

Perfect Matching Map Recovery Under Unknown Scalar Affine Transformation

Tigran Galstyan^{1, 2}

Avetik Karagulyan^{1, 2}

Arshak Minasyan¹

¹Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

²Centre National de la Recherche Scientifique (CNRS), Gif-sur-Yvette, France

Abstract

We study the problem of matching two sets of noisy feature vectors when underlying true features are related by an unknown scalar affine transformation. Our method comprises two primary steps. First, we standardize the feature vectors to estimate the unknown scalar affine transformation. Subsequently, we estimate the permutation by minimizing the Least Sum of Logarithms (LSL) between two sets of observations using the estimated transformation. Our main result shows that the unknown permutation can be perfectly recovered given that the minimal separation distance of true feature vectors scales as $\sqrt{\rho_\sigma} \vee (d \log n)^{1/4} \vee \sqrt{\log n}$, where d is the ambient dimension, n is the sample size, and ρ_σ is the maximal ratio of noise magnitudes. Interestingly, the obtained rate, under mild heteroscedasticity, coincides with that of the non-affine setting. We additionally demonstrate that there exist configurations requiring a larger minimal separation distance for perfect recovery. The latter makes the matching problem more challenging from minimax perspective compared to the non-affine setting. Consequently, we show that in the problem of feature matching, standardizing the data implicitly estimates the scalar affine parameters. As part of our analysis, we prove non-asymptotic concentration bounds for the affine parameter estimators in the presence of heterogeneous noise magnitudes.

1 INTRODUCTION

The problem of recovering the latent correspondence between two sets of noisy observations is a fundamental problem in high-dimensional statistics and data analysis, with applications ranging from point cloud registration in computer vision to sequence alignment in bioinformatics [Liu

et al., 2024, Shi et al., 2023, Barkas et al., 2019]. While the statistical limits of the matching problem have been extensively studied under simple additive noise models and the presence of outliers [Bakirtas and Erkip, 2024, Minasyan et al., 2023], practical scenarios often present more complex challenges. Specifically, the observations can be subject to an unknown affine transformation, caused by factors such as different coordinate systems, sensor calibration [Zhu et al., 2023], measurement scales, or other environmental conditions [Cai et al., 2023]. This work makes a step towards addressing this gap by modeling the unknown scale and translation parameters alongside the unknown permutation and analyzing the statistical limits of distinguishing the true matching from spurious alignments.

The general formulation of the matching problem is as follows. Suppose we observe two independent sequences of d -dimensional random vectors $\mathbf{X} = \{X_i\}_{i=1}^n$ and $\mathbf{X}^\# = \{X_i^\#\}_{i=1}^n$. In addition, we suppose that under some unknown permutation $\pi^* \in \mathfrak{S}_n$, $X_i^\# = X_{\pi^*(i)}$ holds for any $i \in [n] := \{1, \dots, n\}$. The objective of the statistical matching problem is to recover the latent feature matching map π^* from the noisy observations $\mathbf{X}$ and $\mathbf{X}^\#$.

This paper studies the statistical matching problem under unknown scalar affine transformation. In particular, we observe two sequences of d -dimensional random vectors $\mathbf{X} = \{X_i\}_{i=1}^n$ and $\mathbf{X}^\# = \{X_j^\#\}_{j=1}^n$ sampled according to the following model:

$$X_i = \theta_i + \sigma_i \xi_i, \quad X_j^\# = \theta_j^\# + \sigma_j^\# \xi_j^\#, \quad \forall i, j \in [n]. \quad (1)$$

Here, $\boldsymbol{\theta} := \{\theta_i\}_{i=1}^n$ and $\boldsymbol{\theta}^\# := \{\theta_j^\#\}_{j=1}^n$ are d -dimensional unknown feature vectors, while $\boldsymbol{\sigma} := (\sigma_1, \dots, \sigma_n)^\top$ and $\boldsymbol{\sigma}^\# := (\sigma_1^\#, \dots, \sigma_n^\#)^\top$ are unknown noise magnitudes. The sequences $\xi_1, \dots, \xi_n$, and $\xi_1^\#, \dots, \xi_n^\#$ are assumed to be i.i.d. standard Gaussian vectors. The key structural assumption is that there exists an unknown permutation $\pi^* \in \mathfrak{S}_n$, an unknown scaling factor $\tau^* > 0$, and an unknown translation vector $\beta^* \in \mathbb{R}^d$ such that

$$\theta_i = \tau^* \theta_{\pi^*(i)}^\# + \beta^*, \quad \sigma_i = \tau^* \sigma_{\pi^*(i)}^\#, \quad \forall i \in [n]. \quad (2)$$

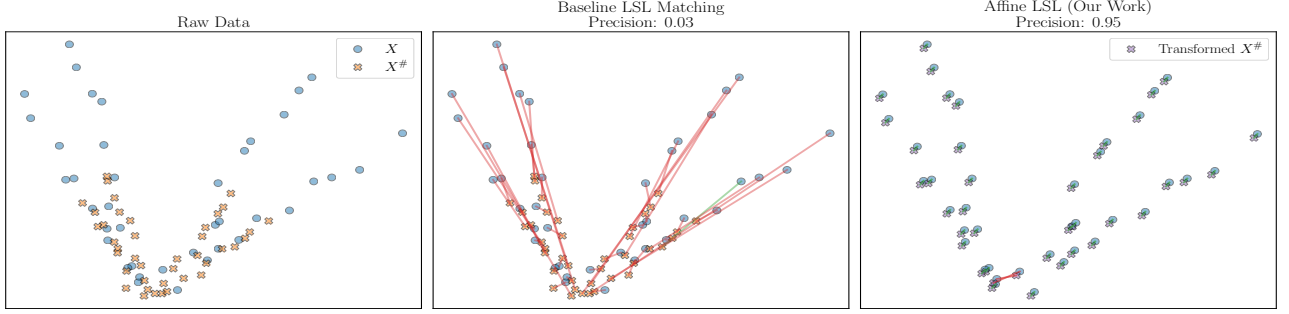


Figure 1: **Step-by-step affine matching process** (Left) PCA projection of the raw data generated according the model (1), showing the affine gap between $\mathbf{X}$ and $\mathbf{X}^\#$. (Middle) The failure of standard LSL, which ignores the transformation and results in a precision of only 0.03. (Right) Our Affine LSL method estimates the transformation to align the $\mathbf{X}^\#$ (purple crosses) with $\mathbf{X}$, achieving 0.95 precision.

Our goal is to jointly estimate π^* , τ^* , and β^* from the noisy observations $\mathbf{X}$ and $\mathbf{X}^\#$.

Known affine transformation. When the parameters τ^* and β^* are known, the problem can be directly solved using existing methods. Specifically, observe that the sets $\mathbf{X}$ and $\mathbf{Y}^* := \{\tau^* X_i^\# + \beta^*\}_{i \in [n]}$ share the same distribution up to the permutation π^* . Thus, estimating π^* reduces to a standard permutation estimation problem, which was studied in Collier and Dalalyan [2016]. To briefly present their results, we define the minimal separation distance.

Definition 1. For independent random vectors $\mathbf{V} = \{V_i\}_{i \in [n]}$ the separation distance κ is a function defined as

$$\kappa(\mathbf{V}) := \min_{j \neq i} \|(\text{var}[V_i] + \text{var}[V_j])^{-\frac{1}{2}} (\mathbf{E}[V_i] - \mathbf{E}[V_j])\|_2.$$

In the non-affine setting, and in the presence of Gaussian isotropic noise, Collier and Dalalyan [2016, Theorem 3] assert that if

$$\kappa(\mathbf{X}) \gtrsim (d \log(n^2/\delta))^{1/4} \vee \sqrt{\log(n^2/\delta)},$$

then the unknown permutation π^* can be perfectly recovered with probability at least $1 - \delta$. Moreover, Galstyan et al. [2022, Theorem 4] assert that this rate is minimax optimal up to constant factors, providing the corresponding lower bound on the minimal separation distance.

Estimation procedures of unknown parameters. In general, one cannot know in advance whether there is an underlying affine transformation applied to one of the datasets or not. Thus, a naive approach of directly matching the observations $\mathbf{X}$ and $\mathbf{X}^\#$ may result in failure to recover the unknown permutation π^* , while the minimal separation distance $\kappa(\mathbf{X})$ being arbitrarily large. Importantly, this does not contradict the previously established minimax rates, as their analysis assumes the absence of any underlying affine transformation. In other words, under model misspecification of permutation estimation the minimax rate on the

minimal separation distance presented above is no longer valid.

In what follows, we propose a sequential estimation procedure: first, we estimate the unknown parameters τ^* and β^* , and then use these estimators to design an estimator for the unknown permutation π^* . In particular, we propose the following estimator for τ^*

$$\hat{\tau}_n := \left(\frac{\sum_{i=1}^n \|X_i - \bar{X}_n\|_2^2}{\sum_{i=1}^n \|X_i^\# - \bar{X}_n^\#\|_2^2} \right)^{1/2}, \quad (3)$$

where $\bar{X}_n$ and $\bar{X}_n^\#$ are the sample mean vectors of the sets $\mathbf{X}$ and $\mathbf{X}^\#$, respectively. Notice that $\hat{\tau}_n$ can be viewed as the ratio of sample standard deviations of $\mathbf{X}$ and $\mathbf{X}^\#$. The translation vector β^* can be estimated by $\hat{\beta}_n := \bar{X}_n - \hat{\tau}_n \bar{X}_n^\#$.

Using these estimators, we show that the problem of recovering the underlying permutation π^* reduces to matching the centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ (scaled by $\hat{\tau}_n$), where $\mathbf{Z} = \{X_i - \bar{X}_n\}_{i=1}^n$ and $\mathbf{Z}^\# = \{X_i^\# - \bar{X}_n^\#\}_{i=1}^n$. Thus, in the presence of an affine transformation, we show that the quantity that plays role for recovering π^* is indeed $\kappa(\mathbf{Z})$. In the non-affine setting when the permutation recovery is guaranteed with high probability, we prove that if the minimal separation distance satisfies $\kappa(\mathbf{Z}) \gtrsim \kappa(\mathbf{X}) \vee \sqrt{\rho_\sigma}$ then the unknown permutation π^* can be perfectly recovered with high probability under unknown scalar affine transformation. Here, $\rho_\sigma := \max_{i,j} (\sigma_j^2/\sigma_i^2)$ is the maximal ratio of noise magnitudes. This quantity arises both in the estimation error of $\hat{\tau}_n$ and in the condition on $\kappa(\mathbf{Z})$. We discuss the dependence on ρ_σ at the end of this section.

The estimation procedure for π^* is rooted in maximizing the profile likelihood function of our model (1), while using permutation invariant estimators $\hat{\tau}_n$ and $\hat{\beta}_n$. The profile MLE approach is common in the literature and has been used in related papers, such as [Han et al., 2025, Galstyan et al., 2022, Collier and Dalalyan, 2016]. In our setting the profile MLE approach leads to the following estimator for

π^* , denoted by $\hat{\pi}_n^{\text{LSL}}$:

$$\hat{\pi}_n^{\text{LSL}} \in \arg \min_{\pi \in \mathfrak{S}_n} \sum_{i=1}^n \log \left(\|Z_i - \hat{\tau}_n Z_{\pi(i)}^\# \|_2^2 \right). \quad (4)$$

We highlight that our choice of $\hat{\tau}_n$ and $\hat{\beta}_n$ implicitly apply standardization to the observations $\mathbf{X}$ and $\mathbf{X}^\#$. Thus, when applying standardization, in the problem of statistical matching, we estimate the unknown transformation of the form $\tau I_d + \beta$, using the aforementioned estimators $\hat{\tau}_n$ and $\hat{\beta}_n$. We defer the detailed discussion of the standardization to Section 3 and Section 4. On the other hand, these estimators are closely connected to the profile MLE in the homogeneous setting. Overall, the contributions of this paper are two-fold: estimation of the scalar affine parameters, and perfect recovery of the underlying permutation. The findings of this paper are briefly described below:

1. Our main result, presented in Theorem 2, establishes an upper bound on the minimal separation distance of the observations $\mathbf{Z}$ so that with high probability the Least Sum of Logarithms estimator $\hat{\pi}_n^{\text{LSL}}$ from (4) perfectly recovers the unknown permutation π^* . Specifically, in the high-dimensional regime $d \geq c \log n$ for some absolute constant c , we show that if

$$\kappa(\mathbf{Z}) \gtrsim (d \log(n/\delta))^{1/4} \vee \sqrt{\rho_\sigma \log(1/\delta)},$$

then our estimator exactly recovers π^* with probability at least $1 - \delta$. We refer to the case of *mild heteroscedasticity* when ρ_σ is of at most of order $(d \log n)^{1/2} \vee \log n$. Interestingly, in this case, centralizing the original observations yields a minimax rate for separation distance that is identical to that of non-affine setting, implying its optimality. Moreover, when the maximal ratio of noise magnitudes $\rho_\sigma \gg (d \log n)^{1/2} \vee \log n$ then the perfect permutation recovery requires the separation distance to scale proportionally to $\sqrt{\rho_\sigma}$.

2. The estimation error of the scaling factor τ^* is presented in Theorem 1, where we establish a high-probability concentration bound of $\hat{\tau}_n$ around τ^* : under the model (1), the ratio $\hat{\tau}_n/\tau^*$ concentrates around 1. Up to logarithmic factors, the relative error $|\hat{\tau}_n^2/\tau^{*2} - 1|$ scales as $\alpha_\sigma/\sqrt{\lambda^2 + d}$, where λ is the *distortion-to-noise ratio* and $\alpha_\sigma := \sigma_{\max}/\|\sigma\|_2 \in [n^{-1/2}, 1]$ is the *relative noise magnitude*, both defined in Section 3.

Notation. We use $[n] = \{1, \dots, n\}$ to denote the set of integers from 1 to n . The symmetric group of all permutations on $[n]$ is denoted by $\mathfrak{S}_n$. For a vector $v = (v_1, \dots, v_n)^\top \in \mathbb{R}^n$, we write $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$ for the ℓ_p norm. If the order is not explicitly written, then it is the Euclidean ℓ_2 -norm. The notation $\stackrel{d}{=}$ denotes equality in distribution. We denote by χ_d^2 the chi-squared distribution with d degrees of freedom. We write $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

We use the following notation for empirical averages $\bar{\theta}_n, \bar{\theta}_n^\#$ of the true features θ and $\theta^\#$, respectively. We use $\gtrsim, \lesssim$ for inequalities up to numerical constants.

