DP-EM: Differentially Private Expectation Maximization

Mijung Park[1] **James Foulds**[2] **Kamalika Chaudhuri**[3] **Max Welling**[1]
[1]QUvA lab, Informatics Institute, University of Amsterdam
[2]California Institute for Telecommunications and Information Technology, University of California, San Diego
[3] Department of Computer Science, University of California, San Diego

Abstract

The iterative nature of the expectation maximization (EM) algorithm presents a challenge for privacy-preserving estimation, as each iteration increases the amount of noise needed. We propose a *practical* private EM algorithm that overcomes this challenge using two innovations: (1) a novel moment perturbation formulation for differentially private EM (DP-EM), and (2) the use of two recently developed composition methods to bound the privacy "cost" of multiple EM iterations: the *moments accountant (MA)* and *zero-mean concentrated differential privacy (zCDP)*. Both MA and zCDP bound the moment generating function of the privacy loss random variable and achieve a refined tail bound, which effectively decrease the amount of additive noise. We present empirical results showing the benefits of our approach, as well as similar performance between these two composition methods in the DP-EM setting for Gaussian mixture models. Our approach can be readily extended to many iterative learning algorithms, opening up various exciting future directions.

1 Introduction

Data on all aspects of our daily lives, such as behavioural, health and financial data, are increasingly collected, stored and analyzed by corporations and government agencies, and there is a dire need for developing machine learning tools that can analyze these data while still guaranteeing the privacy of individuals. Much progress has been made recently in developing privacy-preserving methods [1, 2] and *differential privacy*, in particular, is emerging as the dominant notion of algorithmic privacy [1].

In this paper, we derive differentially private variants of the expectation maximization (EM) algorithm which has been widely used to solve statistical problems in many areas of science including bioinformatics [3], neuroscience [4], and computer vision [5]. Expectation maximization iteratively estimates the parameters of models with unobserved variables. We present a very general privacy-preserving EM algorithm which can be used for any model with a complete-data likelihood in the exponential family. We then apply our algorithm to the mixture of Gaussians (MoG) density estimation model and the factor analysis (FA) model. Having access to a private density estimator is particularly valuable because it provides a means to anonymize the data in a principled way, by simply sampling a dataset from the model and replacing the original data with this sampled data.

Since differentially private machine learning algorithms usually achieve privacy by adding noise to perturb the output of the algorithm or its intermediate stages, the main challenge in developing privacy-preserving algorithms is in controlling the associated loss in *statistical efficiency* or utility per sample. This problem is particularly exacerbated for iterative algorithms such as EM. For example, recent work on the k-means algorithm, a variant of EM for mixture of Gaussians, requires adding noise to the parameters where the noise standard deviation is on the order of the input dimension *times the number of iterations* [6], which may necessitate early termination. To avoid this, more recent work proposes to apply a standard k-means clustering algorithm to a privatized synopsis of the data [7]. Their synopsis generation method consists of putting rectangular bounding boxes in the data space and counting how many data points are in each box. However, this method applies mainly to the clustering task and for low-dimensional data.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

Instead, we propose to resolve the privacy-utility dilemma using two key innovations: a private EM formulation based on moment perturbation for sensible use of the privacy budget *per iteration*, and recently proposed composition methods to improve the privacy cost *across many iterations*. Our moment perturbation approach is applicable for any model in which the complete-data likelihood is in the exponential family. In such cases, the EM parameters are functions of moments of latent and observed variables, which we perturb for privacy. Moment perturbation for differentially private estimators is not a new concept (see [8, 9]). However, unlike [9], we do not require subsampling of the data.

Furthermore, our algorithm calculates the cumulative privacy cost using two refined composition methods, the *moments accountant* and *zCDP*. The moments accountant [10] bounds the moments of the privacy loss random variable. Inspired by CDP [11], zCDP [12] formulates the moments of the privacy loss random variable in terms of the Rényi divergence between the output distributions obtained by running an algorithm on two datasets that differ in the value of a single individual. In both cases, the moments bound yields a tighter tail bound, and consequently, for a given total privacy budget, allows for a higher per-iteration budget than standard methods. Our experimental results show that by combining our moment perturbation formulation of privacy-preserving EM with refined composition methods, we obtain a practical and effective algorithm for privately estimating the parameters of latent variable models.

