NOTES ON ALGORITHM DESIGN FOR PROBABILISTIC CROSS-MATCHING

SAMMY KHALIFE

Abstract. The problem of *cross-matching* naturally intervenes in modern astronomy to identify different observations of the same celestial objects [2, 3, 5]. Despite its simplicity, finding the most relevant algorithm for a given distribution of observations remains challenging. Specifically, for a given a distribution, is there an algorithm (taking as input only the observations and oblivious to their distribution) maximizing the probability of producing the correct matching? These notes present formally the problem and compile partial answers, in particular for Gaussian distributions. We focus our attention in the case of two catalogues, with a family of algorithms solving the *assignment problem*, parameterized by the choice of a distance function. In the one-dimensional case, the choice of the distance function does not impact the solution. In higher dimensions, we prove the existence of a best configuration for the distance function under mild assumptions. We also provide lower bounds of success for all dimensions, and compare with numerical simulations.

1. Introduction

Suppose we are given distinct images of the same region of the sky using telescopes from different locations, each containing thousands or millions of stars. The process to group observations by the celestial object they represent is crucial to fully utilize these observations, namely to enhance the chances of discovering new types of sources [3]. Two main reasons make this problem challenging from a computational perspective. First, as the number of observations grows, there is exponential number of pairings – or groupings in the case of several telescopes – that are admissible. Second, the uncertainty from that arises from the measurements (change of conditions, calibration, ...) may impact differently the probability of success to find the right groups from an algorithm to another. We are then confronted with the choice of "the" right algorithm to choose given some minimal assumptions on the data distribution. In these notes we present some observations and partial results about towards an answer to the problem of algorithm design. For the sake of exposition, we consider that the observations can be modelled as

$$x = \mu + \epsilon$$

where $\mu \in \mathbb{R}^n$ represent the "true" positions of the celestial objects, and $\epsilon \sim \mathcal{N}(0, \Sigma)$ is some centered, gaussian noise. Suppose we are given other set of observations obtained from another set of measurements, but without labelling, i.e:

$$y = \tau^* \mu + \epsilon'$$

where τ^* is an unknown permutation matrix, and $\epsilon' \sim \mathcal{N}(0, \Sigma)$, where ϵ' is independent from ϵ. The problem of cross-matching is to devise an algorithm (as efficient

Johns Hopkins University, Department of Applied Mathematics and Statistics

as possible) to make a reasonable guess on τ^* only based on observations $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$.

To further simplify the presentation, we here suppose that $\Sigma = \mathsf{Id}$. Interpreting μ and τ as *parameters*, a first approach consists in maximizing the likelihood over these parameters. Given the assumptions on the observations, the log-likelihood is proportional to $\mathcal{L}(x, y, \tau, \mu) = \sum_{i=1}^{n} \|x_i - \mu_i\|^2 + \|y_{\tau(i)} - \mu_i\|^2$. Hence one aims to solve

$$\max_{\tau,\mu} \mathcal{L}(x, y, \tau, \mu) = \max_{\tau} \left(\max_{\mu} \mathcal{L}(x, y, \tau, \mu) \right)$$

One can maximize each term of $\mathcal{L}(x, y, \tau, \mu)$ to get

$$\arg\max_{\mu} \left(\|x_i - \mu_i\|^2 + \|y_{\tau^*(i)} - \mu_i\|^2 \right) = \frac{x_i + y_{\tau^*(i)}}{2}$$

Yielding

$$\begin{aligned} \arg\max_{\tau,\mu} \mathcal{L}(x, y, \tau, \mu) &= \arg\max_{\tau} \sum_{i=1}^{n} \|x_i - (\frac{x_i + y_{\tau^*(i)}}{2})\|^2 + \|y_{\tau^*(i)} - (\frac{x_i + y_{\tau^*(i)}}{2})\|^2 \\ &= \arg\max_{\tau} \sum_{i=1}^{n} \|x_i - y_{\tau^*(i)}\|^2 \end{aligned}$$

The solution of the above optimization problem can then be interpreted as the *most probable permutation* given the empirical observations (and the assumptions made on their distribution). This search can be formulated as solution Integer Program (IP):

$$\min_{X \in \mathcal{U}} (\mathsf{Tr}(W^t X)) \tag{1}$$

where $\mathcal{U}$ is the set of matrices of size $n \times n$ with entries in $\{0, 1\}$ such that every row and every column has at most one non zero entry, and W is the matrix of distances. Problem 1 can be relaxed (i.e. consider real values in $[0, 1]$ for the coefficients of X), and the solutions would still be integral because the constraints are unimodular, and solving the assignment problem can be performed in $O(n^3)$, with the Hungarian algorithm.

