Fast and Flexible Inference of Joint Distributions from their Marginals

Charlie Frogner [1] **Tomaso Poggio** [2]

Abstract

Across the social sciences and elsewhere, practitioners frequently have to reason about relationships between random variables, despite lacking joint observations of the variables. This is sometimes called an "ecological" inference; given samples from the marginal distributions of the variables, one attempts to infer their joint distribution. The problem is inherently ill-posed, yet only a few models have been proposed for bringing prior information into the problem, often relying on restrictive or unrealistic assumptions and lacking a unified approach. In this paper, we treat the inference problem generally and propose a unified class of models that encompasses some of those previously proposed while including many new ones. Previous work has relied on either relaxation or approximate inference via MCMC, with the latter known to mix prohibitively slowly for this type of problem. Here we instead give a single exact inference algorithm that works for the entire model class via an efficient fixed point iteration called Dykstra's method. We investigate empirically both the computational cost of our algorithm and the accuracy of the new models on real datasets, showing favorable performance in both cases and illustrating the impact of increased flexibility in modeling enabled by this work.

1. Introduction

Reasoning about relationships between random variables is fundamental in the sciences and in machine learning and typically relies on joint observations of the variables simultaneously; supervised learning, for example, relates features to labels using samples from their joint distribution. There are settings, however, in which we do not have access to joint observations and instead must rely on observations of the variables taken separately. Political scientists, for example, often attempt to estimate the impact of demographics on voting behavior, despite census data and vote counts being collected separately. Unsurprisingly, this inference is ill-posed: many possible relationships exist that can account for the observed data. Yet, by bringing prior information into the problem, it is still possible to make meaningful inferences. This is sometimes called **cross-level** or **ecological** inference.

This type of inference appears in a variety of fields, sometimes with substantial social impact. Federal voting rights cases in the U.S., which prevent the drawing of voting districts that dilute a minority's vote, depend critically on establishing a relationship between race and political preference, by solving exactly this inference problem (King, 1997; Greiner, 2006). Such inference is also carried out by epidemiologists (Morgenstern, 1995), biostatisticians (Jackson et al., 2006; Wakefield, 2008), geographers (Johnston & Pattie, 2006), ecologists (Martin et al., 2005), economists (Honaker, 2008), and climate scientists (Piguet, 2010).

Despite its prevalence, only a handful of methods have been proposed for solving the inference problem. This may be due to its inherent computational challenges: It is, in fact, a strict generalization of **optimal transport** (Villani, 2003), which arises when one assumes perfectly-observed marginal distributions and a particular linear cost criterion. Beyond optimal transport, the few proposed methods rely either on relaxing the problem (so that the result need not be a probability distribution, for example) or on approximate inference via MCMC.

In this paper, we expand the set of available methods substantially, by defining general-purpose inference algorithms that work across a broad class of probability models. This model class, in fact, includes some of the previous work as special cases. Importantly, we show that maximum a posteriori inference within this class requires neither relaxation nor MCMC, but rather can be done using a single exact algorithm called Dykstra's method.

The inference methods we propose are both computationally efficient and applicable to real data, with new models enabled by this work sometimes achieving more accurate inferences than those previously available (Section 6).

[1] CSAIL, Massachusetts Institute of Technology, Cambridge, Massachusetts [2] Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, Massachusetts. Correspondence to: Charlie Frogner <frogner@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

2. Preliminaries

2.1. The inference problem

Our goal is to estimate the joint distribution of two categorical random variables $X : \Omega \to \{1, \ldots, m\}$ and $Y : \Omega \to \{1, \ldots, n\}$, given data consisting of samples from their marginal distributions, $\{\mathbf{x}^{(k)}\}_{k=1}^{N_X} \sim \mathcal{P}_X$, $\{\mathbf{y}^{(\ell)}\}_{\ell=1}^{N_Y} \sim \mathcal{P}_Y$.

We refer to the estimation target as the **table** of joint probabilities, being a matrix $\pi \in \mathbb{R}_+^{m \times n}$ whose row and column marginals are the marginal densities of X and Y, respectively. In the simplest case, we assume we are given perfect observations of the two marginal distributions: vectors $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$ whose elements give the exact probabilities of the categories over which X and Y are defined. In this case, estimation entails finding a nonnegative matrix π satisfying marginal constraints $\pi\mathbf{1} = \mathbf{u}$ and $\pi^\intercal\mathbf{1} = \mathbf{v}$. This is an ill-posed problem: there are in general many possible joint distributions that match any given pair of marginals. Prior information is needed to identify a unique solution.

Sections 3 and 4 explore the perfectly-observed setting. Section 5.2 discusses generalizations to noisy observations.

2.2. Existing methods

The simplest model for inferring a joint distribution from marginals is the **independent model**, which assumes independence of the two random variables: given perfectly-observed vectors of probabilities $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$, this estimates their joint table as $\pi = \mathbf{u}\mathbf{v}^\intercal$.

Other existing models either are constrained to 2×2 tables (notably King's method (King, 1997)) or assume that one of the two marginals is observed inexactly, as via limited samples from the marginal distribution or with noise. The latter may be due to the **difficulty of MCMC inference** on the set of joint distributions with perfectly-observed marginals: the fastest mixing known algorithm is the Vaidya random walk (Chen et al., 2017), whose complexity scales as $\mathcal{O}(n^9)$ (n being the cardinality of the larger of the two categorical spaces) – this is the subject of some recent interest (Kannan & Narayanan, 2012; Lee & Vempala, 2017).

Note that a number of models for joint distributions, including maximum entropy models (Della Pietra et al., 1997; Dudík & Schapire, 2006), copulas (Sklar, 1959; Nelsen, 2007), and low-rank tensors (Kargas & Sidiropoulos, 2017), have never (to our knowledge) been adapted to the problem studied here, which assumes we can only access the marginal distributions of the variables.

For tables of general size, we note three methods: Goodman's regression, the multinomial-Dirichlet model, and regularized optimal transport. The latter two are special cases of the model class proposed in this paper.