We start by reviewing differential privacy, the moments accountant, and EM in Sec. 2. In Sec. 3, we introduce our general DP-EM framework. We then derive the DP-EM algorithm for mixture of Gaussians in Sec. 4. In Sec. 5, we construct the MA and zCDP formulation for EM under MoGs. In Sec. 6, we provide the DP-EM algorithm for factor analysis, and we illustrate the effectiveness of our algorithms in Sec. 7.

2 Background

In this section, we provide background information on the definitions of algorithmic privacy that we use, the MA and zCDP formulations which provide a refined privacy analysis, as well as the general EM algorithm.

Differential privacy. Differential privacy (DP) is a formal definition of the privacy properties of data analysis algorithms [1]. Given an algorithm $\mathcal{M}$ and datasets $\mathbf{X}$, $\mathbf{X}'$ differing by a single entry, the *privacy loss* random variable of an outcome o is

$$L^{(o)} = \log \frac{Pr(\mathcal{M}_{(\mathbf{X})} = o)}{Pr(\mathcal{M}_{(\mathbf{X}')} = o)} . \tag{1}$$

$\mathcal{M}$ is ϵ-DP if and only if $|L^{(o)}| \leq \epsilon, \forall o$. Intuitively, the definition states that the output probabilities must not change very much when a single individual's data is modified, thereby limiting the amount of information that the algorithm reveals about any one individual. An approximate version is (ϵ, δ)-DP, defined to hold if and only if $|L^{(o)}| \leq \epsilon$, with probability at least $1 - \delta$.

Concentrated differential privacy. Concentrated differential privacy (CDP) is a recently proposed relaxation of differential privacy which aims to make privacy-preserving iterative algorithms more practical than for DP while still providing strong privacy guarantees. There are two variants of CDP. First, in (μ, τ)-*mCDP* [11], $L^{(o)}$ subtracted by its mean μ is subgaussian with standard deviation τ: $E[e^{\lambda(L^{(o)} - \mu)}] \leq e^{\lambda^2 \tau^2/2}, \forall \lambda \in \mathbb{R}$. Second, in τ-*zCDP* [12], that arises from a connection between the moment generating function of $L^{(o)}$ and the Rényi divergence between the distributions of $\mathcal{M}_{(\mathbf{X})}$ and that of $\mathcal{M}_{(\mathbf{X}')}$, we require: $e^{(\alpha-1)\mathrm{D}_\alpha} = E[e^{(\alpha-1)L^{(o)}}] \leq e^{(\alpha-1)\alpha\tau}, \forall \alpha \in (1, \infty)$, where the α-Rényi divergence is denoted by $\mathrm{D}_\alpha = \mathrm{D}_\alpha(Pr(\mathcal{M}_{(\mathbf{X})}) || Pr(\mathcal{M}_{(\mathbf{X}')}))$. Observe that in this case $L^{(o)}$ is also subgaussian but zero-mean. In zCDP, composition is straightfoward since the Rényi divergence between two product distributions is simply the sum of the Rényi divergences of the marginals.

We will use zCDP rather than mCDP, since many DP and approximate DP mechanisms can be characterised in terms of zCDP, but not in terms of mCDP without a large loss in privacy parameters. This correspondence will allow us to use zCDP as a tool for analyzing *composition under the* (ϵ, δ)-*DP privacy definition*, for a fair comparison between CDP and DP analyses.[1]

Moments accountant. The moments accountant calculates a privacy budget by bounding the moments of $L^{(o)}$, where the λ-th moment is defined as the log of the moment generating function evaluated at λ [10]:

$$\alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}') = \log \mathbb{E}_{o \sim \mathcal{M}(\mathcal{D})} \left[e^{\lambda L^{(o)}} \right] . \tag{2}$$

The worst case over all the neighbouring databases $\alpha_{\mathcal{M}}(\lambda)$ is defined as $\alpha_{\mathcal{M}}(\lambda) = \max_{\mathcal{D}, \mathcal{D}'} \alpha_{\mathcal{M}}(\lambda; \mathcal{D}, \mathcal{D}')$.[2]

Using Markov's inequality, for any $\epsilon > 0$, the λ-th moment is converted to the (ϵ, δ)-DP guarantee by[3]

$$\delta = \min_{\lambda} \exp \left[\alpha_{\mathcal{M}}(\lambda) - \lambda \epsilon \right] . \tag{3}$$

[1] See Sec. 4 in [12] for a detailed explanation.