Such formulation can be extended to several catalogues of observations [5], where now the MLE aims to maximize the probability over partitions of observations into disjoint subsets. A subset with multiple elements is interpreted as the hypothesis that these elements in the subset are observations of the same object in the sky. Namely, a subset containing a single element is to be thought of as the only observation amongst all the catalogs for that object. Despite desirable theoretical properties of the MLE estimator (e.g. efficiency, and consistency, ...), we see two at least two reasons that could be argued against its dominance over other algorithms from the standpoint of obtained a correct matching. First, the problem of cross matching is performed for a fixed set of catalogues: the number of observations per celestial object is low, and all the aforementioned properties of the MLE are asymptotic. Second, for a fixed number of observations, some algorithms could have better probability of success (even among parametrized algorithms of similar complexity). Depending on the distribution priors, some algorithms may be more desirable than the MLE[1].

[1]Our investigation builds upon numerical experiments following the work exposed in [2, 3].

From a more concrete standpoint, suppose we start from formulation 1. One may choose to replace the weights by $W_{ij} = |x_i - y_j|^k$ where k is a positive integer (or more generally $W_{ij} = J(x_i - y_j)$ where J is a reasonable function, e.g.. convex and symmetric). Does the choice of k (resp. J) have an impact on the quality of the solution, i.e. the probability that the obtained permutation is τ^*? If it does, which value of k (resp. J) should one prefer?

2. Example with one-dimensional observations

To illustrate the problem, consider the simple setup where the observations $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ are reals lying in a given interval $I \subseteq \mathbb{R}$. We suppose that each x_i is an observation of a random variable of mean μ_i and y_i of mean $\mu_{\tau^*(i)}$. The cross-matching problems consists in estimating τ^* from these observations, to pair the observations of the same object together. The design of the algorithm only has only partial information about their distribution (e.g., we suppose we know the observations are gaussians, but the algorithm does not have access to the means μ_i's).

A possible formulation to estimate τ^* is the following surrogate problem

$$\min_{\tau \in \mathfrak{S}_n} \sum_{i=1}^{n} J(x_i - y_{\tau(i)}) \tag{2}$$

where $J : \mathbb{R} \to \mathbb{R}$ is a reasonable function, namely convex and centered ($J(0) = 0$). As mentioned in the previous section, note that for $J : x \mapsto x^2$, this formulation is equivalent to maximizing the log-likelihood for measurement noise that are i.i.d normal gaussians, and Problem 2 corresponds to the change to $W_{ij} = J(x_i - y_j)$ in Problem 1. In the case of one-dimensional observations, it turns out the solution of Problem 2 can be obtained simply by sorting the lists $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$, and matching them by rank. In particular, the choice of the "parameter" J has no impact on the solution: any instance gives the same solution for every choice of J, the one returned by the ranking algorithm. This follows from a simple generalization of the rearrangement inequality that we state below.

Proposition 1. *If $J : \mathbb{R} \to \mathbb{R}^+$ is convex and $J(0) = 0$ then solving Problem 2 matching is the same as sorting the observations and matching them by their rank. In these conditions, the choice of a function J verifying these assumptions has no impact on the solution(s) obtained by solving Problem 2, and can be obtained in $O(n \log n)$.*

Lemma 1. *Let $J : \mathbb{R} \to \mathbb{R}$ be a convex function such that $J(0) = 0$. Then J is superlinear, i.e for any $(x, y) \in (\mathbb{R}_{>0})^2, J(x + y) \geq J(x) + J(y)$.*

Proof. Since $J(0) = 0$, by convexity we have.

$$\forall t \in [0, 1], \forall x \in \mathbb{R}, J(tx) \leq tJ(x) \tag{3}$$

Now, let x and y be positive reals.

$$\begin{aligned} J(x) + J(y) &= J(\frac{x}{x+y}(x+y)) + J(\frac{y}{x+y}(x+y)) \\ &\leq \frac{x}{x+y} J(x+y) + \frac{y}{x+y} J(x+y) \\ &\leq J(x+y) \end{aligned}$$

where the second inequality follows from Eq. 3. □

Lemma 2 (Generalization of the rearrangement inequality)**.** *Let $J : \mathbb{R} \to \mathbb{R}^+$ be a convex function such that $J(0) = 0$. Let $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ be ordered reals, i.e. $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$. Then for any permutation $\tau \in \mathfrak{S}(n)$:*

$$\sum_{i=1}^{n} J(x_i - y_i) \leq \sum_{i=1}^{n} J(x_i - y_{\tau(i)})$$

Proof. Let $J : \mathbb{R} \to \mathbb{R}^+$ be a convex function such that $J(0) = 0$. For any positive integer n, and given reals $x_1, \cdots, x_n, y_1, \cdots, y_n$ and permutation $\tau \in \mathfrak{S}(n)$, let $S_n(\tau) := \sum_{i=1}^{n} J(x_i - y_{\tau(i)})$. We give a proof by induction on n.

