On model-based clustering with entropic optimal transport

Gonzalo Mena
Department of Statistics and Data Science, Carnegie Mellon University *

Abstract. We develop a new methodology for model-based clustering. Based on optimizing the log-likelihood, the standard approach provides a principled statistical framework for clustering where solutions are found via the EM algorithm. However, as the log-likelihood is nonconvex, convergence to only local optima can be guaranteed, and practitioners rely on several starting points with the hope that one of them will converge to the global solution. We consider a new loss function based on entropic optimal transport with the same global optimum as the log-likelihood but a much better-behaved landscape so that spurious local optima configurations known to be pervasive for the log-likelihood are avoided. Similar to the EM algorithm for the log-likelihood, this new loss can be optimized by the so-called Sinkhorn-EM algorithm, which we show to enjoy similar convergence guarantees to EM. By analyzing extensive numerical experiments and two real-world applications on image segmentation in C.elegans microscopy and clustering in spatial transcriptomics experiments, we show that this new loss improves upon log-likelihood optimization, indicating it represents a valuable clustering methodology for practitioners.

AMS 2000 subject classifications: Statistics.
Key words and phrases: EM Algorithm, Mixture of Gaussians, Optimal Transport, Entropic Regularization..

CONTENTS

*Preliminary draft, Jan 19th, 2024

1. INTRODUCTION

Cluster analysis is one of the most ubiquitous tasks in the practice of statistics and data science as the necessity to determine latent group structure arises in a myriad of applications such as bioinformatics Lo et al. (2008), social network analysis (Handcock et al., 2007), text analysis (Blei et al., 2003), economics (Wooldridge, 2003), image processing (Houdard et al., 2018), among others. The mainstream statistical approach for clustering is the so-called model-based (Bouveyron et al., 2019; McLachlan, 1982; McNicholas, 2016), that hinges on the inference of an underlying probabilistic generative model that typically takes the form of a finite *mixture* of distributions, where each component represents an observation model for samples within a cluster.

The primary inferential approach for model-based clustering is to perform maximum likelihood estimation on the mixture model, and the Expectation Maximization (EM) algorithm (Dempster et al., 1977) is the most widely used computational tool for such inference as it exploits the inherent latent structure in mixture models. However, the log-likelihood in mixture models is typically a non-convex function, and often, the EM algorithm will converge to spurious local optima (Biernacki et al., 2003). To deal with this issue, practitioners resort to heuristics such as choosing sensible initializations (e.g., k-means++) or running the algorithm on several seeds and keeping the one with the highest likelihood (Biernacki et al., 2003). While practical, these heuristics are not guaranteed to work, and they could lead to a systematic choice of sub-optimal solutions (Fränti and Sieranoja, 2019; Steinley, 2003).

We propose a new methodology for model-based clustering by importing tools from Optimal Transport (OT) and show they mitigate the risk of getting trapped in sub-optimal solutions. Optimal transport (Villani, 2008) is a mathematical framework that provides us with a rich way of measuring the distance between distributions, and it has enjoyed much attention in statistics, computer vision, machine learning, and related fields (see, e.g., Kolouri et al. (2017); Peyré et al. (2019) for surveys). We focus on Entropic Optimal Transport (EOT), a variation of the original OT problem that includes an entropic penalization term. This formulation possesses computational (Cuturi, 2013) and statistical advantages over its unregularized version (Genevay et al., 2019a; Mena and Niles-Weed, 2019).

This paper is organized as follows: in Section 2, we develop a parameter estimation framework for mixture models based on optimizing an entropic optimal transport-derived loss and describe some basic properties. In Section 3, we describe an algorithm for optimization of this loss, paralleling the EM algorithm for the log-likelihood and establishing elementary convergence properties. In Section 4, we show that our methodology can avoid bad local optima in cases where the log-likelihood cannot. In Section 5, we give an in-depth convergence and local-optima analysis in the elementary mixture of two Gaussians case, showing that the performance of our method at least matches the known convergence rate for the usual EM algorithm. In Section 6, we illustrate the benefits of our method in large-scale experiments on simulated data. Then, in Sections 7 and 8, we describe two real-world examples in image segmentation for C.elegans microscopy and clustering in Spatial Transcriptomics experiments, where our proposed methodology leads to tangible benefits. Finally, in Section 9 we sketch future directions and comments on the limitations of our methodology.

1.1 Related Work

Our work contributes to scholarship on model-based clustering methodology (Bouveyron et al., 2019; McLachlan, 1982; McLachlan et al., 2002). Although not the main focus, our theoretical results connect to recent theoretical analysis of convergence for EM algorithm seeking to obtain rigorous convergence guarantees in simplified settings (Balakrishnan et al., 2017; Daskalakis et al., 2017a,b; Dwivedi et al., 2020; Kwon et al., 2020, 2019; Wu and Zhou, 2019; Xu et al., 2016, 2018). Our local minima analysis is heavily inspired by the recent discussion of the structure of spurious local optima for the log-likelihood (Chen and Xi, 2020; Jin et al., 2016).

The use of optimal transport-based losses for parameter estimation has been advocated in Bassetti et al. (2006); Bernton et al. (2019). Closer to our work are Dessein et al. (2017); Rigollet and Weed (2018), who show that under some conditions, maximizing the log-likelihood on a deconvolution model (e.g., a mixture) yields *the same* solution as solving an entropic optimal transport problem. This observation motivates the question of which scenarios we may expect benefits from solving the EOT problem instead of maximizing the likelihood, and our methodology gives a concrete example of a setup where we obtain benefits.

Several papers have advocated the use of optimal transport for clustering (Canas and Rosasco, 2012; Cuturi and Doucet, 2014; Fettal et al., 2022; Genevay et al., 2019b; Huizing et al., 2022; Kolouri et al., 2018; Laclau et al., 2017; Nejatbakhsh et al., 2020; Pollard, 1982; Titouan et al., 2020), demonstrating benefits

over the likelihood-based approach. In particular, Nejatbakhsh et al. (2020) first introduced Sinkhorn-EM, an algorithm that we study in detail here. Some of these papers have suggested that the benefits of optimal transport are related to avoiding spurious local optima (Kolouri et al., 2018; Yan et al., 2023), but the question hasn't been explored in mathematical detail. Our work provides a solid, unifying perspective for explaining these phenomena.

1.2 Preliminaries

We take the following conventions and definitions: we consider a parametric family on $\mathbb{R}^d$ given by a mixture of K components, defined as follows: Each mixture component is parameterized through a location parameter θ and perhaps other parameters ν (such a scale or variance parameter). Abusing notation, we may ignore ν or collapse it with θ to refer to a unique parameter. The parameterized family of densities q_θ (w.r.t. the Lebesgue measure $\mathcal{L}$) expresses in terms of the template density q as $q_\theta(y) = q^\nu(y-\theta) = q(y-\theta)$. Most of our specialized theoretical results focus on the Gaussian Mixture Model (GMM), i.e., $q_\theta(Y) = \mathcal{N}(Y;\theta,\Sigma)$ (where Σ can be treated as fixed or as a parameter itself), although our methodology is general enough to accommodate other cases such as mixtures of t distributions McLachlan and Peel (1998), etc. To define the mixture, for a vector α of weights in the probability simplex, we write our model as the density q_θ

$$q_\theta(y) = \sum_{k=1}^{K} \alpha_k q_{\theta_k}(y). \tag{1}$$

Note that the mixture density (1) is a marginal density for a certain joint model: let Q_θ be the distribution associated with q_θ and let P_θ be the measure concentrated on the $\theta_k' s$ so that it parameterizes the location parameters,

$$P_\theta = \sum_{k=1}^{K} \alpha_k \delta(\theta_k). \tag{2}$$

Then, Q_θ is the marginal distribution of the joint $Q_\theta^{X,Y}$ with density

$$dQ_\theta^{X,Y}(x,y) = q_x(y) dP_\theta(x) \mathrm{d}y = q(y-x) dP_\theta(x) \mathrm{d}y. \tag{3}$$

For the above model we define the negative log-likelihood as

$$\ell(\theta) = -E\left(\log q_\theta(Y)\right), \tag{4}$$

where the expectation is taken with respect to the true distribution of data, i.e. $Y \sim Q_*$. In the well-specified case where data was generated according to mixture (1) with true parameter θ^* we write Q_{θ^*}. We will sometimes write $q_{\theta,\alpha}(y), P_{\theta,\alpha}$ and $\ell(\theta,\alpha)$ to make the dependence in α explicit, and α can either be considered a parameter to be optimized (as in Section 3.2), although otherwise stated α is fixed. In some cases it will be convenient to see P_θ as a measure p_θ on the set $[K] = \{1,\dots,K\}$ such that $p_\theta(k) = P(X = \theta_k) = \alpha_k$. Likewise, we may see $Q_\theta^{X,Y}$ as a measure on $[K] \times \mathbb{R}^d$.

In practice, we work with a sample $Y_1,\dots,Y_n$ from Q_* or Q_{θ^*}, so we will consider empirical versions of population quantities such as ℓ where the expectation

is taken with respect to the sample. Most of the elementary results grounding our methodology are valid in the population and sample case, while others only hold in the population case. We will make the distinction clear. Likewise, some of our theoretical results (e.g., local optima, convergence) are based on the analysis of the *population* landscape of ℓ and related loss functions. Still, they don't necessarily hold in the sample case.

2. THE ENTROPIC OPTIMAL TRANSPORT LOSS AS AN ALTERNATIVE TO THE LOG-LIKELIHOOD

In this section, we will define a new loss function for estimation in the model (1) based on entropic Optimal Transport. We must first provide some elementary definitions but refer the reader to, e.g., Peyré et al. (2019) for a comprehensive treatment of the topic.

Let P and Q be two probability measures on $\mathbb{R}^d$. Given a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define the entropy-regularized optimal transport loss between P and Q as

$$S(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \Big[\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\,\mathrm{d}\pi(x,y) + H(\pi|\mu\otimes\nu)\Big]. \tag{5}$$

where $\Pi(\mu,\nu)$ is the set of all joint distributions with marginals equal to μ and ν, respectively, and $H(\alpha|\beta)$ denotes the relative entropy between measures α and β defined as $\int \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\mathrm{d}\alpha(x)$ if $\alpha \ll \beta$ and $+\infty$ otherwise. A useful alternative representation of $S(\mu,\nu)$ is in terms of a reference Gibbs kernel with density proportional to $\exp(-c(x,y))\mathrm{d}\mu(x)\mathrm{d}y$ (see Mena and Niles-Weed (2019) for details):

$$S(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} H(\pi|e^{-c}\mu\otimes\mathcal{L}) - H(\nu|\mathcal{L}). \tag{6}$$

We define the entropic OT loss function for parameter estimation in model (1) as $L(\theta) = S(P_\theta, Q_*)$. To make the correspondence precise with model (1) we must restrict to cost functions of the form $c(x,y) = -\log q_x(y) = -\log q(y-x)$ so that we recover the joint densities $Q_\theta^{X,Y}$ in (3) as the set of Gibbs kernels in (6), i.e.,

$$L(\theta) := S(P_\theta, Q_*) = \inf_{\pi\in\Pi(P_\theta,Q_*)} H\left(\pi|Q_\theta^{X,Y}\right) - H(Q_*|\mathcal{L}). \tag{7}$$

For example, in the simplest Gaussian case we let $c(x,y) = \frac{||x-y||^2}{2\sigma^2}$, and for a mixture of Laplace distributions we make $c(x,y) = \frac{|x-y|}{b}$. As the last term in (7) doesn't depend on θ, $L(\theta)$ is indeed a discrepancy between the model (3) and data (roughly understood as a coupling π between P_θ and Q_*). In this definition, the dependence on the location parameters is fully encapsulated in the base measure P_θ, and by making the cost function vary, we can also represent dependency on other parameters such as variances and scales.

Although perhaps not obvious, L is intimately related to ℓ. First, in the well-specified case, we have that $L(\theta^*) = \ell(\theta^*)$ at the true θ^*. To see this, we note that if $\theta = \theta^*$ the coupling π achieving the infimum is exactly the joint $Q_{\theta^*}^{X,Y}$ in (3), so the relative entropy term is zero. Second, we have that $L(\theta) \geq \ell(\theta)$. This follows from another useful alternative representation of the entropic OT loss, the so-called semi-dual formulation Peyré et al. (2019):

$$S(\mu,\nu) = \sup_{\omega} \int \omega(x)d\mu(x) - \int \log\left(\int e^{\omega(x)-c(x,y)}d\mu(x)\right)d\nu(y),$$

where the semi-dual potential ω is a bounded function. In our context ω is a K-dimensional vector so

$$L(\theta) = \sup_{\omega \in \mathbb{R}^K} \left[\sum_{k=1}^{K} \alpha_k \omega_k - E\left(\log \left(\sum_{k=1}^{K} \alpha_k e^{\omega_k} q_{\theta_k}(y) \right) \right) \right], \tag{8}$$

and since optimization over ω includes $\omega = 0$ it follows that $L(\theta) \geq \ell(\theta)$.

The semi-dual potentials can be interpreted as a way of *tilting* the original weights α: for each θ denote $\omega(\theta)$ the one maximizing (8) and define the vector $\alpha(\theta) \in \mathbb{R}^K$ as

$$\alpha(\theta)_k = \frac{\alpha_k e^{\omega(\theta)_k}}{\sum_{k'=1}^{K} \alpha_{k'} e^{\omega(\theta)_{k'}}}. \tag{9}$$

Then

$$L(\theta) = H\left(\alpha | \alpha(\theta)\right) - E\left(\log \left(\sum_{k=1}^{K} \alpha_k(\theta) q_{\theta_k}(Y) \right) \right). \tag{10}$$

Therefore, the computation of $L(\theta)$ amounts to the computation of $\ell(\theta)$ for a *tilted* model with weights $\alpha(\theta)$, plus a relative entropy term. In particular, since this relative entropy doesn't depend on θ, we have by the envelope theorem that

$$\nabla L(\theta) = \nabla \ell(\theta, \alpha(\theta)). \tag{11}$$

While the analysis of the second derivatives of L is more complicated, it is possible to show that L has more strictly more curvature. We summarize all this discussion in the following proposition. A complete proof appears in Appendix A.1.

PROPOSITION 1. *Let ℓ and L be as in (4) and (7), respectively. The following statements are true*

(a) $L(\theta) \geq \ell(\theta)$.
(b) $L(\theta^*) = \ell(\theta^*)$ *if* $Q_* = Q_{\theta^*}$.
(c) $\nabla L(\theta) = \nabla \ell(\theta, \alpha(\theta))$, *where* $\alpha(\theta)$ *is as defined in (9)*
(d) L *has more curvature than* ℓ *around* θ^* *if* $Q_* = Q_{\theta^*}$. *Specifically,*

$$\nabla^2 L(\theta^*) = \nabla^2 \ell(\theta^*) + B^\top(\theta^*) A^{-1}(\theta^*) B(\theta^*),$$

where $A(\theta^*)$ *is a* $K-1 \times K-1$ *definite positive matrix. Explicit expressions for* $A(\theta^*)$, $B(\theta^*)$ *are given in Appendix A.1.*

(a) and (c) are valid both in the population and sample versions of L and ℓ, while (b) and (d) are valid only in the population versions, in the well-specified setup ($Q_* = Q_{\theta^*}$).

Fig. 1A illustrates Proposition 1. As a consequence of (a) and (b), in the population limit, L is also minimized at θ^*, and the curvature statement (d) suggests that L might be a better optimization objective than ℓ. However, since both L and ℓ are typically non-convex, many local optima may exist for both functions and the local statement of Proposition 1 doesn't say much about the global convergence of first-order methods. In Section 4, we show that for a mixture of Gaussians, L will typically avoid bad-local optima configurations that are pervasive for ℓ, making a much stronger case in favor of L as an optimization objective. Before that, in the following section, we describe an algorithm for optimizing L and make a parallel with the usual EM algorithm.

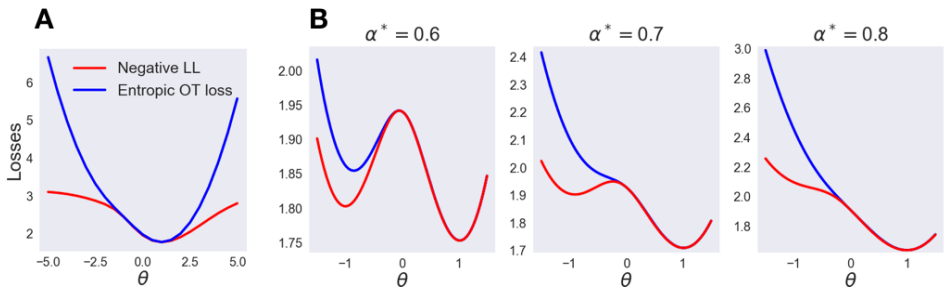

Figure 1: Qualitative comparison between the log-likelihood and entropic OT loss. **A** By Proposition 1, the entropic OT loss dominates the negative log-likelihood but has a larger curvature around the minimum. **B** By Theorem 4, in the model (24) L has fewer bad local minima than ℓ for some values of α^*.

3. ALGORITHMIC ASPECTS: SINKHORN-EM

In principle, we could consider any first-order method to find the local optima of L and ℓ. In the case of ℓ practitioners typically appeal to the EM algorithm (an instance of a first-order method Xu and Jordan (1996)). This algorithm exploits the underlying latent structure of the mixture model (1), and its appeal is ultimately justified by a rich body of work establishing its theoretical guarantees (e.g., Balakrishnan et al. (2017); Daskalakis et al. (2017b); Redner and Walker (1984)). Pivoting on the relationship between ℓ and L from the previous section, we describe an EM-type algorithm for optimizing L, which we name Sinkhorn-EM, and establish some elementary convergence guarantees.

The main observation is that as a consequence of the well-known variational representation of $\log q_\theta$, we can write:

$$\ell(\theta) = \inf_{\pi \in \Pi(\cdot, Q_*)} H\left(\pi | Q_\theta^{X,Y}\right) - H\left(Q_* | \mathcal{L}\right), \tag{12}$$

where the set $\Pi(\cdot, Q_*)$ is the set of joint distributions with arbitrary first marginal and second marginal Q_*. By disintegration, this set can be represented by Q_* along with a set of conditionals $\pi(\cdot|y)$. The EM algorithm exploits this variational representation as it can be understood as coordinate descent on θ and π to minimize ℓ in (12) (Csiszár and Tusnády, 1984; Neal and Hinton, 1998).

By comparing (7) and (12) we see that L and ℓ only differ in that the variational representation of the former has an additional constraint. This observation motivates the definition of Sinkhorn-EM (SEM) as the algorithm performing coordinate ascent on θ and π to minimize the variational representation (7).

Sinkhorn-EM only differs from EM in the E-step. For the EM algorithm, this step corresponds to the computation of the so-called *responsibilities*, the set of conditionals of X given y at the current θ in the joint model (3), i.e., the set of measures

$$\Psi_k\left(y, \theta, \alpha\right) := \mathrm{d}Q_{\theta,\alpha}^{X,Y}(k|y) = \frac{\alpha_k q_{\theta_k}(y)}{\sum_{k=1}^K \alpha_k q_{\theta_k}(y)}. \tag{13}$$

Instead, Sinkhorn-EM solves an entropic Optimal Transport problem on the E-

step. If π_θ is the optimal coupling to (7) we have

$$\Psi_k\left(y,\theta,\alpha(\theta)\right)=\pi_\theta(k|y)=\mathrm{d}Q^{X,Y}_{\theta,\alpha(\theta)}(k|y)=\frac{\alpha(\theta)_k q_{\theta_k}(y)}{\sum_{k=1}^K \alpha(\theta)_k q_{\theta_k}(y)}. \tag{14}$$

where $\alpha(\theta)$ is the *tilted* version of α defined in (9) so that the constraint on the first marginal expresses as

$$E\left(\Psi_k\left(Y,\theta,\alpha(\theta)\right)\right)=\alpha_k. \tag{15}$$

Therefore, the familiar understanding of the E-step as the computation of *responsibilities* is retained, although in a somewhat different sense. This signifies an extra computational burden compared to (13): while each E-step of the standard EM algorithm takes time $O(n\cdot K)$, where n, the E-step of Sinkhorn-EM involves a convex optimization problem and can be solved by an efficient a celebrated algorithm due to Sinkhorn and Knopp (1967) which converges in near-linear time, i.e. $\tilde{O}(n\cdot K)$ (Altschuler et al., 2017). As we will see in Section 4, this mild overhead in operation complexity is easily compensated for in practice because SEM often avoids bad solutions.

We can now state the Sinkhorn-EM algorithm and establish an elementary convergence property.