[2] The form of $\alpha_{\mathcal{M}}(\lambda)$ is determined by the mechanism.

[3] See Appendix A in [10] for the proof.

The λ-th moment in Eq (2) composes linearly, which yields the composability theorem (Theorem 2.1 in [10]). An immediate result from the composibility theorem is that the sum of each upper bound on $\alpha_{\mathcal{M}_j}$ is an upper bound on the total λth moment after J compositions,

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{j=1}^{J} \alpha_{\mathcal{M}_j}(\lambda). \tag{4}$$

The general EM algorithm. Given N *i.i.d.* observations $X := \{\mathbf{x}_i\}_{i=1}^N$, with each observation $\mathbf{x}_i \in \mathbb{R}^d$, and hidden variables $Z := \{\mathbf{z}_i\}_{i=1}^N$, computing the maximum likelihood estimator of a vector of model parameters $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_L]$ is analytically intractable, due to the integral or summation inside the logarithm,

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(X|\boldsymbol{\theta}) = \log \int dZ \; p(X, Z|\boldsymbol{\theta}). \tag{5}$$

Instead, one can lower-bound $\mathcal{L}(\boldsymbol{\theta})$ using the posterior distribution over latent variables $q(Z)$ [13],

$$\mathcal{L}(\boldsymbol{\theta}) \geq \int dZ \; q(Z) \log \tfrac{p(X,Z|\boldsymbol{\theta})}{q(Z)} \overset{\text{def}}{=} \mathcal{F}(q, \boldsymbol{\theta}), \tag{6}$$

where the lower bound is often called *free energy* [14], $\mathcal{F}(q, \boldsymbol{\theta}) = \langle \log p(X, Z|\boldsymbol{\theta}) \rangle_{q(Z)} + H(q)$, where $H(q)$ is the entropy of $q(Z)$. EM alternates between: (1) the E-step: optimizing $\mathcal{F}$ wrt distribution over unobserved variables holding parameters fixed, i.e., $q^{(j)}(Z) = \arg\max_{q(Z)} \mathcal{F}(q(Z), \boldsymbol{\theta}^{(j-1)})$, and (2) the M-step: maximizing $\mathcal{F}$ wrt parameters holding the latent distribution fixed

$$\boldsymbol{\theta}^{(j)} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q^{(j)}(Z), \boldsymbol{\theta}) \tag{7}$$

where $\mathcal{F}(q^{(j)}(Z), \boldsymbol{\theta}) = \langle \log p(X, Z|\boldsymbol{\theta}) \rangle_{q^{(j)}(Z)} + \text{const}$ since $H(q)$ does not directly depend on $\boldsymbol{\theta}$.

To understand what EM does, one can rewrite the free energy in terms of the log-likelihood and the KL divergence terms, $\mathcal{F}(q, \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) - D_{KL}\left[q(Z)||p(Z|X, \boldsymbol{\theta})\right]$. During the E-step, we set $q^{(j)}(Z) = p(Z|X, \boldsymbol{\theta}^{j-1})$, which makes the second term zero and the free energy equals the likelihood. Then, in the M-step, we get the maximum likelihood estimate (MLE). For the maximum a posteriori (MAP) estimate, we add the log prior for the parameters $\log p(\boldsymbol{\theta})$ to the right hand side of Eq (7).