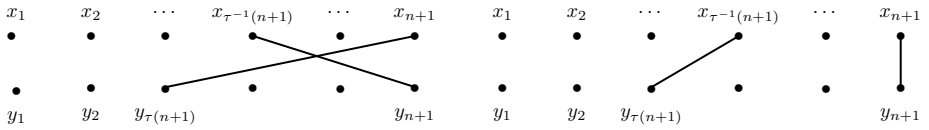
Base case: $(n = 1)$. The property is verified as $J(x_1 - y_1) \leq J(x_1 - y_1)$.

Induction step: Suppose that at rank $n \geq 1$, for any ordered reals $x_1 \leq \cdots \leq x_n \leq x_{n+1}$ and $y_1 \leq y_2 \leq \cdots \leq y_n \leq y_{n+1}$, for any $\tau \in \mathfrak{S}(n)$, $S_n(\mathsf{Id}) \leq S_n(\tau)$. We will prove by contradiction that for every $\tau \in \mathfrak{S}_{n+1}, S_{n+1}(\mathsf{Id}) \leq S_{n+1}(\tau)$. Suppose now that there exists $\tau \in \mathfrak{S}_{n+1}$ such that $S_{n+1}(\mathsf{Id}) > S_{n+1}(\tau)$. We then consider the two following cases:

- if $\tau(n+1) = n+1$ then we have $S_n(\tau) < S_n(\mathsf{Id})$ which is incompatible with the induction hypothesis.

- if $\tau(n+1) \neq n+1$. Then let $d = x_{n+1}$, $b = y_{n+1}$, $c = y_{\tau(n+1)}$, $a = x_{\tau^{-1}(n+1)}$ Using Lemma 1 $J(a-b) = J(a-c+c-b) \geq J(a-c) + J(c-b)$, and $J(d-c) \geq J(c-b) + J(d-c)$. Summing both inequalities and using the nonnegativity of J gives:

$$\begin{aligned} J(a-b) + J(d-c) &\geq J(a-c) + J(c-b) + J(d-c) \\ &\geq J(a-c) + J(c-b) + J(d-b) + J(b-c) \\ &\geq J(d-b) + J(a-c) \end{aligned} \tag{4}$$



(A) Permutation τ (B) Permutation τ' with better cost

Since $\tau(n+1) \neq n+1$, then $n+1 \neq \tau^{-1}(n+1)$ (τ^{-1} is injective). Let $\tau' \in \mathfrak{S}_{n+1}$ such that $\tau'(i) = \tau(i)$ for $i \in \{1, \cdots, n\} - \{\tau^{-1}(n+1)\}$, $\tau'(n+1) = n+1$ and $\tau'(\tau^{-1}(n+1)) = \tau(n+1)$.

Inequality 4 gives

$$J(x_{\tau^{-1}(n+1)} - y_{\underbrace{\tau(n+1)}_{\tau'(\tau^{-1}(n+1))}}) + J(x_{n+1} - y_{\underbrace{n+1}_{\tau'(n+1)}}) \leq J(x_{\tau^{-1}(n+1)} - y_{n+1}) + J(x_{\tau^{-1}(n+1)} - y_{n+1})$$

As τ and τ' are equal on $\{1, \cdots, n\} - \{\tau^{-1}(n+1)\}$

$$S_{n+1}(\tau') \leq S_{n+1}(\tau) < S_{n+1}(\mathsf{Id})$$

But since $\tau'(n+1) = n+1$, we would obtain $S_n(\tau') < S_n(\mathsf{Id})$ after substracting $J(x_{n+1} - y_{n+1})$, which is a contradiction with the induction hypothesis. □

Remark 1. *This result can be extended to observations of points on the one-dimensional sphere instead on the real line, by considering angular positions. An alternate proof can be given as follows, using simple properties of strictly convex sets.*

Without loss of generality, suppose that the true directions of observations (expected directions) can be matched with the identity, i.e. $\tau^* = \mathsf{Id}$. *The event that* $\tau = \mathsf{Id}$ *can be written as:*

$$\psi_{11}^2 + \psi_{22}^2 \leq \psi_{12}^2 + \psi_{21}^2$$

First note that for observation on the 1D sphere, $\psi_{21} = \psi_{11} - \psi_{12} + \psi_{22}$ *(relation between angles of observations).*

Let us consider the set Ω_k *defined as*

$$\Omega_k = \{(x, y, z) \in \mathbb{R}^3 : |x|^k + |y|^k \leq |z|^k + |x + y - z|^k\}$$

We prove the following property on the sets Ω_k*:* $\forall k > 1$, $\Omega_k = \Omega_2$.

