Discrete Probabilistic Inverse Optimal Transport

Wei-Ting Chiu [*1] **Pei Wang** [*1] **Patrick Shafto** [1 2]

Abstract

Inverse Optimal Transport (IOT) studies the problem of inferring the underlying cost that gives rise to an observation on coupling two probability measures. Couplings appear as the outcome of matching sets (e.g. dating) and moving distributions (e.g. transportation). Compared to Optimal transport (OT), the mathematical theory of IOT is undeveloped. We formalize and systematically analyze the properties of IOT using tools from the study of entropy-regularized OT. Theoretical contributions include characterization of the manifold of cross-ratio equivalent costs, the implications of model priors, and derivation of an MCMC sampler. Empirical contributions include visualizations of cross-ratio equivalent effect on basic examples, simulations validating theoretical results and experiments on real world data.

1. Introduction

Inverse Optimal Transport (IOT) involves inferring latent costs from stochastic couplings, a problem that arises across many domains where matching or movement of distributions arises (Ma et al., 2020; Stuart & Wolfram, 2020; Li et al., 2019; Paty & Cuturi, 2020). Unlike (Entropy-regularized) Optimal Transport (Villani, 2008; Cuturi, 2013), mathematical foundations and computational implications have not been systematically investigated. IOT inherits interesting geometric structure from Sinkhorn scaling (Sinkhorn, 1964), which gives rise to natural questions about the nature of posterior distributions, the existence of samplers, and feasibility of inferences from observed couplings to underlying costs.

IOT problems has been applied in many areas. Practical examples include: (1) inferring the cost criterion of international migration (Stuart & Wolfram, 2020); (2) estimating the marriage matching affinity (Dupuy et al., 2019; Ma et al., 2020); (3) learning interaction cost function from noisy and incomplete matching matrix and make prediction on new matching (Li et al., 2019). These works provide methodologies to solve *specific* IOT problems with extra hard constraints on latent costs. In this paper we provide a framework to analyze discrete probabilistic inverse optimal transport (PIOT) for *general* problems, and to understand the geometric properties of the space of latent costs.

Our contributions include: (1) formulating the PIOT problem and identify the support manifold for latent costs; (2) characterizing the geometry of the support manifold along with the implications of prior distributions on posterior inference; (3) analyzing the support manifold for noisy and incomplete observations. (4) developing MCMC algorithms for posterior inference which derives directly from geometric properties of Sinkhorn scaling; (5) demonstrating PIOT through simulations on general examples including visualizations of latent cost manifold and missing/noisy value predictions for migration data.

Notation: $\mathbb{R}^+$ denotes positive reals, $\mathbb{R}^*$ denotes non-negative reals. $\overline{\mathbb{R}^*}$ denotes the extended non-negative reals $\mathbb{R}^* \cup \{+\infty\}$. Δ_k is the k-dimensional simplex. For a space X, $\mathcal{P}(X)$ denotes the set of distributions over X. Matrices are in uppercase and their elements are in the corresponding lowercase. For a matrix $A_{m\times n}$, $\mathbf{a}_1, \ldots, \mathbf{a}_n$ denotes columns of A. Column normalizing a matrix $A_{m\times n}$ with respect to a n-dim row vector ν is defined as $\text{Col}(A, \nu) = A\,\text{diag}(\frac{\nu}{\mathbf{1}_m A})$, where $\mathbf{1}_m$ is a m-dim row vector with each element equals to 1. 'with respect to ν' is omitted when $\nu = \mathbf{1}_n$. Row normalization is defined in the same fashion.

2. Background and Related Work

Given two spaces $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, let $T = (t_{ij})_{m\times n} \in (\mathbb{R}^*)^{m\times n}$ be a *coupling* that records the co-occurrence distribution between X and Y. To avoid trivial cases, we assume $m, n \geq 2$. t_{ij} is allowed to be zero here. Each element t_{ij} represents the probability of observing x_i and y_j simultaneously. The interaction between X and Y can be captured by a *cost matrix* $C = (c_{ij})_{m\times n} \in (\overline{\mathbb{R}^*})^{m\times n}$, where c_{ij} measures the underlying cost of coupling x_i and y_j. Hence, the probabilistic

[*]Equal contribution [1]Department of Mathematics and Computer Science, Rutgers University Newark, NJ [2]School of Mathematics, Institute for Advanced Study (IAS), Princeton NJ. Correspondence to: Patrick Shafto <patrick.shafto@gmail.com>, Wei-Ting Chiu <wc570@rutgers.edu>, Pei Wang <peiwang425@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

inference on the underlying structure over X and Y is then to obtain $P(C|T)$, the distribution over all possible C given an observed coupling T.

Entropy regularized optimal transport is an fundamental building block in modelling $P(C|T)$. Given a cost matrix C and distributions μ, ν over X and Y respectively, *Entropy regularized optimal transport* (EOT) (Cuturi, 2013; Peyré et al., 2019) solves the optimal coupling T^λ that minimizes the entropy regularized cost of transfer X with distribution μ into Y with distribution ν. Thus for a parameter $\lambda > 0$:

$$T^\lambda = \text{argmin}_{T \in U(\mu,\nu)} \{\langle C, T\rangle - \frac{1}{\lambda} H(T)\}, \qquad (1)$$

where $U(\mu, \nu)$ is the set of all couplings between μ and ν (i.e. joint distributions with marginals μ and ν), $\langle C, T\rangle = \sum_{i\in X, j\in Y} c_{ij} t_{ij}$ is the inner product between C and T, and $H(T) := -\sum_{i=1}^m \sum_{j=1}^n t_{ij} \log t_{ij}$ is the *entropy* of T. T^λ is called an *optimal coupling* with parameter λ.

Sinkhorn scaling is used to efficiently compute optimal couplings. (μ, ν)-*Sinkhorn Scaling (SK)* (Sinkhorn, 1964) of a matrix M is the iterated alternation between row normalization of M with respect to μ and column normalization of M with respect to ν. (Cuturi, 2013) proved that: given a cost matrix C, the optimal coupling T^λ between μ and ν can be solved by applying (μ, ν)-Sinkhorn scaling on the *negative exponential cost matrix* K^λ, where $K^\lambda = e^{-\lambda C} = (e^{-\lambda \cdot c_{ij}})_{m\times n}$ (See Example A.1).

Given a coupling T, the inference on the cost matrix C can be formulated as an inverse entropy regularized optimal transport problem. Given λ, EOT is the map Φ:$(\overline{\mathbb{R}^*})^{m\times n} \times \mathcal{P}(\mathbb{R}^m) \times \mathcal{P}(\mathbb{R}^n) \to \mathcal{P}((\mathbb{R}^*)^{m\times n})$ with $\Phi(C, \mu, \nu) = T^\lambda$. Therefore **inverse entropy regularized optimal transport (IOT)** can be defined as $\Phi^{-1} : T^\lambda \to \{(C, \mu, \nu)\}$, where $\{(C, \mu, \nu)\}$ represents the preimage set of T^λ under Φ.

Remark 2.1. In IOT, when T^λ is observed without any noise, T^λ completely determines μ and ν. Indeed $(C, \mu, \nu) \in \Phi^{-1}(T^\lambda)$ implies that $T^\lambda \in U(\mu, \nu)$. Thus μ and ν must equal the row and column marginals of T^λ. Hence we may reduce IOT into $\Phi^{-1} : T^\lambda \to \{C\}$.

Further, note that $\Phi^{-1} : T^\lambda \to \{C\}$ is not well-defined as a function, since T^λ does not uniquely determine cost C. For instance, any two cost matrices differ by an additive constant output the same optimal coupling.

Remark 2.2. The inverse problem is constructed to entropy regularized OT rather than OT because: OT plan is general sparse (deterministic), and zero elements do not carry sufficient information to apply any meaningful inference on the cost. In particular, zero in the plan does not necessarily imply that the corresponding cost is infinity. Moreover, EOT takes into account of the uncertainty and incompleteness of observed data, and is faster (Courty et al., 2016). In addition, when λ goes to 0, EOT recovers OT.

Related Work. The forward OT problem has a long history and is an active field of research (Kantorovich, 1942; Villani, 2008; Peyré et al., 2019; Genevay et al., 2016). Applications in machine learning include: supervised learning (Frogner et al., 2015), Bayesian inference (El Moselhy & Marzouk, 2012), generative models (Salimans et al., 2018), Transfer learning (Courty et al., 2016), and NLP (Alaux et al., 2018).

However, the inverse OT problem is much less studied. The few existing inverse OT papers (Dupuy et al., 2019; Li et al., 2019; Stuart & Wolfram, 2020; Ma et al., 2020) to our best knowledge, all focused on obtaining a unique approximation of the 'ground truth' C^*, even though $\Phi^{-1}(T)$ contains a set of cost matrices. To output an estimate of C^*, existing approaches place constraints on the cost matrix. For example in Stuart & Wolfram (2020), C has to be either a Toeplitz matrix or determined by a given graph structure. In Li et al. (2019), C must be constructed from a metric over a space where both X and Y are able to embedded in, plus X and Y need to have the same dimensionality. In Dupuy et al. (2019), the affinity matrix has a low-rank constraint. In Ma et al. (2020), a regularization term on the cost is added to the objective to express the constraints such as symmetric costs. However, these assumptions do not hold in general.

Rather than adding another constraint, we aim to analyze the intrinsic properties and the underlying geometric structure of the entire set $\{C\} = \Phi^{-1}(T^\lambda)$ for the discrete case in the following sections. To simplify the notation, we will omit the superscript λ, and fix $\lambda = 1$ unless otherwise stated.

3. Probabilistic Inverse Optimal Transport

Definition 3.1. Let T be a noisy observation of the optimal coupling T^*. **Probabilistic inverse optimal transport** is defined as:

$$P(C|T) = \int_{T^* \in (\mathbb{R}^*)^{m\times n}} P(C|T^*) P(T^*|T) dT^*, \qquad (2)$$

where $P(C|T^*)$ is obtained through Bayes' Rule:

$$P(C|T^*) = \frac{P(T^*|C) P_0(C)}{P(T^*)}, \qquad (3)$$

$P_0(C)$ is the prior of cost C; Likelihood $P(T|C) = 1$ if $C \in \Phi^{-1}(T)$, otherwise $P(T|C) = 0$; $P(T^*) = \int P(T^*|C) P_0(C) \mathrm{d}C$ is the normalizing constant; $P(K|T)$ is used to denote the induced posterior on $K = e^{-C}$.

3.1. No observation noise

In this section, we study the case when $T^* = T$. As a direct application of Definition 3.1, we have:

Proposition 3.2. [1] *$P(C|T)$ is supported on the intersection*

[1] Proofs for all results are included in Appendix A.

between $\Phi^{-1}(T)$ and the domain of $P_0(C)$. Moreover, we have that $P(C|T) = \frac{P_0(C)}{\int_{\Phi^{-1}(T)} P_0(C)dC}$.

Therefore, characterization of $\Phi^{-1}(T)$ is essential to understand the manifold on which the posterior distribution $P(C|T)$ is supported.

Notice that for any $C \in \Phi^{-1}(T)$, T is obtained from a Sinkhorn scaling of $K = e^{-C}$, and each step of Sinkhorn Scaling is a normalization which equivalent to multiplication of a diagonal matrix. We show that:

Proposition 3.3. *Let T be a non-negative optimal coupling of dimension $m \times n$. $C \in \Phi^{-1}(T)$ if and only if for every $\epsilon > 0$, there exist two positive diagonal matrices $D^r = diag\{d_1^r, \dots, d_m^r\}$ and $D^c = diag\{d_1^c, \dots, d_n^c\}$ such that: $|D^r K D^c - T| < \epsilon$, where $K = e^{-C}$ and $|\cdot|$ is the L^1 norm. In particular, if T is a positive matrix, then $C \in \Phi^{-1}(T)$ if and only if there exist positive diagonal matrices D^r, D^c such that $D^r K D^c = T$, i.e.*

$$\Phi^{-1}(T) = \{C | \exists D^r \text{ and } D^c \text{ s.t. } K = D^r T D^c\}. \quad (4)$$

Proposition 3.3 identifies a key feature of matrices in $\Phi^{-1}(T)$. However, it is not easy to verify whether an arbitrary matrix belongs to $\Phi^{-1}(T)$ nor reveal any underlying geometric structure of $\Phi^{-1}(T)$.

Towards these goals, we now introduce an equivalent condition: *cross ratio equivalence* between *positive* matrices, and show that the IOT set $\Phi^{-1}(T)$ can be completely characterized by T's *cross ratios*.

Definition 3.4. Let A, B be positive $m \times n$ matrices. A is **cross ratio equivalent** to B, denoted by $A \overset{c.r.}{\sim} B$, if all cross ratios of A and B are the same, i.e.

$$r_{ijkl}(A) := \frac{a_{ik}a_{jl}}{a_{il}a_{jk}} = \frac{b_{ik}b_{jl}}{b_{il}b_{jk}} := r_{ijkl}(B)$$

holds for any $i, j \in \{1, \dots, m\}$ and $k, l \in \{1, \dots, n\}$.

Lemma 3.5. *For two positive matrices A, B, $A \overset{c.r.}{\sim} B$ if and only if there exist positive diagonal matrices D^r and D^c such that $A = D^r B D^c$.*

Theorem 3.6. *Let T be an observed positive optimal coupling of dimension $m \times n$. Then $\Phi^{-1}(T)$ is a hyperplane of dimension $m+n-1$ embedded in $(\mathbb{R}^*)^{m \times n}$, which consists all the cost matrices that of the form:*

$$\Phi^{-1}(T) = \{C \in (\mathbb{R}^*)^{m \times n} | K = e^{-C} \overset{c.r.}{\sim} T\}. \quad (5)$$

Cross ratio equivalence provides a strong connection between IOT and algebraic geometry. Algebraically, cross ratios generate the set of algebraic invariants of matrix scaling. The set of all scale reachable matrices from a coupling T, i.e. $\Phi_K^{-1}(T) := \{K | D^r T D^c\} = \{K = e^{-C} | C \in \Phi^{-1}(T)\}$ forms an algebraic variety defined by the cross-ratios of T. Geometrically, the set $\Phi_K^{-1}(T)$ forms a special manifold (Fienberg, 1968). T is the unique intersection between $\Phi_K^{-1}(T)$ and the hyperplane determined by the linear marginal conditions.