2 RELATED WORK

The problem of map recovery through matching has been widely studied in various settings, such as the geometric planted matching problem [Kunisky and Niles-Weed, 2022], feature alignment [Han et al., 2025], and graph alignment [Ganassali and Massoulié, 2020, Wang et al., 2022, Hall and Massoulié, 2023]. Some variation of robust to outliers versions of feature matching problem are studied in [Galstyan et al., 2022, Minasyan et al., 2023].

Additionally, permutation estimation and related problems have recently been investigated in the context of statistical seriation [Flammarion et al., 2019, Issartel et al., 2024], clustering [Davies and Bouldin, 1979, Giraud and Verzelen, 2019], noisy sorting [Mao et al., 2018], crowd labeling [Shah et al., 2021], and variable selection [Ndaoud and Tsybakov, 2020, Comminges and Dalalyan, 2012].

The problem of permutation recovery has also been in a slightly different setting of unlinked or unmatched regression. In the latter setting, the response variables and their corresponding features are permuted. While the permutation recovery is not the main objective, it remains a crucial part of the problem. The case of linear regression was studied in originating in [Pananjady et al., 2017b] and [Pananjady et al., 2017a]. This line of work has recently gained attention [Balabdaoui et al., 2021, Azadkia and Balabdaoui, 2024, Slawski and Sen, 2024]. Inference from the anonymous and/or privatized data falls inside the framework of unlinked regression.

The separation number plays a crucial role in statistical matching. However, its applications extend far beyond this area. In hypothesis testing, the minimal separation distance has emerged as the standard measure for distinguishing between null and alternative hypotheses, as illustrated in the seminal paper [Ingster, 1982] and the monograph [Juditsky and Nemirovski, 2020]. This concept has since found numerous applications in the recent literature [Wolfer and Kontorovich, 2020, Blanchard et al., 2018, Wei et al., 2019].

The problem of matching under unknown affine transformations has been studied in natural language processing by Grave et al. [2019], Pumir et al. [2021]. Specifically, Grave et al. [2019] show that simultaneously estimating an unknown permutation and affine transformation is non-convex. To address this, they propose a convex relaxation with alternating minimization of the objective function. However, to our knowledge, the statistical limitations from the minimax point of view for the matching recovery were not studied.

3 MAIN THEORETICAL RESULTS

This section contains the main theoretical findings of the paper. We focus on the case where the affine parameters are unknown. We first present a finite-sample high-probability concentration bound for the estimator $\hat{\tau}_n$ around the true parameter τ^* . As for the translation parameter β^* , we do not assess the estimation error of $\hat{\beta}_n$ explicitly. Instead, it is done implicitly when considering the centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$. Given that the norm of the empirical averages of the features $\|\bar{\theta}_n\|_2$ is large, the consistent estimation of β^* is not possible since the estimation error of τ^* propagates. At the same time, the estimator $\hat{\tau}_n$ cannot compensate for this effect, since it is invariant to translations. However, we show that this does not affect permutation recovery. We then present the main result of this paper, that is, we present the condition on the minimal separation distance between the centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ such that $\hat{\pi}_n^{\text{LSL}}$ perfectly recovers the unknown permutation π^* with high probability.

3.1 PERFORMANCE ANALYSIS OF AFFINE PARAMETER ESTIMATORS

Recall that centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ are defined as $Z_i = X_i - \bar{X}_n$, $Z_i^\# = X_i^\# - \bar{X}_n^\#$ for all $i \in [n]$. Notice that for any pairs $(i, j) \in [n]^2$ the random vectors Z_i and $Z_j^\#$ are independent. Moreover, for each $i \in [n]$, we have $Z_i \stackrel{d}{=} \tau^* Z_{\pi^*(i)}^\# \sim \mathcal{N}(\mu_i, s_i^2 I_d)$ with μ_i and s_i defined as

$$\mu_i := \theta_i - \bar{\theta}_n, \quad s_i^2 := \frac{n-2}{n} \sigma_i^2 + \frac{1}{n^2} \|\sigma\|_2^2. \quad (5)$$

We first establish two-sided high-probability bounds on the sum of squared ℓ_2 -norms $\sum_{i=1}^n \|Z_i\|_2^2$ of the centralized observations $Z_1, \dots, Z_n$. This result will then be used to prove that the estimator $\hat{\tau}_n$, defined in (3), accurately estimates the unknown parameter τ^* .

Proposition 1. *Let $Z_1, \dots, Z_n$ be d -dimensional Gaussian vectors with mean vectors $\mu_1, \dots, \mu_n$ and isotropic covariance matrices $s_1^2 I_d, \dots, s_n^2 I_d$, respectively, according to (5). Denote $\sigma_{\max} := \max_{i \in [n]} \sigma_i$. Then, for any $\delta \in (0, 1)$, the following inequality holds with probability at least $1 - \delta$:*

$$\begin{aligned} \sum_{i=1}^n \|Z_i\|_2^2 - \|\mu\|_2^2 - d\|\sigma\|_2^2 &\leq 2\sigma_{\max} \|\mu\|_2 \sqrt{2\log(2/\delta)} \\ &\quad + 2\|\sigma\|_4^2 \sqrt{d\log(2/\delta)} + 2\sigma_{\max}^2 \log(2/\delta). \end{aligned}$$

Similarly, with probability at least $1 - \delta$, we have

$$\begin{aligned} \sum_{i=1}^n \|Z_i\|_2^2 - \|\mu\|_2^2 - d\|\sigma\|_2^2 &\geq -2\sigma_{\max} \|\mu\|_2 \sqrt{2\log(2/\delta)} \\ &\quad - 2\|\sigma\|_4^2 \sqrt{d\log(2/\delta)} - \frac{d}{n} \|\sigma\|_2^2. \end{aligned}$$

The proof is deferred to Section A.1. Proposition 1 is related to the Hanson-Wright inequality [Rudelson and Vershynin, 2013], which provides a general framework for bounding the quadratic form $X^\top A X$, where $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ is a d -dimensional sub-Gaussian vector with independent coordinates and $A \in \mathbb{R}^{d \times d}$ is an arbitrary matrix. In contrast, our setting involves correlated random vectors Z_i , arising from the centering step that projects the observations onto the subspace orthogonal to the all-ones direction. Despite this dependence, we still obtain a comparable concentration bound: as detailed in Section A.1, the proof recasts the noise contribution to $\sum_{i=1}^n \|Z_i\|_2^2$ as a quadratic form in an isotropic Gaussian vector, for which a Hanson-Wright-type bound is available. Our result corresponds to the special case $A = I_d$; we believe Proposition 1 can be generalized to an arbitrary matrix A , but we do not pursue this direction here. Additionally, Proposition 1 can be recast as a concentration inequality for the squared Frobenius norm of the Gaussian matrix $(Z_1, \dots, Z_n) \in \mathbb{R}^{d \times n}$ around its mean $\mathbf{E}[\|\mathbf{Z}\|_{\text{Fr}}^2] = \|\mu\|_2^2 + d \frac{n-1}{n} \|\sigma\|_2^2$, which may be of independent interest. Formally, each of the following inequalities holds with probability at least $1 - \delta$:

$$\|\mathbf{Z}\|_{\text{Fr}}^2 - \mathbf{E}[\|\mathbf{Z}\|_{\text{Fr}}^2] \leq \bar{r}_{\text{ub}}(\sigma, d, \delta)$$

and

$$\|\mathbf{Z}\|_{\text{Fr}}^2 - \mathbf{E}[\|\mathbf{Z}\|_{\text{Fr}}^2] \geq \bar{r}_{\text{lb}}(\sigma, d, \delta),$$

where $\bar{r}_{\text{ub}}(\sigma, d, \delta)$ and $\bar{r}_{\text{lb}}(\sigma, d, \delta)$ are the corresponding upper and lower bounds from Proposition 1, respectively. The next result proves that the estimator $\hat{\tau}_n$ from (3) of the scaling factor concentrates around τ^* . The proof is postponed to Section A.2 and relies on Proposition 1.

Theorem 1 (Scaling factor estimation). *Consider the observations $\mathbf{X}$ and $\mathbf{X}^\#$ following the model in (1) and satisfying the affine conditions from (2). Let $\hat{\tau}_n$ be the estimator defined in (3). Let $\alpha_\sigma := \sigma_{\max}/\|\sigma\|_2$ be the relative noise magnitude, and $\lambda = \|\mu\|_2/\|\sigma\|_2$ be the distortion-to-noise ratio. Fix $\delta \in [4e^{-d/(224\alpha_\sigma^2)}, 1)$ and assume that $n \geq 8$. Then, for any $\tau^* > 0$, with probability at least $1 - 4\delta$, we have*

$$\left| \frac{\hat{\tau}_n^2}{\tau^{*2}} - 1 \right| \leq 12 \sqrt{\frac{\alpha_\sigma^2 \log(4/\delta)}{\lambda^2 + d}} + \frac{2\alpha_\sigma^2 \log(4/\delta)}{\lambda^2 + d} + \frac{2d}{n(\lambda^2 + d)}.$$

Let us now comment on the estimation error of the scaling factor τ^* . The bound consists of several components. The first term, proportional to $\alpha_\sigma/\sqrt{\lambda^2 + d}$, up to $\sqrt{\log(1/\delta)}$ corresponds to the variance of $\hat{\tau}_n^2$ in the small-deviation regime. Similarly, in the large-deviation regime, the dependence scales as $\alpha_\sigma^2/(\lambda^2 + d)$ up to a logarithmic factor in the confidence level δ , that is $\log(1/\delta)$. This division into small- and large-deviation regimes reflects the sub-exponential character of the distribution of $\hat{\tau}_n$, see, for example, [Vershynin, 2018, Section 2.8]. The remainder of the upper

bound of Theorem 1 scales as $d/(n(\lambda^2 + d))$, which corresponds to the bias of the estimator $\hat{\tau}_n^2$. Indeed, it is positively biased, which can be seen by applying Jensen's inequality to the mapping $t \rightarrow 1/t$ and the independence of $\mathbf{Z}$ and $\mathbf{Z}^\#$. It is also worth mentioning that the bias term vanishes as n grows.

The parameter α_σ present in the deviation terms admits a natural interpretation as a noise-concentration coefficient: it measures how much of the total noise variance $\|\sigma\|_2^2$ is carried by the largest component $\sigma_{\max}^2$. The range of possible values of α_σ is on the closed interval $[n^{-1/2}, 1]$, with the lower endpoint $\alpha_\sigma = n^{-1/2}$ attained in the case of homogeneous noise, that is $\sigma_i = \sigma_j = \sigma$ for any $(i, j) \in [n]^2$. Thus, in the homogeneous case the estimation rate scales as $1/\sqrt{n(\lambda^2 + d)} \vee 1/(n(\lambda^2 + d))$, which matches the optimal parametric rate of estimating a scalar parameter given $2nd$ noisy observations. Additionally, if the distortion-to-noise ratio is large enough, e.g., $\lambda^2 \gtrsim n$ we get fast rates of order $1/n$. At the opposite extreme, when α_σ approaches 1, the estimation rate scales as $1/\sqrt{\lambda^2 + d}$. This effectively means that without the knowledge of π^* , only d coordinates carry information about τ^* . A simple example of such configuration is when $\sigma_1 = 1$ and $\sigma_2 = \dots = \sigma_n = 1/n^2$ making $\alpha_\sigma \asymp 1/(1 + n^{-1})$. Overall, the parameter α_σ describes the noise magnitude configuration and its effect on the estimation rate. In the homogeneous case, the estimation rate become faster, while in a single dominated noise configuration it deteriorates. It is also worth mentioning that the result of Theorem 1 is generic and holds true for any configuration of true unavailable feature vectors θ and noise magnitudes σ . The underlying structure of the configuration of mean vectors and noise magnitudes can make the statistical problem of estimation of scaling factor easier (fast rates) or harder (slow rates).

Moreover, the obtained high-probability bound due to its nature also provides us with an in-expectation bound. Indeed, integrating both sides of the bound from Theorem 1 with respect to δ , the proposed estimator $\hat{\tau}_n$ satisfies

$$\sup_{\tau^* > 0} \frac{1}{\tau^{*2}} \mathbf{E}[\|\hat{\tau}_n^2 - \tau^{*2}\|] \lesssim \sqrt{\frac{\alpha_\sigma^2}{\lambda^2 + d}} \vee \frac{\alpha_\sigma^2}{\lambda^2 + d}.$$

The setting studied in this paper is closely related to the geometric planted matching problem studied in [Kunisky and Niles-Weed, 2022]. The latter assumes that $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and $X_i^\# = X_{\pi^*(i)} + \xi_i$ for all $i \in [n]$, where $\xi_1, \dots, \xi_n$ are independent Gaussian vectors with zero mean and identical covariance matrices proportional to I_d , i.e., $\xi_1, \dots, \xi_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$. They showed that, in the high-dimensional regime ($d \gg \log n$), the theoretical limit for exact recovery of the underlying matching π^* is when $\sigma^2 \lesssim d/\log n$. The key difference with our model is that we assume that the observations $X_1, \dots, X_n$ and $X_1^\#, \dots, X_n^\#$ are independent. Since independence induces less structure

in the data, perfect recovery in our model is not possible unless the minimal separation distance is large enough. We quantify the order of the minimal separation distance in Theorem 2, and show that in the case of mild heteroscedasticity it coincides with the minimax lower bound established in [Galstyan et al., 2022, Theorem 4] for known $\tau^* = 1$ and $\beta^* = 0$. This lower bound directly extends to the setting with unknown parameters τ^* and β^* .

3.2 EXACT PERMUTATION RECOVERY OF THE UNDERLYING PERMUTATION

We are now ready to present our main result on the exact permutation recovery of the unknown permutation π^* . To ease the presentation the proof of the theorem is deferred to Section A.3.