Let us have a look at the border of Ω_k*, i.e*

$$\partial\Omega_k = \{(x, y, z) \in \mathbb{R}^3 : |x|^k + |y|^k = |z|^k + |x + y - z|^k\}$$

Then description of Ω_k *can be re-parametrized: Let* x, y *be fixed. Then we consider the change of variable:* $z \mapsto \eta = (x - z)$ *so that:*

$$\partial\Omega_k = \{(x, y, \eta) \in \mathbb{R}^3 : \|(x, y)\|_k = \|(x - \eta, y + \eta)\|_k\}$$

Now again, let (x, y) *be fixed in* $\mathbb{R}^2$*. Then the equation* $\|(x, y)\|_k = \|(x - \eta, y + \eta)\|_k$ *can be interpreted as the intersection between the* k*-ball of radius* $\|(x, y)\|_k$ *and the line directed by the vector* $(-1, 1)$*. When* $k > 1$*, the* k*-ball is strictly convex and this problem has only two solutions given by* $\eta = 0$ *and* $\eta = y - x$*. This shows that all the sets* Ω_k *are equal when* $k > 1$.

The previous analysis does not extend to the multidmensional case ($d > 1$). We will demonstrate this using the toy example of two observations by catalogue. Suppose, without loss of generality for the analysis, that $\tau^* = \mathsf{Id}$. In these conditions, the event that the permutation τ obtained by an algorithm solving Problem 1 is

$$(x, y) \in \Omega_{\|.\|} \iff \|x_1 - y_1\| + \|x_2 - y_2\| \leq \|x_1 - y_2\| + \|x_2 - y_1\| \tag{5}$$

Hence, $\mathbb{P}_{\mathcal{D}}(\Omega_{\|.\|})$ represents the probability of recovering the true permutation, when the observations follow a the distribution $\mathcal{D}$. Motivated by a pertinent choice of a distance function for the algorithm, this suggests to consider

$$\sup_{p \in \mathbb{N}} \mathbb{P}_{\mathcal{D}}(\Omega_{\|.\|_p}) \tag{6}$$

We observe that Problem 6 has a solution among p-norms (including ∞-norm). This can be seen by setting $u_p := \mathbb{P}_{\mathcal{D}}(\Omega_{\|.\|_p})$. u_p is a bounded sequence, so it either has a maximum, or has a subsequence that converges to its supremum. In the latter case, this subsequence is of the form $\|.\|_{\phi(p)}$ where ϕ is strictly increasing. As $\|.\|_p$ converges to $\|.\|_\infty$, the solution of Problem 6 is either some $\|.\|_{p_0}$ for a positive integer p_0, or it is the ∞-norm. In the next section we prove the existence of a solution among a more general class of function.

3. Existence of an optimal solution in the general case

We present in this section a result of existence of an optimal solution that maximizes the probability to recover the true permutation matrix in the case of *bounded support distributions*. In the following, we adopt the following notations:

- $\mathcal{H}$ refers to the set of non-negative symmetric functions from $\mathbb{R}^d$ to $\mathbb{R}^+$ which are strictly increasing on $\mathbb{R}^+$.
- $\mathcal{L}_K$ the subset of $\mathcal{H}$ of functions with Lipschitz constant bounded by K.
- $\mathcal{U}_{K,L}$ the subset of functions of $\mathcal{L}_K$ with a uniform lower bound L on the increasing ratio.

In the following let ψ_{ij} refer to the vector of differences $x_i - y_j \in \mathbb{R}^d$. For a function $h \in \mathcal{H}$, let $\tau_{h(\psi)}$ be the solution of Problem 1 with $W_{ij} = h(\psi_{ij})$. In order to formulate the search of a good candidate h, we consider the following optimization problem

$$\sup_{h \in \mathcal{U}_{K,L}} \mathbb{P}_\psi(\tau_{h(\psi)} = \tau^*) \tag{7}$$

Proposition 2. *For all fixed $K \geq 0$ and $L \geq 0$, Problem 7 has a solution in the case of bounded distributions, i.e. $\exists h^* \in \mathcal{U}_{L,K}$*

$$\mathbb{P}_\psi(\tau_{h^*(\psi)} = \tau^*) = \max_{h \in \mathcal{U}_{K,L}} \mathbb{P}_\psi(\tau_{h(\psi)} = \tau^*) = \sup_{h \in \mathcal{U}_{K,L}} \mathbb{P}_\psi(\tau_{h(\psi)} = \tau^*)$$