Remark 3.7. Since a matrix T and its normalization $T/\sum_{ij}(t_{ij})$ have the same cross ratios (hence have the same image under IOT), we may relax the distribution constraint on the observed coupling T. In this case, instead of probability, t_{ij} represents the frequency of observing x_i and y_i simultaneously.

Example 3.8. Let $T = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$. T has three cross ratios $r_{1212}(T) = \frac{3}{4}, r_{1213}(T) = \frac{1}{6}, r_{1223} = \frac{2}{9}$. Only two are independent as $r_{1213}(T) = r_{1212}(T) * r_{1223}(T)$. Hence $\Phi_K^{-1}(T)$ is the solution set of the algebraic equation:

$$\begin{cases} k_{11}k_{22}/k_{12}k_{21} = 3/4 \\ k_{12}k_{23}/k_{13}k_{22} = 2/9 \end{cases} \iff \begin{cases} 4k_{11}k_{22} - 3k_{12}k_{21} = 0 \\ 9k_{12}k_{23} - 2k_{13}k_{22} = 0. \end{cases}$$

Further as $k_{ij} = e^{-c_{ij}}$, we obtain that $\Phi^{-1}(T)$ is the 4-dim hyperplane defined by the linear equations:

$$\begin{cases} c_{12} + c_{21} - c_{11} - c_{22} = \ln(3/4) & (6a) \\ c_{13} + c_{22} - c_{12} - c_{23} = \ln(2/9). & (6b) \end{cases}$$

Remark 3.9. For a $m \times n$ matrix A, although there are $\binom{m}{2}\binom{n}{2}$ ways to choose i, j, k, l to form cross ratios of A, not all are independent as seen in Example 3.8. A set of independent cross ratios that generates the entire collection of cross ratios is called a **basis**. For example, it is easy to check that $\mathcal{B} = \{r_{1j1k}(A) | j = 2, \dots, m, k = 2, \dots, n\}$ forms a basis for A.

Observed T is assumed to be a positive matrix for the rest of the paper, the case when T is non-negative can be derived using machinery of *maximal partial pattern* developed in (Wang et al., 2019).

Support manifold. Since $\text{Supp}[P(C|T)] = \Phi^{-1}(T) \cap \text{Domain}[P_0(C)]$, different priors would add different constraints on C, which will result different submanifolds embedded in the hyperplane formed by $\Phi^{-1}(T)$.

For instance, suppose $P_0(C)$ follows a Dirichlet distribution over the entire matrix ($\mathbf{P_1}$). Then

$$\text{supp}[P(C|T)] = \{C \in (\mathbb{R}^*)^{m \times n} | e^{-C} \overset{c.r}{\sim} T, \sum_{ij} c_{ij} = 1\}.$$

Notice that K encodes the cross ratios of T, and to avoid the estimation of λ, it is handy to directly put prior over $K = e^{-C}$. For example, assume that $P_0(K)$ is the distribution where each column follows an independent Dirichlet distribution ($\mathbf{P_2}$) with hyperparameter α. Then the domain

of $P_0(K)$ is $\mathcal{P}(\mathbb{R}^n)^m$. Therefore, based on Eq. (4), we have the intersection is of the form:

$$\text{supp}[P(C|T)] = \{C|\exists D^r \text{ s.t. } K = \text{Col}(D^r T)\}. \quad (7)$$

Remark 3.10. Eq. (7) indicates that: **(i)** a known column of K uniquely determines all the other columns. Indeed, denote the j-th column of K, T by $\mathbf{k}_j, \mathbf{t}_j$. For a known column $\mathbf{k}_j$, let $d_i = k_{ij}/t_{ij}$ and $D^r = \text{diag}\{d_1, \dots, d_m\}$. Then the l-th column $\mathbf{k}_l$ equals to the column normalization of $D^r \mathbf{t}_l$. **(ii)** for any $\mathbf{v} \in \mathcal{P}((\mathbb{R}^+)^m)$, there exists a $K \in \text{supp}[P(K|T)]$ such that $\mathbf{k}_j = \mathbf{v}$. Therefore, we have:

Corollary 3.11. *Under prior* $\mathbf{P_2}$*, the projection of supp*$[P(K|T)]$ *onto each column is a* $(m-1)$*-dimensional manifold that is homeomorphic to the simplex* Δ_{m-1}.

Subspace of Supp$[P(C|T)]$**.** Let C_s be the submatrix of the cost C that corresponds to $X_s \times Y_s$, where $X_s = \{x_1, \dots, x_{s_1}\} \subset X$, $Y_s = \{y_1, \dots, y_{s_2}\} \subset Y$. In this section, we characterize the support of $P(C_s|T)$ as a subspace of Supp$[P(C|T)]$.

Denote the sub-coupling of T corresponding to C_s by T_s. According to Eq. (4) and Eq. (5), the projection of $\Phi^{-1}(T)$ onto the cost over $X_s \times Y_s$ is in form of:

$$\begin{aligned}\Phi_s^{-1}(T) &:= \{C_s|C_s \text{ is submatrix of a } C \in \Phi^{-1}(T)\} \\ &= \{C_s|\exists D_s^r, D_s^c \text{ s.t. } K_s = D_s^r T_s D_s^c\} \\ &= \{C_s|K_s = e^{-C_s}, K_s \overset{r.c.}{\sim} T_s\}.\end{aligned}$$

Supp$[P(C_s|T)]$ is determined by $\Phi_s^{-1}(T)$ and $P_0(C)$. Intuitively, supp$[P(C_s|T)]$ contains all the matrices $C_s \in \Phi_s^{-1}(T)$ such that there exists a proper extension of D_s^r, D_s^c to D^r, D^c such that $D^r T D^c$ is in the domain of $P_0(C)$.

Take prior $\mathbf{P_2}$ as an example. Since the domain for a column of K is a copy of Δ_{m-1} as shown in Corollary 3.11, the domain for a column of $K_s = e^{-C_s}$ is then a copy of $\Sigma_{s_1} = \{\mathbf{k} = (k_1, \dots, k_{s_1}) \in (\mathbb{R}^+)^{s_1} | k_1 + \cdots + k_{s_1} < 1\}$. Hence, supp$[P(C_s|T)]$ should be contained in the set :

$$\begin{aligned}\mathcal{W} = \{C^s|&\exists \text{ diagonal matrices } D_s^r, D_s^c \text{ s.t.} \\ &K_s = D_s^r T_s D_s^c \text{ and } \mathbf{1}_{s_1} K_s = (\alpha_1, \dots, \alpha_{s_2}), \alpha_i < 1\}.\end{aligned}$$

It is important to note that supp$[P(C_s|T)]$ is a proper subset of $\mathcal{W}$. As K_s is a submatrix of K, the choice of D_s^c must form a proper column normalizing constant of a $D^r T$. Thus, the choice of $D_s^c = \text{diag}\{d_1^c, \dots, d_{s_2}^c\}$ must guarantee that there exist an extension of D_s^r, denoted by $D'^r = \text{diag}\{D_s^r, d_{s_1+1}^r, \dots, d_m^r\}$, with $d_{s_1+1}^r, \dots, d_m^r > 0$, such that the j-th column sum of $D'^r T$ is $1/d_j^c$. Let T_{m-s} be the submatrix of T corresponding to $\{x_{s_1+1}, \dots, x_m\} \times Y^s$.

Proposition 3.12. $C_s \in$ *supp*$[P(C_s|T)]$ *if and only if there exists positive diagonal matrices* D_s^c, D_s^r *such that* $K_s = D_s^c T_s D_s^r$ *and the system of equations* $(x_1, \dots, x_{m-s_1}) T_{m-s} = (1/d_1^c, \dots, 1/d_{s_2}^c) - \mathbf{1}_s D_s^r T_s$ *have a positive solution for* $\{x_1, \dots, x_{m-s_1}\}$.

Incomplete Observation. Say T is observed with t_{m1} missing. Then generically, we have:

$$\text{supp}[P(C|T)] = \cup_{\widehat{T} \in U(T)} \text{supp}[P(C|\widehat{T})],$$

where $U(T)$ is the set of matrices with T completed by a choice of $t_{m1} > 0$. Given a prior, say **P1**, we have:

$$\text{supp}[P(C|T)] = \{C|\exists t_{m1} > 0 \text{ s.t. } K \overset{c.r}{\sim} \widehat{T}, \textstyle\sum_{ij} c_{ij} = 1\},$$

where $\widehat{T}$ is T with t_{m1} as its $m1$-th element. With prior **P2**, according to (7) we have:

$$\text{supp}[P(C|T)] = \{C|\exists D^r, t_{m1} > 0, \text{ s.t. } K = \text{Col}(D^r \widehat{T})\}.$$

Moreover, according to Remark 3.10, a known column $\mathbf{k}_l$ of K completely determines other fully observed columns. Whereas, for a column with a missing element, say the first column, $\mathbf{k}_l$ determines an 1-dim set $\mathcal{K}_1$ for $\mathbf{k}_1$. In particular, let $D^r = \text{diag}\{d_1, \dots, d_m\}$, where $d_i = k_{il}/t_{il}$.

Corollary 3.13. *Under prior* $\mathbf{P_2}$*,* $\mathcal{K}_1$ *is a line segment in* Δ_{m-1} *that can be parameterized as:* $\mathcal{K}_1 = \{(d_1 t_{11}, \dots, d_m t_{m1}) / \sum_{i=1}^m d_i t_{i1} | t_{m1} \in (0, \infty)\}$.

Note that for different choice of priors, $\mathcal{K}_1$ could be a curve instead of a line segment. Denote the prior imposed constraint on $\mathbf{k}_1$ by $f_1(\mathbf{k}_1) = 0$, the ratio imposed linear constraint on $\mathbf{k}_1$ by $f_i(\mathbf{k}_1) = 0$, where $f_i(\mathbf{k}_1) = k_{11}/k_{i1} - d_1 * t_{11}/(d_i * t_{i1})$, for $i = 2, \dots, m-1$. Then $\mathcal{K}_1$ is the solution set for system of equations $\{f_1, f_2, \dots, f_{m-1}\}$. With Dirichlet prior, f_1 is also linear, so $\mathcal{K}_1$ is a segment.

3.2. With observation noise

For any two observed couplings with the same dimension, we introduce a natural distance between their images under IOT.

Corollary 3.14. *Let* T_1, T_2 *be two positive matrices of dimension* $m \times n$*. The hyperplanes* $\Phi^{-1}(T_1)$ *and* $\Phi^{-1}(T_2)$*, have the same normal direction. In particular, if* $T_1 \overset{c.r.}{\sim} T_2$ *then* $\Phi^{-1}(T_1) = \Phi^{-1}(T_2)$*. Otherwise* $\Phi^{-1}(T_1)$ *is parallel to* $\Phi^{-1}(T_2)$.

Therefore, $\Phi^{-1}(T_1), \Phi^{-1}(T_2)$ are the same or parallel hyperplanes embedded in $(\mathbb{R}^+)^{m \times n}$. **The distance between IOT of** T_1 **and** T_2, denoted by $d(\Phi^{-1}(T_1), \Phi^{-1}(T_2))$ is then well-defined to be the Euclidean distance between $\Phi^{-1}(T_1)$ and $\Phi^{-1}(T_2)$.

Moreover, $\Phi^{-1}(T)$ is completely determined by T's cross ratios. So $d(\Phi^{-1}(T_1), \Phi^{-1}(T_2))$ can be expressed in term of cross ratios of T_1 and T_2. We now illustrate how to obtain $d(\Phi^{-1}(T_1), \Phi^{-1}(T_2))$.

Example 3.15. Let $T_1 = T$ in Example 3.8, $T_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$. Then $\Phi^{-1}(T_1)$ is the hyperplane show in Eq. (6). Taking r_{1212}, r_{1213} as T's cross ratio basis, then we have $\Phi^{-1}(T_2)$ is:

$$\begin{cases} c_{12} + c_{21} - c_{11} - c_{22} = \ln(1/3) & \text{(8a)} \\ c_{13} + c_{22} - c_{12} - c_{23} = \ln(1/3). & \text{(8b)} \end{cases}$$

Both $\Phi^{-1}(T_1)$ and $\Phi^{-1}(T_2)$ are intersections of two 5-dim hydroplanes (i.e.(6a) $\cap$ (6b), (8a) $\cap$ (8b)) with the same coefficient, so they are are parallel. Let the Dihedral angle between (6a) and (6b) be θ. It is easy to check $\cos\theta = 1/2$. So we have $d(\Phi^{-1}(T_1), \Phi^{-1}(T_1)) = \sqrt{\ln^2 \frac{4}{9} + \ln^2 \frac{3}{2} - \ln \frac{3}{2} \ln \frac{4}{9}}$.

Remark 3.16. The hyperparameter λ in Eq.(1) can be viewed as a greedy data selection. For a forward EOT problem, λ raises the cross ratios of $K = e^{-C}$ to a power of λ. Hence, λ either exaggerates or suppresses the cross ratios of T, depending on whether λ is greater or less than 1 (Wang et al., 2020). Conversely, given a backward IOT problem, for a fixed coupling T, the larger the λ is assumed, the smaller the cross ratios of K are. Moreover, for a pair of observed couplings T_1 and T_2, the equations for hyperplanes corresponding $\Phi^{-1}(T_1), \Phi^{-1}(T_2)$ with different choices of λ, only differ on the constants by a scalar, i.e. $d_\lambda(\Phi^{-1}(T_1), \Phi^{-1}(T_2)) = d(\Phi^{-1}(T_1), \Phi^{-1}(T_2))/\lambda$. Therefore, the larger λ is assumed, the smaller the distance between $\Phi^{-1}(T_1)$ and $\Phi^{-1}(T_2)$ are.

Assuming uniform prior $P_0(C)$ over $(\mathbb{R}^*)^{m\times n}$, we now investigate $P(C|T)$ under two common noise types. Results for other specific priors can be obtained by restricting the generic results to the prior's domain. To simplify the notation, without loss, we will assume noise only occurs on one element of T, say t_{11}, as the general case may be treated as compositions of such.