Goodman's regression (Goodman, 1953) was among the first proposed methods for this problem and, along with various generalizations (Kousser, 1973; Loewen, 1982; Kleppner, 1985; Grofman et al., 1985; Achen & Shively, 1995), has been widely used in practice (King et al., 2004). The model assumes that one of the marginals is observed perfectly, while the other is observed with additive Gaussian noise. If $\mathbf{u}$ is the perfect observation, the true joint table π decomposes as $\pi = \operatorname{diag}(\mathbf{u})\rho$ with $\rho \in \mathbb{R}_+^{m \times n}$ a row-stochastic matrix giving the conditional probabilities of the column marginal's categories. Under the Gaussian noise assumption, we have a linear regression model for estimating ρ,

$$\mathbf{v} = \rho^\intercal\mathbf{u} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

with $\mathbf{v}$ the observation of the second marginal. This is straightforwardly solved via ordinary least squares. Note that the problem is underdetermined, requiring either an assumption of shared conditional probabilities ρ amongst many observations or a prior on ρ. Note also that the estimate is unconstrained, such that the estimated conditional probabilities in fact need not lie in the range $[0, 1]$, making interpretation difficult.

The **multinomial-Dirichlet model** (Rosen et al., 2001; Wakefield, 2004) also assumes that one marginal is observed perfectly, and defines a hierarchical distribution on the histogram of sample counts for samples drawn from the other marginal distribution. The model has the column marginal histogram $\hat{\mathbf{v}}$ multinomial distributed, with the conditional probabilities ρ governed by a Dirichlet distribution. When the Dirichlet distribution parameters are not specified, the model includes a Gamma prior, yielding

$$\begin{aligned} \hat{\mathbf{v}} &\sim \operatorname{Multinom}(N, \rho^\intercal\mathbf{u}), \\ \rho_{i,\cdot} &\sim \operatorname{Dir}(\alpha_{i,\cdot}), \;\; \forall i \\ \alpha_{i,j} &\sim \operatorname{Gamma}(\lambda_1, \lambda_2), \;\; \forall i, j. \end{aligned}$$

Inference is done with a Metropolis-within-Gibbs algorithm.

Optimal transport (Villani, 2003) is a special case of the problem considered in this paper. The theory studies transport plans that distribute the mass from one marginal distribution to match the other. The optimal transport plan minimizes the total cost of moving the mass. Specifically, for perfectly-observed marginal distributions $\mathbf{u}$ and $\mathbf{v}$, the optimal transport plan is exactly the soft assignment matrix π that solves

$$\underset{\pi \in U(\mathbf{u},\mathbf{v})}{\operatorname{minimize}} \langle \pi, \mathbf{C} \rangle_{\mathcal{F}}, \tag{2}$$

with $\mathbf{C} \in \mathbb{R}_+^{m \times n}$ the cost matrix having $\mathbf{C}_{ij}$ the cost for transporting a unit of mass from the ith to the jth category, and $U(\mathbf{u}, \mathbf{v})$ the polytope of nonnegative matrices having $\mathbf{u}$ and $\mathbf{v}$ as row and column marginals.

A regularized form of optimal transport has, in fact, been applied to the problem considered here (Muzellec et al., 2017)

– this type of regularized transport falls within the model class we propose in Section 3. Note also that Dykstra's method (Section 4) has been applied to optimal transport problems, in (Benamou et al., 2015; Dessein et al., 2018).

3. A General Class of Models

3.1. Well-behaved priors

Our model relies on a prior distribution over tables, which we allow to come from a general class of distributions, being those that are **separable** and **log-concave of Legendre type**, with support containing $\operatorname{int}(\Delta^{m\times n})$. With two additional technical assumptions, these distributions are sufficiently well-behaved to enable the efficient optimization in Section 4.

Let $\mathcal{P}_\pi(\mathbf{C})$ be a family of distributions over $\operatorname{int}(\Delta^{m\times n})$, parameterized by $\mathbf{C} \in \mathbb{R}^{m\times n}$, and let $\Pr(\pi|\mathbf{C})$ denote the density with respect to the Lebesgue measure. Define $Q(\pi) = -\log \Pr(\pi|\mathbf{C})$ the negative log density, for tables $\pi \in \operatorname{dom} Q \subseteq \mathbb{R}^{m\times n}$. Formally, we assume the following.

$$\begin{aligned} &(A1)\ \ Q \text{ is separable and Legendre type.}\\ &(A2)\ \ \operatorname{int}(\Delta^{m\times n}) \subseteq \operatorname{dom} Q.\\ &(A3)\ \ \operatorname{dom} Q^* \text{ is open.}\\ &(A4)\ \ \mathbf{0} \in \operatorname{dom} Q^*. \end{aligned} \tag{3}$$

Here Q^* is the convex conjugate [1]. For certain priors (such as the Dirichlet prior, Section 5.1) we will drop the assumption A4.

We say $\Pr(\pi|\mathbf{C})$ is **separable** if it decomposes as $\Pr(\pi|\mathbf{C}) \propto \prod_{ij} f(\pi_{ij}|\mathbf{C}_{ij})$, with $f : \mathbb{R}\times\mathbb{R} \to [0, +\infty]$ a one-dimensional density.

$\Pr(\pi|\mathbf{C})$ is **log-concave of Legendre type** if its negative log is convex of Legendre type. Namely, the negative log is closed, proper, essentially smooth and strictly convex on the interior of its domain [2]. Separability and log-Legendreness of $\Pr(\pi|\mathbf{C})$ are satisfied by a number of common distributions, including the component-wise **normal**, **gamma**, **beta**, **chi-square**, **logistic**, and **Weibull distributions**.

One property of Legendre-type functions we will exploit is duality between the domain and range of the gradient ∇Q. Specifically, for a Legendre-type Q, the gradient of Q and that of the convex conjugate Q^* define a bijection between $\operatorname{int}(\operatorname{dom} Q)$ and $\operatorname{int}(\operatorname{dom} Q^*)$, with $\nabla Q^* = (\nabla Q)^{-1}$. We formulate, for example, the MAP estimation procedure in Section 4 as an optimization over the dual space $\operatorname{dom} Q^*$, and recover the primal solution via the map ∇Q^*.

3.2. Probability model

We assume a common prior distribution $\mathcal{P}_\pi(\mathbf{C})$ for N tables $\pi^{(i)}$, which satisfies the regularity properties (3). The tables represent different but related instances of the problem, corresponding for example to different geographic regions in the voter preference example. We assume the distribution's parameters $\mathbf{C} \in \mathbb{R}^{m\times n}$ are shared across instances. The model specifies

$$\begin{aligned} &\pi^{(i)} \sim \mathcal{P}_\pi(\mathbf{C}),\\ &\pi^{(i)} \perp\!\!\!\perp \pi^{(j)} \mid \mathbf{C}, \forall i \neq j. \end{aligned} \tag{4}$$

Conditioned on the observed vectors $\mathbf{u}^{(i)} \in \Delta^m$ and $\mathbf{v}^{(i)} \in \Delta^n$, we will draw inferences about the posterior density $\Pr(\pi|\mathbf{u}^{(i)}, \mathbf{v}^{(i)}, \mathbf{C})$. With perfect observations, the posterior is the truncation of the prior to the polytope of nonnegative tables that are exactly consistent with the observations.