Proof. The set $\mathcal{U}_{K,L}$ is compact as a consequence of the Arzelà-Ascoli Theorem. The event $\tau_{h(\psi)} = \tau^*$ can be described as:

$$\Omega_h = \bigcap_{\tau \in \mathfrak{S}_n} \{\psi \in \mathbb{R}^{n \times n \times d} : \sum_{i=1}^n h(\psi_{i\tau^*(i)}) \leq \sum_{i=1}^n h(\psi_{i\tau(i)})\} \tag{8}$$

Remark that it is sufficient to prove that the volume map $h \mapsto \int_{\Omega_h} d\psi$ is continuous, since for any distribution ϕ,

$$|\int_{\Omega_h} \phi d\psi - \int_{\Omega_{h'}} \phi d\psi| \leq \int_{\Omega_h - \Omega_{h'}} |\phi| d\psi + \int_{\Omega_{h'} - \Omega_h} |\phi| d\psi \leq \int_{\Omega_h \Delta \Omega_{h'}} d\psi$$

because by assumption the right hand side integral could be made arbitrarily small when h is close to h'. Let $\Omega_{h,\sigma}$ be defined as:

$$\Omega_{h,\tau} = \{\psi \in \mathbb{R}^{n \times n \times d} : \sum_{i=1}^n h(\psi_{i\tau^*(i)}) \leq \sum_{i=1}^n h(\psi_{i\tau(i)})\} \tag{9}$$

such that $\Omega_h = \bigcap_{\tau \in \mathfrak{S}_n} \Omega_{h,\tau}$. To prove that the volume of Ω_h is continuous, we will restrict the study to one $\Omega_{h,\sigma}$ using the following claim:

Claim 1. *If $h \mapsto \int_{\Omega_{h,\tau}} d\psi$ is continuous for any $\tau \in \mathfrak{S}_n$ then $h \mapsto \int_{\Omega_h} d\psi$ is continuous.*

Let $\sigma \in \mathfrak{S}_n$, we want to show that $h \mapsto (\int_{\Omega_{h,\tau}} d\psi = \mathsf{vol}(\Omega_{h,\tau}))$ is continuous. Let g be a function in $\mathcal{H} \cap \mathcal{L}_K$. It suffices to prove that $|\mathsf{vol}(\Omega_{h+\epsilon g,\tau}) - \mathsf{vol}(\Omega_{h,\tau})|$ can be made arbitrarily small if ϵ is small enough. Note that $|\mathsf{vol}(\Omega_{h+\epsilon g,\tau}) - \mathsf{vol}(\Omega_{h,\tau})| \leq \mathsf{vol}(\Omega_{h+\epsilon g,\tau} \Delta \Omega_{h,\tau})$, where Δ is the symmetric difference operator between two sets.

By symmetry, it is enough to look at the volume of $\Omega_{h+\epsilon\psi,\tau} - \Omega_{h,\tau}$, which is the set of points of $\mathbb{R}^{n\times n\times d}$ such that both of the following inequalities hold:

$$\sum_{i=1}^{n} h(\psi_{i\tau^*(i)}) + \epsilon g(\psi_{i\tau^*(i)}) \leq \sum_{i=1}^{n} h(\psi_{i\tau(i)}) + \epsilon g(\psi_{i\tau(i)})$$
$$\sum_{i=1}^{n} h(\psi_{i\tau^*(i)}) \geq \sum_{i=1}^{n} h(\psi_{i\tau(i)})$$

Therefore, it is included in the set of points ψ verifying:

$$0 \leq \sum_{i=1}^{n} h(\psi_{i\tau^*(i)}) - h(\psi_{i\tau(i)}) \leq \epsilon \left[\sum_{i=1}^{n} g(\psi_{i\tau(i)}) - g(\psi_{i\tau^*(i)})\right]$$

Since the variables ψ_{ij} are bounded, we can further limit the study to the set:

$$0 \leq \sum_{i=1}^{n} h(\psi_{i\tau^*(i)}) - h(\psi_{i\tau(i)}) \leq \epsilon$$

To simplify the notations, let $u_i = \psi_{i\tau^*(i)}$, $v_i = \psi_{i\tau(i)}$, and § be the support of the distribution of the ψ (S is bounded since the observations have bounded observations). Using this parametrization let $\mathcal{C}$ be the set defined as:

$$\mathcal{C} := \{(u, v) \in S^{n\times d\times 2} : 0 \leq \sum_{i=1}^{n} h(u_i) - h(v_i) \leq \epsilon\}$$