3 The general DP-EM algorithm

The EM algorithm is frequently used for models whose joint distribution over observed and unobserved variables remains in the exponential family: $p(X, Z) = h(X, Z) \exp(\boldsymbol{\theta}^\top T(X, Z))/A(\boldsymbol{\theta})$, while the marginal $p(X)$ does not. In this case, the free energy can be rewritten as

$$\mathcal{F}(q, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \langle T(X, Z) \rangle_{q(Z)} - N \log A(\boldsymbol{\theta}) + c, \tag{8}$$

where c is some constant wrt $\boldsymbol{\theta}$, and $\boldsymbol{\theta}^\top \langle T(X, Z) \rangle_{q(Z)} = \sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}_i)} \sum_{l=1}^L \theta_l T_l(\mathbf{x}_i, \mathbf{z}_i)$. In the E-step, we compute the expected sufficient statistics under q, i.e., $\langle T(X, Z) \rangle_{q(Z)}$. Then, in the M-step, we compute partial derivatives wrt each parameter,

$$\tfrac{\partial}{\partial \theta_l} \mathcal{F}(q, \boldsymbol{\theta}) = \tfrac{1}{N} \sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}_i)} T_l(\mathbf{x}_i, \mathbf{z}_i) - \tfrac{\partial}{\partial \theta_l} \log A(\boldsymbol{\theta}) = 0.$$

Although it is not straightforward to derive a closed-form expression for each parameter update due to the dependence on other parameters in $A(\boldsymbol{\theta})$, it is easy to see that each parameter update depends on each expected sufficient statistics, i.e., moments, denoted by $M_l = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}_i)} T_l(\mathbf{x}_i, \mathbf{z}_i)$. So, to output privatized parameters, all we need is to perturb the moments to compensate any single data point's change. The sensitivity of the expected sufficient statistics is given by

$$\begin{aligned} &\Delta M_l \qquad\qquad (9)\\ &= \max_{|\mathcal{D} - \tilde{\mathcal{D}}|_1 = 1} |M_l(\mathcal{D}) - \tilde{M}_l(\tilde{\mathcal{D}})|, \\ &= \max_{\mathbf{x}_j, \mathbf{x}'_j} \tfrac{1}{N} |\mathbb{E}_{q(\mathbf{z}_j)} T_l(\mathbf{x}_j, \mathbf{z}_j) - \mathbb{E}_{q(\mathbf{z}'_j)} T_l(\mathbf{x}'_j, \mathbf{z}'_j)|, \\ &\leq \max_{\mathbf{x}_j, \mathbf{x}'_j} \tfrac{1}{N} |\langle T_l(\mathbf{x}_j, \mathbf{z}_j) \rangle_{q(\mathbf{z}_j)}| + \tfrac{1}{N} |\langle T_l(\mathbf{x}'_j, \mathbf{z}'_j) \rangle_{q(\mathbf{z}'_j)}|, \end{aligned}$$

where the last line is due to the triangle inequality. The expectation over $\mathbf{z}$ can be rewritten as an inner product, and using Hölder's inequality: $|\langle T_l(\mathbf{x}_j, \mathbf{z}_j) \rangle_{q(\mathbf{z}_j)}| = |\langle q(\mathbf{z}_j), T_l(\mathbf{x}_j, \mathbf{z}_j) \rangle| \leq |q(\mathbf{z}_j)|_1 |T_l(\mathbf{x}_j, \mathbf{z}_j)|_\infty$, where $|q(\mathbf{z}_j)|_1 = 1$ and $|T_l(\mathbf{x}_j, \mathbf{z}_j)|_\infty$ is maximum over all $(\mathbf{x}_j, \mathbf{z}_j)$. As in many existing works (e.g., [15, 16] among many others), we also assume that datasets are pre-processed such that the L_2 norm of any $\mathbf{x}_i$ is less than 1, meaning that any $\mathbf{x}_i$ stays within a unit ball. Furthermore, we assume that $q(Z)$ has a bounded support of Z denoted by $\mathcal{Z}$. Under these assumptions, the sensitivity is given by $\Delta M_l = \max_{(\mathbf{x}_j, \mathbf{z}_j) \in (B_1(\mathcal{X}), \, \mathcal{Z})} \frac{2}{N} |T_l(\mathbf{x}_j, \mathbf{z}_j)|$. Using this sensitivity, we add noise to each moment and the perturbed moments are mapped by a model-specific deterministic function g to the vector of privatized parameters, given as $\tilde{\boldsymbol{\theta}}^* = g(\{\tilde{M}_l\}_{l=1,\cdots,L})$, where $\tilde{M}_{l=1,\cdots,L}$ are perturbed moments. Using this general framework, we derive the differentially private EM algorithm for mixture of Gaussians and factor analysis in the following.