3.3. Specifying the model in practice

In practice, we are combining two datasets, often collected separately but describing the same population; these specify the marginals $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$ above. Political scientists, for example, relate demographics to voting behavior by combining census data and vote counts. Much recent work attempts to use additional measurements of the population to inform the inference problem: These may be voter registration records (Imai & Khanna, 2016) or exit polls (Greiner & Quinn, 2010), for example, which measure a proxy for the underlying joint distribution that is our target. In our framework, the prior $\mathcal{P}_\pi(\mathbf{C})$ may be fit to such proxy data by maximum likelihood:

$$\hat{\mathbf{C}} = \max_{\mathbf{C}} \prod_{i=1}^{N} \Pr(\hat{\pi}^{(i)}|\mathbf{C}),$$

with $\hat{\pi}^{(i)}$ the estimated proxy for the joint distribution in the ith instance (e.g. precinct).

4. Maximum A Posteriori Inference

In the perfectly-observed case, each observed vector defines an affine constraint on the table $\pi^{(i)}$; for marginals $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$, we have $\pi^{(i)}\mathbf{1} = \mathbf{u}^{(i)}$ and $\pi^{(i)\intercal}\mathbf{1} = \mathbf{v}^{(i)}$. Together with the constraints that the entries of the table be nonnegative, these affine constraints define a convex polytope $U(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})$ of tables consistent with the observations. Maximum a posteriori estimation reduces to finding a table $\pi_*^{(i)}$ that maximizes the prior over this polytope:

$$\pi_*^{(i)} = \operatorname*{argmax}_{\pi \in U(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})} \Pr(\pi|\mathbf{C}).$$

[1] The convex conjugate $Q^* : \operatorname{dom} Q^* \to \mathbb{R}$ is defined

$$Q^*(\mathbf{u}) = \sup_{\mathbf{x}\in\operatorname{int}(\operatorname{dom} Q)} \langle \mathbf{u}, \mathbf{x}\rangle - Q(\mathbf{x}).$$

[2] Bauschke and Borwein (Bauschke et al., 1997) Def. 2.8 and surrounding gives a formal treatment of Legendre type functions.

4.1. MAP estimation is a Bregman projection

Key to the tractability of MAP estimation for priors satisfying the assumptions (3) is the fact that it can be formulated as minimization of a Bregman divergence over the polytope of marginal constraints. This is stated in Proposition 1.

Proposition 1 (MAP is divergence minimization)**.** *Let* $\Pr(\pi|\mathbf{C})$ *be a prior density over tables* $\pi \in \mathbb{R}^{m\times n}$*, satisfying the regularity properties* (3)*. Define* $Q(\pi) \triangleq -\log \Pr(\pi|\mathbf{C})$ *the negative log density. Then the posterior density* $\Pr(\pi|\mathbf{u},\mathbf{v},\mathbf{C})$ *has a unique maximum* π_* *which satisfies*

$$\pi_* = \operatorname*{argmin}_{\pi \in U(\mathbf{u},\mathbf{v})} \mathcal{D}_Q(\pi, \nabla Q^*(\mathbf{0})) \tag{5}$$

with Q^* *the convex conjugate and* $\mathcal{D}_Q$ *the Bregman divergence with respect to* Q*.*

In other words, the MAP estimate exists and is unique, and is a Bregman projection onto the set of constraints $U(\mathbf{u},\mathbf{v})$.

4.2. Dykstra's method

Casting MAP estimation as a Bregman projection (Proposition 1) suggests that we can apply efficient general methods for Bregman projections to compute the solution. In particular, we will use the Dykstra-Bregman ("Dykstra's") method (Bregman, 1967) of alternating projections, which has been applied in the matrix balancing (Sinkhorn & Knopp, 1967) and optimal transport (Cuturi, 2013; Benamou et al., 2015; Dessein et al., 2018) settings to obtain fast-converging iterative solvers.

Dykstra's method (Bregman, 1967) obtains the Bregman projection (5) onto $U(\mathbf{u},\mathbf{v})$ by decomposing the polytope into the intersection of three convex sets defined by the constraints,

$$\begin{aligned} \mathcal{C}_+ &= \mathbb{R}_+^{m\times n}, \\ \mathcal{C}_\mathbf{u} &= \{\pi \in \mathbb{R}^{m\times n} : \pi\mathbf{1} = \mathbf{u}\}, \\ \mathcal{C}_\mathbf{v} &= \{\pi \in \mathbb{R}^{m\times n} : \pi^\intercal\mathbf{1} = \mathbf{v}\}, \end{aligned} \tag{6}$$

such that $U(\mathbf{u},\mathbf{v}) = \mathcal{C}_+ \cap \mathcal{C}_\mathbf{u} \cap \mathcal{C}_\mathbf{v}$. The method alternates Bregman projections onto the constraints $\mathcal{C}_+$, $\mathcal{C}_\mathbf{u}$ and $\mathcal{C}_\mathbf{v}$ taken individually. In this case, a theorem of Bauschke and Lewis (Bauschke & Lewis, 2000) guarantees the alternating projections converge linearly to the Bregman projection onto $U(\mathbf{u},\mathbf{v})$. Algorithm 1 gives the generic form of Dykstra's method for this problem, with $\mathrm{P}_\mathcal{C}^Q$ indicating the Bregman projection onto $\mathcal{C}$. Note that the initial table (whose projection we are computing) depends on the particular prior density used: this is $\nabla Q^*(\mathbf{0})$ from (5).