Now, let $\alpha_i(x_i, y_i) = h(u_i) - h(v_i)$ so that

$$\mathcal{C} := \{(u, v) \in S^{n\times d\times 2} : 0 \leq \sum_{i=1}^{n} \alpha(u_i, v_i) \leq \epsilon\}$$

We now claim that $\mathcal{H}^{n\times 2d}(\mathcal{C}) = \mathcal{O}(\epsilon)$. Let M be a real such that the 2-ball of radius M contains $\mathcal{C}$ and let $A := \{(\alpha_1, \cdots, \alpha_d) \in [0, M]^d : 0 \leq \sum_{i=1}^{d} \alpha_i \leq \epsilon\}$ then clearly $\mathcal{H}^d(A) \leq \epsilon\, C(M, d)$ where $C(M, d)$ is a function of M and d only.
Now, let $\mathcal{B}^i_{\alpha_i} := \{(u_i, v_i) \in S^2 : h(u_i) - h(v_i) = \alpha_i\}$. By assumption on h, $\mathcal{H}^1(\mathcal{B}^i_{\alpha_i}) \leq \mathcal{H}^1(S)$. Hence the counter-image:

$$\mathcal{C} = \bigcup_{(\alpha_1,\cdots\alpha_d)\in A} \mathcal{B}^1_{\alpha_1} \times \cdots \times \mathcal{B}^d_{\alpha_d}$$

verifies:

$$\mathcal{H}^{2d}(\mathcal{C}) \leq \mathcal{H}^d(A).\mathcal{H}^1(\mathcal{B}^1_{\alpha_1}).\cdots.\mathcal{H}^1(\mathcal{B}^d_{\alpha_d})$$
$$\leq \epsilon\, C'(M, \mathcal{H}^1(S), d)$$

□

4. Correct matching: lower bounds in the Gaussian case

We provide here some guarantee of obtaining the correct matching solving Problem 1 as a function of p that parametrizes the choice of the p-norm. Recall that $W_{ij} = |x_i - y_j|^p$ is the matrix of weights in Formulation 2.

Notations. For $z \in \mathbb{R}_+$ and $n \in \mathbb{N}$, $z^{\bar{n}} := z(z+1)\cdots(z+n-1)$, with $z^{\bar{0}} := z$. erf is the function $\mathsf{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. For a matrix M, $\|M\|_\infty := \max_{ij} |M_{ij}|$ and $\|M\|_1 := \sum_{ij} |M_{ij}|$. We say that a $n \times n$ matrix M with is diagonally dominated provided for every $(i,j) \in [n]^2$, $M_{ii} \leq M_{ij}$.

Lemma 3. *Given the assumptions on the distribution of Section 1 with $\Sigma = \sigma \mathsf{Id}$,*

$$\mathbb{E}[W_{ij}] = \sum_{k=1}^{d} \frac{2^{p/2} \hat{\sigma}^p}{\sqrt{\pi}} \Gamma(\frac{p+1}{2}) \exp(-r_{ij}^{(k)}) \Phi(\frac{p+1}{2}, \frac{1}{2}, r_{ij}^k)$$

where

- Γ *is the gamma function*
- $\Phi : (a,b,z) \mapsto \sum_{n=0}^{+\infty} \frac{a^{\bar{n}}}{b^{\bar{n}} n!} z^n$ *the hypergeometric function.*
- $r_{ij}^{(k)} = \frac{(\mu_i^{(k)} - \mu_j^{(k)})^2}{2\hat{\sigma}^2}$
- $\hat{\sigma} = 2\sigma$

Proof. $\mathbb{E}[W_{ij}]$ is the p-th raw absolute moment of a gaussian of mean $\mu_i^{(k)} - \mu_j^{(k)}$ and variance $\hat{\sigma} = 2\sigma$. Similar manipulations as in [6, Page 3] can be performed to reach the claimed result. □

We now turn our attention to the Linear Program 1, and confirm that by replacing the matrices of distances between their expectation, the solution is given by τ^*. In the vocabulary of linear programming, this is equivalent to saying that the objective vector $\mathbb{E}[W]$ is in the normal cone at τ^*. We prove this by showing that such matrix is diagonally dominated.

Proposition 3. $\mathbb{E}[W]$ *is in the normal cone at τ^* for Problem 1.*

Proof. We can suppose without loss of generality for the analysis that $\tau^* = \mathsf{Id}$, as the following reasoning can be extended to any τ^* after composition in the search space by $(\tau^*)^{-1}$. We prove that $\mathbb{E}[W]$ is a diagonally dominated matrix, which is trivially in the cone at the identity.

