbf{P}_{\mu,id}) \le \frac{8d\epsilon^4}{\sigma^4}.$$

Furthermore, if $\epsilon = \kappa/\sqrt{d}$, then

$$\mu(\mathbb{R}^{nd}\backslash\Theta_\kappa) \le \frac{n(n-1)}{2}\ e^{-d/8}.$$

Using the prior μ and the value of $\epsilon \le \sigma/2$ defined in the previous lemma, we get

$$\frac{1}{M}\sum_{i=1}^{M} K(\mathbf{Q}_i, \mathbf{Q}_0) \le \frac{8d\epsilon^4}{\sigma^4} = \frac{8\kappa^4}{d\sigma^4} \le \frac{1}{8}\log n.$$

This implies that the minimum risk is larger than

$$\frac{\sqrt{3}}{\sqrt{3}+1}\Big(\frac{3}{4} - \frac{1}{2\sqrt{\log 3}}\Big) - \frac{n^2}{2}e^{-d/8}.$$

Finally, remembering that $d \ge \frac{1}{c^4}\log n$, we have

$$(n^2/2)e^{-d/8} \le \frac{1}{2}n^{2-1/8c^4}.$$

Taking c small enough, we get the desired result.

8.4 Proof of Theorem 4

It is clear that $\{\pi^{\mathrm{gr}} \neq \pi^*\} \supset \{\|X_1 - X_1^{\#}\|^2 > \|X_1 - X_2^{\#}\|^2\} := \Omega_2$. In the following, we lower bound the probability of the event Ω_2. Let us choose any $\boldsymbol\theta$ from $\mathbb{R}^{n\times d}$ satisfying $\|\theta_1 - \theta_2\| = 2\tilde\kappa$.

One easily checks that for suitably chosen random variables $\eta_1 \sim \chi^2_d$, $\eta_2 \sim \chi^2_d$ and $\zeta_3 \sim \mathcal{N}(0,1)$ it holds that

$$\|X_1 - X_1^{\#}\|^2 - \|X_1 - X_2^{\#}\|^2 = 6\eta_1 - 4\tilde\kappa^2 - 8\tilde\kappa\zeta_3 - 4\eta_2.$$

According to Lemma 4 stated in the supplementary material, for every $x > 0$, each one of the following three inequalities holds true with probability at least $1 - e^{-x^2}$:

$$\begin{aligned}
\eta_1 &\ge d - 2\sqrt{d}x,\\
\eta_2 &\le d + 2\sqrt{d}x + 2x^2,\\
\zeta_3 &\le \sqrt{2}x.
\end{aligned}$$

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

$$\begin{aligned}
\|X_1 - X_1^{\#}\|^2 - \|X_1 - X_2^{\#}\|^2 \\
\ge 2d - 20\sqrt{d}x - 4(\tilde\kappa + \sqrt{2}x)^2.
\end{aligned}$$

If $x = \sqrt{\log 6}$, then the conditions imposed in Theorem 4 on $\tilde\kappa$ and d ensure that the right-hand side of the last inequality is positive. Therefore, $\mathbf{P}(\bar\Omega) \ge 1 - 3e^{-x^2} = 1/2$.

References

[1] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding*, 110: 346–359, 2008.

[2] E. Budish, Y.K. Che, F. Kojima and P. Milgrom. Designing Random Allocation Mechanisms: Theory and Applications. *American Economic Review*, to appear, 2012.

[3] O. Collier. Minimax hypothesis testing for curve registration. *Electron. J. Statist.*, 6:1129–1154, 2012.

[4] O. Collier and A. Dalalyan. Wilks' phenomenon and penalized likelihood-ratio test for nonparametric curve registration. *Journal of Machine Learning Research - Proceedings Track (AISTATS 2012)*, 22: 264-272, 2012.

[5] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision. Second edition.* Cambridge University Press, Cambridge, 2003.

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

[7] Jebara T. Images as Bags of Pixels. Presented at *IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

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

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

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

[11] R. Szeliski. *Computer Vision: Algorithms and Applications.* Springer, New York, 2010.

[12] A.B. Tsybakov. Introduction to Nonparametric Estimation. *Springer Verlag*, 2008.

Appendix

Lemma 3. *For every $x \in \mathbb{R}$, it holds that*

$$\frac{x^2}{2} - \frac{x^4}{12} \le \log\cosh(x) \le \frac{x^2}{2}.$$

Lemma 4 (Laurent and Massart [8])**.** *Let $Y \sim \chi^2(D)$, where $D \in \mathbb{N}^*$. Then, for every $x > 0$,*

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

As a consequence, for every $y > 0$,

$$\mathbf{P}\big(D^{-1/2}|Y - D| \ge y\big) \le 2\exp\Big\{-\tfrac{1}{8}y(y \wedge \sqrt{D})\Big\}.$$

Proof of Lemma 2. Let π be the transposition (i, j). Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ be randomly drawn from μ. We define $\boldsymbol{\theta}'$ by $\theta'_i = \theta'_j = 0$ and $\theta'_k = \theta_k$ if $k \notin \{i, j\}$. Let us denote by $\tilde{\mu}$ the probability distribution of this random matrix on $\mathbb{R}^{n\times d}$ and set $\mathbf{P}_{\tilde{\mu},\pi} = \int_\Theta \mathbf{P}_{\boldsymbol{\theta},\pi}\, \tilde{\mu}(d\boldsymbol{\theta})$.