Bregman projections are rarely computable in closed form, but for affine constraints they have a form suitable for iterative optimization. Write the Lagrangians for the projections of a table π' onto $\mathcal{C}_\mathbf{u}$ and $\mathcal{C}_\mathbf{v}$,

$$\mathcal{L}_\mathbf{u}(\pi,\alpha) = Q(\pi) - \langle \nabla Q(\pi'), \pi\rangle + \alpha^\intercal(\pi\mathbf{1} - \mathbf{u}), \tag{7}$$

$$\mathcal{L}_\mathbf{v}(\pi,\beta) = Q(\pi) - \langle \nabla Q(\pi'), \pi\rangle + \beta^\intercal(\pi^\intercal\mathbf{1} - \mathbf{v}). \tag{8}$$

For Q that is convex of Legendre type, the gradient map $\nabla Q : \mathrm{int}(\mathrm{dom}\, Q) \to \mathrm{int}(\mathrm{dom}\, Q^*)$ is a bijection, with the gradient of the conjugate ∇Q^* being the inverse of ∇Q. Applied to the first order conditions for (7) and (8), we get

$$\pi_\mathbf{u} = \mathrm{P}_{\mathcal{C}_\mathbf{u}}^Q\, \pi' = \nabla Q^*\left(\nabla Q(\pi') - \alpha\mathbf{1}^\intercal\right), \tag{9}$$

$$\pi_\mathbf{v} = \mathrm{P}_{\mathcal{C}_\mathbf{v}}^Q\, \pi' = \nabla Q^*\left(\nabla Q(\pi') - \mathbf{1}\beta^\intercal\right), \tag{10}$$

for $\pi_\mathbf{u}$ the projection of π' onto $\mathcal{C}_\mathbf{u}$, and $\pi_\mathbf{v}$ that onto $\mathcal{C}_\mathbf{v}$. Computing $\pi_\mathbf{u}$ reduces to finding α such that the original constraint holds,

$$\nabla Q^*\left(\nabla Q(\pi') - \alpha\mathbf{1}^\intercal\right)\mathbf{1} = \mathbf{u}, \tag{11}$$

and analogously for $\pi_\mathbf{v}$ and β. Dhillon and Tropp (Dhillon & Tropp, 2007) suggest a method for finding α. As Q is Legendre, Q^* is strictly convex, and α satisfying (11) is the unique optimum for a strictly convex problem,

$$\alpha_* = \operatorname*{argmin}_{\alpha\in\mathbb{R}^m} J_\mathbf{u}(\alpha) = Q^*\left(\nabla Q(\pi') - \alpha\mathbf{1}^\intercal\right) + \mathbf{u}^\intercal\alpha, \tag{12}$$

which can be addressed by standard iterative methods. The gradient of (12) exists, and the first order condition is exactly (11). When the Hessian of Q^* is available, for Q that is separable we have

$$\nabla^2 J_\mathbf{u}(\alpha) = \mathrm{diag}\left(\nabla^2 Q^*\left(\nabla Q(\pi') - \alpha\mathbf{1}^\intercal\right)\mathbf{1}\right). \tag{13}$$

The analogous equations hold for optimizing β.

The projection onto nonnegativity constraints $\mathcal{C}_+$, for Q separable, has a simple form. Projecting π' results in

$$\mathrm{P}_{\mathcal{C}_+}\, \pi' = \max\{0, \pi'\}. \tag{14}$$

Algorithm 2 gives a realization of Dykstra's method as an iteration on the dual variable $\Theta = \nabla Q(\pi)$, alternating projections onto the affine constraints $\mathcal{C}_\mathbf{u}$, $\mathcal{C}_\mathbf{v}$ with the nonnegativity constraint $\mathcal{C}_+$. Note that the projections (9) and (10) can be viewed as linearly updating the dual representation of the original table π'. We therefore need only represent Θ to compute the Dykstra iterations.

For solving (12), Algorithm 3 gives a Newton-Raphson method to compute the projection onto $\mathcal{C}_\mathbf{u}$. An analogous method works for $\mathcal{C}_\mathbf{v}$. Note that for some priors $\mathrm{int}(\mathrm{dom}\, Q^*)$ is a bounded subset of $\mathbb{R}^{m\times n}$, in which case backtracking can be used to ensure the bounds are respected.

Algorithm 1 Dykstra's method for MAP estimation

```
Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\pi_0 \in \operatorname{int}(\operatorname{dom} Q)$
$\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}(\pi_0)$
repeat
    $\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}\left(\mathrm{P}^Q_{\mathcal{C}_\mathbf{u}} \pi\right)$
    $\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}\left(\mathrm{P}^Q_{\mathcal{C}_\mathbf{v}} \pi\right)$
until $\pi$ converges
```

Algorithm 2 Dykstra's method for MAP estimation, dual parameterization

```
Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\Theta_0 \in \operatorname{int}(\operatorname{dom} Q^*)$
$\Theta \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta_0\}$
repeat
    $\alpha_* \leftarrow \operatorname{argmin}_{\alpha \in \mathbb{R}^m} Q^*(\Theta - \alpha \mathbf{1}^\intercal) + \mathbf{u}^\intercal \alpha$
    $\Theta \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta - \alpha_* \mathbf{1}^\intercal\}$
    $\beta_* \leftarrow \operatorname{argmin}_{\beta \in \mathbb{R}^n} Q^*(\Theta - \mathbf{1}\beta^\intercal) + \mathbf{v}^\intercal \beta$
    $\Theta \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta - \mathbf{1}\beta_*^\intercal\}$
until $\Theta$ converges
$\pi_* \leftarrow \nabla Q^*(\Theta)$
```

5. Expanding the Model Class

5.1. Expanding the class of priors: ϵ-MAP estimation

Example: Dirichlet prior. The Dirichlet distribution is a natural prior to use in the setting of probability table estimation, as it is supported on the simplex $\Delta^{m \times n}$, which is exactly the set of valid probability tables.

The Dirichlet distribution, however, is not quite regular: it fails assumption A4 from Section 3.1, and so Proposition 1 does not apply. The problem is that the negative log density (which has domain $\mathbb{R}^{m\times n}_{++}$) does not attain its global optimum – there is no finite $\pi \in \mathbb{R}^{m\times n}_{++}$ such that $\nabla Q(\pi) = \mathbf{0}$. We therefore have no starting point for the Bregman projection in Proposition 1.

The Dirichlet distribution, along with certain others failing A4, is still tractable in the following sense. For any small $\epsilon > 0$, we can find a table $\pi_\epsilon \in \operatorname{int}(\operatorname{dom} Q)$ such that $\|\nabla Q(\pi_\epsilon)\|_2 < \epsilon$. This, it turns out, is sufficient to guarantee that the Bregman projection of π_ϵ onto $U(\mathbf{u}, \mathbf{v})$ is $\varepsilon-$close to the MAP solution, for an appropriate ε. This is stated formally in Proposition 2.