To show that for every (i,j), $\mathbb{E}[W_{ii}] \leq \mathbb{E}[W_{ij}]$, let $\Delta_{ij} := \exp(-rij)\Phi(\frac{p+1}{2}, \frac{1}{2}, r_{ij})$. Then it is clear, using Lemma 3, that Δ is diagonally dominated iff $\mathbb{E}[W]$ is. To prove that Δ_{ij} is diagonally dominated, we next show that the function:

$$F_{a,b}(z) : z \mapsto exp(-z)\Phi(a,b,z)$$

is decreasing for some values of a and b. By definition, $F_{a,b}(.)$ is smooth and:

$$F'_{a,b}(z) = exp(-z) \sum_{n=0}^{\infty} V_n \frac{z^n}{n!}$$

where $V_n := \frac{a^{\overline{n+1}}}{b^{\overline{n+1}}} - \frac{a^{\bar{n}}}{b^{\bar{n}}}$. Hence

$$\begin{aligned} V_n \geq 0 &\iff a^{\overline{n+1}} - a^{\bar{n}}(b+n) \geq 0 \\ &\iff a^{\bar{n}}((a+n)-(b+n)) \geq 0 \\ &\iff a^{\bar{n}}(a-b) \geq 0 \end{aligned}$$

So for $a = \frac{p+1}{2}$ and $b = \frac{1}{2}$, $a - b \geq 0$ which proves that $V_n \geq 0$.

We supposed that $\tau^* = \mathsf{Id}$ hence $r_{ii} = 0 \leq r_{ij}$ for all $\{i, j\}$. Due to $F_{\frac{k+1}{2},\frac{1}{2}}$ being increasing, Δ is diagonally dominated. □

Lemma 4. *Let $r > 0$ and $\sigma > 0$. The function $f : \mathbb{R} \to \mathbb{R}$*

$$p \mapsto \frac{2^{p/2}\hat{\sigma}^p}{\sqrt{\pi}}\Gamma(\frac{p+1}{2})\exp(-r)\Phi(\frac{p+1}{2}, \frac{1}{2}, r)$$

is strictly increasing.

Theorem 1 ([4])**.** *Let d be a positive integer and let μ be a centered Gaussian measure on $\mathbb{R}^d$. Let $A \subseteq \mathbb{R}^d$ and $B \subseteq \mathbb{R}^d$ be two convex sets that are symmetric about the origin. Then*

$$\mu(A \cap B) \geq \mu(A)\mu(B)$$

The following proposition presents an example of lower-bound for the probability of obtaining the correct matching as a function of p. This bound is quite pessimistic, as it measures the probability of the distance matrix to stay in a cone of diagonally dominated matrices.

Proposition 4. *There exists a constant C depending only on the distance pairs $(|\mu_i - \mu_j\|_2)_{(i,j)\in[n]^2}$ such that for any $\sigma \leq C$ the probability that the solution of Problem 1 is the correct matching is bounded from below by $(1 - \frac{4\max_{i,j}\mathsf{Var}(W_{i,j})}{\min_{i\neq j}(\mathbb{E}[W_{ij}]-\mathbb{E}[W_{ii}])^2})^{n^2}$.*

Proof. Again, we can suppose w.l.o.g for the analysis. that $\tau^* = \mathsf{Id}$. For convenience, let $U_p := \frac{2^{p/2}\hat{\sigma}^p}{\sqrt{\pi}}$ and consider

$$\epsilon = \min_{p\in\mathbb{N}, i\neq j}(\mathbb{E}[W_{ij}] - \mathbb{E}[W_{ii}])/2$$

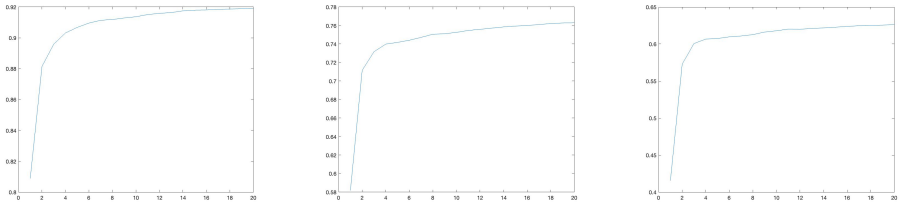
Recall that $W_{ij} = \|x_i - y_j\|_p^p$ and that $\mathbb{E}[W]$ is diagonally dominated: ϵ is chosen such that i) ϵ is positive and independent of p, ii) any perturbation (coefficient-wise) of $\mathbb{E}[W]$ by ϵ is still diagonally dominated. In other words $\max_{i,j}|W_{ij} - \mathbb{E}[W_{ij}]| \leq \epsilon \implies W \in \mathcal{N}(\mathsf{Id})$, giving in probability

$$\mathbb{P}(W \in \mathcal{N}(\mathsf{Id})) \geq \mathbb{P}(\max_{i,j}|W_{ij} - \mathbb{E}[W_{ij}]| \leq \epsilon)$$