Theorem 2 (Exact permutation recovery). *Consider the centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ of $\mathbf{X}$ and $\mathbf{X}^\#$, respectively. The latter sets follow the model (1), and satisfy the affine assumption from (2). Let $\hat{\tau}_n$ be the estimator of τ^* as in (3), and $\hat{\pi}_n^{\text{LSL}}$ the LSL estimator from (4). Let $\rho_\sigma = \max_{i,j} \sigma_j^2/\sigma_i^2$ be the maximal ratio of noise magnitudes. Fix $\delta \in [4e^{-d/(1024\alpha_\sigma^2)}, 1)$, and assume that $n \geq d$, and $n \geq 8$. Then, if the separation distance $\kappa(\mathbf{Z})$ is not smaller than*

$$5(d \log(12n^3/\delta))^{1/4} + 17\sqrt{\log(24n^3/\delta)} + 52\sqrt{\rho_\sigma \log(4/\delta)}, \quad (6)$$

the estimator $\hat{\pi}_n^{\text{LSL}}$ coincides with π^ with probability at least $1 - 4\delta$, i.e. $\mathbf{P}(\forall i \in [n] : \hat{\pi}_n^{\text{LSL}}(i) = \pi^*(i)) \geq 1 - 4\delta$.*

The sufficient condition for exact recovery established above consists of three terms. Interestingly, these terms are of different nature. The first two terms depend only on sample size n , ambient dimension d , and confidence level δ . These terms essentially come from controlling the supremum of Gaussian and χ_d^2 random variables over a segment of constant length and maximum over $2n$ pairs. Unsurprisingly, compared to the non-affine setting, it results in the same minimax separation distance of order $(d \log n)^{1/4} \vee \sqrt{\log n}$. The last term, depends on the noise configuration, namely, the maximal ratio of noise magnitudes. This term comes from the application of Theorem 1, when we estimate the unknown scaling factor τ^* . Interestingly, in the bound of Theorem 1 only the relative noise magnitude $\alpha_\sigma \in [n^{-1/2}, 1]$ appears. One would expect to have the same quantity appearing in the sufficient condition for the separation distance $\kappa(\mathbf{Z})$. However, careful inspection of the proof of Theorem 2 shows that for the matching problem the key quantity is the maximal ratio ρ_σ . Thus the entire dependence on noise heterogeneity is encoded in the ratio $\sigma_{\max}/\sigma_{\min} \equiv \sqrt{\rho_\sigma}$. Overall, we get that as long as $\kappa(\mathbf{Z})$ is at least of order

$$\max \left\{ (d \log n)^{1/4}, \sqrt{\log n}, \frac{\sigma_{\max}}{\sigma_{\min}} \right\}$$

then the exact permutation recovery under an unknown scalar affine transformation is possible with high probability. We further discuss the optimality of the result in the following paragraph.

Lower bounds. Theorem 2 implies that the minimal separation distance between the centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ that suffices for the perfect recovery of π^* is at most of order $(d \log n)^{1/4} \vee \sqrt{\log n} \vee \sqrt{\rho_\sigma}$. A natural question is whether this rate is optimal in the minimax sense. Under mild heteroscedasticity, that is when $\rho_\sigma \lesssim \sqrt{d \log n} \vee \log n$, the rate is indeed minimax optimal, since the third term is dominated by the first two and the minimax optimality follows from [Galstyan et al., 2022, Theorem 4]. Under strong heteroscedasticity, that is when $\rho_\sigma \gg \sqrt{d \log n} \vee \log n$, there is a gap between the minimax lower bound (non-affine setting), and our upper bound. It remains open whether the lower bound for the non-affine setting is loose for our model, or our upper bound is not minimax optimal under strong heteroscedasticity. Additionally, the optimal dependence on ρ_σ , if any, remains open.

We perform numerical experiments to support the findings of Theorem 2, see Section 5 for more detailed explanations of the experiments. In Figure 3, we experimentally illustrate that the permutation recovery problem does become harder as ρ_σ grows. That is, our algorithm starts to exactly recover the permutation when the separation distance becomes larger. For smaller values of the separation distance the estimator $\hat{\pi}_n^{\text{LSL}}$ only partially recovers π^* . The problem of partial permutation recovery is out of scope of this paper.

Standardization cost. In practice, it is typically unknown whether an underlying affine transformation exists in the permutation model.

If there is no affine transformation, that is $\tau^* = 1$, $\beta^* = 0$, and the maximal noise ratio ρ_σ is large, our procedure becomes suboptimal due to error propagation in estimating τ^* . For large ρ_σ , the precision of $\hat{\tau}_n$ deteriorates, introducing additional error when matching $\mathbf{Z}$ and $\hat{\tau}_n \mathbf{Z}^\#$. Consequently, the LSL method becomes more susceptible to noise, leading to failures in perfect permutation recovery. In such cases, data standardization should be avoided.

If an affine transformation is present and the heteroscedasticity is mild, that is ρ_σ is of constant order relative to other parameters, our procedure is preferable. It achieves the minimax optimal rate for permutation recovery. According to Theorem 1, $\hat{\tau}_n$ accurately estimates τ^* , and provided that $\kappa(\mathbf{Z}) \gtrsim (d \log n)^{1/4} \vee \sqrt{\log n}$, the LSL method recovers π^* perfectly with high probability. In this scenario, data standardization implicitly estimates the unknown transformation, enabling perfect recovery via LSL.

4 MOTIVATION AND DISCUSSION

In this section, we describe the approach based on the profile Maximum Likelihood Estimation (MLE) both in heterogeneous and homogeneous noise settings. In the latter case, we derive the estimators for parameters τ^* and β^* showing that they coincide with $\hat{\tau}_n$ and $\hat{\beta}_n$, respectively. Lastly, we draw the connection of these estimators to data standardization.

Profile MLE. Recall that our observations follow the statistical model (1), under the scalar affine transformation described in (2). In the context of statistical matching, the main parameter of interest is the unknown permutation π^* , while the others are treated as nuisance parameters. Below, we derive an estimator for π^* based on maximizing the profile maximum likelihood function of the model (1). Let us define by Θ the set of all nuisance parameters, that is $\{\theta^\#, \sigma^\#\}$. Notice that $\mathbf{X} \cup \mathbf{X}^\#$ is a set of $2n$ independent Gaussians. Thus, for any values of the parameters π, τ, β , the negative logarithm of the full maximum likelihood function is up to some additive and multiplicative constants equal to

$$\ell_n(\pi, \tau, \beta, \Theta; \{\mathbf{X}, \mathbf{X}^\#\}) = \sum_{i=1}^n \left(\frac{\|X_i - \tau \theta_{\pi(i)}^\# - \beta\|_2^2}{\tau^2 \sigma_{\pi(i)}^{\#2}} + \frac{\|X_{\pi(i)}^\# - \theta_{\pi(i)}^\#\|_2^2}{\sigma_{\pi(i)}^{\#2}} + d \log(\tau \sigma_{\pi(i)}^{\#2}) \right),$$

where we have already taken into account the scalar affine assumptions from (2), and got rid of the dependence on parameters θ and σ . The set of parameters Θ is unknown and our goal is to minimize the negative log-likelihood function ℓ_n with respect to them. Since, the function ℓ_n is not convex in Θ we first minimize it with respect to $\theta^\#$, then with respect to $\sigma^\#$. Notice further that the log-likelihood is separable with respect to the variables $\theta_i^\#$. Thus, for fixed π, τ, β and $\sigma^\#$, minimizing with respect to $\theta^\#$ yields the following profile likelihood function up to an additive constant:

$$\ell_n(\pi, \tau, \beta, \sigma^\#; \{\mathbf{X}, \mathbf{X}^\#\}) = \sum_{i=1}^n \left(\frac{\|X_i - \tau X_{\pi(i)}^\# - \beta\|_2^2}{2\tau^2 \sigma_{\pi(i)}^{\#2}} + d \log(\tau \sigma_{\pi(i)}^{\#2}) \right).$$

Further minimizing the expression from the previous display with respect to $\sigma_{\pi(i)}^\#$ for all i , under the condition that $\min_{i \in [n]} \sigma_i^\# := \sigma_{\min}^\# > 0$, we obtain the following expression for ℓ_n :

$$\ell_n(\pi, \tau, \beta; \{\mathbf{X}, \mathbf{X}^\#\}) = \sum_{i=1}^n \log \left(\frac{\|X_i - \tau X_{\pi(i)}^\# - \beta\|_2^2}{\tau} \right).$$

First, notice that the likelihood is non-convex with respect to τ . Second, for any $i \in [n]$ we have

$$\lim_{\beta \rightarrow X_i - \tau X_{\pi(i)}^\#} \ell_n(\pi, \tau, \beta, \mathbf{X}, \mathbf{X}^\#) = -\infty.$$

Thus, there is no closed form solution for the minimization of ℓ_n in terms of β and τ . To circumvent this issue, we plug in the permutation invariant estimators $\hat{\tau}_n$ and $\hat{\beta}_n$ to estimate the unknown permutation π^* . Therefore, using the definition of centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$, the minimization of the quasi log-likelihood function yields the Least Sum of Logarithms estimator $\hat{\pi}_n^{\text{LSL}}$, defined as per (4).

The optimization problem that yields $\hat{\pi}_n^{\text{LSL}}$ is a computationally tractable linear assignment problem with a cost matrix $M_{ij} = \log(\|Z_i - \hat{\tau}_n Z_j^\#\|_2^2)$ for all $(i, j) \in [n]^2$. In particular, it can be efficiently solved using the Hungarian algorithm [Kuhn, 1955]. The idea of the Hungarian algorithm is based on the relaxation of the set of permutation and Birkhoff-von-Neumann theorem [Budish et al., 2013]. In particular, the permutation estimation reduces to

$$\hat{\pi}_n^{\text{LSL}} \in \arg \min_{P_n \in \mathcal{P}} \text{tr}(M \cdot P_n),$$

where $\mathcal{P}$ is the set of all doubly stochastic matrices of size n . We refer the reader to [Collier and Dalalyan, 2016, Galstyan et al., 2022] for more details on this relaxation and the linear assignment problem in the context of matching. To further justify the choice of $\hat{\tau}_n$ and $\hat{\beta}_n$, let us show their connection to the profile MLE estimators in the homogeneous noise setting.

Homogeneous setting. The statistical problem of matching in the case of homogeneous noise case, where $\sigma_i^\# = \sigma^\#$ for all $i \in [n]$ does not pose the difficulties discussed earlier. To this end, we depart from the negative logarithm of the full maximum likelihood function considering that the noise magnitudes are equal. This simplifies the negative log-likelihood to:

$$\ell_n(\pi, \tau, \beta; \{\mathbf{X}, \mathbf{X}^\#\}) = \sum_{i=1}^n \|X_i - \tau X_{\pi(i)}^\# - \beta\|_2^2.$$

Now for any fixed $\tau = \tau_0$ the minimization with respect to β yields the following permutation invariant estimator of β^* , denoted by $\beta_0 := \bar{X}_n - \tau_0 \bar{X}_n^\#$. Then, plugging back in the expression for β_0 and using the notation for the centralized observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ we arrive at

$$\ell_n(\pi, \tau_0; \{\mathbf{X}, \mathbf{X}^\#\}) = \sum_{i=1}^n \|Z_i - \tau_0 Z_{\pi(i)}^\#\|_2^2. \quad (7)$$

By expanding the norms and removing all terms independent of π , we observe that the above function can be minimized without knowledge of τ_0 . Thus, minimizing (7) leads to the

Least Sum of Squares (LSS) estimator, defined as

$$\hat{\pi}_n^{\text{LSS}} \in \arg \max_{\pi \in \mathfrak{S}_n} \sum_{i=1}^n Z_i^\top Z_{\pi(i)}^\#. \quad (8)$$

Thus, in the homogeneous case, we showed that the knowledge of τ^* is not necessary to be able to estimate the unknown underlying permutation π^* . The maximization problem (8) is again a linear assignment problem, and can be solved efficiently with time complexity $\mathcal{O}(n^3)$ using the Hungarian algorithm. Notice that the optimization problem (7) is quadratic with respect to τ . Since the permutation recovery in the homogeneous case is independent of τ^* one can estimate τ^* after recovering the unknown permutation. The value that minimizes (7) reads as

$$\hat{\tau}_0 = \left(\frac{\sum_{i=1}^n Z_i^\top Z_{\pi(i)}^\#}{\sum_{i=1}^n \|Z_{\pi(i)}^\#\|_2^2} \right)^{\frac{1}{2}}.$$

Now, we recognize that on the event $\Omega = \{\hat{\pi}_n^{\text{LSS}} = \pi^*\}$ the denominator is close to the sum of squared ℓ_2 norms of centralized observations $Z_1, \dots, Z_n$. Formally, let $\hat{\pi} \equiv \hat{\pi}_n^{\text{LSS}}$, then conditionally on Ω we have

$$\sum_{i=1}^n \mathbf{E}[Z_i^\top Z_{\hat{\pi}(i)}^\#] = \sum_{i=1}^n \|\mathbf{E}[Z_i]\|_2^2,$$

where the right hand side is permutation invariant. Therefore, this motivates us to use the estimator for τ^* that is defined as per (3). Even though this estimator does not maximize the profile likelihood function, it has a crucial property of being permutation invariant. In Section 3 we show that $\hat{\tau}_n$ accurately estimates τ^* achieving a minimax optimal rate even in the heterogeneous noise setting.

Link to standardization. Let us now demonstrate that the estimators $\hat{\tau}_n$ and $\hat{\beta}_n$ implicitly apply standardization to the observations $\mathbf{X}$ and $\mathbf{X}^\#$. To this end, let $\tilde{\mathbf{X}} = \{\tilde{X}_i\}_{i=1}^n$ and $\tilde{\mathbf{X}}^\# = \{\tilde{X}_i^\#\}_{i=1}^n$ denote the standardized versions of $\mathbf{X}$ and $\mathbf{X}^\#$, respectively. For each $i \in [n]$ define

$$\tilde{X}_i := \frac{X_i - \bar{X}_n}{\sqrt{\sum_{i=1}^n \|X_i - \bar{X}_n\|_2^2}},$$

$$\tilde{X}_i^\# := \frac{X_i^\# - \bar{X}_n^\#}{\sqrt{\sum_{i=1}^n \|X_i^\# - \bar{X}_n^\#\|_2^2}}.$$

Notice that the LSL estimator for π^* for the standardized observations $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}^\#$ is defined as follows:

$$\hat{\pi}_n^{\text{LSL}} \in \arg \min_{\pi \in \mathfrak{S}_n} \sum_{i=1}^n \log(\|\tilde{X}_i - \tilde{X}_{\pi(i)}^\#\|_2^2).$$

Since $\sqrt{\sum_{i=1}^n \|X_i - \bar{X}_n\|_2^2}$ is permutation invariant, the latter optimization problem is equivalent to (4).

5 NUMERICAL EXPERIMENTS

In this section, we present numerical experiments on synthetic data to validate the main theoretical findings of our work. Figure 1 visualizes two 100-dimensional synthetic datasets projected onto a shared 2D space using PCA. The underlying model for these datasets follows (1) with the affine conditions specified in (2). The results demonstrate that the baseline Least Sum of Logarithms (LSL) method, which does not account for the scalar affine transformation, fails to recover the correct matching. In contrast, our proposed method estimates the unknown permutation with high accuracy.

The remainder of this section is organized as follows. In Section 5.1, we discuss the estimation procedure for τ^* using $\hat{\tau}_n$. Subsequently, Section 5.2 presents numerical results that illustrate the findings of Theorem 2, with a particular focus on the error introduced by the affine transformation - the component not present in prior work.