Bounded noise. Suppose that t_{11} is perturbed by a uniform noise from t^*_{11}, i.e. $t^*_{11} = t_{11} + \epsilon$, where ϵ is a random variable with uniform distribution over $[-a, a]$, for $a > 0$.

Theorem 3.6 and Corollary 3.14 imply that $P(C|T)$ is supported on a collection of parallel hyperplanes within bounded distance from $\Phi^{-1}(T)$.

Proposition 3.17. *For a coupling T, assume uniform observation noise on t_{11} with bounded size a, then*

$$supp[P(C|T)] = \cup_{T'\in\mathbb{B}_a(T)} \Phi^{-1}(T'),$$

where $\mathbb{B}_a(T)$ is the set of matrices T' of the same dimension as T with the property that: $t'_{11} > 0$, $|t'_{11} - t_{11}| \le a$ and $t'_{ij} = t_{ij}$ for other i, j. Moreover, $\Phi^{-1}(T')$ can be expressed as intersection of two hyperplanes (may be in different dimensions): one with equation: $c_{11} + c_{22} - c_{21} - c_{12} = -\ln \frac{t'_{11} t_{22}}{t_{21} t_{12}}$, and the other equation does not depend on the value of t'_{11}. Assume the angle between these two hyperplanes is θ. Then $d(\Phi^{-1}(T'_1), \Phi^{-1}(T'_2)) \le \ln \frac{t_{11}+a}{t_{11}-a} / \sin\theta$, for $T'_1, T'_2 \in \mathbb{B}_a(T)$.

Example 3.18. Let T be the same as Example 3.8. Suppose there is a bounded noise ϵ on t_{11} of size $a > 0$. Hence, supp$[P(C|T)]$ is the union of a collection of hyperplanes determined by $\mathbb{B}_a(T)$ of the following form, where $\epsilon \in [-a, a]$.

$$\begin{cases} c_{12} + c_{21} - c_{11} - c_{22} = \ln[(3 + 3\epsilon)/4] & \text{(9a)} \\ c_{13} + c_{22} - c_{12} - c_{23} = \ln(2/9) & \text{(9b)} \end{cases}$$

Let the angle between (9a) and (9b) be θ. Computation shows that $\sin\theta = \sqrt{3}/2$, Hence for $T'_1, T'_2 \in \mathbb{B}_a(T)$, $d(\Phi^{-1}(T'_1), \Phi^{-1}(T'_2)) \le \sqrt{3}\ln(1 + a)$.

Gaussian noise. Suppose that t_{11} is perturbed by a Gaussian noise from the ground truth t^*_{11}, i.e. $t^*_{11} = t_{11} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with standard deviation $\sigma > 0$.

Gaussian distribution is defined over the entire real line, so supp$[P(C|T)] = \cup_{T'\in\mathbb{L}_a(T)} \Phi^{-1}(T')$, where $\mathbb{L}(T)$ is the set of matrices T' of the same dimension as T with the property that: $t_{11} > 0$ and $t'_{ij} = t_{ij}$ for $(i, j) \ne (1, 1)$. Therefore, cross ratios of T' that depend on t_{11} can be any arbitrary number. So similar to the case of missing elements in T, $P(C|T)$ is supported on a hyperplane that is one dimensional higher than $P(C|T^*)$.

Proposition 3.19. *Let T be an observed coupling of dimension $m \times n$ with Gaussian noise on t_{11}. Further, let $\mathcal{B}$ be a basis for cross ratios of $m \times n$ matrices, that contains only one cross ratio depending on t_{11}. Eliminate the cross ratio depending on t_{11} in $\mathcal{B}$, denote the new set by $\mathcal{B}^-$. Then:*

$$supp[P(C|T)] = \{C | r(K) = r(T) \text{ for } r \in \mathcal{B}^-\}$$

In particular, $P(C|T)$ is supported on a hyperplane that is one dimensional higher than $P(C|T^*)$.

Example 3.20. Let T be the same as Example 3.8. Suppose there is a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ on t_{11}. Hence, supp$[P(C|T)]$ is the union of a collection of hyperplanes determined by $\mathbb{L}(T)$, which are in the same form as Eq 9, with constraint on ϵ being: $3 + 3\epsilon > 0$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Thus, $\epsilon \in (-1, \infty)$, which further suggests no restriction is put on $c_{12} + c_{21} - c_{11} - c_{22}$. Hence:

$$\text{supp}[P(C|T)] = \{C | c_{13} + c_{22} - c_{12} - c_{23} = \ln \tfrac{2}{9}\}.$$

Remark 3.21. In this case, $P(C|T)$ is not uniform over its support. According to Eq (2), $P(C|T) \propto P(T_C|T)$ where $T_C \in \mathbb{L}(T)$ and $K = e^{-C} \overset{c.r.}{\sim} T_C$. In particular, $P(C|T)$ has the highest probability on the hyperplane $\Phi^{-1}(T)$. If further constraints are added to the cost matrix C such that there is a unique $C \in \Phi^{-1}(T)$ that meets such constraints, then one may obtain a maximum a posteriori estimation of C. This explains why the simulations in (Stuart & Wolfram, 2020), failed to identify a unique cost matrix for the general case: $P(C|T)$ is supported on an entire hyperplane.

Algorithm 1 MetroMC

input : coupling T, variance σ, prior parameters β, uniform random variable $u \in [0, 1)$
set : $C^{(0)} = -\ln(K^{(0)})$; $K^{(0)} = T_{m\times n}$
for i = 1 to ITERMAX **do**
  $D^r = \exp(N(0, \sigma * \mathbb{I}_m))$
  $D^c = \exp(N(0, \sigma * \mathbb{I}_n))$
  $K^{(i+1)} \leftarrow \text{diag}(D^r)\, K^{(i)}\, \text{diag}(D^c)$
  $C^{(i+1)} = -\ln(K^{(i+1)})$
  $a \leftarrow P(C^{(i+1)}|T)/P(C^{(i)}|T)$
  if a > u **then**
    Accept $C^{(i+1)}$
  else
    Reject $C^{(i+1)}$
    $C^{(i+1)} \leftarrow C^{(i)}$
  end if
end for
output : C

Whereas, maximum a posteriori estimations of C were successfully obtained in both the graph and Toeplitz cases.

4. Algorithm

Approximate inference via MCMC. Given a perfectly observed T and $P_0(C)$, we propose a Markov Chain Monte Carlo (MCMC) method for sampling $P(C|T)$ (see Algorithm 1). At iteration i, denote the current cost C by $C^{(i)}$, $K^{(i)} = e^{-C^{(i)}}$. We generate two vectors D^r and D^c with each element sampled from an exponential of a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. $K^{(i+1)}$ is then obtained by $K^{(i+1)} = \text{diag}D^r K^{(i)} \text{diag}D^c$. The acceptance ratio, $r = \frac{P(-\ln(K^{(i+1)})|T)}{P(-\ln(K^{(i)})|T)} \overset{(*)}{=} \frac{P_0(C^{i+1},\beta)}{P_0(C^i,\beta)}$ where β is the hyper-parameter of prior $P_0(C)$. Equality $(*)$ holds because our proposal preserves the cross-ratios, so we are guaranteed to stay on the manifold of support. Therefore, $P(T|C)$ is always 1 and is omitted (see Proposition 3.2). We accept or reject the move by comparing the acceptance ratio to an uniform random variable, u, which concludes a single step of the MCMC algorithm. We refer this method as MetroMC.

Metropolis-Hastings Monte Carlo method. We can avoid dealing with the parameter λ by putting a prior on K. However, the MetroMC has long autocorrelation time for such priors. In each iteration, $K^{(i+1)}$ is obtained by *multiplying* random factors on $K^{(i)}$, whereas $C^{(i+1)}$ is obtained by *adding* random factors on $C^{(i)}$. In more detail, rows of $K^{(i)}$ are scaled by exponential of Gaussian variables. A lot of time only a few of rows are at the same scale. After the columns are normalized, the rows that have large scale will have most of the weight in a column, which means we will sample a lot on the boundaries of the support. In order to improve the efficiency, we utilize the Metropolis-Hastings (Hastings, 1970) sampling method in below.

We propose the following algorithm to sample from $P(K|T)$ with a prior on K. The sufficient condition for the ergodicity of the MCMC method is detailed balance, which is defined as $\pi(K)Q(K'|K)A(K', K) = \pi(K')Q(K|K')A(K, K')$, where $\pi(K)$ is the stationary distribution, $Q(K'|K)$ is the proposal transition probability from a state K to K' and $A(K', K)$ is the acceptance probability which we choose

$$A(K', K) = \min\left(1, \frac{P(K')}{P(K)} \frac{Q(K|K'))}{Q(K'|K)}\right). \tag{10}$$

We replace the acceptance probability a in Alg. 1 with Eq. (10). We refer this algorithm as the MHMC.

The proposal transition probability $Q(K'|K)$ is defined as follows. In each iteration of MHMC, we draw a random number ϵ_i from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where σ depends on the row sums of K. More precisely, $\sigma = \sigma_0 * s_i^\gamma + \delta$, where σ_0, γ, and δ are hyper-parameters for the model and $s_i = \sum_j K_{ij}$. $K' = \text{diag}(D_i^r)K$, where $D_i^r = e^{[0,..,0,\epsilon_i,0,...,0]}$.

We scale a row at once, so in each step the acceptance ratio is reduced to $A(s_i'|s_i)$. To compute it we need $Q(s_i'|s_i) = F_N(x = \ln(s_i'/s_i), \mu, \sigma)$ and $Q(s_i|s_i') = F_N(x = \ln(s_i/s_i'), \mu', \sigma')$, where $F_N(x, \mu, \sigma)$ is the Gaussian probability density function.

5. Experiments

The experiments are performed on computing clusters with nodes equipped with 40 Intel Xeon 2.10 GHz CPUs and 192 GB memory. Note that the PIOT inference has a linear computational complexity of $\mathcal{O}(mn)$. (See appendix for further details.)

Visualizing subspaces of supp$[P(K|T)]$**.** Fig 1(a-c) illustrates support of 2×2 subspaces supp$[P(K^s|T)]$, for a 4×4 coupling T under prior $\mathbf{P}_2$. In each plot, a T is sampled from a Dirichlet distribution [2]. Consider submatrix $K_s = [[k_{11}, k_{12}], [k_{21}, k_{22}]]$ of K. According to section 3.1, both columns $\mathbf{k}_1, \mathbf{k}_2$ of K_s lie in a copy of Σ_2. Each colored triangle mark represents a uniformly sampled $\mathbf{k}_1$ from Σ_2. A sampled $\mathbf{k}_1$ determines a $D_s^r = \text{diag}\{k_{11}/t_{11}, k_{21}/t_{21}\}$, which further determines a set $\mathcal{K}_2$ for $\mathbf{k}_2$ shown by solid segment with the same color. Projection of $\mathcal{W}$ into the simplex corresponding to $\mathbf{k}_2$, denoted by $\mathcal{W}_2$, is shown by the dashed ray. Here, $\mathcal{K}_2$ is obtained by: fix $\mathbf{k}_1 = (k_{11}, k_{21})^T$, further sample k_{31} and k_{41} to form the first column of K. Then compute and plot the uniquely determined $\mathbf{k}_2$. As predicted in section 3.1, $\mathcal{K}_2$ is only a subset of $\mathcal{W}_2$. The

[2] All coupling Ts in this section are reported in Appendix B.9.

sizes and locations of $\mathcal{K}_2$ in $\mathcal{W}_2$ vary according to cross ratios of T.

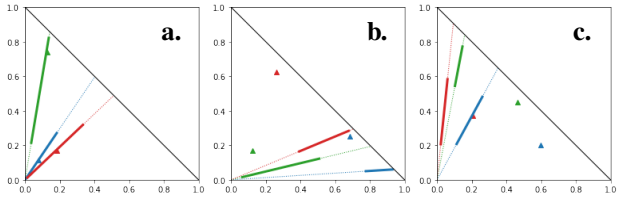

Figure 1: Visualization of 2×2 subspaces of supp$[P(K|T)]$. In each plot, a T is sampled from a Dirichlet distribution. Each colored mark represents a uniformly sampled $\mathbf{k}_1 = (k_{11}, k_{21})^T$. Solid segment with the same color plots the corresponding set $\mathcal{K}_2$ for the second column, and the dashed ray plots $\mathcal{W}_2$.

Visualizing supp$[P(K|T)]$ through MHMC. Now we utilize MHMC to visualize supp$[P(K|T)]$ and see the effect of cross-ratios. Under prior $\mathbf{P}_2$ with $\alpha = 1$, Fig. 2 illustrates the posterior distributions $P(K|T)$ of three 3×3 couplings.

We choose $\alpha = 1$ for the Dirichlet prior $P_0(K)$, $\sigma_0 = 0.5$, $\gamma = 3$, and $\delta = 1.0$. We run for 10,000 burn-in steps and take 10,000 samples with lags of 100. The choices of these numbers are validated in Appendix B.4. Three columns (red, green, and blue colors, respectively) of each T are plotted in the simplex on the left of Fig. 2. Locations of these points illustrate the relations between each column encoded by the cross ratios of T. The right of Fig. 2 plots columns of sampled K from $P(K|T)$ with corresponding T to the left. We clearly see how the cross-ratio of T affects $P(K|T)$. If one column of T is close to one edge, then the posterior distribution of that column will be denser on that edge; on the other hand, if a column of T is located near the center, then the posterior is more evenly distributed on a simplex. E.g. in the top case, the effects on red and blue are exaggerated because the effect depends on relative position (and similarly for the red in the middle panel).

Visualization of supp$[P(K|T)]$ with incomplete observations is included in Appendix B.6.