Algorithm 3 Newton-Raphson method for the projection $\mathrm{P}_{\mathcal{C}_\mathbf{u}}$ onto a marginal constraint

```
Input: $\mathbf{u} \in \Delta^m$, $\Theta \in \operatorname{int}(\operatorname{dom} Q^*)^{m\times n}$, $\kappa > 0$
$\alpha \leftarrow \mathbf{0}$
repeat
    $\alpha \leftarrow \alpha - \kappa\left(\mathbf{u} - \nabla Q^*(\Theta - \alpha\mathbf{1}^\intercal)\mathbf{1}\right) \oslash \nabla^2 Q^*(\Theta - \alpha\mathbf{1}^\intercal)\mathbf{1}$
until $\alpha$ converges
$\Theta_* \leftarrow \Theta - \alpha\mathbf{1}^\intercal$
```

Proposition 2 ($\epsilon-$MAP estimation)**.** *Let* $\Pr(\pi|\mathbf{C})$ *be a prior density over tables* $\pi \in \mathbb{R}^{m\times n}$*, satisfying the regularity properties A1, A2 and A3 from* (3)*. Define* $Q(\pi) \triangleq -\log \Pr(\pi|\mathbf{C})$ *the negative log density, and suppose there exists* $\pi_\epsilon \in \operatorname{int}(\operatorname{dom} Q)$ *such that* $\|\nabla Q(\pi_\epsilon)\|_2 < \epsilon$*. Let* π'_ϵ *be its Bregman projection,*

$$\pi'_\epsilon = \operatorname*{argmin}_{\pi \in U(\mathbf{u},\mathbf{v})} \mathcal{D}_Q(\pi, \pi_\epsilon). \tag{15}$$

Then the posterior density $\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C})$ *has a unique maximum* π_* *that satisfies*

$$Q(\pi'_\epsilon) - Q(\pi_*) < \sqrt{2}\epsilon. \tag{16}$$

We can therefore do MAP inference by Dykstra's method (Algorithm 1), using π_ϵ as the initial table.

5.2. Extending to noisy observations

When we observe finite samples $\{\mathbf{x}^{(i)}\}_{i=1}^m$ and $\{\mathbf{y}^{(j)}\}_{j=1}^n$ from the marginal distributions of X and Y, they specify only imperfectly the true marginal distributions. Moreover, there may be additional noise – errors in counting votes, for example, in the voting preference case. To model both of these effects, we replace the hard marginal constraints $\pi^{(i)}\mathbf{1} = \mathbf{u}^{(i)}$ and $\pi^{(i)\intercal}\mathbf{1} = \mathbf{v}^{(i)}$ by **noise models**,

$$\mathbf{u}^{(i)} \sim \mathcal{P}_\mathbf{u}(\pi^{(i)}\mathbf{1}), \quad \mathbf{v}^{(i)} \sim \mathcal{P}_\mathbf{v}(\pi^{(i)\intercal}\mathbf{1}), \tag{17}$$

whose densities we assume to be *log-concave* and *log-coercive*, with domain containing $\operatorname{int}(\Delta^{m\times n})$. [3]

With noise included in our model, the maximum a posteriori table is now characterized by

$$\pi^{(i)} = \operatorname*{argmax}_{\pi \in \Delta^{m\times n}} \Pr(\mathbf{u}^{(i)}|\pi\mathbf{1}) \Pr(\mathbf{v}^{(i)}|\pi^\intercal\mathbf{1}) \Pr(\pi|\mathbf{C}). \tag{18}$$

Although the objective is more complex, MAP inference nevertheless retains a geometric interpretation: It is a **generalized Bregman projection** onto the probability simplex.

Proposition 3 (MAP with noise is a generalized projection)**.** *Let* $Q(\pi) \triangleq -\log \Pr(\pi|\mathbf{C})$, $\psi_\mathbf{u}(\pi) \triangleq -\log \Pr(\mathbf{u}|\pi\mathbf{1})$*, and* $\psi_\mathbf{v}(\pi) = -\log \Pr(\mathbf{v}|\pi^\intercal\mathbf{1})$ *be the negative log densities. Then the posterior density* $\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C})$ *has a unique global maximum which satisfies*

$$\pi_* = \operatorname*{argmin}_{\pi \in \Delta^{m\times n}} \psi_\mathbf{u}(\pi) + \psi_\mathbf{v}(\pi) + \mathcal{D}_Q(\pi, \nabla Q^*(\mathbf{0})), \tag{19}$$

with Q^* *the convex conjugate of* Q *and* $\mathcal{D}_Q$ *the Bregman divergence with respect to* Q*.*

[3] Note that, rather than being normalized to lie in the simplex, the observed vectors $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$ here might contain the original sample counts for each category, as the noise model can link the marginal probabilities $\pi^{(i)}\mathbf{1}$ and $\pi^{(i)\intercal}\mathbf{1}$ to the observed counts.

5.2.1. Dykstra's method for generalized projections

Peyré (Peyré, 2015) has recently shown that generalized projections of the form (19) can be solved by a **generalized Dykstra's method**. Analogous to the case of hard projections (Section 4), this involves alternating projections onto the two soft constraints defined by $\psi_\mathbf{u}$ and $\psi_\mathbf{v}$ together with the simplex constraint $\pi \in \Delta^{m\times n}$. We define each of these projections by a proximal operator,

$$\text{prox}_{\psi}^{\mathcal{D}_Q}(\xi) = \operatorname*{argmin}_{\pi\in\Delta^{m\times n}} \psi(\pi) + \mathcal{D}_Q(\pi, \xi). \tag{20}$$

The resulting generalized Dykstra's algorithm is shown in Algorithm 4.

(20) is evaluated similarly to the projections in Section 4. For simplicity, we will separate out projection onto the simplex, which can be alternated with the generalized projections onto $\psi_\mathbf{u}$ and $\psi_\mathbf{v}$. For soft constraint $\psi_\mathbf{u}$, the generalized projection of table $\xi \in \mathbb{R}^{m\times n}$ is

$$\pi_\mathbf{u} = \operatorname*{argmin}_{\pi\in\mathbb{R}^{m\times n}} Q(\pi) - \langle \nabla Q(\xi), \pi\rangle + \psi_\mathbf{u}(\pi), \tag{21}$$

and analogously for $\psi_\mathbf{v}$. This is a strictly convex, although possibly nonsmooth (depending on $\psi_\mathbf{u}$), problem, with a unique minimum. For smooth models, we can solve it via a Newton's method analogous to Algorithm 3.