THEOREM 1. *Let Sinkhorn-EM be the algorithm defined as the sequence (π^t,θ^t) from the E and M steps below:*

$$\textbf{\textit{E-step}}: \qquad \pi^{t+1}=\pi_{\theta^t}=\operatorname*{argmin}_{\pi\in\Pi\left(P_{\theta^t},Q_*\right)} H\left(\pi|Q^{X,Y}_{\theta^t}\right) \tag{16a}$$

$$\begin{aligned}\textbf{\textit{M-step}}: \qquad \theta^{t+1}&=\operatorname*{argmax}_\theta E\left(\sum_{k=1}^K \pi^{t+1}(k|Y)\log\left(\alpha_k q_{\theta_k}(Y)\right)\right)\\ &=\operatorname*{argmax}_\theta E\left(\sum_{k=1}^K \Psi_k(Y,\theta^t,\alpha(\theta^t))\log\left(\alpha_k q_{\theta_k}(Y)\right)\right),\end{aligned} \tag{16b}$$

where θ^0 is arbitrary. Then, the sequence $\{L(\theta^t)\}$ is nonincreasing and $L(\theta^{t+1})<L(\theta^t)$ if θ^t is not a stationary point of L. If θ^t converges, then the limit is a stationary point of L.

Although the convergence of θ^t is not guaranteed in general, by trivially extending the results of Wu (1983), this can be established if L satisfies some regularity conditions (such as continuity, radial unboundedness) which typically hold whenever they hold for ℓ as well.

3.1 The Gaussian Mixture Model

Our more specialized theoretical results will be stated in the following GMM

$$q_\theta(Y)=\sum_{k=1}^K \alpha_k\mathcal{N}(Y;\theta_k,I_d\sigma^2)\quad 0<\alpha_k<1,\sigma^2>0, \tag{17}$$

with true parameter θ^*. In that case, M-step (16b) reads

$$\theta^{t+1} = F(\theta^t, \alpha(\theta^t)),$$

with

$$F(\theta, \alpha)_k := \frac{E\left(Y\left(\Psi_k(Y, \theta, \alpha)\right)\right)}{E\left(\Psi_k\left(Y, \theta, \alpha\right)\right)} = \frac{E\left(Y\left(\Psi_k(Y, \theta, \alpha)\right)\right)}{\alpha_k}. \tag{18}$$

Note that for the usual EM algorithm, we would otherwise consider the simpler $\theta^{t+1} = F(\theta, \alpha)$ iterations.

3.2 Weights update

Sinkhorn-EM can, in principle, only be deployed in fixed-weights setups: as the E-step imposes a marginal constraint on weights α, it is bound to remain on those initial weights. However, nothing prevents us from considering L as a function of α as well, as it is customary for mixture models with unknown weights. To enable weight inference we consider simple exponentiated gradient (or mirror descent with relative entropy as Bregman divergence, Kivinen and Warmuth (1997)) updates for α. The gradient of L with respect to α is given by

$$\nabla_\alpha L(\theta)_k = \omega(\theta)_k - E\left(\frac{\Psi_k(Y, \theta, \alpha(\theta))}{\alpha_k}\right). \tag{19}$$

Then, for a step-size $\eta > 0$, current α and fixed θ, an update for α reads:

$$\alpha_k^{new} = \frac{\alpha_k \exp\left(-\eta \nabla_\alpha L(\theta)_k\right)}{\sum_{k=1}^K \alpha_k \exp\left(-\eta \nabla_\alpha L(\theta)_k\right)}. \tag{20}$$

Whenever inference for α is required, we will couple usual Sinkhorn-EM updates (16a),(16b) for θ along with updates for α.

4. ON LOCAL OPTIMA

Theorem 1 in the previous section provides us with machinery to efficiently optimize L, a function that in Proposition 1 was shown to enjoy a better curvature property than ℓ, around θ^*. However, since both EM and SEM can only shown to converge to stationary points of the respective ℓ and L, and since these stationary points may abound, this local property says little about the success or failure of these algorithms in practice.

In Theorem 2 below, we make a much stronger case favoring L as optimization objective: when q_θ is a spherical mixture of Gaussians with equal weights and variances σ^2, then L won't possess a type of bad-local minima that are pervasive for the log-likelihood if some (mild) separation condition is met. These spurious local minima are the so-called *many-fit-one* configurations, where a subset of fitted mixture components are close to a single true component θ. These structures were anticipated in Jin et al. (2016), which gave an explicit construction of a bad local minima for ℓ in one dimension consisting of two nearby components that are both far away from a third isolated component. Here, in Proposition 2 below, we also provide a simple extension of this result to the bi-variate case, showing that the negative log-likelihood associated to mixture of three Gaussians with true centers at $\theta_1^* = (0, D), \theta_2^* = (0, -D)$ and $\theta_3^* = (R, 0)$ has a bad local minimum with one component close to $(0, 0)$ (the average of θ_1^* and θ_2^*) and the two others close to θ_3^*. This illustrates the potentially catastrophic effect of bad

local minima for the negative log-likelihood: although the component $(0, R)$ is identified, the signal on the y axis is entirely destroyed by averaging (see Fig 2A for a depiction).

PROPOSITION 2. *Consider a mixture of three Gaussians in* $\mathbb{R}^2$ *with equal weights and variances* $\sigma^2 = 1$ *and true locations* $\theta_1^* = (0, D)$ *,*$\theta_2^* = (0, -D)$ *and* $\theta_3^* = (R, 0)$*. For* $\varepsilon > 0$ *define the region* $\mathcal{R}^\varepsilon$ *in* $\mathbb{R}^6$

$$\mathcal{R}^\varepsilon := \left\{ \theta = (\theta_1^1, \theta_1^2, \theta_2^1, \theta_2^2, \theta_3^1, \theta_3^2) : \theta_1^1 < \frac{R}{3}, \theta_2^1 > \frac{2R}{3}, \theta_3^1 > \frac{2R}{3}, ||\theta^2|| < \varepsilon \right\}.$$

If R *is large enough, then* $\ell(\cdot)$ *has a bad local minima in* $\theta \in \mathcal{R}^\varepsilon$*. Informally,* θ *is on a neighborhood of the configuration* $\theta_1 \approx \frac{1}{2}(\theta_1^* + \theta_2^*) = (0, 0)$ *and both* $\theta_2 \approx \theta_3 \approx \theta_3^* = (R, 0)$*.*

The existence of these structures has been recently studied in much more detail in Chen and Xi (2020), showing that local minima of the negative log-likelihood (with equal weights and fixed variance) roughly correspond to generalized forms of the above prototypical local minima. They show that if all components are all well separated from each other (i.e. if $\Delta_{min} := \min_{k_1, k_2 \leq n} \left\| \theta_{k_1}^* - \theta_{k_2}^* \right\|$ is large enough) then the components of any spurious local minima θ of $\ell(\cdot)$ partition into groups forming either a *many-fit-one* configuration, or a *one-fits-many* configuration, where a θ_k is placed near to the *average* of a group of some true θ_k^*.

Although appealing, the result of Chen and Xi (2020) suffers from three drawbacks: first, the required lower bound on Δ_{min} is too stringent to provide any insight in practical setups ($\Delta_{min} \geq 18(\sqrt{2\pi} + 1)K^5\sigma$). Also, it only provides necessary conditions for local optimality, but it does not indicate which such configurations eventually realize as local minima among all possible. Third, other components of any such local minimum may satisfy a degenerate so-called "near empty association" condition distinct from the above *many-fit-one* and *one-fit-many*.

Even if a complete characterization of local minima of ℓ is far from complete, we can take the above discussion as a starting point to frame our main result. In Theorem 2, we show that *many-fit-one* configurations where the rightmost true θ_k^*'s are fitted by a bigger number of components θ_k's are not possible for stationary points θ_k of L whenever certain separation condition is met.

THEOREM 2. *Suppose that data has been generated by an equally weighted GMM of* K *components in* $\mathbb{R}^d$ *with variance* σ^2 *and means* θ_k^* *that we fit with the mixture* q_θ *in* (17) *with the same weights and variances but varying parameters* θ_k*. Consider an arbitrary dimension and sort the true component means* θ_k^* *from lowest to highest in that direction. For any given* k_1, k_2 *such that* $k_1 + k_2 = K$ *we define two groups of components,* G_1 *given by the* k_1 *leftmost components and* G_2 *by the* k_2 *rightmost. Let* Δ *be the distance between the leftmost component in* G_2 *and the rightmost in* G_1*.*

Suppose θ *is a stationary point of* L*, and that a group* O *of* $\tilde{k} > k_2$ *distinct components of* θ *cover the* k_2 *rightmost true components in the sense that for some* $\delta > 0$ *and every true component* θ_k^* *with* $k \in G_2$ *there is at least one fitted component* $\theta_{k'}$ *with* $k' \in O$ *such that*

$$\theta_k^* - \theta_{k'} \leq \delta. \tag{21}$$

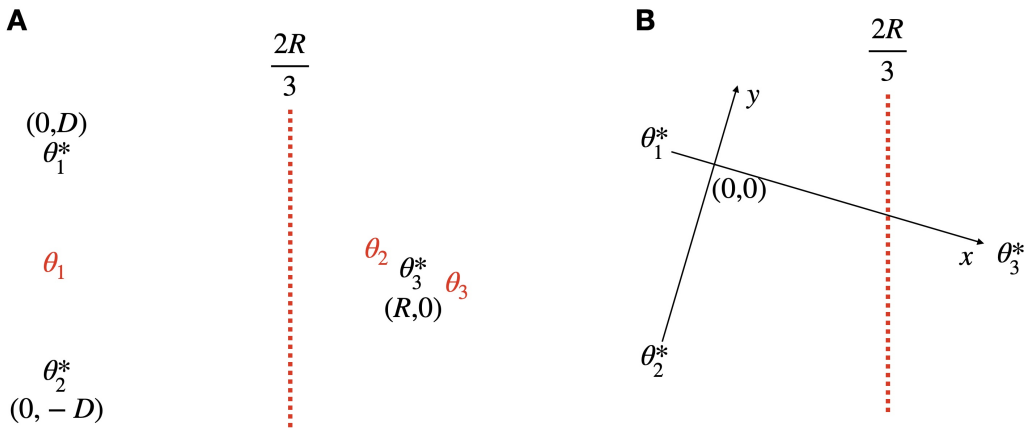

Figure 2: Local optima in the example of Proposition 2. **A**. By Proposition 2, the negative log-likelihood has a bad local minimum around the configuration θ, i.e. within the region $\mathcal{R}^{\varepsilon}$. **B** by rotating axes we can apply Theorem 2 to rule out any stationary point for L in such region if R is large enough.

Take any tolerance level $0 < \alpha < 1 - k_2/\tilde{k}$. *If the following separation condition holds*

$$\Delta \geq \sigma \left((2\pi) \left((1-\alpha)\tilde{k} - k_2 \right) \right)^{-1} K. \tag{22}$$

Then

$$\delta > \alpha\Delta > 0. \tag{23}$$

Intuitively, minimizing L imposes an additional constrain in the way that mass is split over the space: that for a stationary point θ of L, it must be the case that $\sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \theta_k^*$. This balance condition (that doesn't hold for stationary points of the log-likelihood) cannot occur if many more fitted components are placed near a lower number of rightmost true components θ^*.

In particular, from Theorem 2 we can rule out the bad local minimum in $\mathcal{R}^{\varepsilon}$ for L if R is large enough, as shown in Fig. 2B. This is stated as follows (a proof appears in Appendix A.6)

Corollary 1. *If* $R^2 \geq 9D^2$ *and* $D \geq 2$ *then* L *cannot have a stationary point in the region* $\mathcal{R}^{\varepsilon}$.

An important question is whether L can have other types of local minima that don't exist for the log-likelihood, offsetting the benefits of Theorem 2. While we don't attempt to provide a general answer to this question (a characterization of local optima is barely available for the log-likelihood under extremely stringent separation conditions), our results suggest that we should not expect that L has a wildly different set of local minima relative to ℓ. Indeed, from Proposition 1(c), every stationary point of L is a stationary point for the log-likelihood on a GMM for some *tilted* weights $\alpha(\theta)$. Moreover, we have the following corollary of Proposition 1. The proof appears in Appendix A.2.

Corollary 2. *For the GMM in* (17)*, if* θ *is a local minimum for the negative log-likelihood then it must be a local minimum for the entropic OT loss.*

5. ANALYSIS OF A SYMMETRIC MIXTURE OF TWO GAUSSIANS

In this section, we give more specialized results in the following unbalanced (with fixed $\alpha^* < 1$) mixture of two Gaussians in $\mathbb{R}$ with a single one-dimensional parameter θ:

$$q_\theta(y) = \alpha^* \mathcal{N}(y; \theta, 1) + (1 - \alpha^*)\mathcal{N}(y; -\theta, 1). \tag{24}$$

The first result, Theorem 3, complements our results on local minima from Section 4. As illustrated in Fig 1B, we show that compared to ℓ, L will more often have a unique local minima. This result is based on a fine analysis of ℓ and L and their derivatives and doesn't immediately relate to the ideas around Theorem 2.

THEOREM 3. *Consider the unbalanced symmetric model (24). For any $\theta^* > 0$, the set of α^* for which θ^* is the unique stationary point of L is strictly larger than the one for ℓ.*

We also provide an in-depth analysis of the basin of attraction of Sinkhorn-EM towards the global minimum and the convergence speed: we show that as long as SEM is initialized at $\theta^0 > 0$, it enjoys fast (exponential) convergence to the global minimum of L, and there is a large range of initializations on which it never performs worse than EM.

THEOREM 4. *For the model (24), for each $\theta^* > 0$ and initialization $\theta^0 > 0$, the iterates of SEM converge to θ^* exponentially fast:*

$$|\theta^t - \theta^*| \leq \rho^t |\theta^0 - \theta^*|, \text{ with } \rho = \exp\left(-\frac{\min\{\theta^0, \theta^*\}^2}{2}\right). \tag{25}$$

Moreover, there is a $\theta_{fast} \in (0, \theta^*)$ depending only on θ^* and α^* such that if SEM and EM are both initialized at $\theta^0 \in [\theta_{fast}, \infty)$, then

$$|\theta^t - \theta^*| \leq |\theta^t_{EM} - \theta^*| \qquad \forall t \geq 0\,, \tag{26}$$

where θ^t_{EM} are the iterates of EM. In other words, when initialized in this region, SEM never underperforms EM.

Although the model (24) is certainly restrictive, it is one of the few cases where the explicit exponential rate of convergence has been established for the EM algorithm. The fact that SEM matches this rate suggests that replacing the ordinary E-step with the new (16a) doesn't entail a substantial slow-down.

6. EXPERIMENTS ON SIMULATED DATA

To empirically study the benefits of optimization of L, we conducted a series of experiments on simulated mixtures of Gaussians. We compared the result of the optimization of L to two competitive baselines: the EM algorithm and K-means. For initialization, we used the K-means++ method (Arthur and Vassilvitskii, 2007) as implemented in Pedregosa et al. (2011). This initialization samples θ^0_1 at random from data, and subsequently, each θ^0_k is sampled from the empirical distribution of not already chosen data points with probability proportional to D^2, where D is the minimum distance between each remaining data-point and the

previously selected $\theta_1^0, \theta_{k-1}^0$. We used several initial seeds and reported individual realizations or the best among different seeds. To select the best seed, we use the inertia function (Pedregosa et al., 2011) for K-means and the log-likelihood for L and l (we did not observe significant differences by using the same L as selection criteria for L).

We study how results change as a function of the number of components K, the variance of components σ^2 (or Σ in the non-spherical case), dimension d, the number of samples n, and α. We divide the experiments into three parts: known weights and known K (Section 6.1), Unknown weights and known K (Section 6.2), and known weights and unknown K (Section 6.3).

We considered two performance metrics: first, as a direct measure of how well the parameters were inferred, we use the squared average distance between true and fitted centers, which we call simply the *error*:

$$error(\theta, \theta^*) = \min_{\pi \in Perm(K)} \frac{1}{K} \sum_{k=1}^{K} ||\theta_{\pi(k)} - \theta_k^*||^2. \tag{27}$$

Since centers are defined up to permutations, we must search among all possible permutations $Perm(K)$ of K indexes, a computation that we state as an optimal transport problem Flamary et al. (2021). Second, we used the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) to measure the similarity between the actual and fitted clustering solutions. The ARI is a number between 0 and 1, with values closest to one indicating greater similarity.

6.1 Known weights and known K

We study clustering performance as a function of number of components K and a variance parameter σ^2. We consider the following experimental setups: i)spherical variance with known variance σ^2, ii) elliptical variance with diagonal entries sampled from a uniform distribution between $0.5\sigma^2$ and $1.5\sigma^2$, iii) and iv) same as i) and ii) but with unknown variances to be estimated. For each parameter configuration, we sample a number of $n_{exp} = 200$ datasets, and on each of them, we run three methods for a number of $n_{seed} = 5$ different random K-means++ initializations. We considered two sample sizes $n = 200$ and $n = 1000$ and dimensions $d = 2, 5, 10, 20$. In all cases, we use equal weights $\alpha_k = 1/K$. In cases where σ^2 must be inferred, we use the identity matrix as an initial estimate.

Fig. 3 shows a summary for experiment i). Both EM and Sinkhorn-EM typically improve upon the K-means baseline, although the improvement for SEM is the largest (Fig. 3A, B). As a result, Sinkhorn-EM attains the best performance among the three algorithms by far, whether at the level of individual seeds or the best seed. The interplay between relative performance and σ^2 and K is complex: although differences are observed at each K if K is small (e.g., $K \leq 10$), these differences vanish once considering the best seed, indicating that EM and K-means are still capable of finding the global solution. However, for larger values of K, Sinkhorn-EM consistently avoids bad local optima while EM and K-mean++ struggle, even with several seeds. The performance of Sinkhorn-EM also deteriorates for large K, suggesting that it cannot entirely escape bad local optima. Similarly, if σ^2 increases, the performance of all algorithms deteriorates, but Sinkhorn-EM performs consistently better, in line with Theorem 2 which imposes a minimum separation condition so that the guarantee holds.

Also, if σ^2 is too small then all algorithms will recover the true solution. This suggests that there is a mid-range of values for which Sinkhorn-EM is most beneficial. In Appendix B we provide a more detailed account of experimental results by breaking them down into further experimental conditions, and here we give a brief summary: first, by observing that similar results are obtained for $n = 200$ and $n = 1000$, suggesting that we can ignore the finite-sample related errors (Figs B.3 and B.4). Second, perhaps unintuitive, larger values of d are associated with lower error for three algorithms. This is a consequence that the K-means++ initialization consistently produces better estimates in such cases, which is helpful in all algorithms. Again, there is a range of values of d where Sinkhorn-EM is most beneficial, and this range depends on σ^2 as well.

Results for experiments ii), iii) and iv) are also presented in Appendix B. Sinkhorn-EM still attains the lowest regardless of whether the true variance is spherical or diagonal (and different for each cluster) (Fig. B.5) and whether the variance is fixed or has to be inferred (Figs. B.6, B.7).

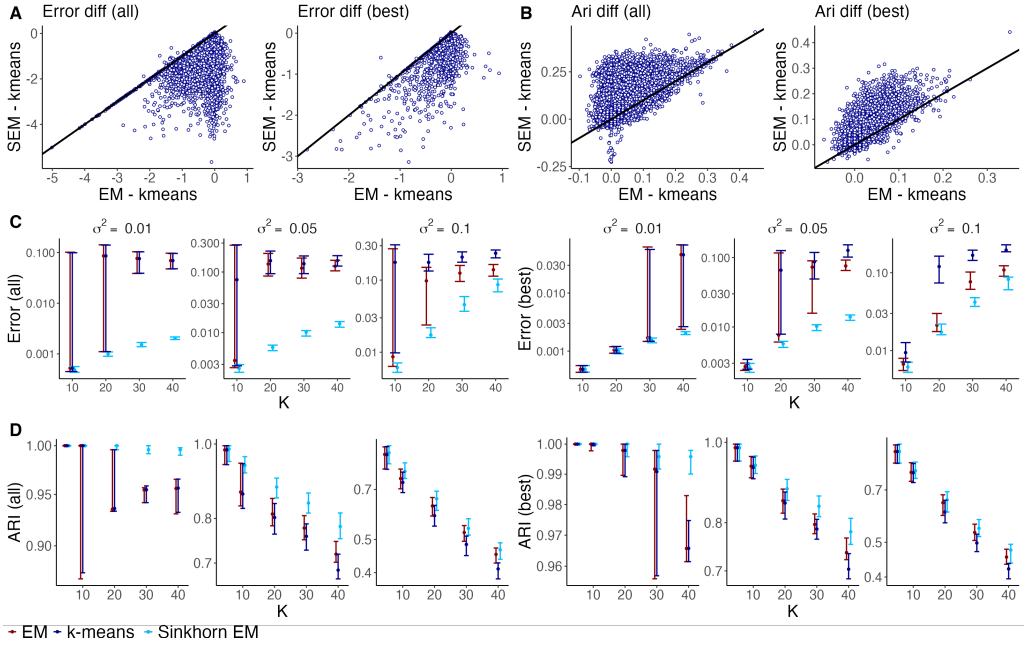

Figure 3: Results for experiment in Section 6.1 with known weights $\alpha_k = 1/K$ and known K, for a mixture of spherical Gaussians with variance σ^2. **A**: Error difference between Sinkhorn-EM and K-means (y-axis) and between EM and K-means (x-axis). Each random seed is considered individually in the left plot, and in the right plot, we consider the best among five seeds. **B**: same as **A** but with ARI score. **C**: Comparison of errors for each algorithm for varying K and σ^2. The error bar represents the interquartile range. **D**: same as **C** but with ARI score.