We first compute the likelihood ratio of $\mathbf{P}_{\mu,\pi}$ and $\mathbf{P}_{\mu,id}$. We get, for every $(X, Y) = (X_1, \dots, X_n, Y_1, \dots, Y_n)$ in $\mathbb{R}^{nd}$,

$$\begin{aligned}
\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\mu,id}}(X,Y) &= \frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\tilde{\mu}}}(X,Y) \times \Big(\frac{d\mathbf{P}_{\mu,id}}{d\mathbf{P}_{\tilde{\mu}}}(X,Y)\Big)^{-1} \\
&= \mathbf{E}_\mu\left[\frac{d\mathbf{P}_{\theta_i}}{d\mathbf{P}_0}(X_i)\frac{d\mathbf{P}_{\theta_j}}{d\mathbf{P}_0}(X_j)\frac{d\mathbf{P}_{\theta_j}}{d\mathbf{P}_0}(Y_i)\frac{d\mathbf{P}_{\theta_i}}{d\mathbf{P}_0}(Y_j)\right] \\
&\times \mathbf{E}_\mu^{-1}\left[\frac{d\mathbf{P}_{\theta_i}}{d\mathbf{P}_0}(X_i)\frac{d\mathbf{P}_{\theta_j}}{d\mathbf{P}_0}(X_j)\frac{d\mathbf{P}_{\theta_i}}{d\mathbf{P}_0}(Y_i)\frac{d\mathbf{P}_{\theta_j}}{d\mathbf{P}_0}(Y_j)\right].
\end{aligned}$$

Now, reminding that for example

$$\frac{d\mathbf{P}_{\theta_i}}{d\mathbf{P}_0}(X_i) = e^{-\frac{\epsilon^2 d}{2\sigma^2} + \frac{1}{\sigma^2}(X_i, \theta_i)},$$

we get that

$$\begin{aligned}
\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\mu,id}}(X,Y) = &\prod_{k=1}^d \frac{\cosh\big(\frac{\epsilon}{\sigma^2}(X_i^{(k)} + Y_j^{(k)})\big)}{\cosh\big(\frac{\epsilon}{\sigma^2}(X_i^{(k)} + Y_i^{(k)})\big)} \\
&\times \prod_{k=1}^d \frac{\cosh\big(\frac{\epsilon}{\sigma^2}(X_j^{(k)} + Y_i^{(k)})\big)}{\cosh\big(\frac{\epsilon}{\sigma^2}(X_j^{(k)} + Y_j^{(k)})\big)}.
\end{aligned}$$

Then, we compute the Kullback-Leibler divergence,

$$\begin{aligned}
K(\mathbf{P}_{\mu,\pi}, \mathbf{P}_{\mu,id}) &= \int \log\Big(\frac{d\mathbf{P}_{\mu,\pi}}{d\mathbf{P}_{\mu,id}}\Big) d\mathbf{P}_{\mu,\pi} \\
&= 2\sum_{k=1}^d \mathbf{E}_\mu\Big[\int \log\cosh\frac{\epsilon}{\sigma^2}(2\theta_{i,k} + \sigma\sqrt{2}X) d\mathbf{Q}\Big] \\
&- 2\sum_{k=1}^d \mathbf{E}_\mu\Big[\int \log\cosh\frac{\epsilon}{\sigma^2}(\theta_{i,k} + \theta_{j,k} + \sigma\sqrt{2}X) d\mathbf{Q}\Big],
\end{aligned}$$

where $\mathbf{Q}$ is a standard Gaussian distribution. Using Lemma 3, we get that the general term in the first sum is smaller than $\frac{\epsilon^2}{\sigma^2} + 2\frac{\epsilon^4}{\sigma^4}$, while the second general term is larger than $\frac{\epsilon^2}{\sigma^2} - 2\frac{\epsilon^6}{\sigma^6} - \frac{2}{3}\frac{\epsilon^8}{\sigma^8}$, whence

$$K(\mathbf{P}_{\mu,\pi}, \mathbf{P}_{\mu,id}) \le 4d\frac{\epsilon^4}{\sigma^4} + 4d\frac{\epsilon^6}{\sigma^6} + \frac{4d}{3}\frac{\epsilon^8}{\sigma^8} \le \frac{8d\epsilon^4}{\sigma^4}.$$

Finally, to upper bound $\mu(\mathbb{R}^{nd}\backslash\Theta_\kappa)$, we notice that

$$\begin{aligned}
\mu(\mathbb{R}^{nd}\backslash\Theta_\kappa) &\le \tfrac{n(n-1)}{2}\mu\Big(\|\theta_1 - \theta_2\|^2 < \kappa^2\Big) \\
&\le \tfrac{n(n-1)}{2}\mu\Big(\sum_{j=1}^d (\zeta_1^{(j)} - \zeta_2^{(j)})^2 < \tfrac{\kappa^2}{\epsilon^2} = d\Big).
\end{aligned}$$

The Hoeffding inequality completes the proof. □