For projecting $\xi \in \mathbb{R}^{m\times n}$ onto the simplex, we can formulate the Lagrangian,

$$\begin{aligned}&\mathcal{L}(\pi, \lambda, \varepsilon) = \\ &\quad Q(\pi) - \langle \nabla Q(\xi), \pi\rangle + \lambda(\sum_{ij} \pi_{ij} - 1) - \langle \varepsilon, \pi\rangle,\end{aligned} \tag{22}$$

which is solved by

$$\pi_* = \nabla Q^* \left(\nabla Q(\xi) - \lambda \mathbf{1}\mathbf{1}^\intercal + \varepsilon\right), \tag{23}$$

with $\varepsilon \geq \mathbf{0}$. λ is found by solving

$$\lambda_* = \operatorname*{argmin}_{\lambda\in\mathbb{R}} Q^* \left(\nabla Q(\xi) - \lambda \mathbf{1}\mathbf{1}^\intercal\right) + \lambda, \tag{24}$$

while ε solves

$$\varepsilon_* = \max\{\nabla Q(\mathbf{0}) - \nabla Q(\xi) + \lambda \mathbf{1}\mathbf{1}^\intercal, \mathbf{0}\}. \tag{25}$$

As in Section 4.2, these generalized projections in practice are computed in terms of a dual variable $\Theta \in \text{int}(Q^*)$, which is related to the primal via $\pi = \nabla Q^*(\Theta)$. Algorithm 5 shows the resulting algorithm.

Algorithm 4 Generalized Dykstra's method for MAP estimation with noisy data

```
Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\pi_0 \in \text{int}(\text{dom}\, Q)$
$\pi^{(0)} \leftarrow \pi_0, \mathbf{Z}_\mathbf{u}, \mathbf{Z}_\mathbf{v} \leftarrow \mathbf{0}$.
repeat
    $\pi^{(\ell+1)} \leftarrow \text{prox}_{\psi_\mathbf{u}}^{\mathcal{D}_Q} \left(\nabla Q^* \left(\nabla Q(\pi^{(\ell)}) + \mathbf{Z}_\mathbf{u}\right)\right)$
    $\mathbf{Z}_\mathbf{u} \leftarrow \mathbf{Z}_\mathbf{u} + \nabla Q(\pi^{(\ell)}) - \nabla Q(\pi^{(\ell+1)})$
    $\pi^{(\ell+2)} \leftarrow \text{prox}_{\psi_\mathbf{v}}^{\mathcal{D}_Q} \left(\nabla Q^* \left(\nabla Q(\pi^{(\ell+1)}) + \mathbf{Z}_\mathbf{v}\right)\right)$
    $\mathbf{Z}_\mathbf{v} \leftarrow \mathbf{Z}_\mathbf{v} + \nabla Q(\pi^{(\ell+1)}) - \nabla Q(\pi^{(\ell+2)})$
until $\pi^{(\ell)}$ converges
```

Algorithm 5 Generalized Dykstra's method for MAP estimation with noisy data, dual parameterization

```
Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\Theta_0 \in \text{int}(\text{dom}\, Q^*)$
$\Theta^{(0)} \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta_0\}, \mathbf{Z}_\mathbf{u}, \mathbf{Z}_\mathbf{v} \leftarrow \mathbf{0}$.
repeat
    (Alternate $(\psi, \mathbf{Z}) = (\psi_\mathbf{u}, \mathbf{Z}_\mathbf{u}), (\psi_\mathbf{v}, \mathbf{Z}_\mathbf{v})$.)
    $\pi_* \leftarrow \text{argmin}_{\pi\in\mathbb{R}^{m\times n}} Q(\pi) - \langle \Theta^{(\ell)} + \mathbf{Z}, \pi\rangle + \psi(\pi)$.
    $\lambda_* \leftarrow \text{argmin}_{\lambda\in\mathbb{R}} Q^*(\nabla Q(\pi_*) - \lambda \mathbf{1}\mathbf{1}^\intercal) + \lambda$.
    $\Theta^{(\ell+1)} \leftarrow \max\{\nabla Q(\mathbf{0}), \nabla Q(\pi_*) - \lambda_* \mathbf{1}\mathbf{1}^\intercal\}$
    $\mathbf{Z} \leftarrow \mathbf{Z} + \Theta^{(\ell+1)} - \Theta^{(\ell)}$
until $\Theta^{(\ell)}$ converges
$\pi_* \leftarrow \nabla Q^*(\Theta_*)$
```

5.3. Tertiary and higher-order relationships

Unlike prior work, the MAP inference methods we have outlined generalize naturally to multidimensional tables relating more than two marginals. We describe this for the perfectly-observed case, with the imperfect case directly analogous. Noting that Dykstra's method (Algorithm 1) alternates projections onto two marginal constraints, we can do the same, cycling through more than two constraints. Let $\{\mathbf{u}_k\}_{k=1}^K$ be a be a set of K marginals given as input data. Each marginal associates to an affine constraint $\mathcal{C}_{\mathbf{u}_k} = \{\pi \in \mathbb{R}^{m_1\times\cdots\times m_K} : \sum_{j_\ell, \ell\neq k} \pi_{j_1,\ldots,j_k,\ldots,j_K} = (\mathbf{u}_k)_{j_k} \forall j_k\}$. With these K constraints, the interior of Algorithm 1 becomes:

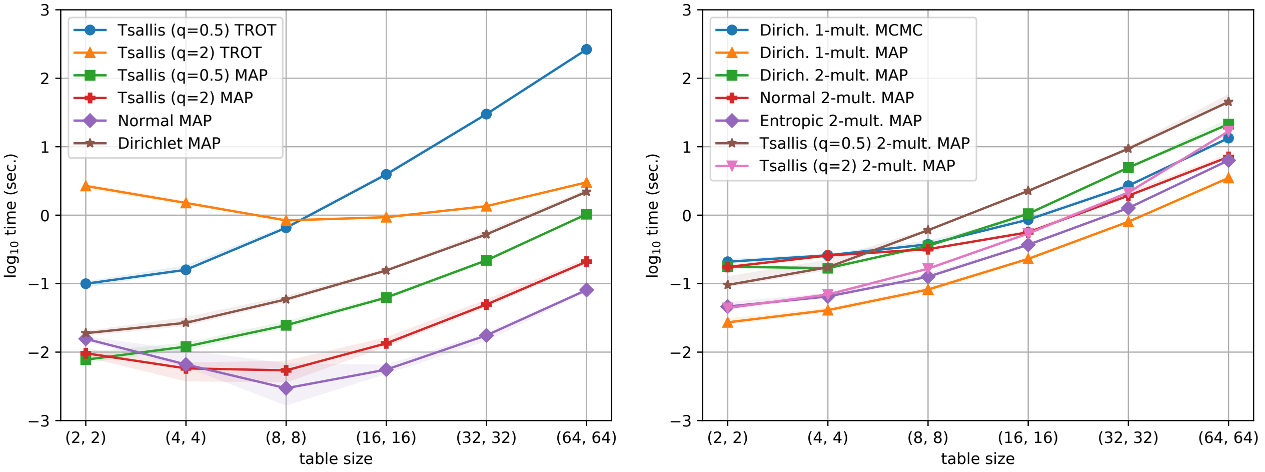
```
repeat
    $\pi \leftarrow \text{P}_{\mathcal{C}_+}\left(\text{P}^Q_{\mathcal{C}_{\mathbf{u}_k}} \pi\right), \quad k \leftarrow 1 + (k+1) \mod K$
until $\pi$ converges
```