6.2 Unknown weights and known K

The setup is the same as experiment i) in Section 6.1, but with unknown (and not necessarily uniform) weights. We treat α as a parameter and update it using (20) by performing coordinate descent on θ and α until convergence. In detail, to update θ we apply Sinkhorn-EM with current α^t until convergence,

leading to θ^{t+1}. To update α we successively apply mirror descent updates (20) with current θ^t until convergence, leading to α^{t+1}. On each update, we start with $\eta = 1$ and make $\eta \leftarrow \eta/2$ until L decreases with respect to initial α^t. We always initialize α as the vector of uniform weights and study performance as a function of deviations from uniformity that we measure with the parameter γ, the concentration of a Dirichlet distribution. For each γ, we consider $n_{exp} = 200$ experiments where weights are sampled from a Dirichlet distribution with parameters γ/K. Smaller values of γ indicate a larger deviation from the uniform distribution. For comparison, we also include the algorithm that doesn't update weights as a baseline, i.e., we perform inference on a model with misspecified weights $\alpha_k = 1/K$.

Results are summarized in Fig. 4. Benefits of SEM algorithm are observed for moderately large values of γ (e.g., $\gamma > 10$) whenever weight updates are applied. If weights are not updated, however, the performance of SEM and EM are similar. This is a remarkable result: although we lack a local optima theory for L in the case of unequal and/or unknown weights, we observe benefits if parameter (θ) updates are coupled with weight updates. The fact that benefits are not observed if weights are not updated indicates that this benefit doesn't come from "continuity".

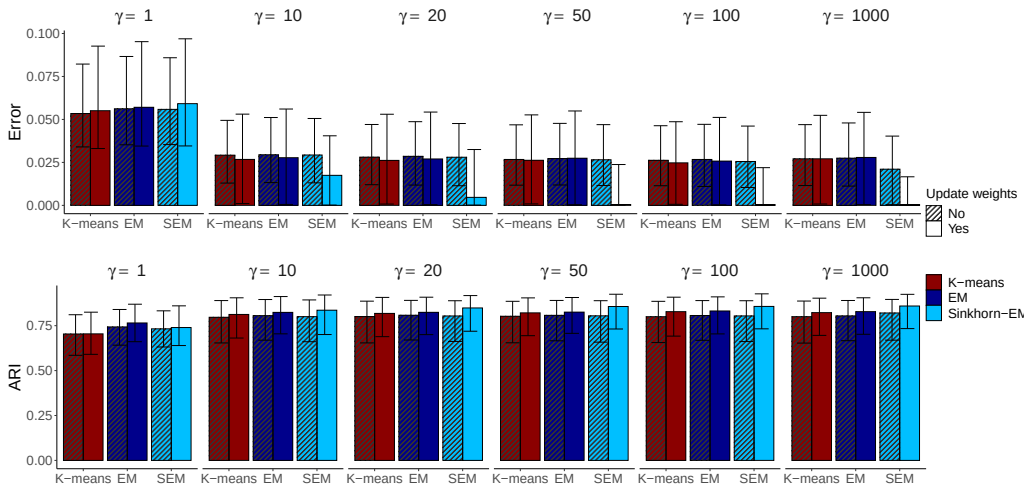

Figure 4: Results for the unknown weights case in Section 6.2. Bars indicate the interquartile range. For mixtures with close-to-uniform weights (γ large) Sinkhorn-EM is the most advantageous algorithm, and the weight update procedure defined in Section 3.2 reduces the error with respect to keeping the (wrong) weights fixed.

6.3 Known weights and unknown K

We treat the unknown number of components case as a model selection problem over K. For each true K, we fit the model using several candidate models with $K_{model} \in \{K-5, K+5\}$ and estimate $\hat{K}$ as customary, via the Bayesian Information Criterion (BIC) (Kass and Wasserman, 1995; Smith and Spiegelhalter, 1980) (not available for K-means). We quantify the estimation error as the difference $K - \hat{K}$ between the actual and inferred number of components. Compared

to EM, SEM recovers the true number of components much more often. Moreover, in our experiment, SEM never overestimates the number of components ($\hat{K} > K$). In Appendix B (Fig. B.8) we provide detailed results for different choices of parameters.

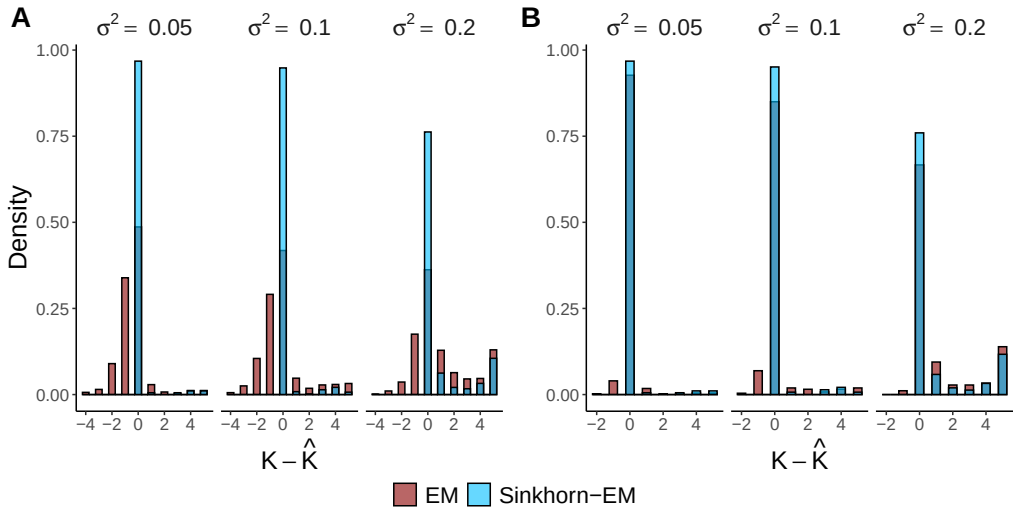

Figure 5: Number of components estimation error histograms for the experiment in Section 6.3. **A**: results for different seeds. **B**: results for the best seed

7. APPLICATION TO IMAGE SEGMENTATION IN C.ELEGANS MICROSCOPY DATA

One of the conclusions of Section 6 is that SEM is most beneficial when the number of clusters K is large. In this Section, we show that this observation materializes on a real-world neural segmentation task that can be framed as the fitting of a mixture of a GMM with multiple components (neurons).

C. elegans is a roundworm used as a model organism in neuroscience for decades due to its stereotypic brain organization and simple structure consisting of 302 neurons. Automated neuron identification and segmentation of *C. elegans* is crucial for conducting high-throughput experiments for many applications, including the analysis of gene expression profiles, cell fate studies (Sulston et al., 1983), stem cell research, and the study of circuit-level neuronal dynamics (Kato et al., 2015). NeuroPAL, Yemini et al. (2021) a recently introduced novel transgenic strain of *C. elegans* has a deterministic coloring of each neuron, enabling the disambiguation of nearby neurons and aiding in their identification Varol et al. (2020). We treat the segmentation problem of neurons in NeuroPAL *C.elegans* images as one of clustering using Gaussian Mixture Models. Given a colorful volume represented by six dimensions (three spatial coordinates and three dimensions for the RGB colors), we aim to recover the locations centers θ_k of the K neurons in the recorded volume, along with their shapes Σ_k. We can then segment the image using the responsibilities Ψ_k, which encode the probabilistic assignment of the pixels to the cells.

In Fig. 6, we compare segmentation using the EM algorithm as a baseline to compare against. Fig. 6B illustrates the primary mode of failure of EM that

explains why SEM outperforms it: EM may typically collapse two nearby cells into a single component, a pathology that SEM most often avoids. This collapse can be directly understood in terms of the "many-fit-one" local optima described in Section 4.

These results indicate that SEM is a valuable alternative to the more traditional EM in this real-world setup characterized by a dense mixture of Gaussians with many ($K > 20$ clusters). Our results were first hinted by Nejatbakhsh et al. (2020), and here are able to further elaborate on them and to provide a theoretical understanding of the supremacy of SEM. Details of our experiments appear in Appendix B.2.

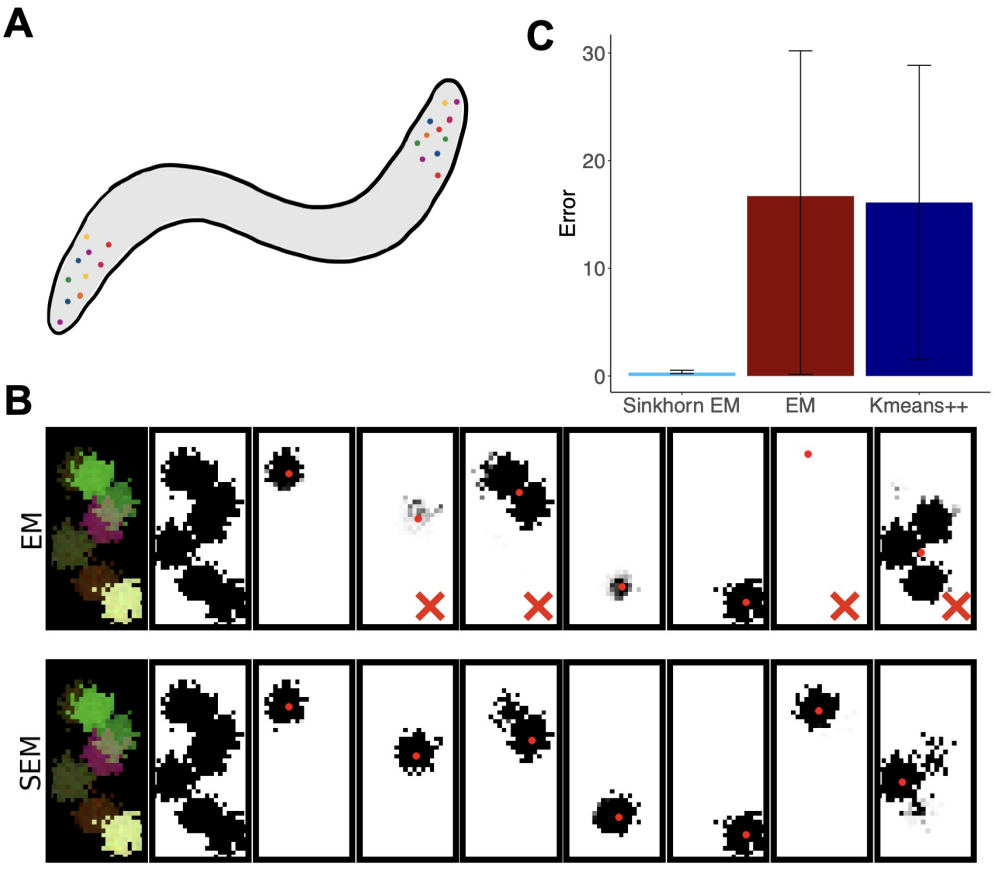

Figure 6: Results on a C.elegans segmentation task. **A**: An illustration of a worm's brain. The task is identifying different locations (coloured dots) from volumetric images. **B**: Comparison between a typical segmentation outcome of EM and Sinkhorn-EM algorithms. The first column shows a true microscopic image containing a subset of neurons. The second column shows pixels containing neurons. All remaining columns show identified neurons for Sinkhorn-EM and EM methods as indicated by the responsibilities Ψ_k for each neuron over pixels (grey scale). EM tends to collapse neuronal shapes, a problem averted by Sinkhorn-EM. Red dots indicate true neural centers and red crosses indicate failures in the identification of individual components. **C**: Sinkhorn-EM consistently leads to better segmentation performance as compared to EM and K-means++ competitors.

8. APPLICATION TO CO-CLUSTERING IN SPATIAL TRANSCRIPTOMICS

Finally, we study an extension of our methodology to the problem of model-based co-clustering and demonstrate benefits on a high-dimensional genomics dataset. Given a matrix Y with dimensions N and M, the co-clustering (or bi-clustering) problem Govaert and Nadif (2013) is the one of how to simultaneously cluster the rows and columns of X, as opposed to the usual setup where only the rows of X (i.e., observations) are clustered. The co-clustering problem has a long history in statistics (see, e.g., Good (1965); Hartigan (1972)) and has found relevant applications in fields such as text analysis (Dhillon, 2001) and bioinformatics (Cheng and Church, 2000; Tan and Witten, 2014). Here, we focus on the model-based formulation (Bouveyron et al., 2019; Govaert and Nadif, 2013) based on the maximization of a loss function (e.g., the log-likelihood) that depends on the density of the observed data given model parameters. Model-based co-clustering stands out as an example where we expect to obtain benefits from entropic OT since likelihood-based approaches are known to be severely affected by bad local-optima (Bouveyron et al., 2019).

The underlying probabilistic model here is the so-called latent block model. For a co-clustering model with K clusters in the row dimension and G clusters in the column dimension, let $z \in \mathbb{R}^{N\times K}$ be a binary matrix representing latent assignments of rows in Y to a class $k \in [K]$ (so that $\sum_k z_{i,j} = 1$). Likewise, we represent latent column assignments with a matrix $w \in \mathbb{R}^{M\times G}$. Then, there are $K\times G$ possible assignments for a particular $Y_{i,j}$ of Y. We assume that these entries are conditionally independent knowing z and w so that for certain parametric family of densities $\phi(\cdot,\theta)$ we express

$$P(Y|z,w,\theta) = \prod_{i,j,k,g} \phi\left(Y_{i,j},\theta_{k,g}\right)^{z_{i,k}w_{j,g}}.$$

If we denote the row and column mixture proportions $\pi_k = P(z_{i,k}=1)$ and $\rho_g = P(w_{j,g}=1)$ (these could also be treated as parameters) then the marginal likelihood of Y writes as the following mixture

$$\begin{aligned} q_\theta(Y) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} p(z)p(w)f(Y|z,w,\theta) \\ &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k}\pi_k^{z_{i,k}} \prod_{j,g}\rho_g^{w_{j,g}} \prod_{i,j,k,g} \phi\left(Y_{i,j},\theta_{k,g}\right)^{z_{i,k}w_{j,g}}. \end{aligned} \tag{28}$$

Evaluation of the above likelihood is intractable, a problem that carries forward to the computation of the E step for the EM algorithm targeting (28): if the above equation corresponded to a usual mixture model (e.g. making $G=1$) then z_i would be uniquely associated to a $Y_{i,}$ and we would be able to express the likelihood as a product over the *sample*. However, the complex simultaneous dependence on z and w in (28) prevents us from achieving this sample representation for the likelihood above, so a sum with exponentially many terms needs to be computed.

It is still possible to deal with intractability with approximate methods: here we consider a Variational EM algorithm (VEM) (Bouveyron et al., 2019; Nadif and

Govaert, 2008) based on a factored approximation for the joint conditional probability $P(z_{i,k}w_{j,g} = 1|Y, \theta) \approx P(z_{i,k} = 1|Y, \theta)P(w_{j,g} = 1|Y, \theta)$. This algorithm, detailed in the Appendix B.3, corresponds to the iterative alternate application of the usual EM algorithm to cluster the rows and columns of Y until convergence. The main drawback of this algorithm is sensitivity to initial values due to pervasive bad-local optima. To deal with this issue, we implement Sinkhorn VEM (SVEM), the algorithm that replaces each call of the EM algorithm with SEM.

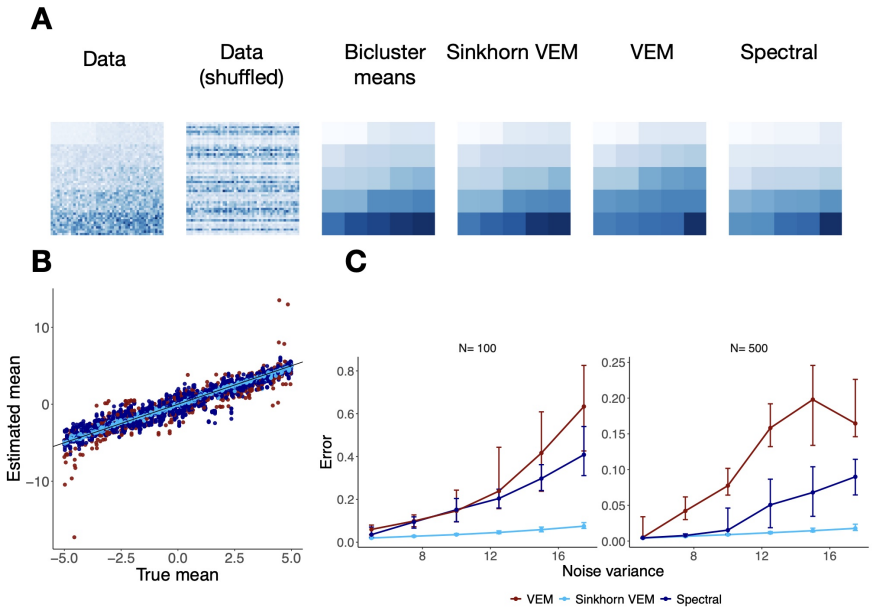

Figure 7: Comparison of three Co-clustering methods based on simulated data. **A**: Setup for each experiment. The first inset contains data, each pixel is a noise-corrupted version of the corresponding co-cluster (each sub-square, also shown in the third subplot). Rows and columns of the data matrix are displayed in increasing order in the co-cluster for visualization purposes. The second inset is the same data matrix in the true shuffled order. The third inset shows the (ordered) co-cluster means. The last three columns are co-clusters recovered by each of the three methods. **B**: Scatterplot of true vs estimated co-cluster of means for three methods over thousands of experimental repetitions. **C**: Dependence of error on noise parameter σ^2 and sample sizes $N = M = 100, 500$ (so that the data matrix Y had dimension $N \times N$). Sinkhorn-VEM consistently produces the most accurate co-clustering estimates.

8.1 Experiments on Synthetic Data

We compare VEM, Sinkhorn-VEM and the competitive spectral co-clustering as implemented in Pedregosa et al. (2011) on simulated data sampled from the simple Gaussian generative model:

$$Y_{i,j}|(z_{i,k} = 1, w_{j,g} = 1, \theta) \sim \mathcal{N}\left(\theta_{k,g}, \sigma^2\right).$$

By performing thousands of experiments we studied differences between estimated and actual $\theta_{k,g}$ for the three methods at different noise levels σ^2 and number of co-clusters K^2. Results are summarized in Fig. 7, Sinkhorn-VEM vastly outperforms both spectral and VEM co-clustering. Details appear in the Appendix B.3.

8.2 Application to Spatial Transcriptomics

Spatial Transcriptomics is an umbrella for the group of technologies enabling the transcriptomic (i.e., gene expression) profiling of samples (e.g., single-cell RNA sequencing) with spatial resolution. It was named method of the year (Marx, 2021) because of its promise to transform our understanding of biology and pathology by providing a more comprehensive molecular characterization of the living tissue (Bressan et al., 2023).

A typical spatial transcriptomics experiment is represented by a matrix $Y \in \mathbb{R}^{N \times M}$ containing gene expression levels of M genes across N locations or *spots*. This type of structure motivates two intertwined research questions: i) whether this high-dimensional expression data clusters coherently in regions in a way that resembles the anatomy of the tissue, and ii) whether there are genes whose expression level depends on space. As suggested by Sottosanti and Risso (2023), it is possible to simultaneously address these two questions by bringing a co-clustering perspective.

We compared the performance of three methods on one dataset from the human dorsolateral prefrontal cortex (DLPFC) from the Liebler Institute for Brain Development (Maynard et al., 2021). Results are shown in Fig. 8. The clustering of spots given by SVEM more faithfully represents the layer organization of the tissue, even if no spatial information has been explicitly encoded in the model.