4 DPEM for mixture of Gaussians

4.1 EM for Mixture of Gaussians

We consider the mixture of Gaussians (MoG) model as a first example to derive the DP-EM algorithm. For K Gaussians and N data points $X := \{\mathbf{x}_i\}_{i=1}^N$, the log-likelihood under MoG is given by $\log p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) =$

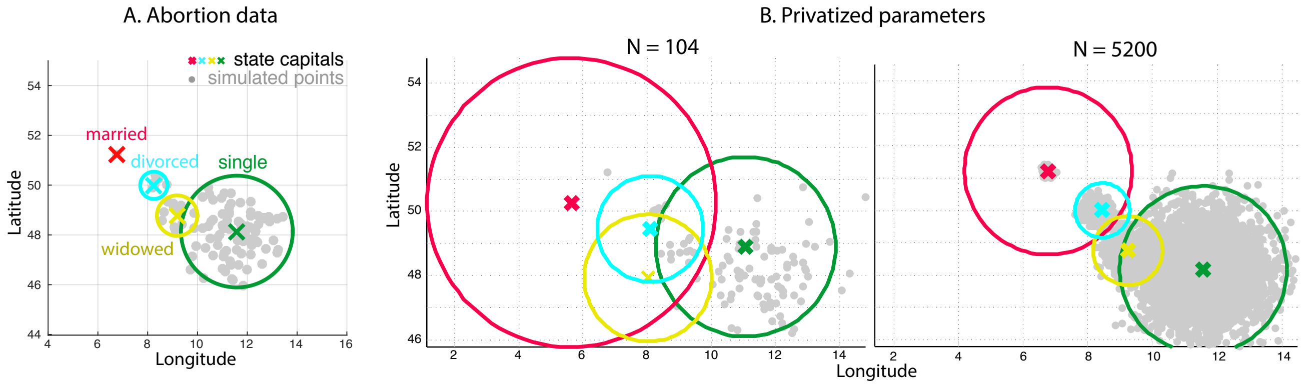

Figure 1: **A.** The abortion dataset (from *destatis.de*) provides per-marital-status abortion rates occurred in the state of Baden-Württemberg in 2015, as well as from which state each individual came from. Due to the lack of exact location, we simulated 104 data points based on the abortion rate in each state (in grey). Notice that there is only one person who is originally from the state of North Rhine-Westphalia (top left, in red) and falls into the 'married' category. Hence, the person's information is completely revealed in the mean parameter if one runs the conventional EM algorithm. **B (Left).** Given the 104 data points, by privatizing the mean and variance parameters as illustrated in Sec. 7, the married person's information (top left, in red) is now not easily inferrable. **B (Right).** When we have 50 times more datapoints, the privatized parameters are closer to those given by the conventional EM algorithm. However, now the mean parameter for the married category provides aggregated information from several people, which makes it hard to infer any individual information.

$\sum_{i=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$, where $\sum_{k=1}^K \pi_k = 1$. We denote the parameters by $\boldsymbol{\theta} := \{\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma\} = \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$.

Introducing a binary vector of length K for each data point, $\mathbf{z}_i \in \mathbb{R}^K$, to represent the membership to which Gaussian each datapoint belongs, e.g., $z_{i,k} \in \{0,1\}$ and $\sum_{k=1}^K z_{i,k} = 1$, the distribution over each $\mathbf{z}_i$ is given by $p(\mathbf{z}_i) := \prod_{k=1}^K \pi_k^{\mathbf{z}_{i,k}}$, and the distribution over all unobserved variables $Z = \{\mathbf{z}_i\}_{i=1}^K$ is given by $p(Z) := \prod_{i=1}^N p(\mathbf{z}_i)$. The joint distribution over observed and unobserved variables, which is in the exponential family, is given by $\log p(X, Z|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{i=1}^N \sum_{k=1}^K z_{i,k}[\log \pi_k + \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)]$. In the E-step, we compute the responsibilities as $\langle \delta_{z_{i,k}=k} \rangle_{q(Z)}$ given the parameters from the previous iteration $\boldsymbol{\theta}^{prev}$