This straightforward extension preserves the efficiency of the two-marginal case, converging linearly to the MAP table (Bauschke & Lewis, 2000).

6. Empirical Evaluation

6.1. Computational cost

The proposed methods are significantly more efficient than prior work. Figure 1 shows wall-clock times for inference in models assuming perfect observations of the marginals and in models assuming multinomial-distributed observations for at least one of the marginals. We compare the proposed method to two baselines, with the shown time being the total for inferring 100 tables from randomly-generated marginals,



(a) Models with perfectly-observed marginals.

(b) Models with multinomial-distributed marginals.

Figure 1. Runtimes, inferring 100 tables. The proposed inference method (MAP) is significantly faster than TROT (Muzellec et al., 2017) and MCMC (Rosen et al., 2001) for equivalent models. Inference for multinomial-distributed models is generally slower than for perfectly-observed models.

using uniform (all 1) values for the prior parameters. The proposed method ("MAP") is run until convergence [4]. Figure 1 shows the median time over 10 runs of each algorithm, with the shaded region the middle 80% [5].

Comparison to TROT. Muzellec et al. (2017) propose two methods for MAP inference, in the particular case of a Tsallis-distributed prior and perfectly-observed marginals. The first is for values of the Tsallis q parameter less than 1, the second for $q > 1$. We try both methods, setting $q = 0.5$ and $q = 2$, and compare to the proposed MAP inference for the same models [6]. In both cases, our proposed method is between one and two orders of magnitude faster.

Comparison to MCMC. Rosen et al. (2001) propose a Metropolis-Hastings sampler for the model having a Dirichlet prior, a single perfectly-observed marginal, and a single marginal observed with multinomial noise ("Dirich. 1-mult." in Figure 1b). We compare MAP inference by our proposed method to MCMC inference for this model, sampling a (relatively short) chain of length 1000, with no thinning and no burn-in. Note that, particularly for larger tables, a responsible application of MCMC will use a much larger number of samples, taking correspondingly longer to execute. Even with this small number of samples, however, our proposed method is faster by roughly a factor of three.

Relative efficiency of MAP inference for perfectly-observed models. The fastest inference methods out of all those evaluated are the proposed MAP algorithms for models with perfectly-observed marginals. For models with both marginals multinomial-distributed ("2-mult." in Figure 1b), the proposed algorithms tend to be slower, due to an additional projection onto the simplex at each iteration, although they are still comparable to (and sometimes faster than) the baselines.

6.2. Inference with real data

Unlike previous work, the inference methods described here apply simultaneously to a broad range of models (outlined in Sections 3 and 5). Here we demonstrate that this flexibility matters for modeling real data, with models newly available in our framework often outperforming those proposed previously.

We use five real datasets, each consisting of some number of joint samples of two categorical variables, conditioned on a geographic variable (such as a state or county). Each geographic location determines a table of counts from these samples, which we treat as the ground truth to be estimated, given the table's marginals as observed data. The datasets:

1. **CDC cause of death** by state, 1999-2016 (CDC, 2018): we relate cause of death amongst males (32 categories) to age group (11 categories).

2. **Indian educational attainment** by district, 2001 (India, 2018): we relate educational attainment (10 categories) to age group (22 categories).

[4] After each outer iteration we check the Frobenius norm of the deviation of the dual variable Θ from its previous value, halting when this deviation is less than `1e-4`. The inner Newton optimization is run for 20 steps within each outer iteration.

[5] All methods were implemented in `Python` using `numpy`, and were run on a MacBook Pro with a dual-core 2.9GHz processor. The TROT implementation is from the original author, available at `https://github.com/BorisMuzellec/TROT`.

[6] We set the parameter λ from Muzellec et al. (2017) to 1.

Table 1. Median absolute error of inferred table entries, real data.

Dataset	Death	Education	Voting	Income	Insurance
N. tables	51	110	68	51	51
Table dim.	11×32	22×10	7×3	11×9	20×4
Score scale	$1e{-}4$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$
Prior work					
Independent	3.94	2.17	1.39	14.0	4.93
Goodman	89.0	11.4	10.5	74.8	5.42
Mult. Dirich. (MCMC)	6.48	49.9	**1.05**	67.7	10.1
MAP, perfect					
Entropic	3.39	0.48	4.40	11.8	5.23
Tsallis ($q = 0.5$)	4.92	1.74	1.34	7.98	1.63
Tsallis ($q = 2$)	4.85	1.68	4.14	12.9	5.83
Normal	**1.58**	0.82	3.54	**5.72**	**0.92**
Dirichlet	6.32	1.56	3.76	51.7	**1.02**
MAP, multinomial					
Entropic	3.57	**0.41**	4.23	9.53	5.51
Tsallis ($q = 0.5$)	15.9	8.63	3.58	**6.17**	2.69
Tsallis ($q = 2$)	32.1	5.28	12.4	79.4	10.2
Normal	2.85	1.14	2.41	**6.15**	1.33
Dirichlet	3.55	45.9	4.23	9.53	**0.92**

3. **Florida voter registration** by county, 2012 (Imai & Khanna, 2016): we relate self-reported race (7 categories) to political affiliation (3 categories).

4. **U.S. total personal income** by state, 2016 (IPUMS, 2018): we relate educational attainment (11 categories) to income (9 categories).

5. **U.S. health insurance coverage** by state, 2016 (IPUMS, 2018): we relate age group (20 categories) to the class of health insurance provider (public vs. private) (4 categories).

Table 1 shows the results of using various models to infer the underlying table from its marginals, in terms of the median absolute error of the inferred table entries with respect to the true table entries [7]. Bolded in each column is the best-performing model, along with any models that were not significantly worse – we discuss tests of significance for differences in performance in Appendix D.