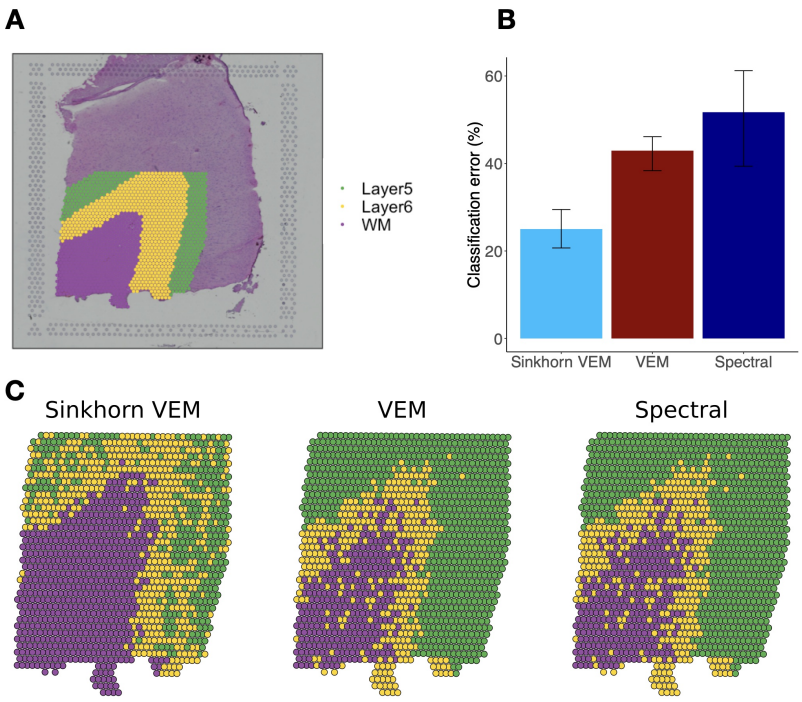

Figure 8: Results for a Spatial Transcriptomic experiment. **A** Tissue sample of DLPFC, with colored dots representing spatial locations where corresponding to three layers (Layer 5, Layer 6 and White Matter) where gene expression vectors were available. **B**. Comparison of clustering performance for the three methods, averaged over a hundred of repetitions. Bars represent 95% confidence intervals. **C**: example of a typical recovered solutions. SVEM most faithfully captures the true underlying histological characterization, even if the model did not include any spatial information.

9. DISCUSSION AND FUTURE DIRECTIONS

Entropic optimal transport endows us with a valuable methodology for model-based clustering, often avoiding the pathologies that the log-likelihood would otherwise suffer. Although our most specialized results pertain to the inference of location parameters in GMMs with equal variances, we stress that the validity of our methodology is more general in many aspects: first, the convergence of Sinkhorn-EM to local optima is still guaranteed even in finite samples. Also, it can be applied to other distributions besides the Gaussian. Third, it is possible to optimize over parameters other than means, such as variances, and these variances need not be equal. Fourth, we showed that just as with the log-likelihood, it is also possible to optimize over weights α although this requires gradient steps that are not understood under the Sinkhorn-EM algorithm.

These promising results motivate three lines of future work. First, our Theorem 2 describes a particular setup where we avoid a type of local optima. However, a characterization of local optima structures for the entropic optimal transport loss is still lacking. One promising approach for such characterization would be extending the results of Chen and Xi (2020) to the unequal case weights.

Second, although the main components of our methodology can be applied in finite-sample and population regimes (Theorem 1 and Proposition 1 (a) and (c)), our most involved theoretical results (Theorems 2, 4,3) are based on a study of the population landscape of L. However, the fact that our successful empirical results are based on samples (sometimes with modest size) motivates a more in-depth analysis of the finite case. Regarding the EM algorithm, some thorough finite-sample convergence analyses have appeared in the last few years (e.g. Balakrishnan et al. (2017); Daskalakis et al. (2017b); Dwivedi et al. (2020)), and they all are based on establishing a non-asymptotic uniform law of large numbers for controlling the difference between the empirical F^n and population F iterates. Unfortunately, establishing such a law for SEM would require a non-asymptotic convergence rate analysis for the empirical transportation plans π^n on a non-compact setup, which is unavailable today. Still, a flurry of recent results in this direction (Groppe and Hundrieser, 2023; Masud et al., 2023; Pooladian et al., 2023; Rigollet and Stromme, 2022) provide us with a novel set of tools for a finite-sample analysis of SEM. Likewise, it may be possible to extend the stability results on empirical risk minimization Mei et al. (2016) to provide finite-sample statements on local optima for L.

The third line consists of identifying the most beneficial setups: we found that SEM works the best whenever there is a sufficiently large number of clusters and if the variance is large enough, but we did not observe an important effect of dimension. However, in some high-dimensional setups, it is reasonable to expect benefits from SEM: for example, if covariances need to be estimated and if $n \approx 2d$ then EM is likely to become unstable since responsibilities can favor one of the clusters, leaving few data points for estimation of a covariance matrix. As the loss L is defined by enforcing a mass splitting, that problem would be entirely avoided.

Our approach has two main limitations. First, although the inferential gains are substantial, computation of the entropic loss L signifies a non-negligible extra computational burden, so the use of our methodology should focus on cases where the gains justify the extra cost; i.e. where we might not get the right optimum even when trying from many different seeds. The second limitation is that L may still have many bad local optima. We observed this problem when fitting mixtures with very unequal weights; the existence of such bad local optima led to an overall performance of the Sinkhorn-EM algorithm similar (but no worse) to EM.

10. ACKNOWLEDGEMENTS

Part of this research was performed while the author was visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation (Grant No. DMS-1929348). This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation MTH230027 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF Grant 2138259, 2138286, 2138307, 2137603, and 2138296. The author expresses gratitude to Jonathan Niles-Weed, Amin Nejatbakhsh, and Erdem Varol for insightful discussions and their contributions to a preliminary version of this work Mena et al. (2020).

APPENDIX A: OMITTED PROOFS

A.1 Proof of Proposition 1

We already showed (a), (b), and (c) in the main text. We note that the main argument in (b), that the coupling π achieving the infimum in (6) is exactly the joint $Q^{X,Y}_{\theta^*}$ in (3) only works in the population case. In the finite-sample case, the empirical measure $Q^n_{\theta^*}$, i.e., the second prescribed marginal in the (6), doesn't have a density with respect to q_{θ^*}.

The proof of (d) is based on the explicit computation of the second derivative of L with respect to θ. We will base heavily on the semi-dual formulation (8) that we here write

$$L(\theta) = \max_{\omega \in \mathbb{R}^K} \tilde{L}(\theta, \omega),$$

where

$$\tilde{L}(\theta, \omega) = \sup_{\omega \in \mathbb{R}^K} \left[\sum_{k=1}^K \alpha_k \omega_k - E\left(\log \left(\sum_{k=1}^K \alpha_k e^{\omega_k} q_{\theta_k}(y) \right) \right) \right] \tag{29}$$

Note that the supremum above is realized for many ω, as one may add an arbitrary constant to any coordinate of ω without changing the right-hand side. Therefore, we can assume $\omega(K) = 0$ and (slightly abusing notation) that $\omega \in \mathbb{R}^{K-1}$. Note that also clearly

$$\tilde{L}(\theta, 0) = -E(\log q_\theta(Y)). \tag{30}$$

Recall that $\omega(\theta)$ the one that achieves the maximum in (8). We will follow an envelope-theorem-like argument: we have $L(\theta) = \tilde{L}(\theta, \omega(\theta))$, and based on this, we may compute first and second derivatives using the chain rule (the notation for the chain rules below is intentionally loose to avoid clutter)

$$\frac{\partial L}{\partial \theta}(\theta) = \frac{\partial \tilde{L}}{\partial \theta}(\theta, \omega(\theta)) + \frac{\partial \tilde{L}}{\partial \omega}(\theta, \omega(\theta)) \frac{\partial \omega(\theta)}{\partial \theta}, \tag{31}$$

where

$$\begin{aligned} \frac{\partial^2 L}{\partial \theta^2}(\theta) \quad = \quad & \frac{\partial^2 \tilde{L}}{\partial \theta^2}(\theta, \omega(\theta)) + \frac{\partial^2 \tilde{L}}{\partial \omega \partial \theta}(\theta, \omega(\theta)) \frac{\partial \omega(\theta)}{\partial \theta} + \left(\frac{\partial^2 \tilde{L}}{\partial \theta \partial \omega}(\theta, \omega(\theta)) \right)^\top \frac{\partial \omega(\theta)}{\partial \theta} \\ & + \frac{\partial \omega(\theta)}{\partial \theta}^\top \frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta, \omega(\theta)) \frac{\partial \omega(\theta)}{\partial \theta} + \frac{\partial \tilde{L}}{\partial \omega}(\theta, \omega(\theta)) \frac{\partial^2 \omega(\theta)}{\partial \theta^2}. \end{aligned} \tag{32}$$

But by optimality of $\omega(\theta)$, for every θ we have

$$0 = \frac{\partial \tilde{L}}{\partial \omega}(\theta, \omega(\theta)) \text{ and } \quad 0 = \frac{\partial^2 \tilde{L}}{\partial \theta \partial \omega}(\theta, \omega(\theta)) + \frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta, \omega(\theta)) \frac{\partial \omega(\theta)}{\partial \theta}. \tag{33}$$

Therefore, by combining (32) and (33), and assuming that the second derivative of $\tilde{L}$ w.r.t. ω is invertible we obtain

$$\frac{\partial^2 L}{\partial \theta^2}(\theta) = \frac{\partial^2 \tilde{L}}{\partial \theta^2}(\theta, \omega(\theta)) - \frac{\partial^2 \tilde{L}}{\partial \omega \partial \theta}(\theta, \omega(\theta)) \left(\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta, \omega(\theta)) \right)^{-1} \frac{\partial^2 \tilde{L}}{\partial \theta \partial \omega}(\theta, \omega(\theta)). \tag{34}$$

We now show that this second derivative is negative definite (hence invertible): note that

$$\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta,\omega) = -E\left(Diag(\Psi(Y,\theta,\alpha(\omega)) - \Psi(Y,\theta,\alpha(\omega))\Psi(Y,\theta,\alpha(\omega))^\top\right), \quad (35)$$

where $\Psi(Y,\theta,\alpha)$ is the vector of responsibilities defined in (13) for $k = 1,\dots,K-1$, and where

$$\alpha(\omega)_k = \frac{\alpha_k e_k^\omega}{\sum_{k=1}^K \alpha_k e_k^\omega},$$

so that $\alpha(\omega(\theta)) = \alpha(\theta)$ as defined in (9). Now, define the (symmetric) matrix $I(\theta,\omega)$ as the extension of $\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta,\omega)$ to the entire range of indexes $k = 1,\dots K$, i.e., $I(\theta,\omega)$ is the right hand side in (35) viewed this time as a $K \times K$ matrix. By definition this matrix coincides with $\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta,\omega)$ for $k \leq K-1$. Note that $-I(\theta,\omega)$ is the Laplacian matrix of a weighted graph since

$$\begin{aligned}
\sum_{k'=1,k'\neq k}^K I(\theta,\omega)_{k,k'} &= \sum_{k'=1,k'\neq k}^K E\left(\Psi_k(Y,\theta,\alpha(\theta))\Psi_{k'}(Y,\theta,\alpha(\theta))\right) \\
&= E\left(\Psi_k(Y,\theta,\alpha(\theta)) \sum_{k'=1,k'\neq k}^K \Psi_{k'}(Y,\theta,\alpha(\theta))\right) \\
&= E\left(\Psi_k(Y,\theta,\alpha(\theta))\left(1 - \Psi_k(Y,\theta,\alpha(\theta))\right)\right) \\
&= -I(\theta,\omega)_{k,k}.
\end{aligned}$$

Then, I is a negative weighted Laplacian matrix and

$$x^\top I x = \frac{1}{2}\sum_{k,k'}^K I(\theta,\omega)_{k,k'}(x_k - x_{k'})^2 \leq 0.$$

The above expression is zero only if x is a constant vector since all entries of $I(\theta,\omega)$ are positive (they are expectations of a strictly positive variable with respect to a Gaussian measure). Since $\frac{\partial^2 \tilde{L}}{\partial w^2}(\theta,\omega(\theta))$ is a submatrix of I, it is also negative semidefinite. Now, suppose $z^\top \frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta,\omega)z = 0$, then, if $x_k = z_k$ for $k \leq K-1$ and $x_K = 0$ we have $z^\top \frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta,\omega)z = x^\top I x = 0$ and since x must be constant, $z = 0$. Therefore, $\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta^*,0)$ is negative definite and the proof is concluded.

We now evaluate at $\theta = \theta^*$ and $\omega = \omega(\theta^*)$. Since $\omega(\theta^*) = 0$ (the same argument as the proof of (b)), we also have

$$\frac{\partial^2 \tilde{L}}{\partial \theta^2}(\theta^*,0) = -\frac{\partial^2}{\partial \theta^2}E(\log q_{\theta^*}(Y)) = \frac{\partial^2}{\partial \theta^2}\ell(\theta^*). \quad (36)$$

Therefore,

$$\nabla^2 L(\theta^*) = \nabla^2 \ell(\theta^*) - \frac{\partial^2 \tilde{L}}{\partial \omega \partial \theta}(\theta,0)\left(\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta^*,0)\right)^{-1}\frac{\partial^2 \tilde{L}}{\partial \theta^* \partial \omega}(\theta^*,0). \quad (37)$$

and we conclude by identifying

$$A(\theta^*) = -\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta^*,0)^{-1}, B(\theta^*) = \frac{\partial^2 \tilde{L}}{\partial \theta^* \partial \omega}(\theta^*,0)$$

A.2 Proof of Corollary 2

Let θ be a stationary point of $L(\theta)$. By Proposition 1(c) it satisfies $\nabla\ell(\theta, \alpha(\theta)) = 0$. In the GMM of equation (17) this means that θ is a stationary point on a GMM with tilted weights $\alpha(\theta)$ instead of α. This statement is equivalent to saying

$$\frac{\partial \tilde{L}}{\partial \theta}(\theta, \omega(\theta)) = \frac{\partial \ell}{\partial \theta}(\theta, \alpha(\theta)).$$

Likewise, by taking derivatives in the definition of $\tilde{L}$ we can verify that

$$\frac{\partial^2 \tilde{L}}{\partial \theta^2}(\theta, \omega(\theta)) = \frac{\partial^2 \ell}{\partial \theta^2}(\theta, \alpha(\theta)),$$

and by (34) this means that

$$\nabla^2 L(\theta) = \frac{\partial^2 \ell}{\partial \theta^2}(\theta, \alpha(\theta)) - \frac{\partial^2 \tilde{L}}{\partial \omega \partial \theta}(\theta, \omega(\theta)) \left(\frac{\partial^2 \tilde{L}}{\partial \omega^2}(\theta, \omega(\theta))\right)^{-1} \frac{\partial^2 \tilde{L}}{\partial \theta \partial \omega}(\theta, \omega(\theta)).$$

So, if θ is a local minimum for ℓ then $\frac{\partial^2 \ell}{\partial \theta^2}(\theta, \alpha(\theta))$ is positive definite. By the same arguments as in the proof of Proposition (1)(d), the second term on the right-hand side above is positive semidefinite. Therefore, in that case, $\nabla^2 L(\theta)$ is positive definite, i.e., θ is a local minimum for L.

A.3 Proof of Theorem 1

Proof. The proof borrows from the one for the EM algorithm introduced in Wu (1983). For an arbitrary coupling π between the set measure α_k in the set $[K]$ and the marginal distribution of $Y \sim Q_*$ we define the function

$$m(\theta, \pi) = -E\left(\sum_{k=1}^{K} \pi(k|Y) \log\left(\alpha_k q_{\theta_k}(Y)\right)\right).$$

By the definition of the M step (16b) after having computed π^{t+1} from the previous E-step (16a) we have, since θ^{t+1} minimizes the function $m(\cdot, \pi^{t+1})$:

$$m(\theta^{t+1}, \pi^{t+1}) \leq m(\theta^t, \pi^{t+1})$$

Here we are implicitly using the fact that since θ^t is fixed, there is one-to-one correspondence between a coupling $\tilde{\pi} \in \Pi(P_{\theta^t}, Q_*)$ (solving the E step) with a coupling $\pi \in \Pi(\alpha, Q_*)$. We can add the relative entropy term $H\left(\pi^{t+1}|\alpha \otimes Q\right)$ to obtain

$$\begin{aligned}
m(\theta^{t+1}, \pi^{t+1}) + H\left(\pi^{t+1}|\alpha \otimes Q_*\right) &\leq m(\theta^t, \pi^{t+1}) + H\left(\pi^{t+1}|\alpha \otimes Q_*\right) \\
&= -E\left(\sum_{k=1}^{K} \pi^{t+1}(k|Y) \log\left(\alpha_k q_{\theta_k^t}(Y)\right)\right) + H\left(\pi^{t+1}|\alpha \otimes Q_*\right) \\
&= H\left(\pi^{t+1}|\alpha \otimes Q_*\right) \\
&= \inf_{\tilde{\pi} \in \Pi\left(P_{\theta^t}, Q_*\right)} H\left(\tilde{\pi}|Q_{\theta^t}^{X,Y}\right) \\
&= L(\theta^t).
\end{aligned}$$

In the second-to-last equality, we used that π^{t+1} solves the previous E step and in the last, we used the definition of $L(\theta^t)$. To conclude, we note that the left-hand side above also expresses as $H\left(\tilde{\pi}^{t+1}|Q^{X,Y}_{\theta^{t+1}}\right)$ but the coupling π^{t+1} may not be optimal for the problem defining $L(\theta^{t+1})$, so

$$L(\theta^{t+1}) \leq H\left(\tilde{\pi}^{t+1}|Q^{X,Y}_{\theta^{t+1}}\right) \leq m(\theta^{t+1}, \pi^{t+1}) + H\left(\pi^{t+1}|\alpha \otimes Q_*\right) \leq L(\theta^t).$$

We now show the inequality is strict if θ^t is not a stationary point. Note that $L(\theta) = m(\theta, \pi_\theta)$ where π_θ minimizes $m(\theta, \cdot)$. By the chain rule and optimality of π_θ we have

$$\frac{\partial L}{\partial \theta}(\theta) = \frac{\partial m}{\partial \theta}(\theta, \pi_\theta) + \frac{\partial m}{\partial \pi}(\theta, \pi_\theta)\frac{\partial \pi_\theta}{\partial \theta} = \frac{\partial m}{\partial \theta}(\theta, \pi_\theta). \tag{38}$$

Since θ^t is not a stationary point for L the above implies that $\frac{\partial m}{\partial \theta}(\theta^t, \pi_{\theta^t}) \neq 0$. Therefore, θ^t cannot globally minimize the (negative of) the function defining the M-step $m(\cdot, \pi_{\theta^t}) = m(\cdot, \pi^{t+1})$ implying that the M step leads to a strict decrease of this function, i.e. $m(\theta^{t+1}, \pi^{t+1}) < m(\theta^t, \pi^{t+1})$. By the same argument as above this implies $L(\theta^{t+1}) < L(\theta^t)$. □

A.4 Proof of Proposition 2

The proof is essentially the same as the one for Theorem 1 in Jin et al. (2016), but a slightly more careful argument is needed for the two-dimensional computations. The negative log-likelihood $\ell(\cdot)$ here writes (θ^j_k represents the $j-$th coordinate of θ_k):

$$\ell(\theta) = -E\left(\log\left(\sum_{k=1}^{3} e^{-\frac{(x-\theta^1_k)^2}{2}} e^{-\frac{(y-\theta^2_k)^2}{2}}\right)\right) + \log(6\pi).$$

Define $\tilde{\theta}_1 = (0,0)$ and $\tilde{\theta}_2 = \tilde{\theta}_3 = (R, 0)$. Clearly, $\theta = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3)$ is in the interior of $\mathcal{R}^\varepsilon$ for every $\varepsilon > 0$. Let's compute the likelihood of this configuration when $R \to \infty$:

$$m_0 := \lim_{R\to\infty} \ell(\tilde{\theta}) = 1 + \frac{D^2}{3} + \log(6\pi) - \frac{\log 2}{3}.$$

For a fixed vector of second coordinates $\bar{\theta}^2$ we consider the regions in R^6

$$\begin{aligned}
R_1(\bar{\theta}^2) &:= \{\theta : \theta^1_1 = \frac{R}{3}, \theta^1_2 \geq \frac{2R}{3}, \theta^1_3 \geq \frac{2R}{3}, \theta^2 = \bar{\theta}^2\},\\
\mathcal{R}_2(\bar{\theta}^2) &:= \{\theta : \theta^1_1 \leq \frac{R}{3}, \theta^1_2 = \frac{2R}{3}, \theta^1_3 \geq \frac{2R}{3}, \theta^2 = \bar{\theta}^2\},\\
\mathcal{R}_3(\bar{\theta}^2) &:= \{\theta : \theta^1_1 \leq \frac{R}{3}, \theta^1_2 \geq \frac{2R}{3}, \theta^1_3 = \frac{2R}{3}, \theta^2 = \bar{\theta}^2\},
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
&\lim_{\mathcal{R}\to\infty} \inf_{\theta\in R_1(\bar{\theta}^2)} \ell(\theta) = \infty,\\
&\lim_{R\to\infty} \inf_{\theta\in R_2(\bar{\theta}^2)} \ell(\theta) = 1 + \frac{1}{6}\left((D-\bar{\theta}^2_1)^2 + (D+\bar{\theta}^2_1)^2\right) + \frac{\bar{\theta}^2_3}{6} + \log(6\pi),\\
&\lim_{R\to\infty} \inf_{\theta\in R_3(\bar{\theta}^2)} \ell(\theta) = 1 + \frac{1}{6}\left((D-\bar{\theta}^2_1)^2 + (D+\bar{\theta}^2_1)^2\right) + \frac{\bar{\theta}^2_2}{6} + \log(6\pi).
\end{aligned}$$

The first line follows from Jensen's inequality and basic moment relationships for a Gaussian distribution. The second and third limits follow from the fact that asymptotically the infimum is attained at $\theta_1^1 = 0, \theta_2^1 = 2R/3, \theta_3^1 = R, \theta^2 = \bar{\theta}$ for $R_2(\bar{\theta}^2)$ and at $\theta_1^1 = 0, \theta_2^1 = R, \theta_3^1 = 2R/3, \theta^2 = \bar{\theta}$ for $R_3(\bar{\theta}^2)$.
Now, for $\varepsilon > 0$ define $\mathcal{R}_k^\varepsilon := \cup_{||\bar{\theta}|| < \varepsilon} \mathcal{R}_k(\bar{\theta})$ for $k = 1, 2, 3$ so that R_k^ε are the three 5-dimensional faces of $\mathcal{R}^\varepsilon$. Call $m_k^\varepsilon = \lim_{R\to\infty} \inf_{\theta \in \mathcal{R}_k^\varepsilon} \ell(\theta)$. Since the above limits are simultaneously minimized with respect to $\bar{\theta}$ if $\bar{\theta}^2 = 0$ we have that

$$m_1^\varepsilon = \infty, m_2^\varepsilon = m_3^\varepsilon = 1 + \frac{L}{3} + \log(6\pi).$$

Therefore, $m_0 < \min\{m_1^\varepsilon, m_2^\varepsilon, m_3^\varepsilon\}$ so, as in Jin et al. (2016) we conclude the existence of a local minimum θ^ε in the interior of $\mathcal{R}^\varepsilon$, whenever $R = R(\varepsilon)$ is sufficiently large. By the same continuity argument as in Jin et al. (2016), this minimum has a smaller likelihood than the global maximizer θ^*.