5.1 ESTIMATION ERROR OF τ^*

We demonstrate the theoretical result in Theorem 1 through synthetic experiments conducted under the affine model (1)–(2). The data is generated with a scale parameter $\tau^* = 3$ and a shift $\beta^* = 0$. The distortion-to-noise ratio, $\lambda = \|\mu\|_2/\|\sigma\|_2$, is controlled by adjusting the spread of the parameters θ , and the noise magnitudes σ are specified separately in each regime below. Beyond the resulting value of λ , the specific configuration of the sets of vectors θ and $\theta^\#$ does not affect the estimation of τ^* . The scale parameter τ^* is estimated using the estimator $\hat{\tau}_n$, as defined in (3). We conduct two experimental regimes, each averaged over 1000 trials.

In the first regime, shown in the top plot of Figure 2, we vary the relative noise magnitude $\alpha_\sigma = \sigma_{\max}/\|\sigma\|_2$ while keeping $(n, d, \lambda) = (2000, 20, 1)$ fixed. We do so by concentrating the total noise budget $\|\sigma\|_2$ on a varying number k of equally noisy coordinates while keeping the remaining magnitudes negligible, which sets $\alpha_\sigma \asymp 1/\sqrt{k}$ and sweeps it across its range $[n^{-1/2}, 1]$; we work in the regime $n \gg d$, so that the bias term $2d/(n(\lambda^2 + d))$ is negligible. The empirical error then grows linearly in α_σ , saturating the leading term $\alpha_\sigma/\sqrt{\lambda^2 + d}$ of Theorem 1. In the second regime, depicted in the bottom plot of Figure 2, we simultaneously increase both d and λ to explore a wide range of values for the joint quantity $\lambda^2 + d$, while fixing the sample size at $n = 100$ and using homogeneous noise (so that $\alpha_\sigma = n^{-1/2}$ is held fixed). Notably, the empirical error closely follows the theoretical reference slope of $1/\sqrt{\lambda^2 + d}$, when plotted against $\lambda^2 + d$. This confirms the sharpness of the bound of Theorem 1, demonstrating that the empirical error is indeed governed by the two quantities α_σ and $\lambda^2 + d$ through its leading term $\alpha_\sigma/\sqrt{\lambda^2 + d}$.

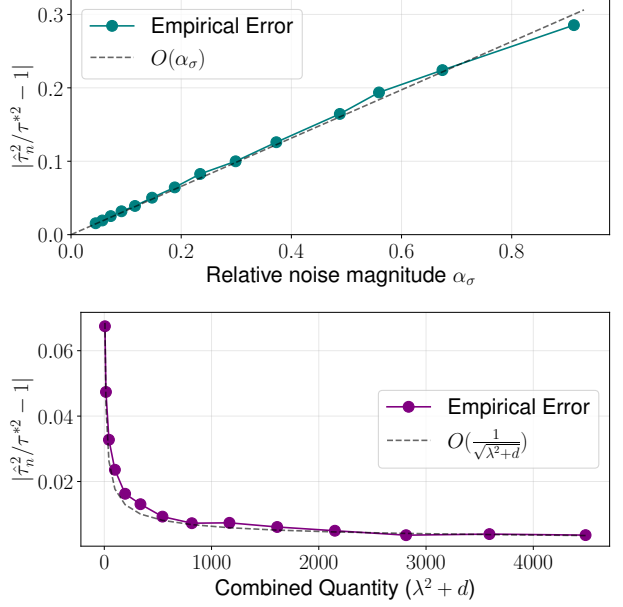


Figure 2: Estimation error versus relative noise magnitude α_σ (Top) and joint quantity $\lambda^2 + d$ (Bottom). The empirical error closely follows the theoretical linear scaling in α_σ and the slope $1/\sqrt{\lambda^2 + d}$, validating the sharpness of the bounds in Theorem 1.

5.2 PERMUTATION RECOVERY AND NECESSITY OF THE $\sqrt{\rho_\sigma}$ TERM

Theorem 2 guarantees perfect recovery once $\kappa(\mathbf{Z})$ exceeds the sum of three contributions. The noise floor $(d \log n)^{1/4} \vee \sqrt{\log n}$, inherent to the Gaussian and χ^2 fluctuations, are standard and consistent with prior work [Collier and Dalalyan, 2016]. The third term $\sim \sqrt{\rho_\sigma}$ is due to the estimation of τ^* and comes downstream as an upper bound of the term $\varepsilon_\tau \max_i \|\mu_i\|_2/s_i$, where ε_τ denotes the relative error of the scale estimate $\hat{\tau}_n$. Below we focus on this term, which is specific to the unknown affine transformation. We ask whether it is an artifact of our proof or an intrinsic limitation of the standardization-based LSL procedure, and we answer the latter with an explicit construction.

Construction. Fix a unit vector $V \in \mathbb{S}^{d-1}$ and set $\tau^* = 1$, $\beta^* = 0$, and $\pi^* = \text{id}$. We place an *adversarial pair* $\theta_1 = MV$, $\theta_2 = (M + \Delta)V$ with $\sigma_1 = \sigma_2 = 1$, a mirror pair $\theta_3 = -MV$, $\theta_4 = -(M + \Delta)V$ that keeps $\bar{\theta}_n = 0$, and $n - 4$ low-signal *inflator* pairs $\theta_{4+2k-1} = ru_k$, $\theta_{4+2k} = -ru_k$ along directions $u_k \in V^\perp$ with large noise $\sigma_i = R$. The inflators carry almost all of the noise mass, so that $\rho_\sigma = R^2$ and $\hat{\tau}_n$ is least accurate precisely along the adversarial direction V . We set $\Delta = \sqrt{2} s_1 \kappa(\mathbf{Z})$ so that the adversarial pair attains the minimum separation $\kappa(\mathbf{Z})$, and $M = CR\sqrt{nd}$ so that $\max_i \|\mu_i\|_2/s_i$ saturates the bound (6), making the term $\varepsilon_\tau \max_i \|\mu_i\|_2/s_i \asymp \sqrt{\rho_\sigma}$ the important quantity.

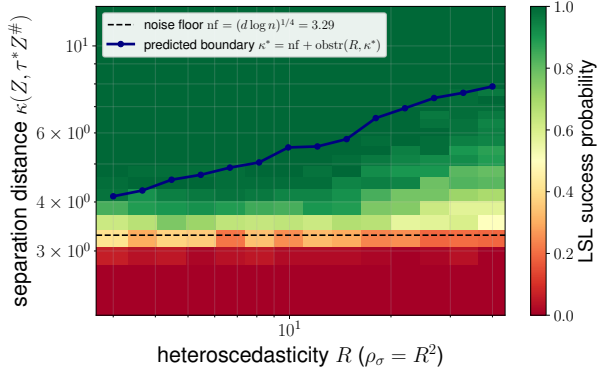


Figure 3: Empirical $\Pr(\hat{\pi}_n^{\text{LSL}} = \pi^*)$ on the adversarial construction, over heteroscedasticity R (with $\rho_\sigma = R^2$) and the separation distance $\kappa(\mathbf{Z})$; here $n = 2500$, $d = 15$, and each cell is averaged over 80 trials. The dashed line marks the noise floor $\text{nf} = (d \log n)^{1/4}$, the separation that already suffices when τ^* is known. The solid curve is the predicted threshold $\kappa^* = \text{nf} + \text{obstr}(R, \kappa^*)$ with $\text{obstr} = \varepsilon_\tau \max_i \|\mu_i\|_2 / s_i$. Recovery fails in a band that rises above the noise floor as ρ_σ grows, tracking the third term of Theorem 2.

Results. Figure 3 reports the empirical recovery probability of $\hat{\pi}_n^{\text{LSL}}$ over the $(R, \kappa(\mathbf{Z}))$ plane. Were the $\sim \sqrt{\rho_\sigma}$ term an artifact, recovery would succeed everywhere above the dashed line; instead a wide band of failure persists above it, and its upper edge climbs steadily with ρ_σ , closely tracked by $\varepsilon_\tau \max_i \|\mu_i\|_2 / s_i$. Figure 4 exposes the mechanism behind this band. Its top panel plots the obstruction $\text{obstr}(\kappa(\mathbf{Z}), R) = \varepsilon_\tau \max_i \|\mu_i\|_2 / s_i$ against the threshold line, and its bottom panel shows the matching success curves on the same $\kappa(\mathbf{Z})$ axis. For each R , success rises to one precisely once the obstruction curve drops below the threshold line; a larger R carries a larger obstruction and therefore fails over a wider range of separations, so the highest- R curve transitions last. Since this obstruction is $\asymp \sqrt{\rho_\sigma}$ in our family, the failure boundary scales as $(d \log n)^{1/4} \vee \sqrt{\rho_\sigma}$: the third term of Theorem 2, cannot be removed by a sharper analysis of LSL.

See more experimental results on real-world translation data from the OPUS-100 dataset [Zhang et al., 2020] in Section C of the supplementary material of this work.

6 CONCLUSION

We have showed that the matching problem under mild heteroscedasticity and a structural assumption that the true feature vectors are connected by an unknown scalar affine transformation is no harder, in a minimax sense, than that of identity transformation. In other words, we showed that it is indeed possible to learn a transformation of the form $\tau I_d + \beta$ between two noisy sequences $\mathbf{X}$ and $\mathbf{X}^\#$ sufficiently well in order to perfectly recover the underlying matching π^* . More-

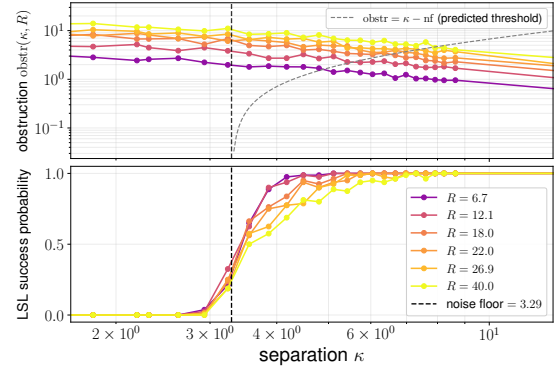


Figure 4: (Top) Empirical obstruction $\text{obstr}(\kappa(\mathbf{Z}), R) = \varepsilon_\tau \max_i \|\mu_i\|_2 / s_i$ for several values of R , against the predicted threshold line (dashed). (Bottom) Corresponding LSL success curves on the same $\kappa(\mathbf{Z})$ axis. For each R , success rises to one just past the $\kappa(\mathbf{Z})$ at which its obstruction curve falls below the threshold line; larger R (hence larger ρ_σ) crosses later, widening the failure region. The vertical dashed line is the noise floor $\text{nf} \approx (d \log n)^{1/4}$.

over, the minimax rate of the minimal separation distance remains the same, matching, up to multiplicative constants, the lower bound in the non-affine setting. However, in the general heterogeneous setting, where the noise magnitudes are allowed to grow with sample size n , we showed that the rate changes and scales linearly with the square root of the maximal ratio of noise variances. This phenomenon is rooted in the estimation error of the scaling factor which depends on the maximal ratio of noise variances. Therefore, a larger minimal separation distance is required to be able to perfectly recover the unknown permutation. Interestingly, learning an unknown scalar affine transformation is equivalent to standardizing the sequences $\mathbf{X}$ and $\mathbf{X}^\#$, that is, subtracting the sample mean and dividing by the square root of the total sample variance.

Acknowledgements

We would like to thank Arnak Dalalyan and Yann Issartel for helpful discussions.

References

- Mona Azadkia and Fadoua Balabdaoui. Linear regression with unmatched data: a deconvolution perspective. *Journal of Machine Learning Research*, 25(197):1–55, 2024.
- Serhat Bakirtas and Elza Erkip. Database matching under noisy synchronization errors. *IEEE Transactions on Information Theory*, 70(6):4335–4367, 2024. doi: 10.1109/TIT.2024.3388990.

- Fadoua Balabdaoui, Charles R Doss, and Cécile Durot. Unlinked monotone regression. *Journal of Machine Learning Research*, 22(172):1–60, 2021.
- Nikolas Barkas, Viktor Petukhov, Daria Nikolaeva, Yaroslav Lozinsky, Samuel Demharter, Konstantin Khodosevich, and Peter V Kharchenko. Joint analysis of heterogeneous single-cell rna-seq dataset collections. *Nature methods*, 16(8):695, 2019.
- Gilles Blanchard, Alexandra Carpentier, and Maurilio Gutzeit. Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electron. J. Stat.*, 12(2):3713–3735, 2018. ISSN 1935-7524/e.
- 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.
- Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snaveley. Doppelgangers: Learning to disambiguate images of similar structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 34–44, 2023.
- Olivier Collier and Arnak S Dalalyan. Minimax rates in permutation estimation for feature matching. *The Journal of Machine Learning Research*, 17(1):162–192, 2016.
- Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Stat.*, 40(5):2667–2696, 2012. ISSN 0090-5364; 2168-8966/e.
- David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227, 1979.
- Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1):623–653, 2019.
- Tigran Galstyan, Arshak Minasyan, and Arnak S. Dalalyan. Optimal detection of the feature matching map in presence of noise and outliers. *Electronic Journal of Statistics*, 16(2):5720–5750, 2022. doi: 10.1214/22-EJS2076. URL <https://doi.org/10.1214/22-EJS2076>.
- Luca Ganassali and Laurent Massoulié. From tree matching to sparse graph alignment. In *Conference on Learning Theory*, pages 1633–1665. PMLR, 2020.
- Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k -means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.
- Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.
- Georgina Hall and Laurent Massoulié. Partial recovery in the graph alignment problem. *Operations Research*, 71(1):259–272, 2023.
- YanJun Han, Philippe Rigollet, and George Stepaniants. Covariance alignment: from maximum likelihood estimation to gromov–wasserstein. *SIAM Journal on Mathematics of Data Science*, 7(3):1491–1513, 2025.
- Yu. I. Ingster. Minimax nonparametric detection of signals in white Gaussian noise. *Probl. Inf. Transm.*, 18:130–140, 1982.
- Yann Issartel, Christophe Giraud, and Nicolas Verzelen. Minimax optimal seriation in polynomial time. *Annals of Statistics*, 2024.
- Anatoli Juditsky and Arkadin Nemirovski. *Statistical inference via convex optimization*. Princeton, NJ: Princeton University Press, 2020. ISBN 978-0-691-19729-6/hbk; 978-0-691-20031-6/ebook.
- Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.
- Dmitriy Kunisky and Jonathan Niles-Weed. Strong recovery of geometric planted matchings. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 834–876. SIAM, 2022.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.
- Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *The Twelfth International Conference on Learning Representations*, 2024.
- Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory*, pages 821–847. PMLR, 2018.
- Arshak Minasyan, Tigran Galstyan, Sona Hunanyan, and Arnak Dalalyan. Matching map recovery with an unknown number of outliers. In *International Conference on Artificial Intelligence and Statistics*, pages 891–906. PMLR, 2023.
- Mohamed Ndaoud and Alexandre B. Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory*, 66(4):2517–2532, 2020. ISSN 0018-9448.

- Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Denoising linear models with permuted data. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 446–450. IEEE, 2017a.
- Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64(5):3286–3300, 2017b.
- Thomas Pumir, Amit Singer, and Nicolas Boumal. The generalized orthogonal procrustes problem in the high noise regime. *Information and Inference: A Journal of the IMA*, 10(3):921–954, 2021.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.
- Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL <https://arxiv.org/abs/2004.09813>.
- Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1 – 9, 2013. doi: 10.1214/ECP.v18-2865.
- Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 67(6):4162–4184, 2021. doi: 10.1109/TIT.2020.3045613.
- Chenghao Shi, Xieyuanli Chen, Huimin Lu, Wenbang Deng, Junhao Xiao, and Bin Dai. Rdmnet: Reliable dense matching based point cloud registration for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 24(10):11372–11383, 2023. doi: 10.1109/TITS.2023.3286464.
- Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in $\mathbb{R}^d$: an approach based on mixture modeling and optimal transport. *Journal of Machine Learning Research*, 25(183):1–57, 2024.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs. In *Conference on Learning Theory*, pages 3441–3488. PMLR, 2022.
- Yuting Wei, Martin J. Wainwright, and Adityanand Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *The Annals of Statistics*, 47(2):994 – 1024, 2019.
- Geoffrey Wolfer and Aryeh Kontorovich. Minimax testing of identity to a reference ergodic markov chain. In *AISTATS 2020*, volume 108, pages 191–201. Proceedings of Machine Learning Research, 2020.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.148.
- Bokai Zhu, Shuxiao Chen, Yunhao Bai, Han Chen, Guanrui Liao, Nilanjan Mukherjee, Gustavo Vazquez, David R McIlwain, Alexandar Tzankov, Ivan T Lee, et al. Robust single-cell matching and multimodal analysis using shared and distinct features. *Nature Methods*, 20(2):304–315, 2023.

Perfect Matching Map Recovery Under Unknown Scalar Affine Transformation (Supplementary Material)

Tigran Galstyan^{1, 2}

Avetik Karagulyan^{1, 2}

Arshak Minasyan¹

¹Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

²Centre National de la Recherche Scientifique (CNRS), Gif-sur-Yvette, France

A POSTPONED PROOFS FROM SECTION 3

We first state a few standard lemmas that are used to control the tails of Gaussian and χ^2 distributions.

Lemma 1. *[Laurent and Massart, 2000, Lemma 4.1] Let $(Y_1, \dots, Y_n)$ be i.i.d. Gaussian random variables with mean 0 and variance 1. Let $\mathbf{a} = (a_1, \dots, a_n)$ be a collection of nonnegative constants. Let T be the weighted sum, defined as*

$$T = \sum_{i=1}^n a_i (Y_i^2 - 1).$$

Then, the following inequalities hold for any positive x

$$\mathbb{P}(T \geq 2\|\mathbf{a}\|_2 \sqrt{x} + 2\|\mathbf{a}\|_\infty x) \leq \exp(-x), \quad \mathbb{P}(T \leq -2\|\mathbf{a}\|_2 \sqrt{x}) \leq \exp(-x).$$

Lemma 2 (Gaussian concentration inequality). *Let $(Y_1, \dots, Y_n)$ be independent Gaussian random vectors in $\mathbb{R}^d$ with mean 0 and covariance matrices $\sigma_i^2 I_d$. Let $\mathbf{a} = (a_1, \dots, a_n)$ be a collection of fixed arbitrary vectors in $\mathbb{R}^d$. Let T be the weighted sum, defined as*

$$T = \sum_{i=1}^n a_i^\top Y_i.$$

Then, for any positive x we have

$$\mathbb{P}(|T| \geq x) \leq 2 \exp(-x^2 / (2 \sum_{i=1}^n \sigma_i^2 \|\mathbf{a}_i\|^2))$$

A.1 PROOF OF PROPOSITION 1

Proof. Recall that Z_i are Gaussian vectors with parameters $(\mu_i, s_i^2 I_d)$, and decompose them into $Z_i = \mu_i + \bar{\xi}_{i,n}$, where

$$\bar{\xi}_{i,n} := \sigma_i \xi_i - \frac{1}{n} \sum_{k=1}^n \sigma_k \xi_k.$$

Then the sum of squared ℓ_2 norms of $Z_1, \dots, Z_n$ can be decomposed as

$$\sum_{i=1}^n \|Z_i\|_2^2 = \|\boldsymbol{\mu}\|_2^2 + 2 \sum_{i=1}^n \sigma_i \mu_i^\top \xi_i + \sum_{i=1}^n \|\bar{\xi}_{i,n}\|_2^2,$$

where we used that summing $\mu_i^\top \bar{\xi}_{i,n}$ over $[n]$ is the same as summing $\sigma_i \mu_i^\top \xi_i$ over the same set, since $\sum_{i=1}^n \mu_i = 0$. We bound the last two terms separately.

The term $\sum_{i=1}^n \sigma_i \mu_i^\top \xi_i$ has a Gaussian distribution with mean zero and variance $\sum_{i=1}^n \sigma_i^2 \|\mu_i\|_2^2 \leq \sigma_{\max}^2 \|\boldsymbol{\mu}\|_2^2$. Thus, by the Gaussian concentration inequality, with probability at least $1 - \delta/2$,

$$\sum_{i=1}^n \sigma_i \mu_i^\top \xi_i \leq \sigma_{\max} \|\boldsymbol{\mu}\|_2 \sqrt{2 \log(2/\delta)}.$$

As for the term $\sum_{i=1}^n \|\bar{\xi}_{i,n}\|_2^2$, we write it as a quadratic form in the underlying noise. Stack the noise vectors into $\boldsymbol{\xi} := (\xi_1^\top, \dots, \xi_n^\top)^\top \sim \mathcal{N}(0, I_{nd})$, and let $D := \text{diag}(\sigma_1, \dots, \sigma_n)$ scale the n blocks while $P := I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$ is the centering projection that subtracts the block average. The two operations act blockwise through the Kronecker product: $(D \otimes I_d)\boldsymbol{\xi}$ stacks the scaled vectors $\sigma_i \xi_i$, and applying $P \otimes I_d$ subtracts their average, so that

$$[(PD \otimes I_d)\boldsymbol{\xi}]_i = \sigma_i \xi_i - \frac{1}{n} \sum_{k=1}^n \sigma_k \xi_k = \bar{\xi}_{i,n}.$$

Therefore $\sum_{i=1}^n \|\bar{\xi}_{i,n}\|_2^2 = \|(PD \otimes I_d)\boldsymbol{\xi}\|_2^2 = \boldsymbol{\xi}^\top (PD \otimes I_d)^\top (PD \otimes I_d) \boldsymbol{\xi}$. By the Kronecker identities $(M \otimes I_d)^\top = M^\top \otimes I_d$ and $(M_1 \otimes I_d)(M_2 \otimes I_d) = (M_1 M_2) \otimes I_d$, together with the symmetry of D and P and the idempotence $P^2 = P$, the inner matrix equals $(PD)^\top (PD) \otimes I_d = (DPD) \otimes I_d$. Moreover, $DPD = D^\top P D$ is a congruence of the projection $P \succeq 0$ and is therefore positive semi-definite. Writing $A := (DPD) \otimes I_d$, we conclude that

$$\sum_{i=1}^n \|\bar{\xi}_{i,n}\|_2^2 = \boldsymbol{\xi}^\top A \boldsymbol{\xi}, \quad A \succeq 0.$$

Using $\text{tr}(M \otimes I_d) = d \text{tr}(M)$, $\|M \otimes I_d\|_F = \sqrt{d} \|M\|_F$ and $\|M \otimes I_d\|_{\text{op}} = \|M\|_{\text{op}}$, a direct computation yields

$$\begin{aligned} \text{tr}(A) &= d \text{tr}(PD^2) = d \frac{n-1}{n} \|\boldsymbol{\sigma}\|_2^2, \\ \|A\|_F^2 &= d \text{tr}((DPD)^2) = d \left(\frac{n-2}{n} \|\boldsymbol{\sigma}\|_4^4 + \frac{1}{n^2} \|\boldsymbol{\sigma}\|_2^4 \right) \leq d \|\boldsymbol{\sigma}\|_4^4, \\ \|A\|_{\text{op}} &= \|DPD\|_{\text{op}} \leq \|D\|_{\text{op}}^2 \|P\|_{\text{op}} = \sigma_{\max}^2, \end{aligned}$$

where the bound on $\|A\|_F^2$ uses $\|\boldsymbol{\sigma}\|_2^4 \leq n \|\boldsymbol{\sigma}\|_4^4$. We now reduce the quadratic form $\boldsymbol{\xi}^\top A \boldsymbol{\xi}$ to a weighted sum of squared Gaussians. Since A is symmetric and positive semi-definite, the spectral theorem gives $A = U \Lambda U^\top$, where U is orthogonal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{nd})$ collects the eigenvalues $\lambda_k \geq 0$. Let $\boldsymbol{\eta} := U^\top \boldsymbol{\xi}$. Since $\boldsymbol{\xi} \sim \mathcal{N}(0, I_{nd})$ and U is orthogonal, $\boldsymbol{\eta}$ is again centered Gaussian with covariance $U^\top U = I_{nd}$, so $\boldsymbol{\eta} \sim \mathcal{N}(0, I_{nd})$ and its coordinates $\eta_1, \dots, \eta_{nd}$ are independent $\mathcal{N}(0, 1)$ variables. In this eigenbasis,

$$\boldsymbol{\xi}^\top A \boldsymbol{\xi} = \boldsymbol{\eta}^\top \Lambda \boldsymbol{\eta} = \sum_{k=1}^{nd} \lambda_k \eta_k^2 \quad \text{and} \quad \text{tr}(A) = \sum_{k=1}^{nd} \lambda_k,$$

so that

$$\boldsymbol{\xi}^\top A \boldsymbol{\xi} - \text{tr}(A) = \sum_{k=1}^{nd} \lambda_k (\eta_k^2 - 1).$$

The right-hand side is a weighted sum of independent centered chi-squared variables with nonnegative weights $a_k = \lambda_k$, so Lemma 1 applies. Its norms are exactly the matrix norms computed above, namely $\sum_k a_k^2 = \sum_k \lambda_k^2 = \|A\|_F^2$ and $\max_k a_k = \max_k \lambda_k = \|A\|_{\text{op}}$. Applying the upper bound of Lemma 1 with $x = \log(2/\delta)$, we have with probability at least $1 - \delta/2$,

$$\sum_{i=1}^n \|\bar{\xi}_{i,n}\|_2^2 \leq d \frac{n-1}{n} \|\boldsymbol{\sigma}\|_2^2 + 2 \|\boldsymbol{\sigma}\|_4^4 \sqrt{d \log(2/\delta)} + 2 \sigma_{\max}^2 \log(2/\delta).$$

Combining these two bounds via the union bound, and using $d \frac{n-1}{n} \|\boldsymbol{\sigma}\|_2^2 \leq d \|\boldsymbol{\sigma}\|_2^2$, we arrive at the following upper bound, which holds with probability at least $1 - \delta$:

$$\sum_{i=1}^n \|Z_i\|_2^2 \leq \|\boldsymbol{\mu}\|_2^2 + d \|\boldsymbol{\sigma}\|_2^2 + 2\sigma_{\max} \|\boldsymbol{\mu}\|_2 \sqrt{2 \log(2/\delta)} + 2 \|\boldsymbol{\sigma}\|_4^2 \sqrt{d \log(2/\delta)} + 2\sigma_{\max}^2 \log(2/\delta),$$

completing the first part of the proof.

To prove the lower bound, we apply the lower-bound parts of the Gaussian concentration inequality and Lemma 1. With probability at least $1 - \delta/2$,

$$\sum_{i=1}^n \sigma_i \mu_i^\top \xi_i \geq -\sigma_{\max} \|\boldsymbol{\mu}\|_2 \sqrt{2 \log(2/\delta)},$$

and also with probability at least $1 - \delta/2$,

$$\sum_{i=1}^n \|\bar{\xi}_{i,n}\|_2^2 \geq d \frac{n-1}{n} \|\boldsymbol{\sigma}\|_2^2 - 2 \|\boldsymbol{\sigma}\|_4^2 \sqrt{d \log(2/\delta)}.$$

Combining these via the union bound and using $d \frac{n-1}{n} \|\boldsymbol{\sigma}\|_2^2 = d \|\boldsymbol{\sigma}\|_2^2 - \frac{d}{n} \|\boldsymbol{\sigma}\|_2^2$, we obtain that with probability at least $1 - \delta$,

$$\sum_{i=1}^n \|Z_i\|_2^2 \geq \|\boldsymbol{\mu}\|_2^2 + d \|\boldsymbol{\sigma}\|_2^2 - 2\sigma_{\max} \|\boldsymbol{\mu}\|_2 \sqrt{2 \log(2/\delta)} - 2 \|\boldsymbol{\sigma}\|_4^2 \sqrt{d \log(2/\delta)} - \frac{d}{n} \|\boldsymbol{\sigma}\|_2^2,$$

which completes the proof. $\square$

A.2 PROOF OF THEOREM 1

Recall the definition of $\hat{\tau}_n$ from (3):

$$\hat{\tau}_n := \left(\frac{\sum_{i=1}^n \|X_i - \bar{X}_n\|_2^2}{\sum_{i=1}^n \|X_i^\# - \bar{X}_n^\#\|_2^2} \right)^{1/2} = \left(\frac{\sum_{i=1}^n \|Z_i\|_2^2}{\sum_{i=1}^n \|Z_i^\#\|_2^2} \right)^{1/2}.$$

Then,

$$\frac{\hat{\tau}_n^2}{\tau^{*2}} - 1 = \frac{\sum_{i=1}^n (\|Z_i\|_2^2 - \|Z_i^\#\|_2^2)}{\tau^{*2} \sum_{i=1}^n \|Z_i^\#\|_2^2}.$$

The proof consists of simultaneously upper and lower bounding the numerator and the denominator, then combining these bounds using the union bound. In what follows we apply Proposition 1 four times: upper and lower bounds for both centralized sequences $\mathbf{Z}$ and $\mathbf{Z}^\#$ in each of the cases $\hat{\tau}_n > \tau^*$ and $\hat{\tau}_n < \tau^*$, yielding an upper bound for the absolute value of the difference between $\hat{\tau}_n^2$ and τ^{*2} . Notice that, for all $i \in [n]$ we have that $Z_i \perp Z_{\pi^*(i)}^\#$, and $Z_i, \tau^* Z_{\pi^*(i)}^\# \sim \mathcal{N}(\mu_i, s_i^2 I_d)$. In what follows, without loss of generality, we assume $\tau^* = 1$ for notational convenience. Using the lower bound of Proposition 1 for $\mathbf{Z}^\#$ we get that with probability at least $1 - \delta$,

$$\sum_{i=1}^n \|Z_i^\#\|_2^2 \geq \|\boldsymbol{\mu}\|_2^2 + d \|\boldsymbol{\sigma}\|_2^2 - 2\sigma_{\max} \|\boldsymbol{\mu}\|_2 \sqrt{2 \log(2/\delta)} - 2 \|\boldsymbol{\sigma}\|_4^2 \sqrt{d \log(2/\delta)} - \frac{d}{n} \|\boldsymbol{\sigma}\|_2^2. \quad (9)$$

Similarly, for the upper bound we have, with probability at least $1 - \delta$,

$$\sum_{i=1}^n \|Z_i\|_2^2 \leq \|\boldsymbol{\mu}\|_2^2 + d \|\boldsymbol{\sigma}\|_2^2 + 2\sigma_{\max} \|\boldsymbol{\mu}\|_2 \sqrt{2 \log(2/\delta)} + 2 \|\boldsymbol{\sigma}\|_4^2 \sqrt{d \log(2/\delta)} + 2\sigma_{\max}^2 \log(2/\delta), \quad (10)$$

Both (9) and (10) follow from Proposition 1 applied at confidence level δ .