We will now lower bound the right hand side by invoking the Gaussian Correlation Inequality (GCE) stated in Theorem 1. Each variable $W_{ij} - \mathbb{E}[W_{ij}]$ is a gaussian centered random variable, so the vector obtained by stacking all the n^2 coordinates still is, and the set $\{x \in \mathbb{R}^{n^2} : \text{for every } i,\ |x_i| \leq \epsilon\}$ is convex and symmetric about the origin. Using the GCE we obtain

$$\mathbb{P}(\max_{i,j}|W_{ij} - \mathbb{E}[W_{ij}]| \leq \epsilon) \geq \prod_{i,j\in[n]^2} p_{i,j,\epsilon}$$

where $p_{i,j,\epsilon} := \mathbb{P}(|W_{ij} - \mathbb{E}[W_{ij}]| \leq \epsilon) = 1 - \mathbb{P}(|W_{ij} - \mathbb{E}[W_{ij}]| \geq \epsilon)$. Chebyshev's inequality $p_{i,j,\epsilon} \geq 1 - \frac{\mathsf{Var}(W_{ij})}{\epsilon^2}$. In order to obtain the final inequality, one comment is



(A) $n = 5$ (B) $n = 10$ (C) $n = 15$

FIGURE 1. Monte-Carlo simulations (10000 samples, $\sigma = 0.2$, $\min_{i \neq j} \|\mu_i - \mu_j\|_2 = 7$) of empirical probability to recover the true permutation, as a function of the parameter p. The empirical probability of success increases fast as a function of k (and decreases fast as a function of n) as suggested by the lower bound.

in order: we can make $\mathsf{Var}(W_{ij})$ arbitrarily small when $\sigma \to 0$ so that $1 - \frac{\mathsf{Var}(W_{ij})}{\epsilon} > 0$ for all pair (i, j) and we can multiply all the inequalities together. This can be seen from an asymptotic development of ϕ for fixed parameters a and b, as stated in [1, Page 508]:

$$\Phi(a, b, z) \underset{z \to +\infty}{\sim} \Gamma(b) \left(\frac{e^z z^{a-b}}{\Gamma(a)} + \frac{(-z)^{-a}}{\Gamma(b-a)} \right)$$

Given

$$\begin{aligned} \mathsf{Var}(W_{ij}) &= \mathbb{E}[W_{ij}^2] - \mathbb{E}[W_{ij}]^2 \\ &= E[\psi_{ij}^{2p}] - E[\psi_{ij}^{p}]^2 \\ &= H(r_{ij}, 2p) - H(r_{ij}, p)^2 \\ &\leq H(r_{ij}, 2p) \\ &\leq \frac{2^p \hat{\sigma}^{2p}}{\sqrt{\pi}} \Gamma(\frac{2p+1}{2}) \sum_{i=1}^{d} \Phi(-p, \frac{1}{2}, -r_{ij}^{(k)}) \end{aligned}$$

it is now clear that $\mathsf{Var}(W_{ij}) \underset{\sigma \to 0}{\to} 0$ as $r_{ij} = \frac{(\mu_i(k) - \mu_j^{(k)})}{\hat{\sigma}^2}$ and $\hat{\sigma} = 2\sigma$. The last inequality is in fact an equality, corresponding to the equality holding for $c > -1$:

$$\exp(-c)\Phi(\frac{a+1}{2}, \frac{1}{2}, c) = \Phi(-\frac{a}{2}, \frac{1}{2}, -c)$$

.

□

The numerical simulations seem to suggest that for a regime of σ that is sufficiently small, the infinity norm dominates among the p-norms to recover the true permutation with largest probability. This is consistent with the increasing nature lower-bounds as a function of p.

References

[1] Abramowitz, M., Stegun, I.A.: Handbook of mathematical functions with formulas, graphs, and mathematical tables, vol. 55. US Government printing office (1948)
[2] Budavári, T., Basu, A.: Probabilistic cross-identification in crowded fields as an assignment problem. The Astronomical Journal **152**(4), 86 (2016)
[3] Budavári, T., Szalay, A.S.: Probabilistic cross-identification of astronomical sources. The Astrophysical Journal **679**(1), 301 (2008)
[4] Royen, T.: A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions. arXiv preprint arXiv:1408.1028 (2014)
[5] Shi, X., Budavári, T., Basu, A.: Probabilistic cross-identification of multiple catalogs in crowded fields. The Astrophysical Journal **870**(1), 51 (2019)
[6] Winkelbauer, A.: Moments and absolute moments of the normal distribution. arXiv preprint arXiv:1209.4340 (2012)