A.5 Proof of Theorem 2

Call

$$\Psi_k(Y, \theta) = \frac{\alpha_k(\theta) e^{-\frac{1}{2\sigma^2}||Y-\theta_k||^2}}{\sum_{k=1}^K \alpha_k(\theta) e^{-\frac{1}{2\sigma^2}||Y-\theta_k||^2}},$$

where $\alpha(\theta)$ is the vector of weights arising in the semi-dual optimal transport formulation. By definition of $\alpha(\theta)$, $E(\Psi_k(Y, \theta)) = \frac{1}{K}$. The first-order optimality conditions for θ read

$$E((Y - \theta_k)\Psi_k(Y, \theta, \alpha(\theta)) = 0,$$

or equivalently,

$$E(Y\Psi_k(Y, \theta, \alpha(\theta))) = \frac{1}{K}\theta_k. \tag{39}$$

We now show that this implies that the sum of the true weights must equal the sum of the model weights:

$$\sum_{k=1}^K \theta_k^* = \sum_{k=1}^K \theta_k. \tag{40}$$

Indeed, since $\sum_{k=1}^K \Psi_k(Y, \theta, \alpha(\theta)) = 1$, by adding over all components in (39) we obtain

$$\frac{1}{K}\sum_{k=1}^K \theta_k^* = E(Y) = E\left(Y\left(\sum_{k=1}^K \Psi_k(Y, \theta, \alpha(\theta))\right)\right) = \frac{1}{K}\sum_{k=1}^K \theta_k.$$

First-order optimality also implies a lower bound for averages of groups of θ_k. Indeed, let $S \subseteq [K]$ be an arbitrary set of indexes. Equation (39) implies that

$$\frac{1}{K}\sum_{k\in S} \theta_k = E\left(Y\left(\sum_{k\in S} \Psi_k(Y, \theta, \alpha(\theta))\right)\right). \tag{41}$$

The right-hand side of (41) can be written as $E(Yf(Y))$ for some $f(\cdot) \in [0, 1]$. Although the above relations are multi-dimensional, hereafter, we only need to look at the first coordinate, which w.l.o.g. is our direction of interest, since the

stationary points of L_R on a rotated Gaussian Mixture Model are the rotations of the stationary points of L for the original Gaussian Mixture Model. Likewise, we can assume that the rightmost component of G_1 equals 0 so that G_1 is the group of negative components, $\theta_k^* \leq 0$ for all $k \in G_1$, and that $\theta_k^* \geq 0$ for $k \in G_2$. If $0 \leq \phi(\cdot) \leq \frac{1}{\sqrt{2\pi}}$ and $0 \leq \Phi(\cdot) \leq 1$ are the standard Gaussian pdf. and cdf., respectively, and if E_k denotes expectation under the k-th component (i.e. $\mathcal{N}(\theta_k^*, \sigma^2)$), we have:

$$\begin{aligned} E(Yf(Y)) &\geq E(Y1_{Y\leq 0}) \\ &= \frac{1}{K}\sum_{k=1}^{K} E_k(Y1_{Y\leq 0}) \\ &= \frac{1}{K}\left(\sum_{k=1}^{K}\theta_k^*\Phi\left(-\frac{\theta_k^*}{\sigma}\right) - \sigma\phi\left(\frac{\theta_k^*}{\sigma}\right)\right) \\ &\geq \frac{1}{K}\left(\sum_{k=1}^{K}\theta_k^*\Phi\left(-\frac{\theta_k^*}{\sigma}\right) - \frac{\sigma}{\sqrt{2\pi}}\right) \\ &\geq \frac{1}{K}\sum_{i\in G_1}\theta_k^* - \frac{\sigma}{\sqrt{2\pi}}, \end{aligned} \tag{42}$$

where in the third line, we used simple properties of the truncated Gaussian distribution, and in the fourth, we have used that $0 \leq \Phi(\cdot) \leq 1$ and that $\theta_k^* \geq 0$ in G_2. By combining (41) and (42) we obtain

$$\frac{1}{K}\sum_{k\in S}\theta_k \geq \frac{1}{K}\sum_{k\in G_1}\theta_k^* - \frac{\sigma}{\sqrt{2\pi}}. \tag{43}$$

To conclude, we will use the above relations to obtain a lower bound on δ. Take $S = O^c$, the complement of the group of fitted components in O is δ-close to true components in G_2. We have

$$\frac{1}{K}\sum_{k\in O}\theta_k + \frac{1}{K}\sum_{k\in O^c}\theta_k \geq \frac{1}{K}\sum_{k\in O}\theta_k + \frac{1}{K}\sum_{k\in G_1}\theta_k^* - \frac{\sigma}{\sqrt{2\pi}}.$$

And by (40):

$$\frac{1}{K}\sum_{k\in G_1}\theta_k^* + \frac{1}{K}\sum_{k\in G_2}\theta_k^* \geq \frac{1}{K}\sum_{k\in O}\theta_k + \frac{1}{K}\sum_{k\in G_1}\theta_k^* - \frac{\sigma}{\sqrt{2\pi}},$$

so that

$$\frac{1}{K}\sum_{k\in O}\theta_k - \frac{\sigma}{\sqrt{2\pi}} \leq \frac{1}{K}\sum_{k\in G_2}\theta_k^*.$$

Now, by definition, each true component θ_k^* in G_2 is δ close to at least one fitted component θ_k in O. Conversely, each fitted component in O is close to some θ^* in G_2. More precisely, if for each θ_k in O we call $\tilde{\theta}_k^*$ the true component in G_2 that is δ close to θ_k, we will exhaust G_2 with the $\tilde{\theta}_k^*$ by enumerating O, and there will be $(\tilde{k} - k_2)$ duplicates in G_2 that we call $\tilde{G}_2$. With this observation, we can write

$$\frac{1}{K}\sum_{k\in O}(\tilde{\theta}_k^* - \theta_k) + \frac{\sigma}{\sqrt{2\pi}} \geq \frac{1}{K}\sum_{k\in O}\tilde{\theta}_k^* - \frac{1}{K}\sum_{k\in G_2}\theta_k^*.$$

Using the fact that fitted components are close to true components in G_2

$$\delta\tilde{k} \geq \sum_{k\in O}(\tilde{\theta}_k^* - \theta_k) \geq \sum_{k\in O}\tilde{\theta}_k^* - \sum_{k\in G_2}\theta_k^* - \frac{K\sigma}{\sqrt{2\pi}}.$$

Abusing notation, since we can express the difference of the sums in the right-hand side above in terms of the duplicates $\tilde{\theta}_k^*$ in $\tilde{G}_2$:

$$\delta\tilde{k} \geq \sum_{k\in O}(\tilde{\theta}_k^* - \theta_k) \geq \sum_{k\in \tilde{G}_2}\tilde{\theta}_k^* - \frac{K\sigma}{\sqrt{2\pi}}.$$

Now, since each of the duplicates is a member of G_2 we have $\tilde{\theta}_k^* \geq \theta_l^*$ where θ_l^* is the leftmost component of G_2, which is at least at a distance Δ above $\theta_r^* = 0$, the rightmost component in G_1. Therefore,

$$\delta\tilde{k} \geq (\tilde{k} - k_2)\Delta - \frac{K\sigma}{\sqrt{2\pi}},$$

and the proof is concluded.

A.6 Proof of Corollary 1

Define new rotated axis as described in Fig. 2B with the origin at θ_2^*. The new coordinates are given by

$$\theta_{1,x}^{*,new} = -\frac{2D^2}{\sqrt{R^2+D^2}}, \theta_{2,x}^{*,new} = 0, \theta_{3,x}^{*,new} = \sqrt{R^2+D^2} - \frac{2D^2}{\sqrt{R^2+D^2}}.$$

We split these components in the two leftmost $\theta_1^{*,new}, \theta_2^{*,new}$ and the rightmost $\theta_3^{*,new}$. The minimum separation between groups is given by

$$\Delta(R) = \theta_{3,x}^{*,new} - \theta_{2,x}^{*,new} = \sqrt{R^2+D^2} - \frac{2D^2}{\sqrt{R^2+D^2}} = D\left(\sqrt{\frac{R^2}{D^2}+1} - \frac{2}{\sqrt{\frac{R^2}{D^2}+1}}\right).$$

Taking $\tilde{k} = 2, k_2 = 1, \alpha = 0.45, K = 3$ and $\sigma^2 = 1$ we obtain, by Theorem 2, that for

$$\Delta(R) \geq \frac{3}{2\pi((1-0.45)\times 2 - 1)} \geq \frac{30}{2\pi} \geq 5 \tag{44}$$

we can rule out stationary points for L with $\delta \geq 0.45\Delta(R)$. Note also that since

$$x - \frac{2}{x} \geq \frac{4}{5}x$$

whenever $x^2 \geq 10$ we have that

$$\Delta(R) \geq \frac{4}{5}D\sqrt{\frac{R^2}{D^2}+1}$$

holds if $R^2 \geq 9D^2$. Therefore,

$$\delta \geq \frac{9}{25}D\sqrt{\frac{R^2}{D^2}+1}$$

for a stationary point if $R^2 \geq 9D^2$ and if (44) holds. To ensure (44) we can additionally impose that $D \geq \frac{5\sqrt{5}}{4\sqrt{2}} > 2$. We must now go back to the original coordinates: this δ breaks down into δ_x and δ_y in the original coordinates, with

$$\delta_x = \frac{R}{\sqrt{R^2+D^2}}\delta \geq \frac{9}{25}\frac{R}{\sqrt{R^2+D^2}}D\sqrt{\frac{R^2}{D^2}+1} \geq \frac{9}{25}R > \frac{R}{3}$$

for a stationary point of L. In contrast, Proposition 2 anticipates a bad local optima for the log-likelihood with $\delta_x \leq R/3$ if R is sufficiently large.

A.7 Proof of Theorems 3 and 4

The proof of Theorem 3 relies on an analysis of the functions $L(\cdot, \alpha^*)$ and $\ell(\cdot, \alpha^*)$ and their derivatives. Let's denote those functions as L_{α^*} ℓ_{α^*} to simplify notation Fig. A.1 depicts the main properties of the functions that will be used in the proofs. The first row shows $L_{\alpha^*}(\theta) \geq \ell_{\alpha^*}(\theta)$, which is the conclusion of Proposition 1. The second through fourth rows illustrate the behavior of the derivatives L' and ℓ'. We show in Proposition 3 that $L'_{\alpha^*}(\theta) \geq \ell'_{\alpha^*}(\theta)$ for all $\theta < 0$, which is clearly visible in the second and fourth row. In the third row, we plot the absolute values of the derivatives, with stationary points visible as cusps. In the last row, we plot an important auxiliary function described in more detail below.

As mentioned in the main text, we assume $\alpha^* > 0.5$ by a simple symmetry argument. The fourth column in Fig. A.1 illustrates this symmetry. Additionally, we exclude the $\alpha^* = 0.5$ from our analyses, as in this case, the entropic OT loss coincides with the negative log-likelihood (last column of Fig. A.1) and SEM and EM define the same algorithm.

We make several additional definitions for the proof of Theorem 4. Let $\omega = \omega(\theta)$ be the semi-dual weight defined in (8). The first-order optimality conditions for ω reads

$$\alpha^* = \int \frac{e^{\omega_1}\alpha^* e^{-(\theta-y)^2/2}}{e^{\omega_1}\alpha^* e^{-(\theta-y)^2/2} + e^{\omega_2}(1-\alpha^*)e^{-(\theta+y)^2/2}} q_{\theta^*}(y)\mathrm{d}y. \tag{45}$$

The above condition can be expressed in terms of the *tilted* $\alpha(\theta)$ introduced in (9): $\alpha(\theta)$ the unique number in $[0,1]$ satisfying

$$\alpha^* = G(\theta, \alpha(\theta)), \tag{46}$$

with $G(\theta, \alpha)$ defined as

$$G(\theta,\alpha) := \int \frac{\alpha e^{-(\theta-y)^2/2}}{\alpha e^{-(\theta-y)^2/2} + (1-\alpha)e^{-(\theta+y)^2/2}} q_{\theta^*}(y)\mathrm{d}y \tag{47}$$

$$= \int \frac{\alpha e^{\theta y}}{\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}} q_{\theta^*}(y)\mathrm{d}y \tag{48}$$

$$= E\left(\Psi_1(Y, \theta, \alpha(\theta))\right). \tag{49}$$

We plot the tilting $\alpha(\theta^*)$ in the last row of Fig. A.1.

To analyze the behavior of Sinkhorn-EM and vanilla EM, we also introduce the auxiliary function $F(\theta, \alpha)$ defined by

$$F(\theta,\alpha) := \int_{\mathbb{R}} y \frac{\alpha e^{\theta y} - (1-\alpha)e^{-\theta y}}{\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}} q_{\theta^*}(y)\mathrm{d}y\,. \tag{50}$$

With this notation, the updates of SEM satisfy

$$\theta_{SEM}^{t+1} = F(\theta_{SEM}^{t}, \alpha(\theta_{SEM}^{t}))\,,$$

where $\alpha(\theta)$ is defined in (46). On the other hand, the updates of EM satisfy

$$\theta_{EM}^{t+1} = F(\theta_{EM}^{t}, \alpha^*).$$

Proof of Theorem 3. We assume, as above, that $\alpha^* > 0.5$. First, we show that $L(\cdot, \alpha^*)$ never has spurious stationary points on $(0, \infty)$. Suppose there was such $\theta' > 0$. Then, the SEM algorithm initialized at that value would stay remain there, by virtue of Theorem 1. Since Theorem 4 guarantees that SEM converges to θ^* for any positive initialization, this implies $\theta' = \theta^*$.

We now show that if L_{α^*} has a spurious stationary point, then so does ℓ_{α^*}. Suppose that L_{α^*} has a stationary point $\theta \in (-\infty, 0]$. In Proposition 3 we show that if $\theta \leq 0$, then $L'_{\alpha^*}(\theta) > \ell'_{\alpha^*}(\theta)$. Therefore, if θ is stationary point of L_{α^*}, then $\ell'_{\alpha^*}(\theta) < 0$. Since ℓ_{α^*} is continuously differentiable and $\ell'_{\alpha^*}(0) = (2\alpha^{*2} - 1)^2 > 0$, there must be a $\theta' \in (\theta, 0)$ such that $\ell'_{\alpha^*}(\theta') = 0$. Therefore ℓ_{α^*} also has a spurious stationary point.

Finally, to show that the set of α^* for which ℓ_{α^*} has a spurious stationary point is strictly larger than the corresponding set for L_{α^*}, we note that the arguments in the proof of Theorem 1 and Lemma 4 in (Xu et al., 2018) establish that there is $\delta > 0$ such that if $\alpha^* = 0.5 + \delta$ then ℓ_{α^*} has a single spurious stationary point on $(-\infty, 0)$, and if $\alpha^* > 0.5 + \delta$, then ℓ_{α^*} does not have any spurious stationary points. Sine $\ell'_{\alpha^*}(\theta)$ is a continuous function of α^*, this implies that $\ell'_{0.5+\delta}$ is nonnegative for all $\theta < 0$. Since $L'_{0.5+\delta}(\theta) > \ell'_{0.5+\delta}(\theta)$ for all $\theta < 0$, we obtain that $L'_{0.5+\delta}$ has no spurious stationary points. □

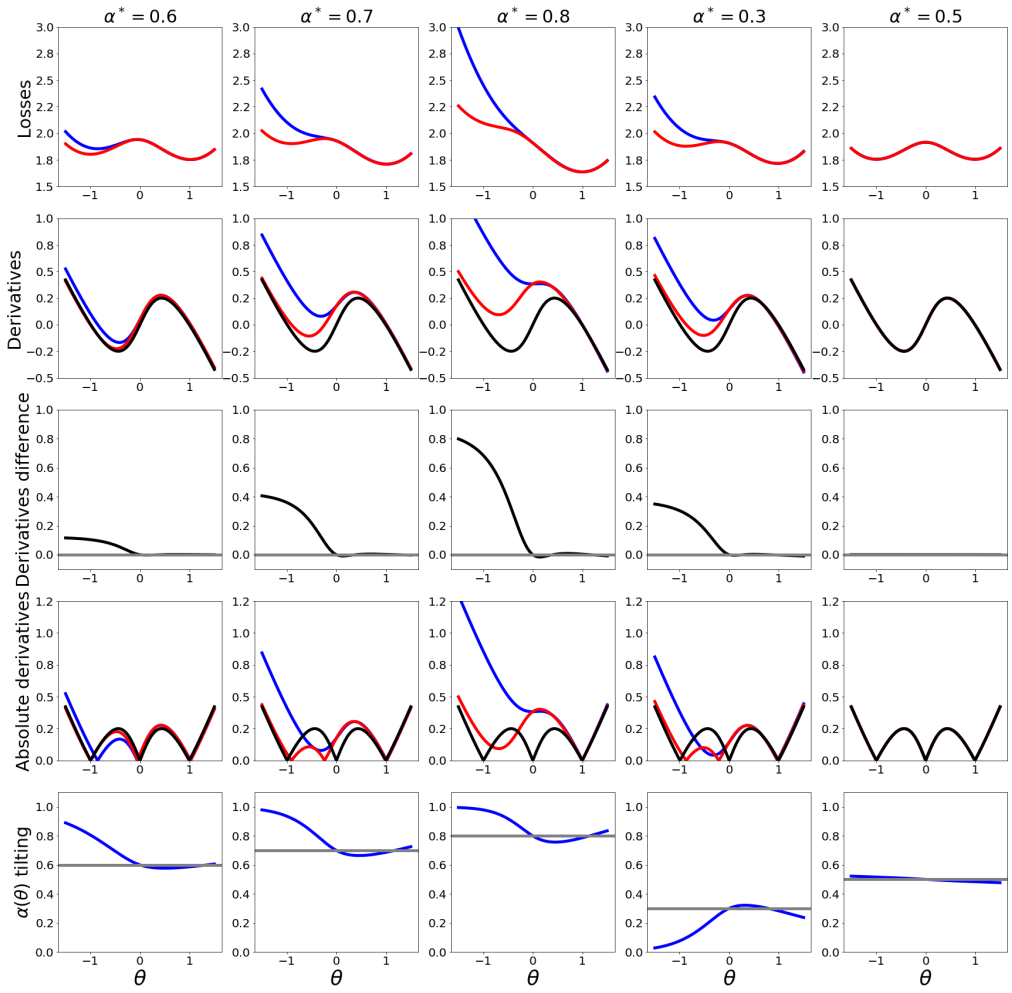

Figure A.1: Behavior of L, ℓ and their derivatives for different values of α^*. Black lines correspond to the reference $\alpha = 0.5$ (also in the last column). **First row** entropic OT (L, blue) and negative log likelihood ℓ (red). **Second row** derivatives of L and ℓ. **Third row** difference between the derivatives L and ℓ. **Fourth row** absolute value of the derivatives. **Fifth row** optimal $\alpha(\theta)$ from the semi-dual entropic OT formulation.