Combining these bounds we get that

$$\sum_{i=1}^n (\|Z_i\|_2^2 - \|Z_i^\# \|_2^2) \leq 4\sigma_{\max}\|\boldsymbol{\mu}\|_2\sqrt{2\log(2/\delta)} + 4\|\boldsymbol{\sigma}\|_4^2\sqrt{d\log(2/\delta)} + 2\sigma_{\max}^2\log(2/\delta) + \frac{d}{n}\|\boldsymbol{\sigma}\|_2^2$$

with probability at least $1 - 2\delta$, holding simultaneously with (9). Thus, with probability at least $1 - 4\delta$,

$$\left| \sum_{i=1}^n (\|Z_i\|_2^2 - \|Z_i^\# \|_2^2) \right| \leq 4\sigma_{\max}\|\boldsymbol{\mu}\|_2\sqrt{2\log(2/\delta)} + 4\|\boldsymbol{\sigma}\|_4^2\sqrt{d\log(2/\delta)} + 2\sigma_{\max}^2\log(2/\delta) + \frac{d}{n}\|\boldsymbol{\sigma}\|_2^2,$$

which holds simultaneously with (9). Recall that $\alpha_\sigma = \sigma_{\max}/\|\boldsymbol{\sigma}\|_2$ and $\lambda = \|\boldsymbol{\mu}\|_2/\|\boldsymbol{\sigma}\|_2$. Dividing the previous display by $\|\boldsymbol{\sigma}\|_2^2$ and using the following identities

$$\frac{\sigma_{\max}\|\boldsymbol{\mu}\|_2}{\|\boldsymbol{\sigma}\|_2^2} = \alpha_\sigma\lambda, \quad \frac{\sigma_{\max}^2}{\|\boldsymbol{\sigma}\|_2^2} = \alpha_\sigma^2, \quad \frac{\|\boldsymbol{\sigma}\|_4^2}{\|\boldsymbol{\sigma}\|_2^2} \leq \alpha_\sigma,$$

where the last follows from $\|\boldsymbol{\sigma}\|_4^4 \leq \sigma_{\max}^2\|\boldsymbol{\sigma}\|_2^2$. Combining the linear cross-terms via $\sqrt{2}\lambda + \sqrt{d} \leq \sqrt{2}(\lambda + \sqrt{d})$, we obtain

$$\left| \sum_{i=1}^n (\|Z_i\|_2^2 - \|Z_i^\# \|_2^2) \right| \leq \|\boldsymbol{\sigma}\|_2^2 \left(4\alpha_\sigma(\lambda + \sqrt{d})\sqrt{2\log(2/\delta)} + 2\alpha_\sigma^2\log(2/\delta) + \frac{d}{n} \right). \quad (11)$$

Applying the same identities to the right-hand side of (9), we get that with probability at least $1 - \delta$

$$\sum_{i=1}^n \|Z_i^\# \|_2^2 \geq \|\boldsymbol{\sigma}\|_2^2 \left(\lambda^2 + d - 2\alpha_\sigma(\lambda + \sqrt{d})\sqrt{2\log(2/\delta)} - \frac{d}{n} \right). \quad (12)$$

The conditions of the theorem ensures that the obtained lower bound (12) is indeed larger than the upper bound from (11) for any $\lambda \geq 0$. Notice that otherwise the consistent estimation of τ^* is impossible. In order for $|\hat{\tau}_n^2/\tau^{*2} - 1|$ to be smaller than one, the right-hand side of (12) must dominate that of (11), which is equivalent to

$$h(\lambda) := \lambda^2 + d - 6\alpha_\sigma(\lambda + \sqrt{d})\sqrt{2\log(2/\delta)} - 2\alpha_\sigma^2\log(2/\delta) - \frac{2d}{n} \geq 0,$$

for every $\lambda \geq 0$. The minimum of this quadratic in λ , attained at $\lambda^* = 3\alpha_\sigma\sqrt{2\log(2/\delta)} \geq 0$, gives

$$h(\lambda^*) = d - 20\alpha_\sigma^2\log(2/\delta) - 6\alpha_\sigma\sqrt{2d\log(2/\delta)} - \frac{2d}{n}.$$

By the AM-GM inequality, $6\alpha_\sigma\sqrt{2d\log(2/\delta)} = \sqrt{d \cdot 72\alpha_\sigma^2\log(2/\delta)} \leq d/2 + 36\alpha_\sigma^2\log(2/\delta)$, hence

$$h(\lambda^*) \geq \frac{d}{2} - 56\alpha_\sigma^2\log(2/\delta) - \frac{2d}{n}.$$

Now given that $n \geq 8$ and $\log(2/\delta) \leq d/(224\alpha_\sigma^2)$, one can check that

$$56\alpha_\sigma^2\log(2/\delta) \leq \frac{d}{4}, \quad \frac{2d}{n} \leq \frac{d}{4},$$

and therefore $h(\lambda^*) \geq 0$. Since for any positive a and b such that $a \leq b$ and $x > 0$ we have $\frac{a}{b} \leq \frac{a+x}{b+x}$, we arrive at the following bound for the estimation error of τ^* , with probability $1 - 4\delta$

$$\begin{aligned} \left| \hat{\tau}_n^2 - 1 \right| &= \frac{\left| \sum_{i=1}^n \|Z_i\|_2^2 - \sum_{i=1}^n \|Z_i^\# \|_2^2 \right|}{\sum_{i=1}^n \|Z_i^\# \|_2^2} \\ &\leq \frac{4\alpha_\sigma(\lambda + \sqrt{d})\sqrt{2\log(2/\delta)} + 2\alpha_\sigma^2\log(2/\delta) + n^{-1}d}{\lambda^2 + d - 2\alpha_\sigma(\lambda + \sqrt{d})\sqrt{2\log(2/\delta)} - n^{-1}d} \\ &\leq \frac{6\alpha_\sigma(\lambda + \sqrt{d})\sqrt{2\log(2/\delta)} + 2\alpha_\sigma^2\log(2/\delta) + 2n^{-1}d}{\lambda^2 + d} \\ &\leq 12\alpha_\sigma\sqrt{\frac{\log(2/\delta)}{\lambda^2 + d}} + \frac{2\alpha_\sigma^2\log(2/\delta) + 2n^{-1}d}{\lambda^2 + d}, \end{aligned}$$

where we used $\lambda + \sqrt{d} \leq \sqrt{2(\lambda^2 + d)}$ in the final step, concluding the proof.

A.3 PROOF OF THEOREM 2

We first state two key lemmas that are used to control the tails of standard Gaussian and normalized χ^2 distributions. Their proofs are postponed to Sections B.1 and B.2, respectively.

Lemma 3. *Let $\eta_1, \dots, \eta_n$ and $\eta'_1, \dots, \eta'_n$ be two mutually independent sets of d -dimensional Gaussian random vectors, with $\eta_i, \eta'_i \sim N(0, \omega_i^2 I_d)$. Notice that η_i and η_j may be correlated. Fix a set of unit vectors $u_{i,j} \in \mathbb{S}^{d-1}$, with $i, j \in [n]$ and $\varepsilon \in (0, 1/2)$. Define*

$$\Phi_1 := \sup_{a \in [1-\varepsilon, 1+\varepsilon]} \max_{i,j} \left| \frac{u_{i,j}^\top (\eta_i - a\eta'_j)}{\sqrt{\omega_i^2 + a^2 \omega_j^2}} \right|.$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\Phi_1 \leq \sqrt{2 \log(12n^3/\delta)} + \frac{2\varepsilon}{n} \sqrt{\log(8n/\delta)}.$$

Lemma 4. *Let $\eta_1, \dots, \eta_n$ and $\eta'_1, \dots, \eta'_n$ be two mutually independent sets of d -dimensional Gaussian random vectors, with $\eta_i, \eta'_i \sim N(0, \omega_i^2 I_d)$. Notice that η_i and η_j may be correlated. For some $\varepsilon \in (0, 1/2)$, define*

$$\Phi_2 := \sup_{a \in [1-\varepsilon, 1+\varepsilon]} \max_{i,j} \left| \frac{\|\eta_i - a\eta'_j\|_2^2}{\omega_i^2 + a^2 \omega_j^2} - d \right|.$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\Phi_2 \leq 2\sqrt{d \log(12n^3/\delta)} + 2 \log(12n^3/\delta) + \frac{4\varepsilon}{n} \left(d + 2\sqrt{d \log(4n/\delta)} + 2 \log(4n/\delta) \right).$$

Proof of Theorem 2. Without loss of generality, we assume π^* is the identity map id and $\tau^* = 1$. Recall that the observations $\mathbf{Z}$ and $\mathbf{Z}^\#$ are the centralized versions of $\mathbf{X}$ and $\mathbf{X}^\#$, respectively. Recall as well the objective function of the LSL estimator from (4):

$$\ell_n(\pi, \mathbf{Z}, \mathbf{Z}^\#) = \sum_{i=1}^n \log \left(\|Z_i - \hat{\pi}_n Z_{\pi(i)}^\#\|_2^2 \right),$$

up to some permutation independent constants. We aim to bound the probability of the event $\Omega = \{\hat{\pi} \neq \text{id}\}$, where $\hat{\pi} = \hat{\pi}_n^{\text{LSL}} = \arg \min_{\pi \in \mathfrak{S}_n} \ell_n(\pi, \mathbf{Z}, \mathbf{Z}^\#)$. Our primary observation is that

$$\Omega \subset \cup_{\pi \neq \text{id}} \Omega_\pi, \quad \text{where} \quad \Omega_\pi = \left\{ \pi \in \arg \min_{\pi \in \mathfrak{S}_n} \ell_n(\pi, \mathbf{Z}, \mathbf{Z}^\#) \right\}.$$

Then, the following chain of inclusions holds:

$$\begin{aligned} \Omega &\subset \bigcup_{\pi \neq \text{id}} \Omega_\pi \subset \bigcup_{\pi \neq \text{id}} \{ \ell_n(\pi, \mathbf{Z}, \mathbf{Z}^\#) \leq \ell_n(\text{id}, \mathbf{Z}, \mathbf{Z}^\#) \} \\ &\subset \bigcup_{\pi \neq \text{id}} \left\{ \sum_{i: \pi(i) \neq i} \log \frac{\|Z_i - \hat{\pi}_n Z_i^\#\|_2^2}{\|Z_i - \hat{\pi}_n Z_{\pi(i)}^\#\|_2^2} \geq 0 \right\}. \end{aligned}$$

Jensen's inequality yields the following for every $t > 0$:

$$\begin{aligned} \sum_{\pi(i) \neq i} \log \left(\frac{(1+t^2)s_i^2}{s_i^2 + t^2 s_{\pi(i)}^2} \right) &= \sum_{i=1}^n \left(\log((1+t^2)s_i^2) - \log(s_i^2 + t^2 s_{\pi(i)}^2) \right) \\ &= \sum_{i=1}^n \left(\frac{\log((1+t^2)s_i^2) + t^2 \log((1+t^2)s_{\pi(i)}^2)}{1+t^2} - \log(s_i^2 + t^2 s_{\pi(i)}^2) \right) \leq 0. \end{aligned}$$

Setting $t = \hat{\tau}_n$ yields

$$\begin{aligned}\Omega &\subset \bigcup_{\pi \neq \text{id}} \left\{ \sum_{i: \pi(i) \neq i} \log \frac{\|Z_i - \hat{\tau}_n Z_i^\# \|_2^2}{\|Z_i - \hat{\tau}_n Z_{\pi(i)}^\# \|_2^2} - \sum_{i: \pi(i) \neq i} \log \left(\frac{(1 + \hat{\tau}_n^2) s_i^2}{s_i^2 + \hat{\tau}_n^2 s_{\pi(i)}^2} \right) \geq 0 \right\} \\ &\subset \bigcup_{i=1}^n \bigcup_{j \neq i} \left\{ \frac{\|Z_i - \hat{\tau}_n Z_i^\# \|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} \geq \frac{\|Z_i - \hat{\tau}_n Z_j^\# \|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} \right\}.\end{aligned}$$

Define the random event $\Omega_\tau := \{\hat{\tau}_n \in \mathcal{E}_\tau := [1 - \varepsilon_\tau, 1 + \varepsilon_\tau]\}$, where ε_τ is the upper bound on the estimation error of $\hat{\tau}_n$ derived in Theorem 1. Then, intersecting with the event Ω_τ we get the following inclusion:

$$\Omega \subseteq \left\{ \bigcup_{i=1}^n \bigcup_{j \neq i} \left\{ \frac{\|Z_i - \hat{\tau}_n Z_i^\# \|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} \geq \frac{\|Z_i - \hat{\tau}_n Z_j^\# \|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} \right\} \cap \Omega_\tau \right\} \cup \Omega_\tau^c. \quad (13)$$