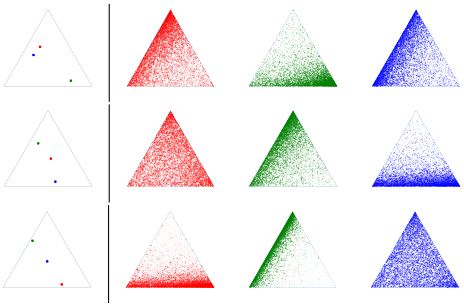

Figure 2: **(Left.)** The columns of three T matrices plotted on the same simplex: $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ are shown in Red, green for , blue. **(Right.)** Samples from posterior distribution $P(K|T)$. Each column of inferred matrices K are plotted in a separate simplex, indicated by corresponding color.

Symmetric cost. In this section, we consider $n \times n$ squared symmetric cost with $C^g_{ij} = \left|\frac{(i-j)}{n}\right|^p$ for $i, j \in [n]$, $n = 10$, and p =0.5, 1, and 2. T^g is obtained by EOT with $\lambda = 10$ and randomly sampled μ and ν. The relative error $\| \hat{C} - C^g \| / \| C^g \|$ is used to evaluate the quality of estimation $\hat{C}$, where $\| \cdot \|$ is the l^2 norm. MetroMC is used.

We compare PIOT to the algorithm in Ma et al. (2020) and find comparable performance despite our more general approach. Ma et al. (2020) projecting the inferred cost to the set of symmetric matrices at every iteration, effectively assuming the ground truth is symmetric. Their algorithm reaches an error below 10^{-5} in 300 iterations for all three p values. The flexibility of PIOT allows us to pick any prior. Here we choose a prior such that $P_0(C)$ follows a Gibbs distribution $\exp(-\beta(\| \gamma(C - C^T) \|))$. The results show that after burn-in, the inferred cost distributed is already concentrated around the ground truth cost, therefore, with 1000 MC samples the PIOT method achieve an error less than 10^{-5} for all the three p cases. ($\hat{C}$ is the median of inferred cost distribution, see simulation details in Appendix B.7.)

Next, we test the robustness of both algorithms by adding random perturbations onto the symmetric costs. Random matrices with all entries in $[0, 10^{-4})$ are sampled, and added to the symmetric costs. C^g is then obtained by normalizing the combined matrices to 1. For PIOT with the uniform prior $\mathbf{P_1}$ described in section 3, the relative errors fall between 10^{-3} to 10^{-4} for all p case, whereas the errors are in the range of 10^{-2} to 10^{-3} using the algorithm from (Ma et al., 2020). Figure 3 shows the results on $C_{1,9}$, other elements behave similarly. (See simulation details in Appendix B.7.)

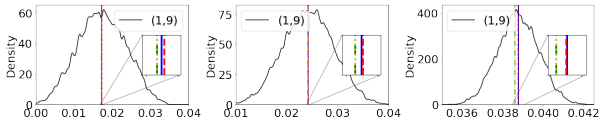

Figure 3: The KDE of the inferred distribution of $C_{1,9}$. The blue solid lines indicate the ground truth cost, the green dash-dotted lines are results using algorithm from (Ma et al., 2020) at 300-iteration, the yellow dotted lines are Ma's results at 1000-iteration (Overlapped with green lines for all cases.), and the red dashed lines are the median of PIOT. (Left.) $p = 0.5$. (Middle.) $p = 1$. (Right.) $p = 2$. The insets zoom-in the range of $C^g_{1,9} \pm 10^{-4}$.

Thus when observed T did come from a cost aligned with the symmetry assumption, both PIOT and (Ma et al., 2020) achieve comparable results. However, even with small perturbations, PIOT attains better accuracy.

Bounded noise on observed plan. Given an observed noisy T, we assume t_{11} was perturbed by a bounded noise ϵ sampled uniformly from distribution $[-a, a]$ with $a = 0.01$. Assuming prior $\mathbf{P}_1$ for $P_0(C)$ with symmetric Dirichlet parameter equals to 1, we use MetroMC to sample supp$[P(C|T)]$.

Notice that initializing $K^{(0)} = T$ may cause the corresponding $C^{(0)}$ lies outside the domain of $P_0(C)$ as $\sum_{ij} c_{ij}^{(0)} = \sum_{ij} -\ln(t_{ij})$ is not guaranteed be 1. This will cause an extremely long burning time. In order fix to this issue, we scale $K^{(0)}$ by a constant when necessary. Scaling does not change the cross-ratios, hence it equivalent to a move within the support of $\Phi^{-1}(T)$.

For each iteration in the MCMC method we generate $m+n-1$ (rather than $m+n$) independent Gaussian random numbers with $\sigma = 0.02$ to construct vectors D^r and D^c satisfying $\sum_{i,j} \ln(d_i^r) + \ln(d_j^c) = 0$. The purpose of the above condition is to ensure that the newly generated sample lies on the support of the posterior. Otherwise, the acceptance rate will be extremely low. Because C is a positive matrix, the MetroMC method rejects any sample of K if $\max_{(i,j)} k_{i,j} \geq 1$.

Fig. 4 shows the MetroMC results on the inferred cost matrices C for a given 3×3 coupling T. For each simulation we burn in 10,000 steps and take 100,000 samples with lags of 100. The acceptance rates range from 0.51 to 0.58. The MCMC method is validated by checking the autocorrelation function and the running averages (see Appendix B.) Noise level ϵ is indicated by colors. The **Left** panel plots components for a cross-ratio depend on t_{11}: $(c_{11}+c_{22})$ by solid curves and $(c_{21}+c_{12})$ by dashed curves. The **Right** panel plots components for a cross-ratio dose not depend on t_{11}: $(c_{12}+c_{23})$ by solid curves and $(c_{13}+c_{22})$ by dashed curves. All the curves are smoothed using the Gaussian kernel density estimation (KDE) with a bandwidth of 0.05. Fig. 4 illustrates that curves for $(c_{11}+c_{22})$ and $(c_{21}+c_{12})$ with the same ϵ (i.e. curves in the same color) are in the same shape. The off-set between curves varies as the noisy level changes. Whereas off-set between curves for $(c_{12}+c_{23})$ and $(c_{13}+c_{22})$ on the right panel remains a constant for any choice of ϵ. This is consistent with our theoretical result in Prop. 3.17 that: $(c_{11}+c_{22})-(c_{21}+c_{12}) = -\ln \frac{t'_{11}t_{22}}{t_{21}t_{12}}$ depends on ϵ, whereas $(c_{12}+c_{23})-(c_{13}+c_{22}) = -\ln \frac{t_{12}t_{23}}{t_{13}t_{22}}$ is a constant.

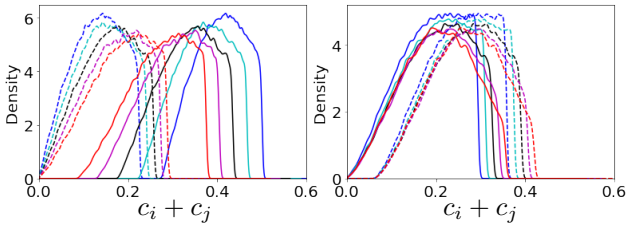

Figure 4: Results on infered cost C given a T with bounded noise ϵ on t_{11}. ϵ varies in $[-0.01, -0.005, 0.0, 0.005, 0.01]$, indicated by [blue, cyan, black, magenta, red] in order. **(Left)** Solid curves are for $(c_{11}+c_{22})$, and dashed for $(c_{21}+c_{12})$. **(Right)** The solid curves are for $(c_{12}+c_{23})$, and dashed for $(c_{13}+c_{22})$.

Gaussian noise on observed plan. We start with a Toeplitz ground truth cost function C^g and obtain an OT plan, T^g, with C^g and uniform μ, ν. A random Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.004^2)$, about 4% of the average value of T, is added to $t^g_{1,2}$, and the corrupted plan $T = T^g + \epsilon$. We then infer the posterior distribution $P(C|T)$. We put a $\mathbf{P}_1$ prior on C with the hyperparameters follows Toeplitz matrix format. Inference is done by first generating 10 random numbers $\epsilon_i^* \sim \mathcal{N}(0, 0.004^2)$ and apply the noise on t_{12} to get $T_i^* = T + \epsilon_i^*$. Then, we utilize MetroMC on every T_i^* to obtain $P(C|T_i^*)$, and finally $P(C|T) \approx \sum_i P(C|T_i^*)$.

For the MetroMC method, we burn in 10,000 steps and then take 10,000 samples with lags of 200 steps for each $P(C|T_i^*)$. For each step, we use the same sampling procedure as described in the bounded noise case but use a Gaussian standard deviation of 0.003. The acceptance rates are around 0.52-0.70.

The posterior distributions of each element of the inferred C are shown in Fig. 5. The inferred distributions are close to the ground truth costs which are indicated by blue dashed lines. This experiment demonstrates that our method successfully make accurate inferences using soft constraints $P_0(C)$, with the results comparable to hard constraints imposed by (Stuart & Wolfram, 2020).

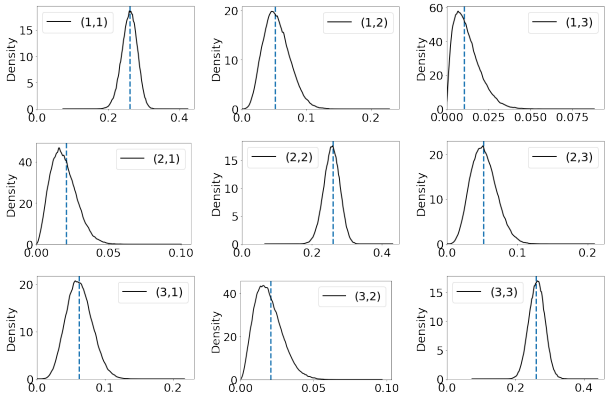

Figure 5: The KDE of each element of inferred cost C with Gaussian noise on t_{12}. The blue dashed vertical lines indicate the ground truth cost C^g.

Application on European Migration data. We apply PIOT to analyze European migration flow. The observed migration data T^g between 9 European countries for the period 2002- 2007 is shown in Fig. 6 top (Raymer et al., 2013). All results in this section are obtained by performing MetroMC with 10,000 burn-in steps, 10,000 samples with lags of 1000, and the sum of cost matrix is assumed to be a constant, see details in Appendix B.10.

First, we demonstrate the ability of PIOT by inferring the latent costs under various priors. Three priors are considered with increasing strength: 1. Fully uniform prior, Dirichlet with all $\alpha = 1$; 2. Semi-uniform prior, diagonal α=25 and 1

elsewhere, as the data does not consider migration within a country itself. 3. Graph-geometrical prior, the diagonals are 25 and α=1 if the two countries are directly connected and 1.5 if not, as comparison to the hard Graph-geometrical constraint imposed by (Stuart & Wolfram, 2020).

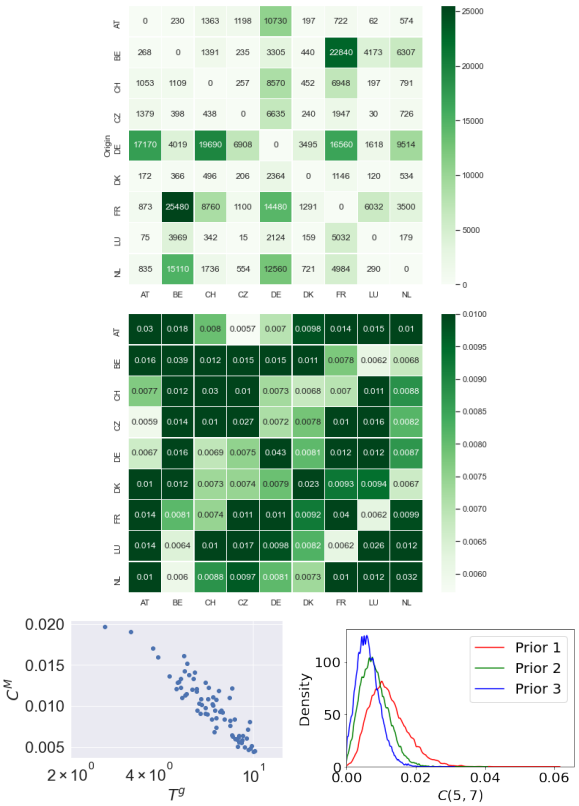

Figure 6: (**Top.**) European migration data. (**Middle.**) Mean of the inferred cost using Prior 2. (**Bottom-left.**) Plots for logarithm of each migration data with its corresponding inferred cost mean. (**Bottom-right.**) The inferred cost distributions for DE→FR under three priors.

Fig. 6 middle plots the mean of the inferred costs under prior 2. Even with a weak prior, PIOT correctly captures the general reciprocal relation between coupling and cost, lower costs lead to higher migration numbers, shown in Fig. 6 bottom-left. Results for the other two priors are similar, see Appendix B. Fig. 6.bottom-right plots the inferred costs from Germany to France (DE→FR). The implications of priors can be observed: the stronger the prior is, the narrower the inferred cost distribution is. In particular, using prior 3, the mass is concentrated between 0 and 0.01 the same as in (Stuart & Wolfram, 2020), where hard graphical constraint on the cost is considered.

Next, we test PIOT's performance on noisy data. We corrupting an element of T^g with a Gaussian noise to obtain T, whose standard deviation is set to be 4% of the value. With prior 2, inference on cost distributions is done in the same fashion as in Fig. 5. We then apply EOT on the cost matrix

Table 1: Comparison between the predicted coupling, T^p, with the ground truth coupling, T^g. $\Delta = (T^p - T^g)/\overline{T}^g$. Corresponding element is corrupted by a 4% Gaussian noise on the left column, and missing on the right column.

	Noise			Missing value		
	(3,6)	(4,5)	(5,7)	(3,6)	(4,5)	(5,7)
T^g	452	6635	16560	452	6635	16560
T^p	454	6593	16515	3699	7462	12349
$\lvert\Delta\rvert$	.45%	.87%	.25%	26%	6.6%	34%

formed by means of inferred distributions with the observed marginals, to predict the true migration data. Left of Table 1 shows the results when noise occurs at $T^g_{36}, T^g_{45}, T^g_{57}$. Our predictions T^p are within 1% to the mean of prior on the ground truth, denoted by $\overline{T}^g$ as shown in the third row.