Proof of Theorem 4, equation (25). Let us fix $\alpha^* > 0.5$. We first recall the results of (Daskalakis et al., 2017a, Theorem 1), where the bound (25) is stated for the EM algorithm in the symmetric mixture ($\alpha^* = 0.5$). Let us denote $\theta^t_{EM_0}$ for the iterates of EM on the symmetric mixture, initialized at $\theta^0 > 0$. We write θ^t_{SEM} for the iterates of SEM on the *asymmetric* mixture. We will show

that, for all $t \geq 0$, θ^t_{SEM} and $\theta^t_{EM_0}$ satisfy

$$\theta^* \leq \theta^t_{SEM} \leq \theta^t_{EM_0} \quad \text{if } \theta^0 \geq \theta^*, \tag{51}$$
$$\theta^* \geq \theta^t_{SEM} \geq \theta^t_{EM_0} \quad \text{if } 0 < \theta^0 \leq \theta^*. \tag{52}$$

This will then prove the claim since it implies

$$|\theta^t_{SEM} - \theta^*| \leq |\theta^t_{EM_0} - \theta^*| \leq \rho^t |\theta_0 - \theta^*|.$$

It remains to prove (51) and (52). Recall the function F defined in (50). We first show that

$$F(\theta, \alpha(\theta)) \begin{cases} \leq \theta^*, & 0 < \theta < \theta^* \\ = \theta^*, & \theta = \theta^* \\ \geq \theta^*, & \theta > \theta^* \end{cases}. \tag{53}$$

This implies the first inequalities of (51) and (52). To show (53), notice first that clearly $F(\theta^*, \alpha(\theta^*)) = F(\theta^*, \alpha^*) = \theta^*$. Therefore, it is enough to establish that $\theta \mapsto F(\theta, \alpha(\theta))$ is non-decreasing. Let us define $f(\theta) = F(\theta, \alpha(\theta))$. We then have

$$f'(\theta) = \frac{\partial F}{\partial \theta}(\theta, \alpha(\theta)) + \frac{\partial F}{\partial \alpha}(\theta, \alpha(\theta))\alpha'(\theta), \tag{54}$$

and

$$\frac{\partial F}{\partial \theta}(\theta, \alpha) = 4\alpha(1-\alpha) \int y^2 \frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2} \mathrm{d}y \geq 0, \tag{55}$$
$$\frac{\partial F}{\partial \alpha}(\theta, \alpha) = 2 \int y \frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2} \mathrm{d}y. \tag{56}$$

Additionally, by taking derivatives with respect to θ in (46) we have

$$\alpha'(\theta) = -\frac{\partial G}{\partial \alpha}(\theta, \alpha(\theta))^{-1} \frac{\partial G}{\partial \theta}(\theta, \alpha(\theta)), \tag{57}$$

and likewise,

$$\frac{\partial G}{\partial \theta}(\theta, \alpha) = 2\alpha(1-\alpha) \int y \frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2} \mathrm{d}y, \tag{58}$$
$$\frac{\partial G}{\partial \alpha}(\theta, \alpha) = \int \frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2} \mathrm{d}y > 0. \tag{59}$$

The conclusion follows by replacing (55),(56),(57),(58) and (59) in (54) and invoking the Cauchy-Schwarz inequality.

We now show the second inequalities in (51) and (52). To this end, we will first show

$$F(\theta, \alpha(\theta)) \begin{cases} \geq F(\theta, 0.5) & 0 \leq \theta \leq \theta^*, \\ \leq F(\theta, 0.5) & \theta \geq \theta^*. \end{cases} \tag{60}$$

Let ϕ denote the density of a standard Gaussian random variable. We can write

$$\frac{F(\theta, \alpha) - F(\theta, 0.5)}{2\alpha - 1} = \int y \cdot \frac{\left(\alpha^* e^{\theta^* y} + (1-\alpha^*)e^{-\theta^* y}\right)}{\left(e^{\theta y} + e^{-\theta y}\right)\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)} \phi(y) e^{-\theta^{*2}/2} \mathrm{d}y.$$
$$=: \int y \cdot \rho_{\theta,\alpha}(y) \mathrm{d}y\,.$$

It is straightforward to verify that for $\alpha, \alpha^* \geq 1/2$, if $\leq \alpha \leq \alpha^*$ and $\theta \leq \theta^*$, then

$$\rho_{\theta,\alpha}(y) \geq \rho_{\theta,\alpha}(-y) \quad \forall y \geq 0.$$

On the other hand, if $\alpha \geq \alpha^*$ and $\theta \geq \theta^*$, then

$$\rho_{\theta,\alpha}(y) \leq \rho_{\theta,\alpha}(-y) \quad \forall y \geq 0.$$

In particular, this yields that for $\alpha, \alpha^* \geq 1/2$,

$$\frac{F(\theta,\alpha) - F(\theta, 0.5)}{2\alpha - 1} \begin{cases} \geq 0 & \text{if } \alpha \leq \alpha^* \text{ and } 0 \leq \theta \leq \theta^* \\ \leq 0 & \text{if } \alpha \geq \alpha^* \text{ and } \theta \geq \theta^*. \end{cases}$$

To complete the proof of (60), we used the facts, proved in Lemma 1 that $\alpha(\theta) \geq 1/2$ and that $\alpha(\theta) \leq \alpha^*$ if $0 \leq \theta \leq \theta^*$ and $\alpha(\theta) \geq \theta^*$ if $\theta \geq \theta^*$.

Now, we use the fact that the iterates $\theta^t_{EM_0}$ satisfy (see for example Daskalakis et al. (2017a))

$$\theta^{t+1}_{EM_0} = F(\theta^t_{EM_0}, 0.5)\,. \tag{61}$$

With this, we can now show the second two inequalities in (51) and (52). We proceed by induction. Let's first suppose $\theta^0 \geq \theta^*$. Then indeed for $t = 0$, we have $\theta^* \leq \theta^t_{SEM} \leq \theta^t_{EM_0}$. If this relation holds for some t, then we have

$$\begin{aligned} \theta^{t+1}_{SEM} &= F(\theta^t_{SEM}, \alpha(\theta^t_{SEM})) \\ &\leq F(\theta^t_{SEM}, 0.5) \\ &\leq F(\theta^t_{EM_0}, 0.5) \\ &= \theta^{t+1}_{EM_0}\,, \end{aligned}$$

where the first inequality uses (60), the second uses the fact that F is an increasing function in its first coordinate (55), and the final equality is (61). The proof of the second inequality in (52) is completely analogous. □

Proof of Theorem (4), equation (26). Suppose first $\theta^0 > \theta^*$. In this case, it suffices to show that $F(\theta,\alpha) \leq F(\theta,\alpha^*)$ for all $\alpha \geq \alpha^*$ for all $\theta \geq \theta^*$. Indeed, we can then appeal to precisely the same argument as in the proof of Theorem (4), equation (25), to compare the iterates of SEM (which satisfy $\theta^{t+1}_{SEM} = F(\theta^t_{SEM}, \alpha(\theta^t_{SEM}))$) to those of EM (which satisfy $\theta^{t+1}_{SEM} = F(\theta^t_{SEM}, \alpha^*)$.)

We have that

$$\begin{aligned} \frac{F(\theta,\alpha) - F(\theta,\alpha^*)}{2(\alpha - \alpha^*)} &= \int y \frac{q_{\theta^*}(y)}{(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y})\,(\alpha^* e^{\theta y} + (1-\alpha^*)e^{-\theta y})} \mathrm{d}y \\ &= \int_{y \geq 0} f_\theta(y) \mathrm{d}y, \end{aligned} \tag{62}$$

where

$$f_\theta(y) := y \frac{g_\theta(y)\phi(y)e^{-\theta^{*2}/2}}{(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y})\,(\alpha^* e^{\theta y} + (1-\alpha^*)e^{-\theta y})\,(\alpha e^{-\theta y} + (1-\alpha)e^{\theta y})\,(\alpha^* e^{-\theta y} + (1-\alpha^*)e^{\theta y})},$$

and

$$\begin{aligned} g_\theta(y) := &\left(\alpha e^{-\theta y} + (1-\alpha)e^{\theta y}\right)\left(\alpha^* e^{-\theta y} + (1-\alpha^*)e^{\theta y}\right) q_{\theta^*}(y) \\ &- \left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)\left(\alpha^* e^{\theta y} + (1-\alpha^*)e^{-\theta y}\right) q_{\theta^*}(-y) \\ =&L\left(e^{y(2\theta-\theta^*)} - e^{-y(2\theta-\theta^*)}\right) + M\left(e^{y\theta^*} - e^{-y\theta^*}\right) + N\left(e^{y(2\theta+\theta^*)} - e^{-y(2\theta+\theta^*)}\right) \end{aligned}$$

with

$$\begin{aligned} L &= (1-\alpha^*)^2(1-\alpha) - \alpha{\alpha^*}^2, \\ M &= (2\alpha^*-1)(\alpha+\alpha^*-2\alpha\alpha^*), \\ N &= \alpha^*(1-\alpha^*)(1-2\alpha). \end{aligned}$$

Notice that for $y \geq 0$ and $\theta > \theta^*$ the three above differences of exponentials are positive, and that $e^{y(2\theta-\theta^*)} - e^{-y(2\theta-\theta^*)} \geq e^{y\theta^*} - e^{-y\theta^*}$. Moreover, if $1/2 \leq \alpha^*, \alpha < 1$, then $N < 0$ and $M > 0$, and if furthermore $\alpha \geq \alpha^*$, then also $L < 0$. Therefore,

$$\begin{aligned} g_\theta(y) &< (L+M)\left(e^{y\theta^*} - e^{-y\theta^*}\right) \\ &= (1-2\alpha)(3{\alpha^*}^2 - 3\alpha^* + 1)\left(e^{y\theta^*} - e^{-y\theta^*}\right) \leq 0. \end{aligned} \tag{63}$$

This proves that when $\alpha \geq \alpha^* \geq 1/2$, we have $F(\theta,\alpha) \geq F(\theta,\alpha^*)$, as claimed.

Now let's show that there exists a $\theta_{\text{fast}} < \theta^*$ such that if $\theta \in [\theta_{\text{fast}}, \theta^*]$, then $F(\theta,\alpha) \geq F(\theta,\alpha^*)$ for all $\alpha^* \geq \alpha > 1/2$. (As above, this will suffice to prove the desired claim by applying the argument in the proof of Theorem (4), equation (25).) It suffice to show that for $\theta \in [\theta_{\text{fast}}, \theta^*]$, we have

$$g_\theta(y) \leq 0 \quad \forall y \geq 0\,.$$

First, we note that, since $N < 0$, for any $\theta > \theta^*/2$, the term $N\left(e^{y(2\theta+\theta^*)} - e^{-y(2\theta+\theta^*)}\right)$ is always eventually dominant, so there exists a y^* such that

$$g_\theta(y) < 0 \quad \forall \theta > \theta^*/2, y > y^*\,.$$

It therefore suffices to focus on the compact interval $[0, y^*]$.

To proceed, let us consider what happens when $\theta = \theta^*$. Carrying out the same argument as above, we obtain that as long as $\alpha > 1/2$, we have

$$g_{\theta^*}(y) \leq (L+M)(e^{y\theta^*} - e^{-y\theta^*}) \quad \forall y \geq 0,$$

where $L+M$ is negative.

Let us examine the derivative $\frac{\partial}{\partial\theta} g_\theta(y)$:

$$\frac{\partial}{\partial\theta} g_\theta(y) = 2yL(e^{y(2\theta-\theta^*)} + e^{-y(2\theta-\theta^*)}) + 2yN(e^{y(2\theta+\theta^*)} + e^{-y(2\theta+\theta^*)})\,.$$

We conclude that if $\theta' < \theta^*$ is such that

$$(\theta^* - \theta')(4|L|ye^{y\theta^*} + 4|N|ye^{3y\theta^*}) \leq -(L+M)(e^{y\theta^*} - e^{-y\theta^*})\,,$$

then

$$\left|\frac{\partial}{\partial\theta}g_\theta(y)\right|\cdot(\theta^*-\theta')\le -g_{\theta^*}(y)\,,$$

which yields

$$g_{\theta'}(y)=g_{\theta^*}(y)-\int_{\theta'}^{\theta^*}\frac{\partial}{\partial\theta}g_\theta(y)\mathrm{d}\theta\le g_{\theta^*}(y)+\left|\frac{\partial}{\partial\theta}g_\theta(y)\right|\cdot(\theta^*-\theta')\le 0\,.$$

Hence, if we define

$$\delta=\inf_{y\in[0,y^*]}\frac{-(L+M)(e^{y\theta^*}-e^{-y\theta^*})}{4|L|ye^{y\theta^*}+4|N|ye^{3y\theta^*}}\,,$$

then as long as this quantity is positive, we can take $\theta_{\text{fast}}=\theta^*-\delta$. But positivity follows immediately from the fact that this function is continuous and positive on $(0,y^*]$ and has a positive limit as $y\to\infty$. □

A.8 Intermediate results for Theorem 3 and 4

Proposition 3. *For the asymmetric mixture of two Gaussians* (24) *we have that for all* $\theta<0$

$$L'_{\alpha^*}(\theta)>\ell'_{\alpha^*}(\theta) \tag{64}$$

Proof. Let us write $\ell(\theta,\alpha)=-E\log q_{\theta,\alpha}(Y)$ the the expected negative log-likelihood function in the overparametrized model

$$q_{\theta,\alpha}=\alpha\mathcal{N}(y;\theta,1)+(1-\alpha)\mathcal{N}(y;-\theta,1)\,.$$

We then have

$$\frac{\partial}{\partial\theta}\ell(\theta,\alpha)=\int y\left[\frac{\alpha e^{-(\theta-y)^2/2}-(1-\alpha)e^{-(\theta+y)^2/2}}{\alpha e^{-(\theta-y)^2/2}+(1-\alpha)e^{-(\theta+y)^2/2}}\right]q_{\theta^*,\alpha^*,}(y)\mathrm{d}y-\theta=F(\theta,\alpha)-\theta\,, \tag{65}$$

where F is defined in (50).

We have $\ell'_{\alpha^*}(\theta)=\frac{\partial}{\partial\theta}\ell(\theta,\alpha^*)=F(\theta,\alpha^*)-\theta$. Likewise, if we recall that $L_{\alpha^*}(\theta)=\tilde{L}(\theta,\omega(\theta))$ where $\omega(\theta)$ solves (8), then we have

$$L'_{\alpha^*}(\theta)=\frac{\partial\tilde{L}}{\partial\theta}(\theta,\omega(\theta))=\frac{\partial}{\partial\theta}\ell(\theta,\alpha(\theta))=F(\theta,\alpha(\theta))-\theta. \tag{66}$$

Then, to establish (64) it suffices to show that for each $\theta<0$,

$$F(\theta,\alpha^*)<F(\theta,\alpha(\theta))\,.$$

Moreover, since Lemma 1 shows that $\alpha(\theta)>\alpha^*>0.5$, it suffices to show that $F(\theta,\alpha)$ is a strictly increasing function of α for $\alpha\ge 0.5$ and an arbitrary $\theta\le 0$.

Recall (56), which shows

$$\frac{\partial F}{\partial\alpha}(\theta,\alpha)=2\int y\frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y}+(1-\alpha)e^{-\theta y}\right)^2}\mathrm{d}y\,.$$

Since $\alpha^* > 0.5$, we have $q_{\theta^*}(y) > q_{\theta^*}(-y)$ for all $y > 0$. Furthermore, for $\theta \leq 0$ and $\alpha \geq 0.5$, it holds

$$\frac{1}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2} \geq \frac{1}{\left(\alpha e^{-\theta y} + (1-\alpha)e^{\theta y}\right)^2} \quad \forall y > 0\,.$$

Therefore

$$\begin{aligned}\frac{\partial F}{\partial \alpha}(\theta,\alpha) &= 2\int y \frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2}\mathrm{d}y \\ &= 2\int_{y>0} y \frac{q_{\theta^*}(y)}{\left(\alpha e^{\theta y} + (1-\alpha)e^{-\theta y}\right)^2}\mathrm{d}y + 2\int_{y>0}(-y)\frac{q_{\theta^*}(-y)}{\left(\alpha e^{-\theta y} + (1-\alpha)e^{\theta y}\right)^2}\mathrm{d}y > 0\,,\end{aligned}$$

which proves the claim. □

LEMMA 1. *Suppose* $\alpha^* > 1/2$

(a) for any θ, $\alpha(\theta) > 1/2$.
(b) $\alpha(\cdot)$ *is decreasing (increasing) whenever* $\theta < 0$ $(\theta > \theta^*)$ *and in either case* $\alpha(\theta) \geq \alpha^*$. *Moreover,* $\lim_{\|\theta\|\to\infty} \alpha(\theta) = 1$.
(c) $\alpha(\theta) \leq \alpha^*$ *whenever* $0 \leq \theta \leq \theta^*$.

PROOF. We begin by recalling the function G, defined in (47). By (59), this function is a strictly increasing function of α and $\alpha(\theta)$ is the unique number in $[0,1]$ satisfying

$$G(\theta, \alpha(\theta)) = \alpha^*,$$

and $\alpha(\theta) > p$ if and only if $G(\theta, p) < \alpha^*$. Let us first prove (a). It suffices to show that $G(\theta, 1/2) < \alpha^*$. Write ϕ_{θ^*} for the density of $\mathcal{N}(\theta^*, 1)$. We then have

$$G(\theta, 1/2) = \int \frac{e^{\theta y}}{e^{\theta y} + e^{-\theta y}} q_{\theta^*}(y)\mathrm{d}y = \int \frac{\alpha^* e^{\theta y} + (1-\alpha^*)e^{-\theta y}}{e^{\theta y} + e^{-\theta y}} \phi_{\theta^*}(y)\mathrm{d}y\,.$$

But if $\alpha^* > 1/2$, then $\frac{\alpha^* e^{\theta y} + (1-\alpha^*)e^{-\theta y}}{e^{\theta y} + e^{-\theta y}} < \alpha^*$ for all θ and y. Since $\phi_{\theta^*}(y)$ is a probability density, we obtain that $G(\theta, 1/2) < \alpha^*$, as desired.

It is straightforward to see that $\alpha(0) = \alpha(\theta^*) = \alpha^*$. To show monotonicity, we rely on the formula (57) for $\alpha'(\theta)$. If $\theta < 0$, the conclusion is a direct consequence of (57) and Lemma 2(b). If $\theta > \theta^*$, the conclusion follows similarly from Lemma 2(c), but the argument is more delicate, as applying this lemma requires that $\alpha(\theta) \geq \alpha^*$. Suppose that there exists a $\theta > \theta^*$ for which $\alpha'(\theta) < 0$. Let us denote by θ_0 the infimum over all such θ. By (57), $\frac{\partial G}{\partial \theta}(\theta_0, \alpha(\theta_0))$ must be therefore nonnegative, which by Lemma 2(c) implies that $\alpha(\theta_0) < \alpha^* = \alpha(\theta^*)$. But since $\alpha'(\theta) \geq 0$ for all $\theta \in [\theta^*, \theta_0)$, this is a contradiction. Therefore $\alpha'(\theta) \geq 0$ for all $\theta \geq \theta^*$, as claimed.

Finally, the limit statement follows from the dominated convergence theorem. Since $\alpha^* = G(\theta, \alpha(\theta))$ for all $\theta \in \mathbb{R}$, it holds

$$\begin{aligned}\alpha^* &= \lim_{\|\theta\|\to\infty} G(\theta, \alpha(\theta)) \\ &= \int \lim_{\|\theta\|\to\infty} \frac{\alpha(\theta)e^{\theta y}}{\alpha(\theta)e^{\theta y} + (1-\alpha(\theta))e^{-\theta y}} q_{\theta^*}(y)\mathrm{d}y\,,\end{aligned}$$

where the second inequality is by the dominated convergence theorem. Since $\alpha(\cdot)$ is monotonic outside the interval $[0, \theta^*]$, as $\alpha(\theta)$ has a limit as $\theta \to +\infty$ or $\theta \to -\infty$. Let us first consider $\theta \to \infty$ (the negative case is exactly analogous). If this limit is different from 1, then

$$\lim_{\theta\to\infty} \frac{\alpha(\theta)e^{\theta y}}{\alpha(\theta)e^{\theta y} + (1-\alpha(\theta))e^{-\theta y}} = \begin{cases} 1 & y > 0 \\ 0 & y < 0 \end{cases}.$$

But this is a contradiction, since $\alpha^* \neq \int_{y\geq 0} q_{\theta^*}(y)\mathrm{d}y$ if $\alpha^* > 1/2$. This proves the claim.