In order to upper bound the probability of the event in (13) we first upper and lower bound the left and right hand sides of the inequality from (13), respectively. In other words, for the correct pair we present an upper bound on the left hand side, while for the incorrect pair we compute bound it from below. From the definitions of Z_i and $Z_j^\#$, for any pair $(i, j) \in [n]^2$ we have the following decomposition

$$Z_i - \hat{\tau}_n Z_j^\# = \mu_i - \hat{\tau}_n \mu_j + \bar{\xi}_{i,n} - \hat{\tau}_n \bar{\xi}_{j,n}^\#,$$

where $\mu_i = \theta_i - \bar{\theta}_n$ is defined in (5) and

$$\bar{\xi}_{i,n} := \sigma_i \xi_i - \frac{1}{n} \sum_{i=1}^n \sigma_i \xi_i, \quad \bar{\xi}_{i,n}^\# := \sigma_i^\# \xi_i^\# - \frac{1}{n} \sum_{i=1}^n \sigma_i^\# \xi_i^\#.$$

We now proceed with upper bounding the normalized distance for the correct pair:

$$\begin{aligned}\frac{\|Z_i - \hat{\tau}_n Z_i^\# \|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} &= \frac{\|\mu_i - \hat{\tau}_n \mu_i\|_2^2 + 2(\mu_i - \hat{\tau}_n \mu_i)^\top (\bar{\xi}_{i,n} - \hat{\tau}_n \bar{\xi}_{i,n}^\#) + \|\bar{\xi}_{i,n} - \hat{\tau}_n \bar{\xi}_{i,n}^\# \|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} \\ &\leq \frac{\|\mu_i - \hat{\tau}_n \mu_i\|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} + \frac{2\zeta_1 \|\mu_i - \hat{\tau}_n \mu_i\|}{s_i \sqrt{1 + \hat{\tau}_n^2}} + d + \sqrt{d} \zeta_2,\end{aligned}$$

where ζ_1 and ζ_2 are defined as

$$\zeta_1 := \sup_{\hat{\tau}_n \in \mathcal{E}_\tau} \max_{i,j} \left| \frac{(\mu_i - \hat{\tau}_n \mu_j)^\top (\bar{\xi}_{i,n} - \hat{\tau}_n \bar{\xi}_{j,n}^\#)}{\|\mu_i - \hat{\tau}_n \mu_j\| \sqrt{s_i^2 + \hat{\tau}_n^2 s_j^2}} \right|, \quad \zeta_2 := d^{-1/2} \sup_{\hat{\tau}_n \in \mathcal{E}_\tau} \max_{i,j} \left| \frac{\|\bar{\xi}_{i,n} - \hat{\tau}_n \bar{\xi}_{j,n}^\# \|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} - d \right|.$$

The lower bound on the distance between incorrect pairs ($i \neq j$) is constructed in a similar manner:

$$\frac{\|Z_i - \hat{\tau}_n Z_j^\# \|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} \geq \frac{\|\mu_i - \hat{\tau}_n \mu_j\|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} - \frac{2\zeta_1 \|\mu_i - \hat{\tau}_n \mu_j\|}{\sqrt{s_i^2 + \hat{\tau}_n^2 s_j^2}} + d - \sqrt{d} \zeta_2.$$

Thus, the first event from (13) satisfies the following chain of inclusions

$$\begin{aligned}&\bigcup_{i=1}^n \bigcup_{j \neq i} \left\{ \frac{\|Z_i - \hat{\tau}_n Z_i^\# \|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} \geq \frac{\|Z_i - \hat{\tau}_n Z_j^\# \|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} \right\} \cap \Omega_\tau \\ &\subseteq \bigcup_{i=1}^n \bigcup_{j \neq i} \left\{ \frac{\|\mu_i - \hat{\tau}_n \mu_i\|_2^2}{(1 + \hat{\tau}_n^2) s_i^2} + \frac{2\zeta_1 \|\mu_i - \hat{\tau}_n \mu_i\|}{s_i \sqrt{1 + \hat{\tau}_n^2}} + 2\sqrt{d} \zeta_2 \geq \frac{\|\mu_i - \hat{\tau}_n \mu_j\|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2} - \frac{2\zeta_1 \|\mu_i - \hat{\tau}_n \mu_j\|}{\sqrt{s_i^2 + \hat{\tau}_n^2 s_j^2}} \right\} \cap \Omega_\tau \\ &\subseteq \bigcup_{i=1}^n \bigcup_{j \neq i} \left\{ \tilde{\kappa}_{i,i}^2 + 2\zeta_1 (\tilde{\kappa}_{i,i} + \tilde{\kappa}_{i,j}) + 2\sqrt{d} \zeta_2 \geq \tilde{\kappa}_{i,j}^2 \right\} \cap \Omega_\tau,\end{aligned}$$

where we define the proxy separation distance $\tilde{\kappa}_{i,j}$ as follows

$$\tilde{\kappa}_{i,j}^2 := \frac{\|\mu_i - \hat{\tau}_n \mu_j\|_2^2}{s_i^2 + \hat{\tau}_n^2 s_j^2}, \quad \forall (i,j) \in [n]^2, \quad \text{and} \quad \tilde{\kappa}_{\min} := \min_{i \neq j} \tilde{\kappa}_{i,j}.$$

On the event Ω_τ , the estimator $\hat{\tau}_n$ is close to $\tau^* = 1$, and thus $\tilde{\kappa}_{\min}$ is close to $\kappa(\mathbf{Z})$, hence its name. Therefore, the probability of the event Ω satisfies

$$\Omega \subseteq \left\{ \bigcup_{i=1}^n \bigcup_{j \neq i} \mathcal{A}_{i,j} \cap \Omega_\tau \right\} \cup \Omega_\tau^c, \quad (14)$$

where $\mathcal{A}_{i,j} := \left\{ \tilde{\kappa}_{i,i}^2 + 2\zeta_1(\tilde{\kappa}_{i,i} + \tilde{\kappa}_{i,j}) + 2\sqrt{d}\zeta_2 \geq \tilde{\kappa}_{i,j}^2 \right\}$. In order to upper bound the right-hand side of (14), we then show that $\mathcal{A}_{i,j} \cap \Omega_\tau$ for all distinct pairs (i,j) is a subset of an event, the probability of which is smaller than 2δ . We first establish the relation between $\tilde{\kappa}_{\min}$ and $\kappa(\mathbf{Z})$. For any $i \neq j$, the reverse triangle inequality gives $\|\mu_i - \hat{\tau}_n \mu_j\| \geq \|\mu_i - \mu_j\| - |1 - \hat{\tau}_n| \|\mu_j\| \geq \|\mu_i - \mu_j\| - \varepsilon_\tau \|\mu_j\|$, while $\sqrt{s_i^2 + \hat{\tau}_n^2 s_j^2} \leq (1 + \varepsilon_\tau) \sqrt{s_i^2 + s_j^2}$ on Ω_τ ; dividing, using $\sqrt{s_i^2 + s_j^2} \geq s_j$, and minimizing over $i \neq j$ yields

$$\tilde{\kappa}_{\min} \geq \frac{1}{1 + \varepsilon_\tau} \left(\kappa(\mathbf{Z}) - \varepsilon_\tau \max_{i \in [n]} \frac{\|\mu_i\|_2}{s_i} \right). \quad (15)$$

Consequently, given $\kappa(\mathbf{Z}) \geq (3 + 2\varepsilon_\tau) \varepsilon_\tau \max_{i \in [n]} \|\mu_i\|_2 / s_i$ we get $\tilde{\kappa}_{\min} \geq 2\varepsilon_\tau \max_{i \in [n]} \|\mu_i\|_2 / s_i$. On the other hand, we have the following upper bound on $\tilde{\kappa}_{i,i}$:

$$\tilde{\kappa}_{i,i} = \frac{|1 - \hat{\tau}_n| \|\mu_i\|_2}{s_i \sqrt{1 + \hat{\tau}_n^2}} \leq \frac{\varepsilon_\tau \|\mu_i\|_2}{s_i}.$$

Hence, we deduce that the following inequality holds for every pair $i \neq j$:

$$\max_{i \in [n]} \tilde{\kappa}_{i,i} \leq \varepsilon_\tau \max_{i \in [n]} \frac{\|\mu_i\|_2}{s_i} \leq \frac{1}{2} \tilde{\kappa}_{\min} \leq \frac{1}{2} \tilde{\kappa}_{i,j}. \quad (16)$$

Direct application of (16) yields

$$\mathcal{A}_{i,j} \cap \Omega_\tau \subseteq \left\{ 3\zeta_1 \tilde{\kappa}_{i,j} + 2\sqrt{d}\zeta_2 \geq \frac{3}{4} \tilde{\kappa}_{i,j}^2 \right\} \cap \Omega_\tau.$$

Notice also that on the event $\Omega_{\zeta_1} := \{\tilde{\kappa}_{\min} \geq 2\zeta_1\}$, the larger $\tilde{\kappa}_{i,j}$ the smaller the probability of the event from the previous display. Therefore, if we show that the probability of this event can be controlled by 2δ for $\tilde{\kappa}_{\min}$ it would follow that it holds for arbitrary distinct pairs $i \neq j$. Hence, we arrive at the following inclusion:

$$\Omega \cap \Omega_{\zeta_1} \subseteq \left[\bigcup_{i=1}^n \bigcup_{j \neq i} \mathcal{A}_{i,j} \cap \Omega_\tau \cap \Omega_{\zeta_1} \right] \cup \Omega_\tau^c \subseteq \left(\left\{ 3\zeta_1 \tilde{\kappa}_{\min} + 2\sqrt{d}\zeta_2 \geq \frac{3}{4} \tilde{\kappa}_{\min}^2 \right\} \cap \Omega_\tau \right) \cup \Omega_\tau^c.$$

Combining above inclusions we can bound the probability of the event Ω as follows:

$$\begin{aligned} \mathbf{P}(\Omega) &\leq \mathbf{P}(\Omega_{\zeta_1}^c) + \mathbf{P}(\Omega_\tau^c) + \mathbf{P}\left(\left\{ 3\zeta_1 \tilde{\kappa}_{\min} + 2\sqrt{d}\zeta_2 \geq \frac{3}{4} \tilde{\kappa}_{\min}^2 \right\} \cap \Omega_\tau\right) \\ &\leq \mathbf{P}(\zeta_1 \geq \tilde{\kappa}_{\min}/2) + \mathbf{P}(\Omega_\tau^c) + \mathbf{P}(\zeta_1 \geq \tilde{\kappa}_{\min}/8) + \mathbf{P}\left(\left\{ \zeta_2 \sqrt{d} \geq \frac{3}{16} \tilde{\kappa}_{\min}^2 \right\} \cap \Omega_\tau\right) \\ &\leq 2\mathbf{P}(\zeta_1 \geq \tilde{\kappa}_{\min}/8) + \mathbf{P}\left(\left\{ \zeta_2 \sqrt{d} \geq \frac{3}{16} \tilde{\kappa}_{\min}^2 \right\} \cap \Omega_\tau\right) + \delta, \end{aligned} \quad (17)$$

where in the last step we used the result of Theorem 1 to bound the probability of the event Ω_τ^c . Now, using Lemma 3 and Lemma 4, we find the threshold for $\tilde{\kappa}_{\min}$ such that both probabilities from (17) are bounded by δ . To keep the expressions concise, we notice that the dominating terms from Lemmas 3 and 4 do not scale with ε_τ . Hence, we replace the terms that do

depend on ε_τ by the corresponding dominating terms. Formally, $\mathbf{P}(\zeta_1 \geq \tilde{\kappa}_{\min}/8) \leq \delta/2$ holds as soon as $\tilde{\kappa}_{\min}/8$ exceeds the bound of Lemma 3, for which it suffices that

$$\tilde{\kappa}_{\min} \geq 8\sqrt{2\log(24n^3/\delta)}, \quad \text{i.e.} \quad \tilde{\kappa}_{\min}^2 \geq 128\log(24n^3/\delta).$$

Similarly, $\mathbf{P}(\{\zeta_2\sqrt{d} \geq \frac{3}{16}\tilde{\kappa}_{\min}^2\} \cap \Omega_\tau) \leq \delta$ holds as soon as $\frac{3}{16}\tilde{\kappa}_{\min}^2$ exceeds the bound of Lemma 4, for which it suffices that

$$\frac{3\tilde{\kappa}_{\min}^2}{16} \geq 2\left(\sqrt{d\log(12n^3/\delta)} + \log(12n^3/\delta)\right), \quad \text{i.e.} \quad \tilde{\kappa}_{\min}^2 \geq \frac{32}{3}\left(\sqrt{d\log(12n^3/\delta)} + \log(12n^3/\delta)\right).$$

Combining these two bounds we get that as soon as $\tilde{\kappa}_{\min}^2 \geq \frac{32}{3}\sqrt{d\log(12n^3/\delta)} + 128\log(12n^3/\delta)$, then the first two terms of (17) are bounded by $\delta/2$ and δ , respectively. Now, using the condition $\tilde{\kappa}_{\min} \geq 2\varepsilon_\tau \max_i \|\mu_i\|_2/s_i$ from (16) along with the triangle inequality, we obtain that a sufficient condition

$$\tilde{\kappa}_{\min} \geq \sqrt{\frac{32}{3}}(d\log(12n^3/\delta))^{1/4} + 8\sqrt{2\log(24n^3/\delta)} + 2\varepsilon_\tau \max_{i \in [n]} \frac{\|\mu_i\|_2}{s_i}, \quad (18)$$

ensures that the probability of recovering the wrong permutation $\pi^* \equiv \text{id}$ is bounded by 3δ , i.e., $\mathbf{P}(\Omega) \leq 3\delta$. Recall the relation between $\tilde{\kappa}_{\min}$ and $\kappa(\mathbf{Z})$ from (15), then using (18) we arrive at the following condition on $\kappa(\mathbf{Z})$:

$$\kappa(\mathbf{Z}) \geq (1 + \varepsilon_\tau) \left(\sqrt{\frac{32}{3}}(d\log(12n^3/\delta))^{1/4} + 8\sqrt{2\log(24n^3/\delta)} \right) + (3 + 2\varepsilon_\tau) \varepsilon_\tau \max_{i \in [n]} \frac{\|\mu_i\|_2}{s_i}.$$