Finally, we explore inference of missing data. We fill in the missing element with 100 samples ranging uniformly over $[100, 25000)$. With semi-uniform prior, we estimate the cost and perform EOT as described above. Right column of Table 1 shows that PIOT successfully recovered the ground truth with estimation error less than 35% of the prior mean.

For the European migration data, the true cost function is not simply the Euclidean distance between countries but could be a combination of complex factors such as ticket prices, international relationships, and languages. To model all the factors is impractical. Strong priors or constraints may simplify the problem (Stuart & Wolfram, 2020), but uniform priors and softer constraints provide less biased results. Our experiments show that one may use PIOT to analyze the underlying geometry, and infer the migration costs quantitatively. The inferred cost is then an estimate of the consequences of the hidden factors. Table 1 additionally demonstrates inference with noise and missing values.

6. Conclusion

We have generalized prior treatments of IOT by defining and studying the underlying support manifold and associated inference problems. We provided MCMC methods for inference over general priors on discrete costs and induced kernels. Simulations illustrate underlying geometric structure and demonstrate the feasibility of the inference.

Acknowledgements

Research was sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) and was accomplished under Cooperative Agreement Numbers W911NF-20- 2-0001 and HR00112020039 to P.S. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or ARO, or the U.S. Government.

References

Alaux, J., Grave, E., Cuturi, M., and Joulin, A. Unsupervised hyperalignment for multilingual word embeddings. *arXiv preprint arXiv:1811.01124*, 2018.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

Dupuy, A., Galichon, A., and Sun, Y. Estimating matching affinity matrices under low-rank constraints. *Information and Inference: A Journal of the IMA*, 8(4):677–689, 2019.

El Moselhy, T. A. and Marzouk, Y. M. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

Fienberg, S. E. The geometry of an r× c contingency table. *The Annals of Mathematical Statistics*, 39(4):1186–1190, 1968.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a wasserstein loss. *arXiv preprint arXiv:1506.05439*, 2015.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. *arXiv preprint arXiv:1605.08527*, 2016.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. 1970.

Idel, M. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pp. 199–201, 1942.

Li, R., Ye, X., Zhou, H., and Zha, H. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20, 2019.

Ma, S., Sun, H., Ye, X., Zha, H., and Zhou, H. Learning cost functions for optimal transport. *arXiv preprint arXiv:2002.09650*, 2020.

Paty, F.-P. and Cuturi, M. Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pp. 7532–7542. PMLR, 2020.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Pretzel, O. Convergence of the iterative scaling procedure for non-negative matrices. *Journal of the London Mathematical Society*, 2(2):379–384, 1980.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., and Bijak, J. Integrated modeling of european migration. *Journal of the American Statistical Association*, 108(503): 801–819, 2013.

Rothblum, U. G. and Schneider, H. Scalings of matrices which have prespecified row sums and column sums via optimization. *Linear Algebra and its Applications*, 114: 737–764, 1989.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Stuart, A. M. and Wolfram, M.-T. Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619, 2020.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wang, P., Paranamana, P., and Shafto, P. Generalizing the theory of cooperative inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1841–1850. PMLR, 2019.

Wang, P., Wang, J., Paranamana, P., and Shafto, P. A mathematical theory of cooperative communication. In *NeurIPS*, 2020.

A. Examples, Definitions and Proofs

Example A.1. Let $\mu = \nu = (\frac{3}{8}, \frac{5}{8})$, $C = \begin{pmatrix} \ln 1 & \ln 2 \\ \ln 4 & \ln 1 \end{pmatrix}$. For $\lambda = 1$, we may obtain T by applying SK scaling on $K^\lambda = e^{-C} = \begin{pmatrix} 1 & 1/2 \\ 1/4 & 1 \end{pmatrix}$: first, row normalize K^λ with respect to μ results: $K_1^\lambda = \begin{pmatrix} 1/4 & 1/8 \\ 1/8 & 1/2 \end{pmatrix}$. Then column normalization of K_1^λ with respect to ν outputs $K_2^\lambda = \begin{pmatrix} 1/4 & 1/8 \\ 1/8 & 1/2 \end{pmatrix}$. As $K_1^\lambda = K_2^\lambda$, the SK scaling has converged with $T = K_1^\lambda$. In general, multiple iterations may be required to reach the limit.

Proposition 3.2. *When $T = T^*$, $P(C|T)$ is supported on the intersection between $\Phi^{-1}(T)$ and the domain of $P_0(C)$, moreover, we have that $P(C|T) = \dfrac{P_0(C)}{\int_{\Phi^{-1}(T)} P_0(C) dC}$.*

Proof. When $T = T^*$, Eq (2) and Eq (3) imply that: $P(C|T) = \frac{P(T|C)P_0(C)}{P(T)}$. $P(T) > 0$ is the normalizing constant, so Supp$[P(C|T)]$ = Supp$[P(T|C)] \cap$ Domain$[P_0(C)]$. Note that $P(T|C)P_0(C) = P_0(C)$ for $C \in \Phi^{-1}(T)$, otherwise $P(T|C)P_0(C) = 0$. Hence the proposition holds. □

Proposition 3.3. *Let T be a non-negative optimal coupling of dimension $m \times n$. $C \in \Phi^{-1}(T)$ if and only if for every $\epsilon > 0$, there exist two positive diagonal matrices $D^r = diag\{d_1^r, \ldots, d_m^r\}$ and $D^c = diag\{d_1^c, \ldots, d_n^c\}$ such that: $|D^r K D^c - T| < \epsilon$, where $K = e^{-C}$ and $|\cdot|$ is the L^1 norm. In particular, if T is a positive matrix, then $C \in \Phi^{-1}(T)$ if and only if there exist positive diagonal matrices D^r, D^c such that $D^r K D^c = T$, i.e.*

$$\Phi^{-1}(T) = \{C | \exists D^r \text{ and } D^c \text{ s.t. } K = D^r T D^c\}$$

Proof. Let the row and column of marginals of T be μ and ν. Then $C \in \Phi^{-1}(T)$ if and only if (μ, ν)-Sinkhorn scaling of K converges to T in L^1 norm (for finite matrices, convergence in all L^k norms are equivalent). Notice that a row (column) normalization step in the Sinkhorn scaling is equivalent to a left(right) matrix multiplication of a positive diagonal matrix. Indeed, let $K_0 = K$, $\mathbf{r}_0 = K_0 \mathbf{I}_n$ (vector for row sums of K_0), $D_0^r = \text{diag}(\mu / \mathbf{r}_0)$. Here $\mu / \mathbf{r}_0$ represents element-wise division. Let K_0' be the matrix obtained by row normalization of K_0 with respect to μ. Then $K_0' = D_0^r K_0$. Similarly, let K_1 be the matrix obtained by column normalization of K_0' with respect to ν. Then $K_1 = K_0' D_0^c$, where $D_0^c = \text{diag}(\nu / \mathbf{I}_m^T K_0')$. Iteratively, we have $K_s = \Pi_{i=1}^s D_i^r \cdot K_0 \cdot \Pi_{i=1}^s D_i^c$.

The **only if direction** [$C \in \Phi^{-1}(T) \implies$ existence of D^r and D^c for any ϵ]: $C \in \Phi^{-1}(T) \implies (\mu, \nu)$-Sinkhorn scaling of K converges to $T \implies$, for any $\epsilon > 0$, there exists $N > 0$ such that for any $s > N$, $|K_s - T| < \epsilon$, where $K_s = \Pi_{i=1}^s D_i^r \cdot K \cdot \Pi_{i=1}^s D_i^c$. Let $D^r = \Pi_{i=1}^s D_i^r$, $D^c = \Pi_{i=1}^s D_i^c$, the only if direction is complete.

The **if direction** [existence of D^r and D^c for any $\epsilon \implies C \in \Phi^{-1}(T)$]: Let the limit of (μ, ν)-Sinkhorn scaling on K be K^*. Hence K^* and T have the same marginals and pattern. According to Lemma A.3 of (Wang et al., 2019), K^* and T are diagonally equivalent. Further by Proposition 1 of (Pretzel, 1980) $K^* = T$.

In particular, when T is a positive matrix, K must also be a positive matrix. Then T and K have the same pattern. $T = \lim_{s\to\infty} \Pi_{i=1}^s D_i^r \cdot K \cdot \Pi_{i=1}^s D_i^c$ implies that $\lim_{s\to\infty} \Pi_{i=1}^s D_i^r$ and $\lim_{s\to\infty} \Pi_{i=1}^s D_i^c$ exist. Hence the claim. (Rothblum & Schneider, 1989; Idel, 2016)

□

Lemma 3.5. *For two positive matrices A, B, $A \overset{c.r.}{\sim} B$ if and only if there exist positive diagonal matrices D^r and D^c such that $A = D^r B D^c$.*

Proof. The **if direction**: $A = D^r B D^c \implies \dfrac{a_{ik} a_{jl}}{a_{il} a_{jk}} = \dfrac{d_i^r b_{ik} d_k^c \cdot d_j^r b_{jl} d_l^c}{d_i^r b_{il} d_l^c \cdot d_j^r b_{jk} d_k^c} = \dfrac{b_{ik} b_{jl}}{b_{il} b_{jk}} \implies A \overset{c.r.}{\sim} B$.

The **only if direction**: Let the dimension of A, B be $m \times n$, we will prove by induction on $m + n$. *Step1.* Assume the dimension of A, B is 2×2. Let the marginal of A be μ_A and ν_A, and $B^* = \Phi(B, \mu_A, \nu_A)$. Then A and B^* have the same marginals and cross ratios, which put four same independent constraints on elements of A and B, hence $A = B^* = D^r B D^c$ for some D^r, D^c. *step2.* Assume the statement holds for $m + n < N$.

Now assume that $m+n=N$. Denote the submatrices of A and B consisted by their first $n-1$ columns by A_1 and B_1. $A \overset{c.r.}{\sim} B \Longrightarrow A_1 \overset{c.r.}{\sim} B_1$. By the inductive assumption, there exist diagonal matrices D_1^r and D_1^c such that $A_1 = D_1^r B_1 D_1^c$. Further $\frac{a_{11}a_{1n}}{a_{i1}}a_{in} = \frac{b_{11}b_{1n}}{b_{i1}}b_{in}$ holds for any $i \in \{1,\dots,m\}$ imply that there exists an $d>0$ such that $\mathbf{a}_n = d \cdot \mathbf{b}_n$. Let $D^r = D_1^r$, $D^c = \text{diag}D_1^c, d$, we have $A = D^r B D^c$ holds. Hence the only if direction is completed. □

Theorem 3.6. *Let T be an observed positive optimal coupling of dimension $m \times n$. Then $\Phi^{-1}(T)$ is a hyperplane of dimension $m+n-1$ embedded in $(\mathbb{R}^*)^{m\times n}$, which consists all the cost matrices that of the form:*

$$\Phi^{-1}(T) = \{C \in (\mathbb{R}^*)^{m\times n} | K = e^{-C} \overset{c.r.}{\sim} T\}. \tag{11}$$

Proof. Combining Proposition 3.3 and Lemma 3.5, we have Eq. (11) holds. Hence $\Phi^{-1}(T)$ contains all the matrices C satisfying the following set of equations: $\frac{k_{ik}k_{jl}}{k_{il}k_{jk}} = \frac{t_{ik}t_{jl}}{t_{il}t_{jk}}$ for any $i,j \in \{1,\dots,m\}$ and $k,l \in \{1,\dots,n\}$, where $K = e^{-C}$, i.e. $k_{ik} = e^{-c_{ik}}$. Thus $\Phi^{-1}(T)$ is the solution set of the system of linear equations:

$$c_{il} + c_{jk} - c_{ik} - c_{jl} = \ln(t_{ik}t_{jl}) - \ln(t_{il}t_{jk}) \tag{12}$$

Further Remark 3.9 shows that there are only $(m-1)(n-1)$ independent equations in the system (12). Hence $\Phi^{-1}(T)$ is a hyperplane of dimension $mn-(m-1)(n-1) = m+n-1$. □

Corollary 3.11. *Under prior $\mathbf{P_2}$, the projection of supp$[P(K|T)]$ onto each column is a $(m-1)$-dimensional manifold that is homeomorphic to the simplex Δ_{m-1}.*

Proof. According to prior $\mathbf{P_2}$, for any $K \in$ supp$[P(K|T)]$, each column of K sums to 1. Hence the projection of supp$[P(K|T)]$ onto any j-th column is embedded in the simplex Δ_{m-1}. To show 'homeomorphic ', we only need to show that for each $v \in \Delta_{m-1}$, there exists a $K \in$ supp$[P(K|T)$ such that $\mathbf{k}_j = v$. As discussed in Remark 3.10, let $D'^r = \text{diag}(v/\mathbf{t}_j)$, $K' = \text{Col}(D'^r T)$. It's easy to check that $K' \in$ supp$[P(K|T)]$, and $\mathbf{k'}_j = v$. Hence the corollary holds. □

Proposition 3.12. *$C_s \in$ supp$[P(C_s|T)]$ if and only if there exists positive diagonal matrices D_s^c, D_s^r such that $K_s = D_s^c T_s D_s^r$ and the system of equations shown below have a positive solution for $\{x_1,\dots,x_{m-s_1}\}$.*

$$(x_1,\dots,x_{m-s_1})T_{m-s} = (1/d_1^c,\dots,1/d_{s_2}^c) - \mathbf{1}_s D_s^r T_s \tag{13}$$

Proof. The **if direction**: Let $(x_1,\dots,x_{m-s_1})$ be a positive solution of Eq. (13), let $D'^r = \text{diag}(D_s^r, x_1,\dots,x_{m-s_1})$ be an extension of D_s^r. Denote the column sum of $D'^r T$ by ν'. Then Eq. (13) implies that $\nu'_i = d_i^c$ for $i \in \{1,\dots,s_2\}$. Let $D'^c = \text{diag}(D_s^c, \nu'_{s_2+1},\dots,\nu'_n)$. It is clear that $K' = D'^r T D'^c \in$ supp$[P(K|T)]$, and K_s is K''s submatrix corresponding to $X_s \times Y_s$.