Let's now prove (c). Notice it suffices to show that (i) $\alpha'(0) < 0$ and (ii) the only solutions to the equation $\alpha(\theta) = \alpha^*$ are $\theta = 0$ and $\theta = \theta^*$. The first claim is a simple consequence of (57) and Lemma 2(b).

The second claim is a bit more involved. Suppose $\alpha(\theta) = \alpha^*$. By simple algebra (as in the proof of theorem 4), it can be shown that the following relation holds

$$\int_{y\geq 0} \frac{2\alpha^*(1-\alpha^*)(2\alpha^*-1)e^{-\theta^{*2}}\left(e^{2\theta y}-1\right)\left(e^{2\theta^* y}-e^{2\theta y}\right)}{e^{(\theta^*+2\theta)y}\left(\alpha^* e^{\theta y}+(1-\alpha^*)e^{-\theta y}\right)\left((1-\alpha^*)e^{\theta y}+\alpha^* e^{-\theta y}\right)}\phi(y)\mathrm{d}y = 0.$$

The integral above can only be zero if $\theta = 0$ or $\theta = \theta^*$; otherwise, the integrand is either positive or negative for each value of $y \geq 0$. This concludes the proof. □

LEMMA 2. *Suppose $\alpha^* > 0.5$. Let*

$$G_\theta(\theta,\alpha) := \frac{1}{2\alpha(1-\alpha)}\frac{\partial G}{\partial \theta}(\theta,\alpha) = \int \frac{y}{\left(\alpha e^{\theta y}+(1-\alpha)e^{-\theta y}\right)^2} q_{\theta^*}(y)\mathrm{d}y.$$

Then,

(a) For each $\theta \geq 0$, G_θ is a decreasing as function of α. Conversely, for each $\theta \leq 0$, G_θ is an increasing function of α.
(b) $G_\theta(\theta,\alpha) \geq 0$ if $\theta \leq 0$ and $\alpha > 1/2$.
(c) $G_\theta(\theta,\alpha) \leq 0$ if $\theta \geq \theta^*$ and $\alpha \geq \alpha^*$.

PROOF. To see (a), notice that

$$\frac{\partial G_\theta}{\partial \alpha}(\theta,\alpha) = -2\int \frac{y\left(e^{\theta y}-e^{-\theta y}\right)}{\left(\alpha e^{y\theta}+(1-\alpha)e^{-y\theta}\right)^3} q_{\alpha^*,\theta^*}(y)\mathrm{d}y.$$

The integrand is either positive (if $\theta > 0$) or negative (if $\theta < 0$) for each y, and the conclusion follows.

To prove (b) and (c), we note that (56) implies that

$$G_\theta(\theta,\alpha) = \frac{1}{2}\frac{\partial F}{\partial \alpha}(\theta,\alpha).$$

But we have already shown in the proof of Proposition 3 that $\frac{\partial F}{\partial \alpha}(\theta,\alpha) > 0$ for all $\theta \leq 0$ and $\alpha > 1/2$. This proves (b).

Likewise, the proof of Theorem (4), equation (26), shows that $F(\theta,\alpha) \leq F(\theta,\alpha^*)$ for all $\theta \geq \theta^*$ and $\alpha \geq \alpha^*$. This proves that $G_\theta(\theta,\alpha^*) = \frac{1}{2}\frac{\partial F}{\partial \alpha}(\theta,\alpha^*) \leq 0$. To conclude, we appeal to part (a): since $\theta \geq \theta^* > 0$, G_θ is decreasing as a function of α, and hence, $G_\theta(\theta,\alpha) \leq G_\theta(\theta,\alpha^*) \leq 0$. □

APPENDIX B: EXPERIMENTAL DETAILS

B.1 Synthetic experiment details

For the four synthetic data experiments of 6.1 (i),(ii),(iii), and (iv), we considered several possible dimension sizes $d = 2, 5, 10, 20$ variances $\sigma^2 = 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5$ and component numbers $K = 10, 20, 30, 40$. (Boerner et al., 2023; Brown et al., 2021). Each true mean θ_k was sampled from a uniform $U(-1, 1)$ distribution. In cases where variances were not spherical, diagonal entries were sampled from a $U(0.5\sigma^2, 1.5\sigma^2)$ distribution (and i.i.d across different clusters). For each run of EM and Sinkhorn-EM, we performed at most 100 iterations with a termination criterion of $\varepsilon = 1e^{-3}$ change in the overall L_1 differences between consecutive parameter values (adding mean and variance differences). For each Sinkhorn-EM iteration, we performed at most 1000 iterations with a termination criterion of $\varepsilon = 1e^{-3}$ for L_1 differences between consecutive entropic optimal transport potentials. For k-means we used the scikit-learn implementation with K-means++ initialization. This initialization was also used on the same seeds for Sinkhorn-EM and EM. Figs. B.2, B.3 and B.4 supplement the results of the main text regarding experiment (i). In Fig. B.2 we compare the performance of three algorithms of each individual experiment, both at the level of individual seed or at the level of best seed. Both EM and Sinkhorn-EM outperform k-means and Sinkhorn-EM outperforms EM. In Figs. B.3 we show performance of each algorithm as a function of d, σ^2 and K for the two sample sizes we considered, $n = 500$ and $n = 1000$. These figures point to the same pattern described in the main text. Figs. B.5, B.6 and B.7 are analogs of B.2 in the main text for experiments (ii) (known unequal diagonal variances), (iii) (unknown spherical variance) and (iv) (unknown unequal variances). Results align with the same pattern favoring Sinkhorn-EM.

For the unknown weights experiments of Section 6.2 we considered parameters $d = 2$, $K = 10, 20, 30, 40$, $n = 1000$, $\sigma^2 = 0.005, 0.001, 0.01$ and $\gamma = 1, 10, 20, 50, 100, 1000$ and all other setups as in the experiments in Section 6.1. We jointly optimized over weights α and means θ by coordinate descent: for optimization over θ we performed Sinkhorn-EM and for optimization over α we used the mirror descent method described in 3.2, with $\eta = 1$. We iterated these two sub-routines until convergence. For the model selection experiments of Section 6.3 we considered several choices of d and K as shown in Fig. B.8, which gives a more detailed account than Fig. 5 in the main text.

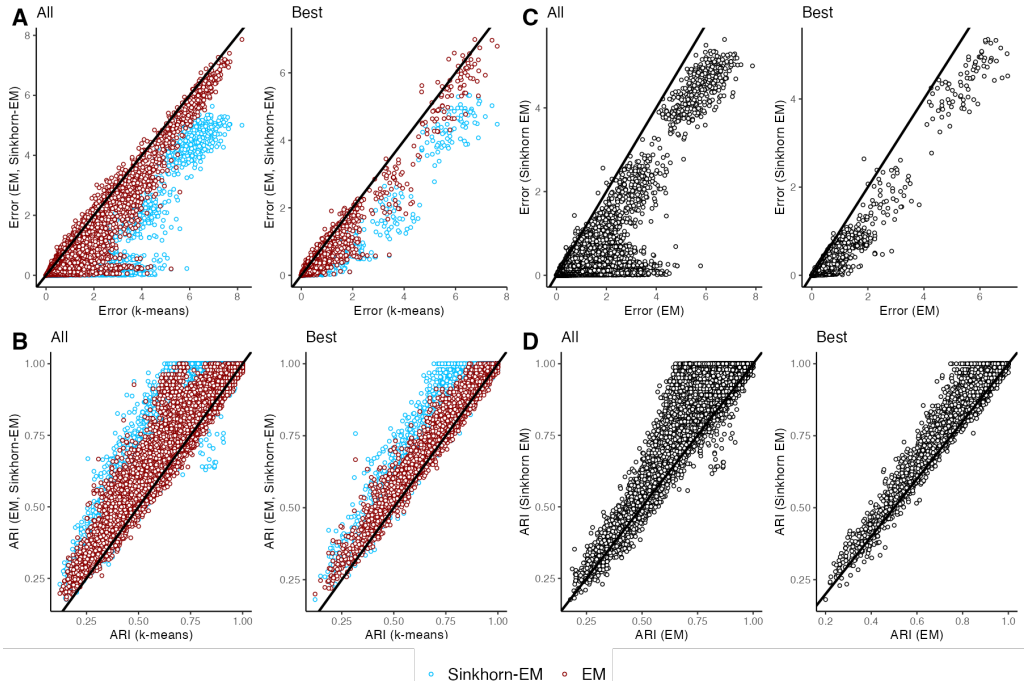

Figure B.2: Performance comparison of three methods in experiment (i) of Section 6.1. **A** Error scatterplots comparing the performance of k-means with EM or Sinkhorn-EM (colors). **B** same as **A** but with ARI score. **C,D** same as **A,B** but comparing EM with Sinkhorn-EM

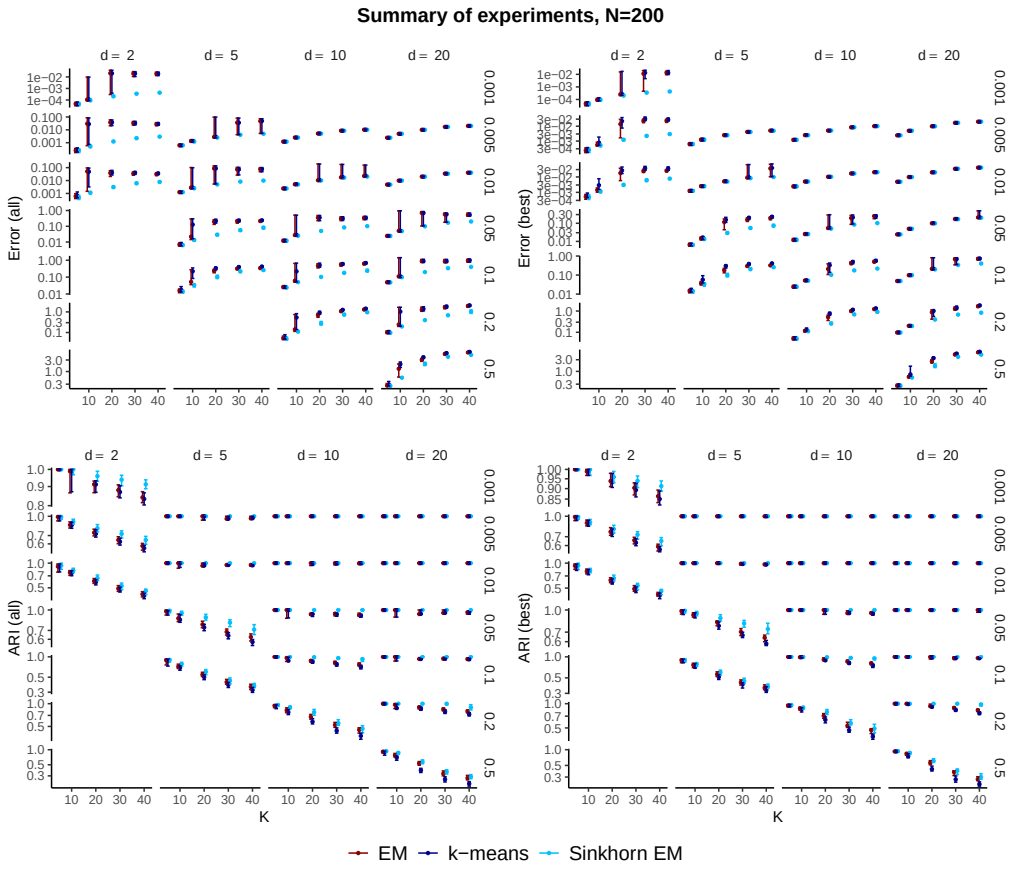

Figure B.3: Performance of three algorithms when the sample size is $N = 200$. Each column represents a different dimension d and each row a difference noise variance σ^2.

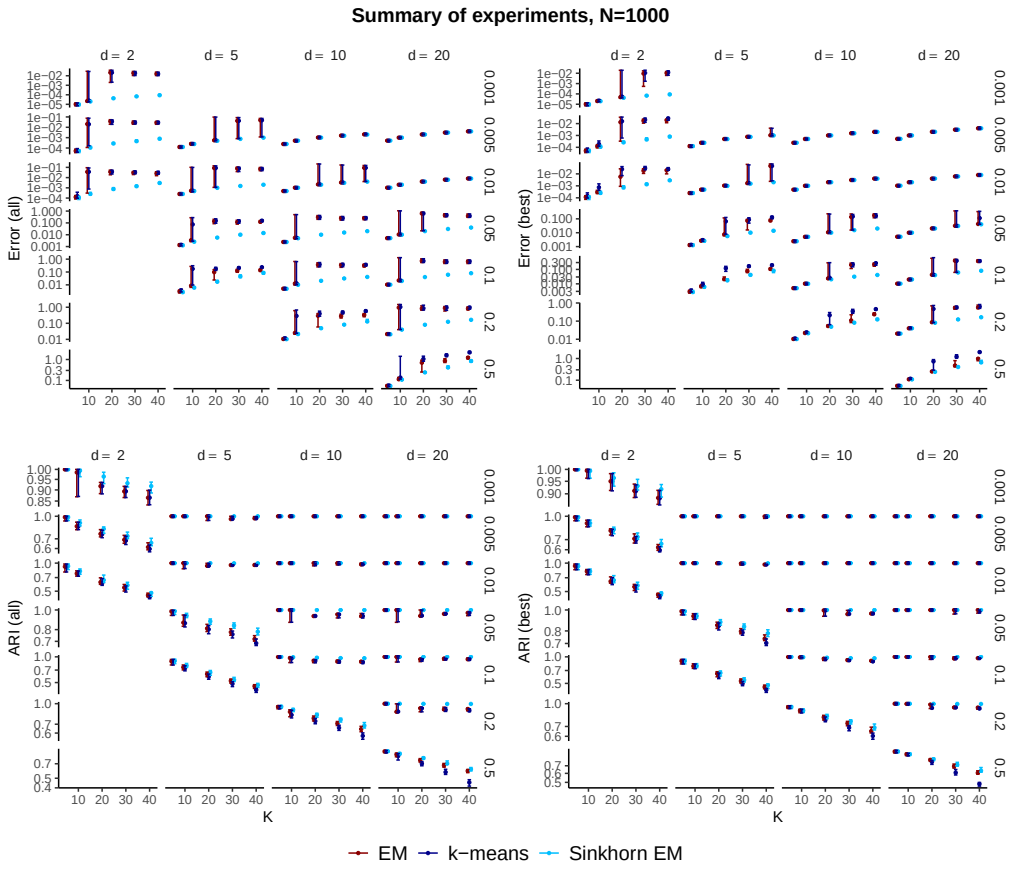

Figure B.4: Performance of three algorithms when the sample size is $N = 1000$. Each column represents a different dimension d and each row a difference noise variance σ^2.

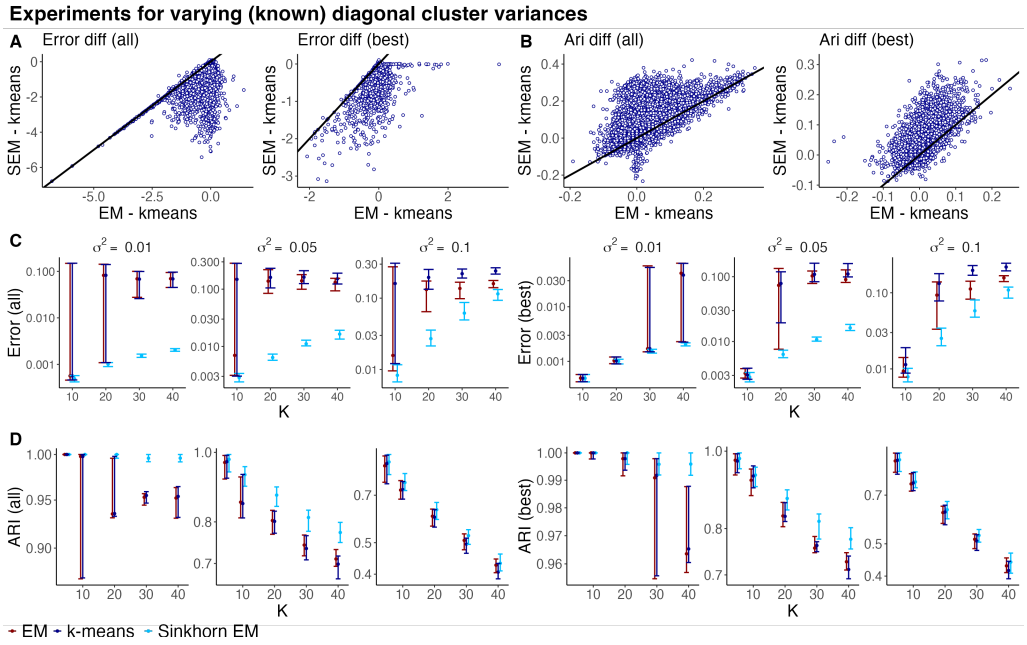

Figure B.5: Results for the experiment (ii) in Section 6.1: each cluster is characterized by a different diagonal covariance matrix, and these variances are assumed known (i.e. not estimated in the M-step)

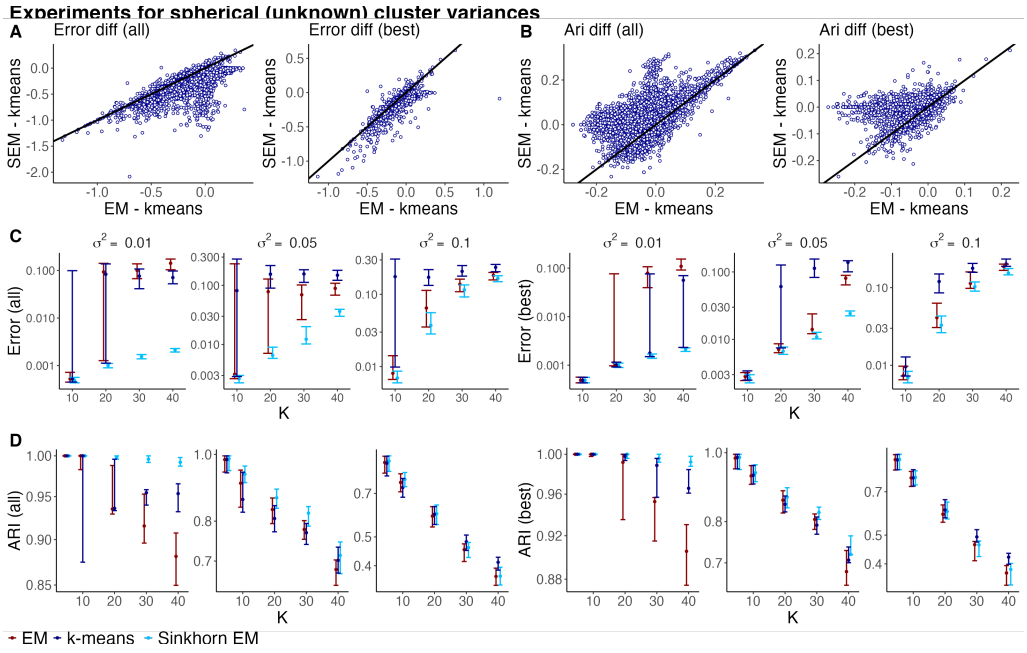

Figure B.6: Results for the experiment (iii) in Section 6.1: each cluster is characterized by the sample spherical covariance matrix with variance $\sigma^2 I_d$, but variances not assumed known and hence estimated in the M-step.