We test two types of models, using the proposed inference methods ("MAP"): those assuming perfect observations of the marginals and those assuming multinomial-distributed observations. We also compare to the existing methods described in Section 2.2 [8]. Note that the "entropic" prior is the natural exponential family whose log-base measure is given by the negative entropy function.

To test the ability of each model to capture the distribution of the data (and put all models on roughly equal footing), we estimate the parameters of each prior distribution by maximum likelihood, given a single table containing the total proportions across each entire dataset. This is quasi-realistic: In the electoral example, this would be equivalent to a statewide poll providing the prior for subsequent inference at the county level.

Comparison to prior work. In only one case – the voter registration dataset – is a model available from prior work among the best-performing models. This is also the only dataset for which the independence model is competitive with the best, indicating that the two variables are relatively well-modeled as being independent. Note that the entropic and Tsallis models with perfectly-observed marginals have previously been suggested for this problem (Muzellec et al., 2017), but in the Tsallis case different inference algorithms were used; one of these is only approximate (for $q < 1$), whereas the method we use here is exact. None of the models with multinomial observations of both marginals was available previous to the current work; neither were the perfectly-observed models with normal and Dirichlet priors. Note that these represent only a small sample of the possible models within our framework (Sections 3 and 5).

7. Conclusion

The general-purpose inference methods presented here apply to a broad class of models, greatly extending the choices available to practitioners. In addition to being computationally efficient, these methods can enable more accurate modeling of real data than existing methods. The proposed methods might be extended further to non-separable priors, which can have correlated components, and to continuous variables, leveraging ideas from regularized optimal transport (Genevay et al., 2016).

[7] We normalize each ground truth table to sum to 1, before computing the absolute error.

[8] For the multinomial-Dirichlet (MCMC) model we sample a chain of length 10^5 with burn-in of 10^4, with no thinning. The estimate is the mean. For the entropic and Tsallis models, we set $\lambda = 1$ (as defined in (Muzellec et al., 2017)). For normal models, we fix $\sigma = 0.1$.

References

Achen, C. H. and Shively, W. P. *Cross-level inference*. University of Chicago Press, 1995.

Bauschke, H. H. and Lewis, A. S. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

Bauschke, H. H., Borwein, J. M., et al. Legendre functions and the method of random bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3): 200–217, 1967.

CDC. Underlying cause of death, 1999-2016. https://wonder.cdc.gov/UCD-ICD10.html, 2018.

Chen, Y., Dwivedi, R., Wainwright, M. J., and Yu, B. Vaidya walk: A sampling algorithm based on the volumetric barrier. In *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*, pp. 1220–1227. IEEE, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Della Pietra, S., Della Pietra, V., and Lafferty, J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

Dessein, A., Papadakis, N., and Rouas, J.-L. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

Dhillon, I. S. and Tropp, J. A. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

Dudík, M. and Schapire, R. E. Maximum entropy distribution estimation with generalized regularization. In *International Conference on Computational Learning Theory*, pp. 123–138. Springer, 2006.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

Goodman, L. Ecological regressions and the behavior of individuals. *American Sociological Review*, 18:663–665, 1953.

Greiner, D. J. Ecological inference in voting rights act disputes: Where are we now, and where do we want to be. *Jurimetrics*, 47:115, 2006.

Greiner, D. J. and Quinn, K. M. Exit polling and racial bloc voting: Combining individual-level and r x c ecological data. *The Annals of Applied Statistics*, pp. 1774–1796, 2010.

Grofman, B., Migalski, M., and Noviello, N. The "totality of circumstances test" in section 2 of the 1982 extension of the voting rights act: A social science perspective. *Law & Policy*, 7(2):199–223, 1985.

Honaker, J. Unemployment and violence in northern ireland: a missing data model for ecological inference. In *Summer Meetings of the Society for Political Methodology*, 2008.

Imai, K. and Khanna, K. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272, 2016.

India. Population attending educational institution by completed education level, age and sex, census 2001. https://data.gov.in/catalog/population-attending-educational-institution-completed-education-level-age-and-sex-census, 2018.

IPUMS. American community survey. https://usa.ipums.org/usa, 2018.

Jackson, C., Best, N., and Richardson, S. Improving ecological inference using individual-level data. *Statistics in medicine*, 25(12):2136–2159, 2006.

Johnston, R. and Pattie, C. *Putting voters in their place: Geography and elections in Great Britain*. Oxford University Press, 2006.

Kannan, R. and Narayanan, H. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

Kargas, N. and Sidiropoulos, N. D. Completing a joint pmf from projections: A low-rank coupled tensor factorization approach. In *2017 Information Theory and Applications Workshop (ITA)*, pp. 1–6. IEEE, 2017.

King, G. *A solution to the ecological inference problem*. Princeton, NJ: Princeton University Press, 1997.

King, G., Tanner, M. A., and Rosen, O. *Ecological inference: New methodological strategies*. Cambridge University Press, 2004.

Kleppner, P. *Chicago divided: The making of a black mayor*. northern illinois University Press, 1985.

Kousser, J. M. Ecological regression and the analysis of past politics. *The Journal of Interdisciplinary History*, 4 (2):237–262, 1973.

Lee, Y. T. and Vempala, S. S. Geodesic walks in polytopes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 927–940. ACM, 2017.

Loewen, J. W. Social science in the courtroom: Statistical techniques and research methods for winning class-action suits. 1982.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters*, 8(11):1235–1246, 2005.

Morgenstern, H. Ecologic studies in epidemiology: concepts, principles, and methods. *Annual review of public health*, 16(1):61–81, 1995.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. Tsallis regularized optimal transport and ecological inference. In *AAAI*, pp. 2387–2393, 2017.

Nelsen, R. B. *An introduction to copulas*. Springer Science & Business Media, 2007.

Peyré, G. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Piguet, E. Linking climate change, environmental degradation, and migration: a methodological overview. *Wiley Interdisciplinary Reviews: Climate Change*, 1(4):517–524, 2010.

Rosen, O., Jiang, W., King, G., and Tanner, M. A. Bayesian and frequentist inference for ecological inference: The r × c case. *Statistica Neerlandica*, 55(2):134–156, 2001.

Sinkhorn, R. and Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

Wakefield, J. Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–445, jul 2004.

Wakefield, J. Ecologic studies revisited. *Annu. Rev. Public Health*, 29:75–90, 2008.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.