The **only if direction**: for $C_s \in$ supp$[P(C_s|T)]$, let $K_s = e^{-C_s}$ be the submatrix of $K' \in$ supp$[P(K|T)$. Then there exists D'^r, D'^c such that $K' = D'^r T D'^c$. Let the corresponding submatrices of D'^r, D'^c be D_s^r, D_s^c. We have $K_s = D_s^c T_s D_s^r$ hold. Further it is easy to verify that $(d'^r_{s_1+1},\dots d'^r_n)$ is a positive solution for Eq. (13). Hence the proof is completed. □

Corollary 3.13. *Under prior $\mathbf{P_2}$, $\mathcal{K}_1$ is a line segment in Δ_{m-1} that can be parameterized as: $\mathcal{K}_1 = \{(d_1 t_{11},\dots,d_m t_{m1})/\sum_{i=1}^m d_i t_{i1} | t_{m1} \in (0,\infty)\}$.*

Proof. Corollary 3.11 implies that $\mathcal{K}_1 \subset \Delta_{m-1}$. Further Remark 3.10 implies that for each choice of $t_{m1} > 1$, a known $\mathbf{k}_l$ uniquely determines a point in form of $(d_1 t_{11},\dots,d_m t_{m1})/\sum_{i=1}^m d_i t_{i1}$ in $\mathcal{K}_1$. Hence the corollary holds. □

Corollary 3.14. *Let T_1, T_2 be two positive matrices of dimension $m \times n$. The hyperplanes $\Phi^{-1}(T_1)$ and $\Phi^{-1}(T_2)$, have the same normal direction. In particular, if $T_1 \overset{c.r.}{\sim} T_2$ then $\Phi^{-1}(T_1) = \Phi^{-1}(T_2)$. Otherwise $\Phi^{-1}(T_1)$ is parallel to $\Phi^{-1}(T_2)$.*

Proof. According to the proof of Theorem 3.6 above, both $\Phi^{-1}(T_1)$ and $\Phi^{-1}(T_2)$ are defined by system of equations in form of Eq. (12). In particular, they have the same coefficients, only the constants on the right side of the equations are different. Hence, $\Phi^{-1}(T_1)$ and $\Phi^{-1}(T_2)$, have the same normal direction. □

Proposition 3.17. *For a coupling T, assume uniform observation noise on t_{11} with bounded size a,*

$$supp[P(C|T)] = \cup_{T' \in \mathbb{B}_a(T)} \Phi^{-1}(T'), \tag{14}$$

where $\mathbb{B}_a(T)$ is the set of matrices T' of the same dimension as T with the property that: $t'_{11} > 0$, $|t'_{11} - t_{11}| \leq a$ and $t'_{ij} = t_{ij}$ for other i, j. Moreover, $\Phi^{-1}(T')$ can be expressed as intersection of two hyperplanes (may be in different dimensions): one with equation: $c_{11} + c_{22} - c_{21} - c_{12} = -\ln \frac{t'_{11} t_{22}}{t_{21} t_{12}}$, and the other equation does not depend on the value of t'_{11}. Assume the angle between these two hyperplanes is θ. Then $d(\Phi^{-1}(T'_1), \Phi^{-1}(T'_2)) \leq \ln \frac{t_{11}+a}{t_{11}-a} / \sin\theta$ for $T'_1, T'_2 \in \mathbb{B}_a(T)$.

Proof. With bounded noise on t_{11}, the domain for t^*_{11} is $[t_{11} - a, t_{11} + a] \cap \{t^*_{11} > 0\}$. Hence Eq. (14) holds. $\Phi^{-1}(T')$ is defined by the cross ratios of T as shown in Eq. 12. Using cross-ratio basis say $\mathcal{B} = \{r_{mjnk}(T') | j = 1, \ldots, m-1, k = 1, \ldots, n-1\}$, Eq. 12 can be simplified into a system with only independent $(m-1)(n-1)$, where only the equation defined by $r_{m1n1}(T')$ involves t'_{11}, the other equations can be combined into one equation by substitutions. Each of these two equations determines a hyperplane, hence $\Phi^{-1}(T')$ can be expressed as the intersection of two hyperplanes. Moreover, let the defining hyperplanes for $T^1, T^2 \in \mathbf{B}_a(T)$ be $\{H^1_1, H^1_2\}$ and $\{H^2_1, H^2_2\}$ respectively, where H^i_1 is the one defined by $c_{11} + c_{mn} - c_{m1} - c_{1n} = \ln(t_{m1} t_{1n}) - \ln(t^i_{11} t_{mn})$, where $i = 1, 2$. Then the Euclidean distance between H^1_1, H^1_1 is $d(H^1_1, H^1_1) = |\ln t^1_{1,1} - \ln t^2_{1,1}| \leq |\ln(t_1 1 + a) - \ln(t_1 1 - a)|$. Notice that $H^1_2 = H^2_2$. So the distance between $\Phi^{-1}(T^1), \Phi^{-1}(T^2)$ is bounded by $(\ln(t_{11} + a) - \ln(t_{11} - a))/\sin\theta$. □

Proposition 3.19. *Let T be an observed coupling of dimension $m \times n$ with Gaussian noise on t_{11}. Further, let $\mathcal{B}$ be a basis for cross ratios of $m \times n$ matrices, that contains only one cross ratio depending on t_{11}. Eliminate the cross ratio depending on t_{11} in $\mathcal{B}$, denote the new set by $\mathcal{B}^-$. Then:*

$$supp[P(C|T)] = \{C | r(K) = r(T) \text{ for } r \in \mathcal{B}^-\}$$

In particular, $P(C|T)$ is supported on a hyperplane that is one dimensional higher than $P(C|T^*)$.

Proof. Since the domain for $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is $(-\infty, \infty)$, hence the domain for possible t^*_{11} is $(0, \infty)$. Thus the observation T essentially put no constraint on the cross ratio of T^* depends on t^*_{11}. Hence $\text{supp}[P(C|T)] = \{C | r(K) = r(T) \text{ for } r \in \mathcal{B}^-\}$. As only one equation was eliminated, so $P(C|T)$ is supported on a hyperplane that is one dimensional higher than $P(C|T^*)$. □

B. Additional details of experiments

B.1. Auto-correlation function.

To monitor the efficiency of our MC methods, we compute the auto-correlation function during the burn-in phase and choose the lags accordingly.

The auto-correlation function is defined as

$$R(t) = \frac{1}{(N-t)\sigma^2} \sum_{l=1}^{N-t} \sum_{i,j} (K^{(l)}_{i,j} - \bar{K}_{i,j}) \cdot (K^{(l+t)}_{i,j} - \bar{K}_{i,j}), \tag{15}$$

where N is the total number of samples generated, (i,j) runs over all indices of the matrices K, $\bar{K}$ is the mean averaging over all the samples, and $\sigma^2_K = \sum_{i,j} \sigma^2_{K,i,j}$ is the variance.

B.2. Re-normalization of T for MetroMC over prior $\mathbf{P}_1$.

For using this method, not every random T satisfies the condition such that $C^{(0)} = -\ln(K^{(0)}) = -\ln(T)$ lies on the support of the posterior, therefore, we need to re-normalize T so that $C^{(0)} - \ln[T/F(T, \lambda)]$ is on the support. $F(T, \lambda)$ is defined as

$$F(T, \lambda) = \exp\left(\frac{\lambda + \sum_{i=1}^m \sum_{j=1}^n \ln(T_{i,j})}{m * n} \right). \tag{16}$$

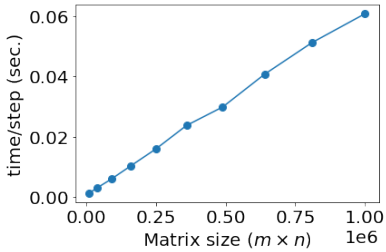

Figure 7: PIOT running time per MC step v.s. matrix size.

B.3. Running time of PIOT method.

The PIOT inference has a linear computational complexity of $\mathcal{O}(mn)$ (m and n are the dimensions of the matrix.) as shown in Fig. 7. The slope can be further reduced by optimizing the implementation. Prediction uses the entropy-regularized OT (EOT). Note that our algorithm is highly parallelizable.

B.4. MHMC on uniform matrix.

In this section, we validate our MHMC algorithm by apply it to an uniform T, i.e $T_{m,n} = 1/m$ under prior $\mathbf{P}_2$. The cross-ratios of T are all 1 and stay the same for all inferred sample $K^{(i)}$. Thus, each $K^{(i)}$ has same copies of n columns. We compare posterior sampled by MHMC to a m-dimensional uniform symmetric Dirichlet distribution generated by the SciPy package (Virtanen et al., 2020).

We perform our MHMC algorithm on a 3×3 uniform matrix with prior P_2 and the Dirichlet concentration parameter α are the same for every element. We choose $\alpha = 1$ for the Dirichlet parameter, $\sigma_0 = 0.5$, $\gamma = 3$, and $\delta = 1.0$. We run for 10,000 burn-in steps and take 10,000 samples with lags of 100. The result is shown in Fig. 8. We choose the lag by observing the auto-correlation function in the burn in, where the chain becomes uncorrelated after ~ 100 steps. Also, the running averages of the row sums are stable at taking 10,000 samples, which means 10,000 samples can effectively represent the posterior distribution.

The posterior distribution (Shown in top right of Fig. 8.) is compared to the distribution $P(K|T)$ of three-dimensional vectors generated by Dirichlet vector sampler (Virtanen et al., 2020) with symmetric $\alpha = 1$ in top left of Fig. 8. The results are comparable, which suggests that our MHMC method is able to generate distributions which well represent the support of the posterior distributions.

B.5. MetroMC on uniform matrix.

We test MetroMC on 3x3 uniform T, where $T_{i,j} = 1/(m * n)$, and assume the prior $\mathbf{P}_1$ is put on C. We take $K^{(0)} = T/F(T, \lambda = 1)$. For each MC iteration we sample $m + n - 1$ Gaussian random numbers with standard deviations of 0.02 for the diagonal matrices D^r and D^c satisfying the condition

$$\sum_{i,j} \ln(d_i^r) + \ln(d_j^c) = 0. \tag{17}$$

We burn in 10,000 steps and take 100,000 samples with 100 lags in between. The Gaussian Kernel density estimation (KDE) of each component of $P(C|T)$ is demonstrated in Fig. 9. We use bandwidths of 0.05 for all the Gaussian KDE in this article. There are a few notable features. First, we assume a Dirichlet prior on the whole C matrix, the distributions will be smaller at large values because each large value will suppress the value of the other elements. Second, since we have cross-ratio = 1 everywhere in T and prior $\mathbf{P}_1$ on C, there will not be a single element with value close to one. Otherwise, the cross-ratio involving that element could never be 1 anymore. Also, for the same reason all distributions should be similar.

B.6. Visualizing supp$[P(K|T)]$ with incomplete observations.

Fig. 10 illustrates the supp$[P(K|T)]$ for a 3×3 observation T with t_{31} missing under prior $\mathbf{P}_2$. Columns of three sampled $K \in$ supp$[P(K|T)]$ are plotted in three colored copies Δ_2 as shown. Here, three $\mathbf{k}_2$ are sampled uniformly from the middle Δ_2. The uniquely determined $\mathbf{k}_3$ are shown in the right Δ_2. The corresponding set for $\mathbf{k}_1$ ($\mathcal{K}_1$) is plotted in the left Δ_2. $\mathcal{K}_1$ is obtained by uniformly sample a thousand t_{31} from $[0, 10]$. Each t_{31} uniquely determines a $\mathbf{k}_1$. As explained in

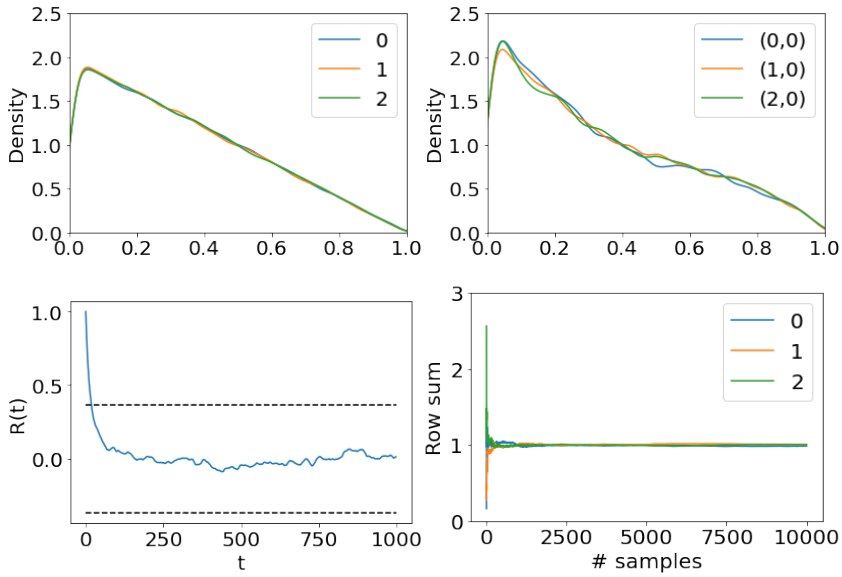

Figure 8: **(Top left.)** The distributions of each component of three-dimensional Dirichlet vectors with symmetric $\alpha = 1$ generated using SciPy package. **(Top right.)** The distributions of each component of the first column of $P(K|T)$ sampled by the MHMC method. **(Lower left.)** The auto-correlation function of the MHMC simulation. The black dashed lines indicate the range $[-1/e, 1/e]$. **(Lower right.)** The running average of the row sums of the MHMC simulation.