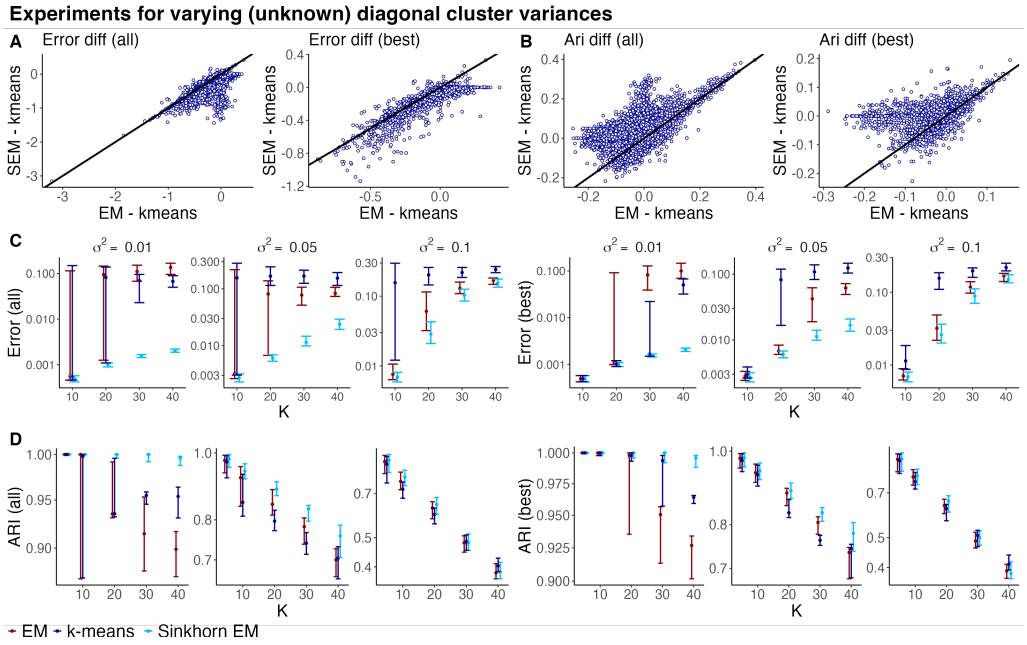

Figure B.7: Results for the experiment (iv) in Section 6.1: each cluster is characterized by a different diagonal covariance matrix and these variances are not assumed known and hence estimated in the M-step.

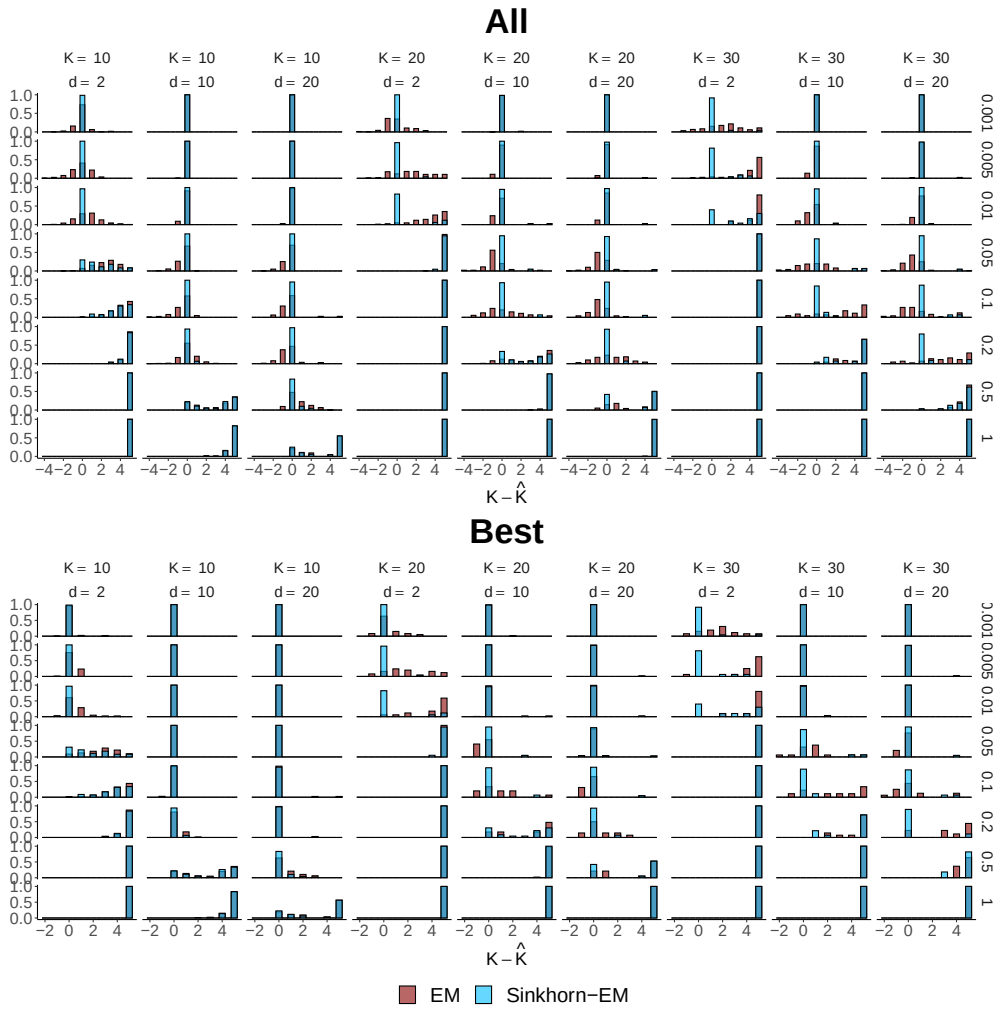

Figure B.8: Number of components estimation error histograms for the experiment in Section 6.3. Each row represents a different variance σ^2. **A**: results for different seeds. **B**: results for the best seed.

B.2 C.elegans experiment details

Results displayed in Fig. 6 in Section 7 correspond to the following semi-synthetic setup: we considered a vector of true neural locations and colors as described in Nejatbakhsh et al. (2020). We considered clustering segmentation tasks in the head (195 neurons) and tail (45 neurons). In each experiment, we sampled 35 neurons at random either from the head or tail. We sampled covariance matrices whose diagonal entries were sampled from a log-normal distribution with location parameter 1 and scale parameter 0.1. We considered GMM with 35 samples centered at sampled neural locations prescribed covariances and uniform weights, and based our inferences on $n = 5000$ samples from that mixture. In this experiment, we assumed weights were fixed and well-specified, but we allowed ourselves to fit the (diagonal) covariance matrices. Results displayed in 6 are summaries of 200×2 (head and tail) experiments, on each of them we kept the best of 10 seeds. As in the experiments of Section 6.1 we compared k-means,

EM, and Sinknorn-EM algorithms, in all cases using the k-means++ initialization method. Examples of segmentation performance in Fig. 6B were plotted with the code released with Nejatbakhsh et al. (2020).

B.3 Co-clustering experiment details

B.3.1 Algorithmic details: Suppose we want to perform co-clustering on a $N \times M$ matrix using $K \times G$ co-clusters. We consider row and column responsibility vectors $z \in \mathbb{R}^{N \times K}$ and $w \in \mathbb{R}^{M \times G}$. We briefly describe the VEM and SVEM algorithms. We implement the VEM algorithm as described in Algorithm 5.3 in Chapter 5 of Govaert and Nadif (2013), that we reproduce in Algorithm 1.

Algorithm 1 *Variational EM (VEM) for co-clustering in model (28)*

1. **Input**: Data matrix Y, number of co-clusters K and G. Initial values for z and w
2. **Initialization**: compute

$$\pi_k = \frac{\sum_k z_{i,k}}{N}, \rho_g = \frac{\sum_j w_{j,g}}{M}, \theta_{k,g} = \frac{\sum_{i,j} z_{i,k} Y_{i,j} w_{j,g}}{\sum_i z_{i,k} \sum_j w_{j,g}}, \sigma^2_{k,g} = \frac{\sum_{i,j} z_{i,k} Y^2_{i,j} w_{j,g}}{\sum_i z_{i,k} \sum_j w_{j,g}} - \theta_{k,g}$$

3. **Outer loop: repeat until convergence in θ,σ^2,π, ρ)**

   3.1 **Define**

$$Y^w_{i,g} = \frac{\sum_j w_{j,g} Y_{i,j}}{\sum_j w_{j,g}}, u^w_{i,g} = \frac{\sum_j w_{j,g} Y^2_{i,j}}{\sum_j w_{j,g}}$$

   3.2 **Inner loop in row (k) coordinates until convergence of π, μ, σ^2**

   3.2.1 **Update**

$$z_{i,k} \propto \pi_k \exp\left(-\frac{1}{2} \sum_g \left(\sum_j w_{j,g}\right) \left(\log \sigma^2_{k,g} + \frac{u^w_{i,g} - 2\mu_{k,g} Y^w_{i,g} + \mu^2_{k,g}}{\sigma^2_{k,g}}\right)\right)$$

   3.2.2 **Update**

$$\pi_k = \frac{\sum_i z_{i,k}}{N}, \mu_{k,g} = \frac{\sum_i z_{i,k} Y^w_{i,g}}{\sum_i z_{i,k}}, \sigma^2_{k,g} = \frac{\sum_i z_{i,k} u^w_{i,g}}{\sum_i z_{i,k}} - \mu^2_{k,g}$$

   3.3 **Define**

$$Y^z_{j,k} = \frac{\sum_i z_{i,k} Y_{i,j}}{\sum_i z_{i,k}}, v^z_{j,k} = \frac{\sum_j z_{i,k} Y^2_{i,j}}{\sum_i z_{i,k}}$$

   3.4 **Inner loop in column (g) coordinates until convergence of ρ, μ, σ^2**

   3.4.1 **Update**

$$w_{j,g} \propto \rho_g \exp\left(-\frac{1}{2} \sum_k \left(\sum_i z_{i,k}\right) \left(\log \sigma^2_{k,g} + \frac{v^z_{j,k} - 2\mu_{k,g} Y^z_{j,k} + \mu^2_{k,g}}{\sigma^2_{k,g}}\right)\right)$$

   3.4.2 **Update**

$$\rho_g = \frac{\sum_j w_{j,g}}{M}, \mu_{k,g} = \frac{\sum_j w_{j,g} Y^z_{j,k}}{\sum_j w_{j,g}}, \sigma^2_{k,g} = \frac{\sum_j w_{j,g} v^z_{j,k}}{\sum_j w_{j,g}} - \mu^2_{k,g}$$

4. **Output** π, ρ, σ^2, μ

In the case that some parameters (e.g. the weights π, ρ or variances σ^2) are known then the algorithm can be simplified by simply skipping the steps that implement updates of those parameters. The SVEM algorithm consists of a small variation over the above routine. In the case where π and ρ are fixed then we

only need to replace steps 3.2.1 and 3.4.1 in Algorithm 1 by the computation of an entropic optimal transport plan between π (resp, ρ) and a uniform measure with weights $1/N$ (resp, $1/M$) and cost function given by what is inside of the exponential term in 3.2.1 (resp, 3.4.1). This computation can be interpreted as transport towards a weighted version of Y, and we omit the details for simplicity. The computed optimal transport plan gives immediately rise to responsibilities z (resp w) as described in the main text.

In the case where ρ and π are also parameters, we considered a variation over the above scheme where on a small proportion of times we carried out usual VEM updates in the inner loops (so that we can update weights) and in the rest of the times we performed the SVEM updates described in the above paragraphs. In our applications, we performed one VEM update every 6 SVEM updates. We could have alternatively considered weight updates as described in 3.2, but we did not deem this necessary.

B.3.2 Synthetic experiments: For the synthetic data experiments in Fig. 7 we considered $K = 5, G = 5$ or $K = 10, G = 10$ co-clusters, noise variances $\sigma^2 = 1.0, 2.5\ , 5.0, 7.5, 10.0, 12.5, 15.0, 17.5$ and sample sizes $N = M = 100, 500$ (i.e. the data matrix has N rows and N columns). Each co-cluster mean $\theta_{k,g}$ was sampled from a uniform distribution in the interval $[-5, 5]$. For each parameter configuration, we considered a number of 40 experiments, and on each experiment, a number of 5 random seeds. To sample data matrix Y we divided the $N \times M$ entries into subsquares of size $N/K \times N/G$ each. We sampled entries $Y_{i,j}$ as a noisy version of the corresponding co-cluster mean, i.e. $Y_{i,j} = \mu_{k,g} + \mathcal{N}\left(0, \sigma^2\right)$ if where (k, g) are indexes for the subsequent where indexes (i, j) belong to. In this experiment we kept weights ρ, π and variances σ^2 and only optimized over mean parameters θ

As initial values for VEM and Sinkhorn-VEM we used the solution provided by the spectral method with random initialization as implemented in Scikit-learn Pedregosa et al. (2011).

B.3.3 Spatial Transcriptomic experiment details We used the Prefrontal Dorsolateral Cortex spatial transcriptomic data from (Maynard et al., 2021) available in the R package Pardo et al. (2022). We considered a subset from the original dataset, sample id 151673: we focused on the expression of the most 1000 genes (out of 33538) from the sample. We focused on the region with X coordinates between 139 and 400 (the original ranges were 139 and 498) and Y coordinates between -521 and -292 (the original ranges were -521 and -109). This subregion contains the majority of measurements anatomically associated to Layers 5, 6 and white matter (WM), and we only consider these layers for clustering purposes. In total, we clustered the resulting $N = 1473$ spatial measurements out of the original 3639.

We compared the performance of VEM (Algorithm 1), SVEM and the spectral algorithm Pedregosa et al. (2011) with 'scale' initialization. In all cases, we used $K = 3$ row (spatial) clusters and $G = 8$ column (gene) clusters. As initialization for VEM and SVEM we used randomly sampled binary assignment matrices z, w. For VEM and SVEM algorithms We optimized over mean θ, variance σ and weights ρ, π. In the case of SVEM, weight updates alternated with

Since solutions from SVEM seemed to be stable to changes in random seed, to

estimate variation in performance, we considered a number of 100 experiments where each time, we subsampled the original setup with a subsampling rate of 0.5. Bar plots in Fig. 8B are averages over these 100 repetitions.

REFERENCES

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1961–1971.

Arthur, D. and Vassilvitskii, S. (2007). K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: from population to sample-based analysis. *Ann. Statist.*, 45(1):77–120.

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation with the wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boerner, T. J., Deems, S., Furlani, T. R., Knuth, S. L., and Towns, J. (2023). Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing*, pages 173–176.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Bressan, D., Battistoni, G., and Hannon, G. J. (2023). The dawn of spatial omics. *Science*, 381(6657):eabq4964.

Brown, S. T., Buitrago, P., Hanna, E., Sanielevici, S., Scibek, R., and Nystrom, N. A. (2021). Bridges-2: A platform for rapidly-evolving and data intensive research. In *Practice and Experience in Advanced Research Computing*, pages 1–4.

Canas, G. D. and Rosasco, L. (2012). Learning probability measures with respect to optimal transport metrics. *arXiv preprint arXiv:1209.1077*.

Chen, Y. and Xi, X. (2020). Likelihood landscape and local minima structures of gaussian mixture models. *arXiv preprint arXiv:2009.13040*.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Ismb*, volume 8, pages 93–103.

Csiszár, I. and Tusnády, G. (1984). Information geonetry and alternating minimization procedures. *Statistics and decisions*, 1:205–237.

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

Cuturi, M. and Doucet, A. (2014). Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR.

Daskalakis, C., Tzamos, C., and Zampetakis, M. (2017a). Ten steps of EM suffice for mixtures of two gaussians. In Kale, S. and Shamir, O., editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 704–710. PMLR.

Daskalakis, C., Tzamos, C., and Zampetakis, M. (2017b). Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dessein, A., Papadakis, N., and Deledalle, C.-A. (2017). Parameter estimation in finite mixture models by regularized optimal transport: A unified framework for hard and soft clustering. *arXiv preprint arXiv:1711.04366*.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., and Yu, B. (2020). Singularity, misspecification and the convergence rate of em.

Fettal, C., Nadif, M., et al. (2022). Efficient and effective optimal transport-based biclustering. *Advances in Neural Information Processing Systems*, 35:32989–33000.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. (2021). Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578.

Fränti, P. and Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112.

Genevay, A., Chizat, L., Bach, F., , Cuturi, M., and Peyré, G. (2019a). Sample complexity of sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Genevay, A., Dulac-Arnold, G., and Vert, J.-P. (2019b). Differentiable deep clustering with cluster size constraints. *arXiv preprint arXiv:1910.09036*.

Good, I. (1965). Categorization of classification. *Mathematics and computer science in medicine and biology*, pages 115–128.

Govaert, G. and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Groppe, M. and Hundrieser, S. (2023). Lower complexity adaptation for empirical entropic optimal transport. *arXiv preprint arXiv:2306.13580*.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129.

Houdard, A., Bouveyron, C., and Delon, J. (2018). High-dimensional mixture models for unsupervised image denoising (hdmi). *SIAM Journal on Imaging Sciences*, 11(4):2815–2846.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.

Huizing, G.-J., Peyré, G., and Cantini, L. (2022). Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*, 38(8):2169–2177.

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and Jordan, M. I. (2016). Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *NeuRIPS*, pages 4116–4124.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934.

Kato, S., Kaplan, H. S., Schrödel, T., Skora, S., Lindsay, T. H., Yemini, E., Lockery, S., and Zimmer, M. (2015). Global brain dynamics embed the motor command sequence of caenorhabditis elegans. *Cell*, 163(3):656–669.

Kivinen, J. and Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.*, 34(4):43–59.

Kolouri, S., Rohde, G. K., and Hoffmann, H. (2018). Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436.

Kwon, J., Ho, N., and Caramanis, C. (2020). On the minimax optimality of the em algorithm for learning two-component mixed linear regression. *arXiv preprint arXiv:2006.02601*.

Kwon, J., Qian, W., Caramanis, C., Chen, Y., and Davis, D. (2019). Global convergence of the em algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pages 2055–2110. PMLR.

Laclau, C., Redko, I., Matei, B., Bennani, Y., and Brault, V. (2017). Co-clustering through optimal transport. In *International conference on machine learning*, pages 1955–1964. PMLR.

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 73(4):321–332.

Marx, V. (2021). Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14.

Masud, S. B., Werenski, M., Murphy, J. M., and Aeron, S. (2023). Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling. *Journal of Machine Learning Research*, 24(160):1–65.

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, J. L., Tran, M. N., Besich, Z., Tippani, M., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436.

McLachlan, G. J. (1982). 9 the classification and mixture maximum likelihood approaches to cluster analysis. *Handbook of statistics*, 2:199–208.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in Pattern Recognition: Joint IAPR International Workshops SSPR'98 and SPR'98 Sydney, Australia, August 11–13, 1998 Proceedings*, pages 658–666. Springer.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33:331–373.

Mei, S., Bai, Y., and Montanari, A. (2016). The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*.

Mena, G., Nejatbakhsh, A., Varol, E., and Niles-Weed, J. (2020). Sinkhorn em: An expectation-maximization algorithm based on entropic optimal transport. *arXiv preprint arXiv:2006.16548*.

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4543–4553.

Nadif, M. and Govaert, G. (2008). Algorithms for model-based block gaussian clustering. *DMIN*, 8:14–17.

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Nejatbakhsh, A., Varol, E., Yemini, E., Hobert, O., and Paninski, L. (2020). Probabilistic joint segmentation and labeling of c. elegans neurons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 130–140. Springer.

Pardo, B., Spangler, A., Weber, L. M., Hicks, S. C., Jaffe, A. E., Martinowich, K., Maynard, K. R., and Collado-Torres, L. (2022). spatiallibd: an r/bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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

Pollard, D. (1982). Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. (2023). Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. *arXiv preprint arXiv:2301.11302*.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239.

Rigollet, P. and Stromme, A. J. (2022). On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*.

Rigollet, P. and Weed, J. (2018). Entropic optimal transport is maximum-likelihood deconvolution. *Comptes rendus Mathématique*, 356(11–12).

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

Smith, A. F. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):213–220.

Sottosanti, A. and Risso, D. (2023). Co-clustering of spatially resolved transcriptomic data. *The Annals of Applied Statistics*, 17(2):1444–1468.

Steinley, D. (2003). Local optima in k-means clustering: what you don't know may hurt you. *Psychological methods*, 8(3):294.

Sulston, J. E., Schierenberg, E., White, J. G., Thomson, J. N., et al. (1983). The embryonic cell lineage of the nematode caenorhabditis elegans. *Developmental biology*, 100(1):64–119.

Tan, K. M. and Witten, D. M. (2014). Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics*, 23(4):985–1008.

Titouan, V., Redko, I., Flamary, R., and Courty, N. (2020). Co-optimal transport. *Advances in neural information processing systems*, 33:17559–17570.

Varol, E., Nejatbakhsh, A., Sun, R., Mena, G., Yemini, E., Hobert, O., and Paninski, L. (2020). Statistical atlas of c. elegans neurons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–129. Springer.

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

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2):133–138.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

Wu, Y. and Zhou, H. H. (2019). Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $o(\sqrt{n})$ $iterations. arXiv preprint arXiv$: 1908.10935.

Xu, J., Hsu, D. J., and Maleki, A. (2016). Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684.

Xu, J., Hsu, D. J., and Maleki, A. (2018). Benefits of over-parameterization with EM. In *Advances in Neural Information Processing Systems*, pages 10662–10672.

Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151.

Yan, Y., Wang, K., and Rigollet, P. (2023). Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow. *arXiv preprint arXiv:2301.01766*.

Yemini, E., Lin, A., Nejatbakhsh, A., Varol, E., Sun, R., Mena, G. E., Samuel, A. D., Paninski, L., Venkatachalam, V., and Hobert, O. (2021). Neuropal: a multicolor atlas for whole-brain neuronal identification in c. elegans. *Cell*, 184(1):272–288.