$$\begin{aligned}\gamma_{i,k} &= p(z_{i,k} = 1|\mathbf{x}_i, \boldsymbol{\theta}^{prev}),\\ &= \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) / \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k), \qquad (10)\end{aligned}$$

and in the M-step, we update the parameters $\boldsymbol{\theta}$ by

$$\pi_k^{MLE} = \frac{N_k}{N}, \quad \boldsymbol{\mu}_k^{MLE} = \frac{1}{N_k} \sum_{i=1}^N \gamma_{i,k} \mathbf{x}_i, \qquad (11)$$

$$\Sigma_k^{MLE} = \frac{1}{N_k} \sum_{i=1}^N \gamma_{i,k} (\mathbf{x}_i - \boldsymbol{\mu}_k^{MLE})(\mathbf{x}_i - \boldsymbol{\mu}_k^{MLE})^\top,$$

where $N_k = \sum_{i=1}^N \gamma_{i,k}$. We provide the formulae for the maximum a posteriori estimate in the supplementary material.

Before moving to the next section, we would like to motivate why it is important to construct a privacy preserving algorithm for MoG. In Fig. 1, we show that if one runs the EM algorithm for the given dataset, an individual's information can be easily revealed by just looking at the EM parameters, while the noised-up parameters obtained by the method, which will be described next, protect private information effectively.[4]

4.2 DPEM for MoG

Under MoG, we plug in the responsibilities given in Eq (10) to the parameter update expressions given in Eq (11). We then perturb each of these by taking into account one datapoint's worst-case difference between two neighboring datasets. We use ϵ_i to denote a privacy budget allocated per iteration.

ϵ_i-DP or (ϵ_i, δ_i)-DP mixing coefficients. For two neighbouring datasets with a single data point difference, the maximum difference in $\boldsymbol{\pi}$ occurs when the data point $\mathbf{x}_j$ is assigned to the k-th Gaussian with $\gamma_{j,k} = 1$ and the altered data point $\mathbf{x}'_j$ is assigned to another, e.g., the k'-th Gaussian, with $\gamma'_{j,k'} = 1$. Hence, we get the following sensitivity:

$$\Delta\boldsymbol{\pi}^{MLE} = \max_{\mathbf{x}_j, \mathbf{x}'_j} \sum_{k=1}^K \tfrac{1}{N} |\gamma_{j,k} - \gamma'_{j,k}| \leq 2/N, \qquad (12)$$

[4] We first pre-processed the data by scaling down the magnitude with the maximum L2 norm of the data points, and then added noise to each parameter following the derivations in Sec. 4. For visualisation, we map the results back to the original latitude/longtitude space.

since $0 \leq \gamma_{j,k} \leq 1$ and $\sum_{k=1}^{K} \gamma_{j,k} = 1$. We add noise to compensate the maximum difference[5]

$$\tilde{\boldsymbol{\pi}}^{MLE} = \boldsymbol{\pi}^{MLE} + (Y_1, \cdots, Y_K), \tag{13}$$

where $Y_i \sim^{i.i.d.} \text{Lap}(\frac{\Delta \boldsymbol{\pi}^{MLE}}{\epsilon'})$ or $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 \geq 2\log(1.25/\delta_i)(\Delta \boldsymbol{\pi}^{MLE})^2/\epsilon_i^2$. For π_k^{MAP}, we do not need any additional sensitivity analysis, since the MAP estimate is a deterministic mapping of the MLE.