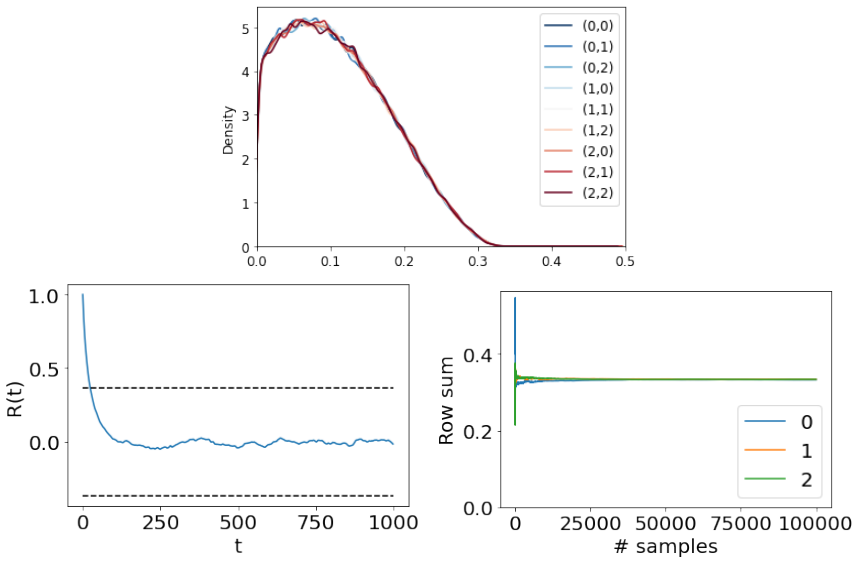

Figure 9: **(Top.)** The distribution of each matrix element of $P(C|T)$ with a uniform T and a prior $\mathbf{P}_1$ over C matrix. **(Bottom left.)** The autocorrelation function of the simulation. **(Bottom right.)** The running averages of the row sums.

Corollary 3.13, each $\mathcal{K}_1$ forms a line segments. The slope of each line is determined by T's cross ratios.

B.7. Monte Carlo results for synthetic symmetric costs.

For the symmetric costs we choose a constrained prior: $P_0(C)$ follows a Gibbs distribution $\exp(-\beta(\| \gamma(C - C^T) \|)$. We take $\beta = 10$ and $\gamma = 10^6$. We use $\lambda = 10$ in this example. Before the MCMC simulation we first pre-process $K^{(0)} = \text{diag}(D^{r_0})F(\frac{T^{(0)}}{\mu\nu^T}, \lambda = 10)\text{diag}(D^{c_0})$, where $\text{diag}(D^{r_0})$ and $\text{diag}(D^{c_0})$ are diagonal matrices such that the diagonals of $K^{(0)}$ are 1 and $F(\cdot)$ is the re-normalization function as described in 16. Then, during MetroMC we sample $D^r = \exp(N(0, \sigma * \mathbb{I}_n))$ and set $D^c = 1/D^r$. We use $\sigma = 10^{-6}$ for the burn-in steps and $\sigma = 10^{-7}$. We burn-in 100,000 steps and take 1,000 samples with lags of 300 steps. The acceptance ratios are 0.6-0.7. The KDE of the sampled posterior

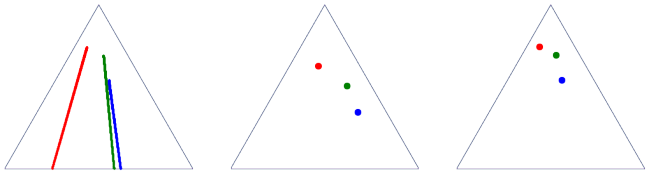

Figure 10: Visualization of supp$[P(K|T)]$ for a 3×3 observed T with t_{31} missing. Three sampled $K \in$ supp$[P(K|T)]$ are shown here with each column is plotted in a copy of Δ_2 as shown. Colored dots in the middle Δ_2 represents uniformly sampled $\mathbf{k}_2$'s. The uniquely determined $\mathbf{k}_3$'s are shown on the right, the set $\mathcal{K}_1$ for the corresponding $\mathbf{k}_1$'s is shown on the left.

distribution $C_{2,7}$ are shown in Fig. 11 for $p = 0.5$ (left), 1 (middle), and 2 (right). The blue vertical lines indicate value of the the ground truth cost and the red dashed lines are the median of the distributions. The averaged inferred $C_{2,7}$ at 300 iteration over 20 instances using algorithm from (Ma et al., 2020) is shown in green dashed lines. At 1000 iteration, their algorithm converge to the ground truth cost. The other elements show similar behaviors. Note that the scale in x-axis is 10^{-7}, so the differences between the ground truth and the inferred costs are small. The PIOT achieves achieve accurate result with relative error falls below 10^{-5} in all cases.

Next, to test the robustness we add $n \times n$ random matrices with all entries in $[0, 10^{-4})$ to the symmetric costs and then normalize the combined matrices to 1 as the cost functions. We use the prior $\mathbf{P_1}$ as described in section 3. The hyper-parameter α is set to 1 for all entries. We pre-process $K^{(0)} = F(\text{diag}(D^{r_0})\frac{T^{(0)}}{\mu\nu^T}\text{diag}(D^{c_0}), \lambda = 10)$, where $\text{diag}(D^{r_0})$ and $\text{diag}(D^{c_0})$ are diagonal matrices such that the diagonals of $\frac{T^{(0)}}{\mu\nu^T}$ are 1 and $F(\cdot)$ is the re-normalization function as described in 16. In MCMC we sample $D^r = \exp(N(0, \sigma * \mathbb{I}_n))$ and set $D^c = 1/D^r$. We choose $\sigma = 5 * 10^{-3}$, $5 * 10^{-3}$, and $1 * 10^{-3}$ for $p = 0.5$, 1, and 2, respectively. We burn-in 100,000 steps and take 10,000 samples with lags of 500 steps between samples. The inferred posterior distribution of $C_{1,9}$ is shown in Fig. 12. The blue solid lines indicate the values of the ground truth, and the red dashed lines are the medians of each distribution, and the green dash-dot lines and yellow dotted lines are averaged results over 20 instances using algorithm from (Ma et al., 2020) at 300 and 1000 iterations , respectively. The green and yellow lines overlap, meaning their algorithm converges at around 300 iteration. The inset shows a zoomed-in region of $C^g_{1,9} \pm 10^{-4}$. Overall, the PIOT accurately infers the cost function with smaller relative errors ($\sim 10^{-3}$ to 10^{-4}) than (Ma et al., 2020) ($\sim 10^{-2}$ to 10^{-3}) for all p.

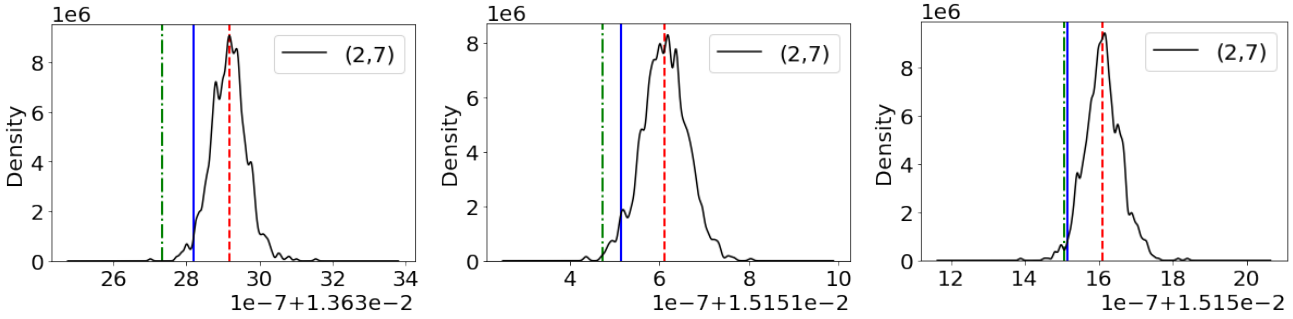

Figure 11: The KDE of the inferred posterior cost function of the matrix element $C_{2,7}$. The blue solid lines indicate the ground truth cost, the green dash-dotted lines are results using algorithm from (Ma et al., 2020) at 300-iteration, and the red dashed lines are the median of each inferred posterior distribution of the cost function. (Left.) $p = 0.5$. (Middle.) $p = 1$. (Right.) $p = 2$. At 1000 iteration, their algorithm converge to the ground truth cost.

B.8. Supporting plots for the MetroMC method on noisy T.

In this section, we provide data for diagnostics of the MetroMC simulations on the noise models in section 5 of the main text.

Bounded noise on observed plan. The autocorrelation function and the running averages of row sums of the samples for each noise are plotted in Fig. 13. In the simulation we set 100 for the lags by observing that the chains become uncorrelated after $\sim$ 100 steps for all cases. We take 100,000 samples in total for all cases since the running averages are stable at (or before) 100,000 MC steps, which means the samples reaches a stationary distribution.

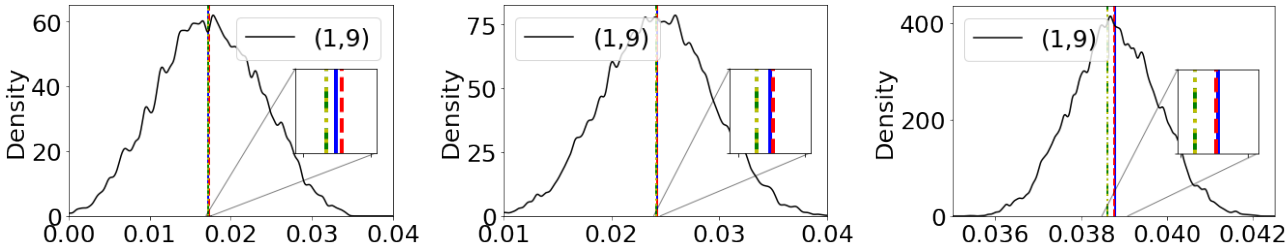

Figure 12: The KDE of the inferred posterior cost function of the matrix element $C_{1,9}$. The blue solid lines indicate the ground truth cost, the green dash-dotted lines are results using algorithm from (Ma et al., 2020) at 300-iteration, the yellow dotted lines are Ma's results at 1000-iteration (Overlapped with green lines for all cases.), and the red dashed lines are the median of each inferred posterior distribution of the cost function. (Left.) $p = 0.5$. (Middle.) $p = 1$. (Right.) $p = 2$. The insets have range of $C^g_{1,9} \pm 10^{-4}$.

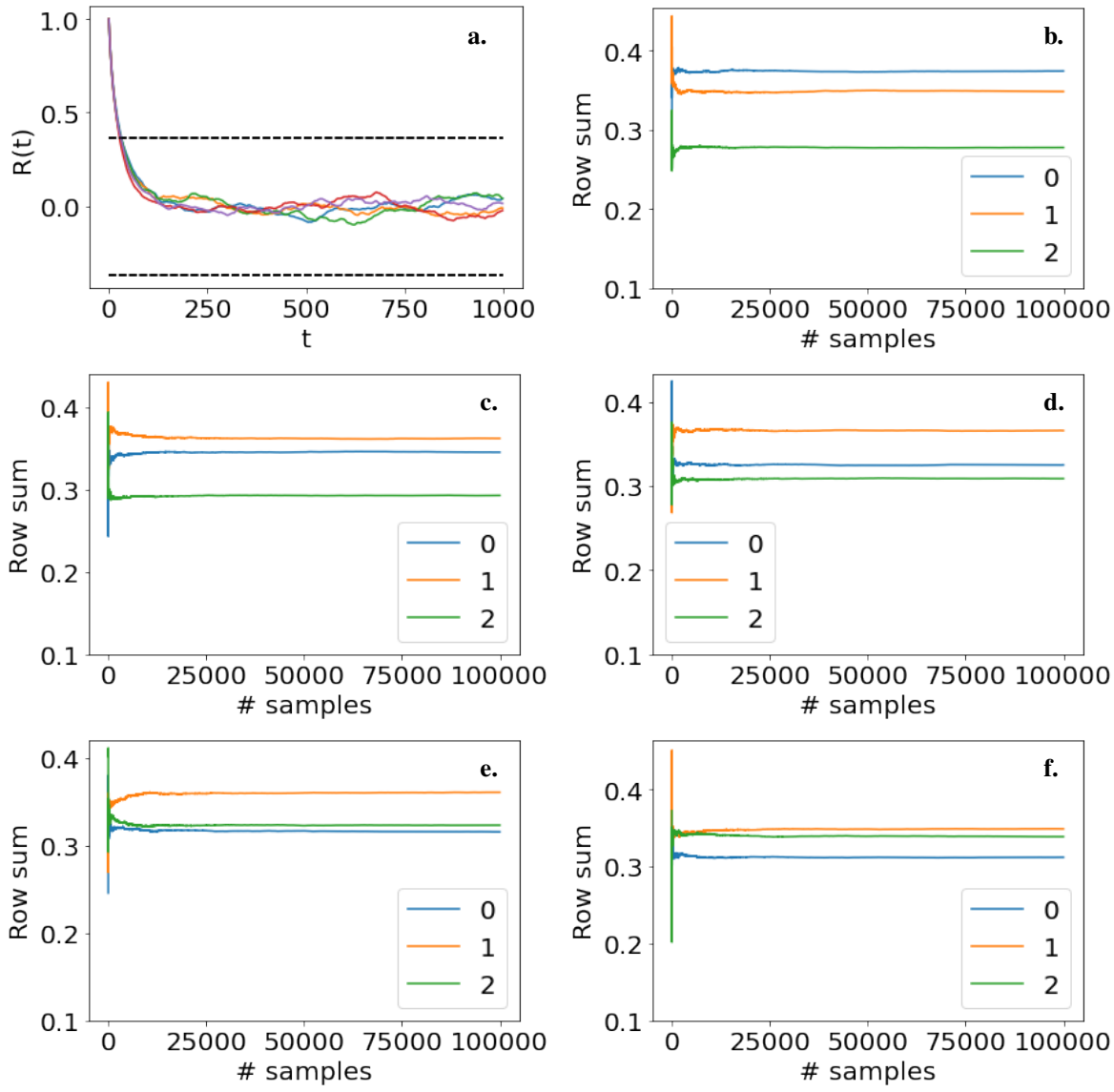

Figure 13: MC diagnostics for the Bounded noise example. **(a.)** The autocorrelation funciton for all noises. **(b.)-(f.)** The running averages of row sums of the samples for noises -0.01, -0.005, 0.0, 0.005, and 0.01 added to t_{11}, respectively.