Using the fact that $\varepsilon_\tau \leq 1/2$ one can verify that the latter reduces to

$$\kappa(\mathbf{Z}) \geq 5(d\log(12n^3/\delta))^{1/4} + 17\sqrt{\log(24n^3/\delta)} + 4\varepsilon_\tau \max_{i \in [n]} \frac{\|\mu_i\|_2}{s_i}, \quad (19)$$

where we didn't make much effort to optimize the numerical constants. To further bound the last term from the previous display we use the result of Theorem 1, and apply the straightforward inequalities $\max_{i \in [n]} \|\mu_i\|_2 \leq \lambda \|\boldsymbol{\sigma}\|_2$ and $\min_{i \in [n]} s_i \geq \sigma_{\min} \sqrt{1 - 1/n}$. Namely, with probability at least $1 - \delta$, we have

$$\varepsilon_\tau \max_{i \in [n]} \frac{\|\mu_i\|_2}{s_i} \leq 12\sqrt{\frac{\alpha_\sigma^2 \log(4/\delta)}{\lambda^2 + d}} \cdot \frac{\lambda \|\boldsymbol{\sigma}\|_2}{\sigma_{\min} \sqrt{1 - 1/n}} = 12\sqrt{\frac{\rho_\sigma \lambda^2 \log(4/\delta)}{(1 - 1/n)(\lambda^2 + d)}} \leq 13\sqrt{\rho_\sigma \log(4/\delta)},$$

where in the last inequality we used $\lambda^2 \leq \lambda^2 + d$ and the assumption $n \geq 8$. Substituting back the upper bound from the previous display into (19) and using the union bound we get that under the required condition on $\kappa(\mathbf{Z})$ we have $\mathbf{P}(\Omega) \leq 4\delta$, concluding the proof. $\square$

B PROOFS OF THE TECHNICAL LEMMAS

B.1 PROOF OF LEMMA 3

Define $g_i = u_{i,j}^\top \eta_i / \omega_i$ and $g'_j = u_{i,j}^\top \eta'_j / \omega_j$. Since $u_{i,j}$ is a unit vector and $\eta_i, \eta'_i \sim N(0, \omega_i^2 I_d)$, the variables $g_i, g'_i \sim N(0, 1)$ are independent scalar Gaussians for all $i \in [n]$. For a fixed pair (i, j) , we define

$$f_{ij}(a) := \frac{\omega_i g_i - a \omega_j g'_j}{\sqrt{\omega_i^2 + a^2 \omega_j^2}} = \frac{u_{i,j}^\top \eta_i - a u_{i,j}^\top \eta'_j}{\sqrt{\omega_i^2 + a^2 \omega_j^2}}.$$

Since g_i and g'_j are independent, for any fixed $a > 0$ we obtain $f_{ij}(a) \sim N(0, 1)$. Thus, we need to bound $\Phi_1 = \sup_{a \in [1-\varepsilon, 1+\varepsilon]} \max_{i,j} |f_{ij}(a)|$. In order to bound this supremum, we construct a net with radius r on the interval $[1-\varepsilon, 1+\varepsilon]$. We first bound the derivative of $f_{ij}(a)$ with respect to a . A direct calculation yields:

$$f'_{ij}(a) = -\frac{\omega_i^2 \omega_j g'_j + a \omega_j^2 \omega_i g_i}{(\omega_i^2 + a^2 \omega_j^2)^{3/2}}.$$

The absolute value of the latter satisfies

$$|f'_{ij}(a)| \leq \frac{\omega_i^2 \omega_j |g'_j| + a \omega_j^2 \omega_i |g_i|}{(\omega_i^2 + a^2 \omega_j^2)^{3/2}} \leq \frac{\omega_i \omega_j (\omega_i + a \omega_j)}{(\omega_i^2 + a^2 \omega_j^2)^{3/2}} \times M_g,$$

where $M_g := \max \{|g_1|, |g'_1|, \dots, |g_n|, |g'_n|\}$. Applying Cauchy-Schwartz we obtain

$$\frac{\omega_i \omega_j (\omega_i + a \omega_j)}{(\omega_i^2 + a^2 \omega_j^2)^{3/2}} \leq \frac{1}{\sqrt{2}a}.$$

Since $a \in [1 - \varepsilon, 1 + \varepsilon]$ and $\varepsilon < 1/2$, we have $a > 1/2$. Thus, for all $i, j \in [n]$ and $a \in [1 - \varepsilon, 1 + \varepsilon]$

$$\sup_{a \in [1 - \varepsilon, 1 + \varepsilon]} |f'_{ij}(a)| \leq \sqrt{2} M_g.$$

On the other hand, M_g is defined as the maximum (of absolute values) of standard Gaussian random variables. Therefore, union bound and standard tail estimates imply that with probability $1 - \delta/2$:

$$M_g \leq \sqrt{2 \log(8n/\delta)}.$$

Thus, the following holds with probability $1 - \delta/2$:

$$\sup_{a \in [1 - \varepsilon, 1 + \varepsilon]} |f'_{ij}(a)| \leq 2\sqrt{\log(8n/\delta)}. \quad (20)$$

Let $\mathcal{N}$ be an r -net of the interval $[1 - \varepsilon, 1 + \varepsilon]$ with spacing $r = \frac{\varepsilon}{n}$. The size of the net is $|\mathcal{N}| \leq \lceil \frac{2\varepsilon}{r} \rceil \leq 2n + 1 \leq 3n$. For a fixed value of a , the total number of random variables $f_{ij}(a)$ on the grid is $N_{tot} = n^2 |\mathcal{N}| \leq 3n^3$. Applying the union bound over all pairs (i, j) and grid points $a_k \in \mathcal{N}$:

$$\mathbf{P} \left(\max_{i,j,k} |f_{ij}(a_k)| > t \right) \leq 2N_{tot} e^{-t^2/2}.$$

The latter implies that with probability $1 - \delta/2$:

$$\max_{i,j,a_k \in \mathcal{N}} |f_{ij}(a_k)| \leq \sqrt{2 \log(12n^3/\delta)}. \quad (21)$$

Combining (20), (21) and mean value theorem:

$$\Phi_1 \leq \max_{i,j,k} |f_{ij}(a_k)| + r \sup_{i,j,a} |f'_{ij}(a)|.$$

Plugging in $r = \frac{\varepsilon}{n}$, we then obtain that with probability at least $1 - \delta$:

$$\Phi_1 \leq \sqrt{2 \log(12n^3/\delta)} + \frac{2\varepsilon}{n} \sqrt{\log(8n/\delta)}.$$

This concludes the proof.

B.2 PROOF OF LEMMA 4

Proof. The proof follows a similar structure to Lemma 3. We define the normalized random vectors $g_i = \eta_i/\omega_i$, $g'_j = \eta'_j/\omega_j$ and

$$f_{ij}(a) = \frac{\omega_i g_i - a \omega_j g'_j}{\sqrt{\omega_i^2 + a^2 \omega_j^2}}.$$

For any fixed a , $f_{ij}(a)$ is a linear combination of independent Gaussians normalized to have unit variance; thus $f_{ij}(a) \sim N(0, I_d)$. Our objective now is to upper bound the quantity $Z = \sup_{a \in [1 - \varepsilon, 1 + \varepsilon]} \max_{i,j} \left| \|f_{ij}(a)\|^2 - d \right|$. Again, as done in

the proof of Lemma 3, we will bound $|\frac{d}{da}\|f_{ij}(a)\|^2|$ uniformly over $[1 - \varepsilon, 1 + \varepsilon]$ and $\|f_{ij}(a)\|^2$ on fixed ε -net of points then employ a Lipschitz bound style argument.

Let us now bound the derivative of $\|f_{ij}(a)\|^2$ with respect to a . Using the chain rule we obtain

$$\left| \frac{d}{da} \|f_{ij}(a)\|^2 \right| = 2 |f_{ij}(a)^T f'_{ij}(a)| \leq 2 \|f_{ij}(a)\| \|f'_{ij}(a)\|.$$

We now give uniform upper bounds for $a \in [1 - \varepsilon, 1 + \varepsilon]$ for each of the norms on the right-hand side. Using the triangle inequality and Cauchy-Schwarz:

$$\|f_{ij}(a)\| \leq \frac{\omega_i \|g_i\| + a\omega_j \|g'_j\|}{\sqrt{\omega_i^2 + a^2\omega_j^2}} \leq \frac{\omega_i + a\omega_j}{\sqrt{\omega_i^2 + a^2\omega_j^2}} M_g \leq \sqrt{2} M_g,$$

where $M_g = \max_k \{\max(\|g_k\|, \|g'_k\|)\}$. Similarly,

$$\|f'_{ij}(a)\| = \left\| \frac{-(\omega_i^2 \omega_j g'_j + a\omega_j^2 \omega_i g_i)}{(\omega_i^2 + a^2\omega_j^2)^{3/2}} \right\| \leq M_g \frac{\omega_i \omega_j (\omega_i + a\omega_j)}{(\omega_i^2 + a^2\omega_j^2)^{3/2}}.$$

As already stated in the proof of the previous lemma:

$$\frac{\omega_i \omega_j (\omega_i + a\omega_j)}{(\omega_i^2 + a^2\omega_j^2)^{3/2}} \leq \frac{1}{\sqrt{2}a}.$$

Since $\varepsilon < 1/2$, $a \geq 1/2$, implying $1/(\sqrt{2}a) \leq \sqrt{2}$. Therefore, $\|f'_{ij}(a)\| \leq \sqrt{2} M_g$. Combining these results, the derivative of the squared norm is bounded by:

$$\left| \frac{d}{da} \|f_{ij}(a)\|^2 \right| \leq 4M_g^2.$$

Next, we bound M_g^2 . Recall, that each g_k and g'_k are standard Gaussians, thus their norms are distributed as χ_d^2 . Using Lemma 1 and taking a union bound over $2n$ variables we get that with probability $1 - \delta/2$:

$$M_g^2 \leq d + 2\sqrt{d \log(4n/\delta)} + 2 \log(4n/\delta).$$

Finally, we apply an ε -net argument. Let $\mathcal{N}$ be an r -net of $[1 - \varepsilon, 1 + \varepsilon]$ with $r = \varepsilon/n$. The net size is bounded by $3n$, resulting in $N_{tot} \leq 3n^3$ points to check for all pairs (i, j) . For any fixed $a_k \in \mathcal{N}$, $\|f_{ij}(a_k)\|^2 \sim \chi_d^2$. Using the two-sided concentration bound $P(|X - d| \geq 2\sqrt{dx} + 2x) \leq 2e^{-x}$ from Lemma 1 and applying the union bound over N_{tot} points with failure probability $\delta/2$:

$$\max_{i,j,k} |\|f_{ij}(a_k)\|^2 - d| \leq 2\sqrt{d \log(12n^3/\delta)} + 2 \log(12n^3/\delta).$$

Combining the discretization error and the Lipschitz bound:

$$Z \leq \max_{i,j,k} |\|f_{ij}(a_k)\|^2 - d| + r \sup_{i,j,a} \left| \frac{d}{da} \|f_{ij}(a)\|^2 \right|.$$

Substituting $r = \varepsilon/n$ and the derived bounds, with probability at least $1 - \delta$:

$$Z \leq 2\sqrt{d \log(12n^3/\delta)} + 2 \log(12n^3/\delta) + \frac{4\varepsilon}{n} \left(d + 2\sqrt{d \log(4n/\delta)} + 2 \log(4n/\delta) \right).$$

□

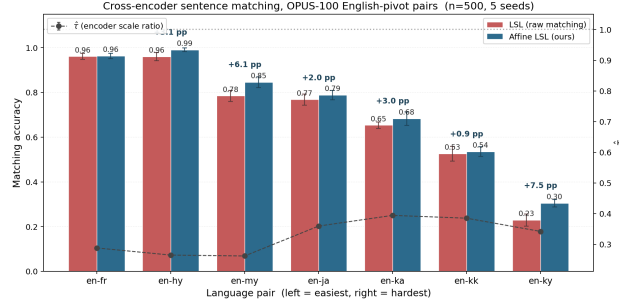


Figure 5: Accuracies of vanilla LSL and Affine LSL (ours) on sentence matching task. For each pair we match 500 English sentences to their translations in 7 pivot languages. As one can see, our approach outperforms vanilla LSL on all, especially harder pairs. Also overlaid is predicted $\hat{r}$ for each pair. All experiments are repeated and averaged across 5 independent trials.

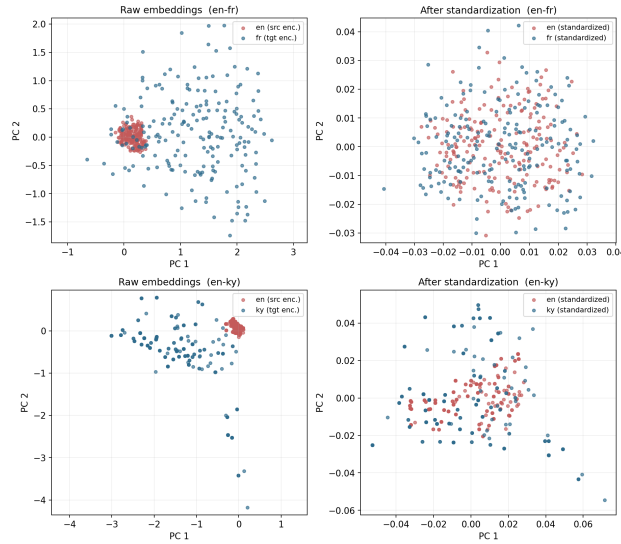


Figure 6: PCA visualization of OPUS-100 en-fr (easy) and en-ky (hard) sentence embeddings. Note that here standardization changes the geometry of the space to better suit the matching problem down the line.

C EXPERIMENTS ON REAL-WORLD TRANSLATION DATA

We validate our method on cross-encoder sentence matching task. This and similar tasks arise in real-world data pipelines, where embeddings or features from different models are compared or analyzed, e.g. multi-vendor aggregation in RAG systems, asymmetric query/document encoders, etc. Concretely, our goal is to match English sentences from OPUS-100 parallel corpora [Zhang et al., 2020] to their corresponding translations across 7 languages (fr, hy, my, kk, ka, ky, ja). Note that the ground-truth is available, as OPUS-100 is sentence-level parallel. English-side sentences are encoded using *all-MiniLM-L6-v2* [Reimers and Gurevych, 2019], while for target-language side we use completely different *paraphrase-multilingual-MiniLM-L12-v2* encoder [Reimers and Gurevych, 2020]. Results are shown in Figure 5.

Two encoders consistently differ in centered scale by a factor of ≈ 3 , with $\hat{r}$ tightly concentrated across seeds in line with the rate predicted by Theorem 1. On well-resourced pairs (en-fr) both methods recover the matching with $\approx 96\%$ accuracy and the gap is negligible, directly illustrating our *standardization cost* caveat. On low-resource pairs, where the multilingual encoder produces noisier embeddings, naive LSL accuracy degrades to as low as 23% (en-ky) while Affine LSL maintains a strict advantage in every pair.

Figure 6 makes this geometric: the raw embedding clouds occupy visibly different regions of PC space, while standardization collapses them onto a shared geometry.