ϵ_i-DP or (ϵ_i, δ_i)-DP mean parameters. Using the noised-up $\tilde{N}_k$ obtained from the noised-up mixing coefficients, i.e., $\tilde{N}_k = N\tilde{\pi}_k$, the maximum difference in mean parameters due to one datapoint's difference is

$$\begin{aligned} \Delta_1 \boldsymbol{\mu}_k^{MLE} &= \max_{\mathbf{x}_j, \mathbf{x}'_j} \tfrac{1}{\tilde{N}_k} \left| (A_k + \gamma_{j,k}\mathbf{x}_j) - (A_k + \gamma'_{j,k}\mathbf{x}'_j) \right|_1, \\ &\leq 2\sqrt{d}/\tilde{N}_k, \end{aligned} \tag{14}$$

where $A_k := \sum_{i=1, i\neq j}^{N} \gamma_{i,k}\mathbf{x}_i$ and the L1 term is bounded by Eq (9). The $\sqrt{d}$ term is from the fact that each input vector is L2-norm bounded by 1.[6] We add noise to the MLE via[7]

$$\tilde{\boldsymbol{\mu}}_k^{MLE} = \boldsymbol{\mu}_k^{MLE} + (Y_1, \cdots, Y_d), \tag{15}$$

where $Y_i \sim^{i.i.d.} \text{Lap}(\Delta_1 \boldsymbol{\mu}_k^{MLE}/\epsilon')$ or $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 \geq 2\log(1.25/\delta_i)(\Delta_2 \boldsymbol{\mu}_k^{MLE})^2/\epsilon_i^2$, where $\Delta_2 \boldsymbol{\mu}_k^{MLE} = 2/\tilde{N}_k$.

(ϵ_i, δ_i)-DP covariance parameters. For covariance perturbation, we follow the Analyze Gauss (AG) algorithm [17], which provides (ϵ_i, δ_i)-DP. We first draw Gaussian random variables

$$\mathbf{z} \sim \mathcal{N}\left(0, \beta I_{d(d+1)/2}\right), \tag{16}$$

where $\beta = 2\log(1.25/\delta_i)(\Delta\Sigma_k^{MLE})^2/(\epsilon_i)^2$ and the sensitivity of the covariance matrix [8] in Frobenius norm is given by

$$\begin{aligned} \Delta\Sigma_k^{MLE} &= \max_{\mathbf{x}_j, \mathbf{x}'_j} \tfrac{1}{\tilde{N}_k} |\text{vec}\{(B_k + \gamma_{j,k}\mathbf{x}_j\mathbf{x}_j{}^\top - \tilde{M}_k) \\ &\qquad - (B_k + \gamma'_{j,k}\mathbf{x}'_j\mathbf{x}'_j{}^\top - \tilde{M}_k)\}|_2, \\ &\leq \tfrac{2}{\tilde{N}_k}\sqrt{\sum_{l=1}^{d}\sum_{l'=1}^{d}(\mathbf{x}_{j,l}\mathbf{x}_{j,l'})^2} \leq \tfrac{2}{\tilde{N}_k} \end{aligned} \tag{17}$$

where $B_k := \sum_{i=1, i\neq j}^{N} \gamma_{i,k}\mathbf{x}_i\mathbf{x}_i{}^\top$, and $\tilde{M}_k = \tilde{N}_k \tilde{\boldsymbol{\mu}}_k^{MLE} \tilde{\boldsymbol{\mu}}_k^{MLE\top}$. Using $\mathbf{z}$, we construct a upper triangular matrix (including diagonal), then copy the upper part to the lower part so that the resulting matrix Z becomes symmetric. Then, we add this noisy matrix to the covariance matrix

$$\tilde{\Sigma}_k^{MLE} := \Sigma_k^{MLE} + Z. \tag{18}$$

The perturbed covariance might not be positive definite. In such case, we project the negative eigenvalues to some value near zero to maintain positive definiteness of the covariance matrix.