Gaussian noise on observed plan. The autocorrelation function and the running averages of row sums of the samples for the 10 Gaussian noises are plotted in Fig. 14. We set 200 for the lags since the chains become uncorrelated after $\sim$ 180 steps

for all cases. 10,000 samples are taken for all cases since the running averages of the row sums are stable at 10,000 MC steps.

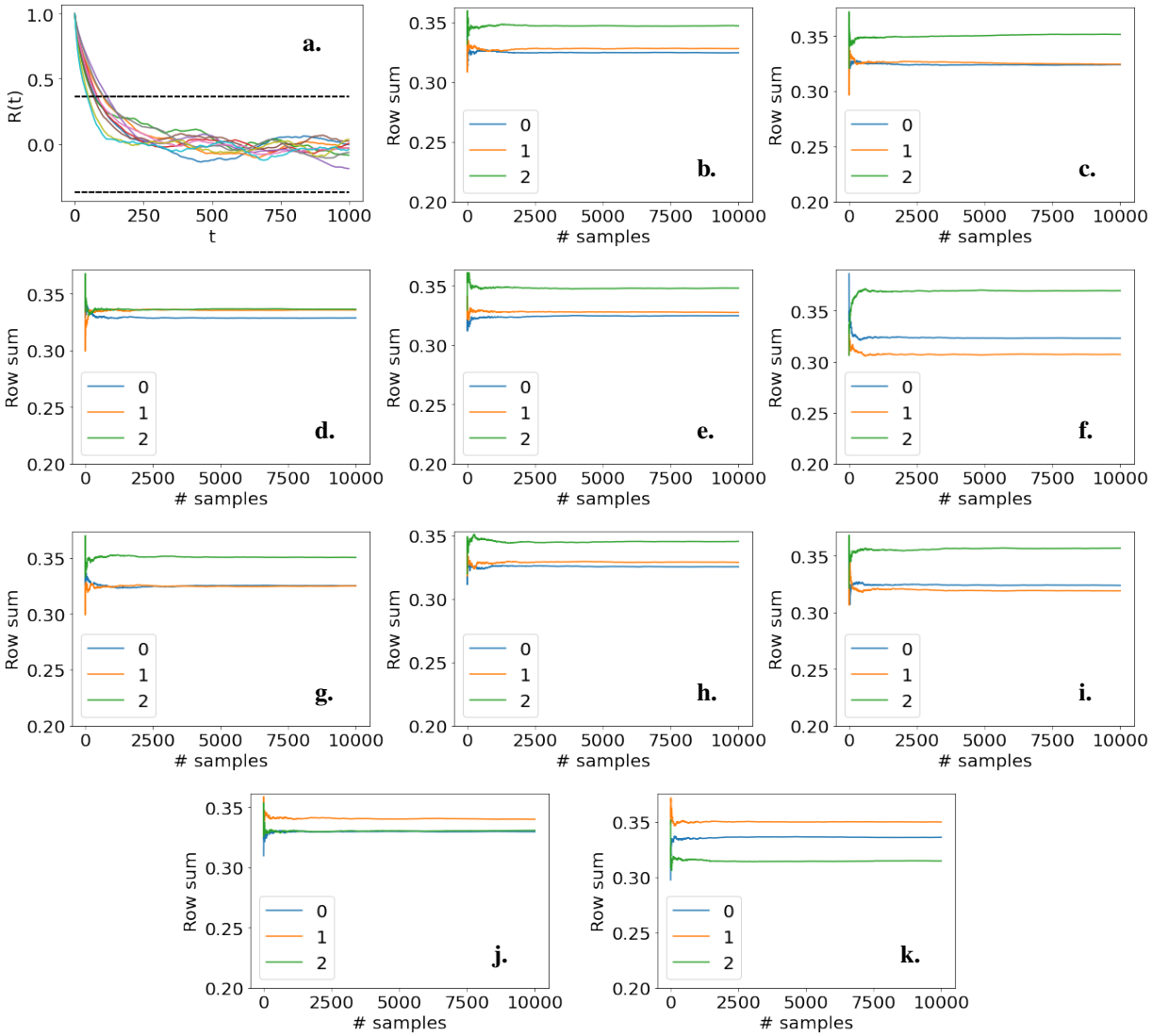

Figure 14: MC diagnostics for the Bounded noise example. **(a.)** The autocorrelation function for all noises. **(b.)-(k.)** The running averages of row sums of the samples for the 10 random Gaussian noises added to t_{12}.

B.9. Matrices.

We list the matrices used in section 5: **Simulations on synthetic data** below.

The three T matrices in Fig. 1 are

$$T_a = \begin{bmatrix} 0.3096 & 0.3785 & 0.0544 & 0.2575 \\ 0.2522 & 0.3203 & 0.1860 & 0.2415 \\ 0.4318 & 0.1433 & 0.4196 & 0.0053 \\ 0.0064 & 0.1579 & 0.3400 & 0.4957 \end{bmatrix},$$

$$T_b = \begin{bmatrix} 0.2532 & 0.4143 & 0.2894 & 0.0431 \\ 0.1925 & 0.0548 & 0.0958 & 0.6569 \\ 0.4459 & 0.0905 & 0.3480 & 0.1156 \\ 0.1083 & 0.4404 & 0.2669 & 0.1844 \end{bmatrix},$$

$$T_c = \begin{bmatrix} 0.4790 & 0.0994 & 0.0838 & 0.3378 \\ 0.1343 & 0.1514 & 0.1920 & 0.5224 \\ 0.1678 & 0.6182 & 0.1963 & 0.0177 \\ 0.2189 & 0.1310 & 0.5279 & 0.1222 \end{bmatrix}.$$

The matrix T in Fig. 10 is

$$T = \begin{bmatrix} 0.4583 & 0.2297 & 0.2633 \\ 0.4631 & 0.4785 & 0.2755 \\ 0.0785 & 0.2919 & 0.4611 \end{bmatrix}.$$

The matrices T in Fig. 2 are

$$T_1 = \begin{bmatrix} 0.1104 & 0.0684 & 0.1545 \\ 0.0505 & 0.2401 & 0.0428 \\ 0.1725 & 0.0249 & 0.1360 \end{bmatrix},$$

$$T_2 = \begin{bmatrix} 0.0950 & 0.1100 & 0.1283 \\ 0.1155 & 0.0343 & 0.1835 \\ 0.1228 & 0.1890 & 0.0215 \end{bmatrix},$$

$$T_3 = \begin{bmatrix} 0.1053 & 0.1193 & 0.1088 \\ 0.2148 & 0.0090 & 0.1096 \\ 0.0133 & 0.2051 & 0.1150 \end{bmatrix}.$$

The matrix T for the bounded noise case is

$$T = \begin{bmatrix} 0.1067 & 0.1141 & 0.1125 \\ 0.1175 & 0.1052 & 0.1106 \\ 0.1092 & 0.1139 & 0.1102 \end{bmatrix}.$$

The ground truth C^g in the Gaussian noise example is

$$C^g = \begin{bmatrix} 0.2604 & 0.0521 & 0.0104 \\ 0.0208 & 0.2604 & 0.0521 \\ 0.0625 & 0.0208 & 0.2604 \end{bmatrix},$$

and the hyperparameter matrix for the prior is

$$\alpha = \begin{bmatrix} 25.0 & 5.0 & 1.9 \\ 3.0 & 25.0 & 5.0 \\ 6.0 & 3.0 & 25.0 \end{bmatrix}.$$

B.10. Details of the MCMC simulations on European migration data.

For this part of simulations we consider the migration flows(Raymer et al., 2013) between the following 9 countries: Austria (AT), Belgium (BE), Switzerland (CH), Czech Republic (CZ), Germany (DE), Denmark (DK), France (FR), Luxembourg (LU) and Netherlands (NL). The migration flow matrix is shown in Fig. 15.

We utilize MetroMC method to infer the cost functions. In this section we assume that the sum of cost function is a constant. We use $\sum_{i,j} C_{i,j} = \sum_{i,j} \lambda C'_{i,j}$=320 and put Dirichlet priors on C' so that there exists cost functions following the Dirichlet priors can be inferred from the ground truth plan. For each MCMC simulation using the three different Dirichlet priors we burn-in 10,000 steps and then take 10,000 samples with lags of 1,000 between. The standard deviations of the independent Gaussian random numbers (Centered at 0.) in each MCMC iteration are 0.1, 0.1, and 0.07 for prior (1), (2), and (3), respectively. The acceptance rate for the MCMC simulations are 0.58-0.68. The autocorrelation function and the running average of each simulation are shown in Fig. 16. 1,000 steps are enough to de-correlate samples and 10,000 samples are able to reach convergence. The kernel density estimations (KDE) of each matrix element of the inferred cost functions are shown in Fig. 17. The mean of each element of the cost for the three priors are shown in Fig. 18.

For the noisy simulations on the EU migration flow we first generate a noise (ϵ) which is 4% of the value of $T^g_{(i,j)}$ and add it to $T^g_{(i,j)}$. Then, we sample ten Gaussian random numbers centered at 0 having a standard deviation equal to 4% of $T^g_{(i,j)}$. The generated numbers are shown in table 2. The MetroMC method is used for cost function inferences. In each Monte Carlo step we sample random variables $x \in \mathcal{N}(0, 0.1^2)$ for the diagonal matrices D^r and D^c. For each MCMC simulation we burn-in 10,000 steps and take 10,000 samples with a lag of 1,000 steps in between samples. The acceptance rates are 0.5-0.7. Once we infer the cost functions we compute the mean of them, C^M. Afterwards, forward OT simulation on C^M using the observed marginals is performed to obtain the predicted coupling, T^p. The results are shown in table 1.

We follow the same procedure and use the same parameters for the MCMC simulations of predicting the missing values, only instead of ten random Gaussian samples we take 100 samples uniformly in range of [100, 25000) to fill the missing value.

Table 2: Sampled Gaussian noises.

(I,J)	$T^g_{(i,j)}$	ϵ	NOISES									
(3,6)	452	3	7	-2	5	22	-22	-6	3	6	-24	5
(4,5)	6635	-131	-404	763	-118	-315	7	422	-16	66	-145	79
(5,7)	16560	-243	785	205	269	525	678	-164	217	-180	-379	-459

Figure 15: European migration flow matrix between the selected 9 countries.

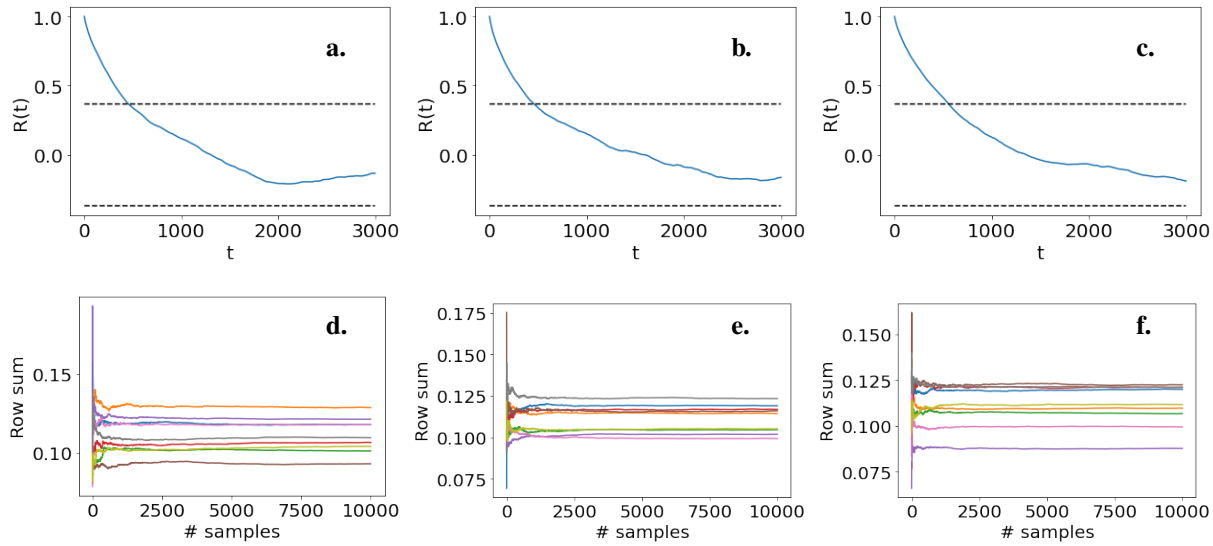

Figure 16: MC diagnostics for the European migration flow data. **(a.)-(c.)** are the autocorrelation function for using prior 1, 2, and 3, respectively. **(d.)-(f.)** are the running averages of row sums of the samples for using prior 1, 2, and 3.

Figure 17: The KDE of the inferred cost functions from the EU migration flow coupling. All 9×9 matrix elements are shown in order. The red curves are for prior (1), green for (2), and blue for (3).

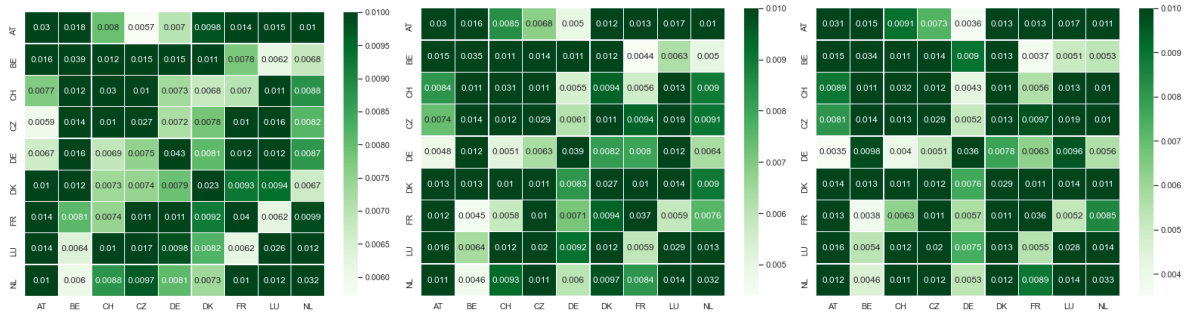

Figure 18: The heatmaps of the mean of the inferred cost functions using **(left)** prior 1, **(middle)** prior 2, and **(right)** prior 3.