Combinations of the perturbations. Among all the possible combinations of these parameter perturbation mechanisms, we focus on two scenarios. Scenario 1 (which we call *LLG*) uses the ϵ_i-DP Laplace mechanism for perturbing mixing coefficients (once) and mean parameters (K times) and the (ϵ_i, δ_i)-DP Gaussian mechanism for perturbing the covariance parameters (K times). Since there are K Gaussians, for J iterations, there will be $J(K+1)$ compositions of ϵ_i-DP mechanism and JK compositions of (ϵ_i, δ_i)-DP mechanisms in total in this scenario. Scenario 2 (which we call *GGG*) uses the (ϵ_i, δ_i)-DP Gaussian mechanism for perturbing all the parameters. For J iterations, there will be $J(2K+1)$ compositions of (ϵ_i, δ_i)-DP mechanism in total in this scenario.

5 Compositions for DP-EM for MoGs

Before describing our method, we first describe the two baseline methods. First, in *Linear* (Lin) composition (Theorem 3.16 [1]), privacy degrades linearly with the number of iterations. This result is from the Max Divergence of the privacy loss random variable being bounded by a total budget. Hence, the linear composition yields $(J(2K+1)\epsilon_i,\ JK\delta_i)$-DP under scenario *LLG* and $(J(2K+1)\epsilon_i,\ J(2K+1)\delta_i)$-DP under scenario *GGG*. Second, *Advanced* (Adv) composition (Theorem 3.20 [1]), resulting from the Max Divergence of the privacy loss random variable being bounded by a total budget including a slack variable δ, yields $(J(2K+1)\epsilon_i(e^{\epsilon_i}-1) + \sqrt{2J(2K+1)\log(1/\delta')}\epsilon_i,\ \delta' + JK\delta_i)$-DP under scenario *LLG* and $(J(2K+1)\epsilon_i(e^{\epsilon_i} - 1) + \sqrt{2J(2K+1)\log(1/\delta')}\epsilon_i,\ \delta' + J(2K+1)\delta_i)$-DP under scenario *GGG*.

Our method calculates the per-iteration budget using the two composition methods below.

zCDP composition (zCDP). *z-CDP* composition yields $(\rho + 2\sqrt{\rho\log(1/\delta)}, \delta)$-DP, where

$$\rho = J(K+1)\epsilon_i^2/2 + JK\Delta\Sigma_k^2/(2\sigma_3^2)$$

[5] To ensure $\tilde{\boldsymbol{\pi}}_k^{MLE} \in [0,1]$, we set $\tilde{\boldsymbol{\pi}}_k^{MLE} = 0$, if $\tilde{\boldsymbol{\pi}}_k^{MLE} < 0$, and $\tilde{\boldsymbol{\pi}}_k^{MLE} = 1$, if $\tilde{\boldsymbol{\pi}}_k^{MLE} > 1$. Then, we re-normalize $\tilde{\pi}^{MLE}$ after the projection to ensure $\sum_{k=1}^{K} \tilde{\pi}_k^{MLE} = 1$.

[6] $\sum_{l=1}^{d} |\mathbf{x}_{i,l}| \leq \left(\sum_{l=1}^{d} |\mathbf{x}_{i,l}|^2\right)^{\frac{1}{2}} \left(\sum_{l=1}^{d} 1\right)^{\frac{1}{2}} \leq \sqrt{d}$.

[7] The MAP estimate only differs from the MLE in the denominator. Hence, we simply replace $\tilde{N}_k$ with $\tilde{N}_k + \kappa_0$ in Eq (14) in the MAP estimation case.

[8] The MAP estimate only differs from the MLE in the denominator. We replace $\tilde{N}_k$ with $\tilde{N}_k + \nu_0 + d + 2$ in Eq (17) in the MAP estimation case.

under scenario LLG and

$$\rho = J\Delta\boldsymbol{\pi}^2/(2\sigma_1^2) + JK\Delta\boldsymbol{\mu}_k^2/(2\sigma_2^2) + JK\Delta\Sigma_k^2/(2\sigma_3^2)$$

under scenario GGG, for sensitivity $\Delta\boldsymbol{\pi}, \Delta\boldsymbol{\mu}_k, \Delta\Sigma_k$ and $\sigma_1^2 \geq 2\log(1.25/\delta_i)\Delta\boldsymbol{\pi}^2/\epsilon_i^2$, $\sigma_2^2 \geq 2\log(1.25/\delta_i)\Delta\boldsymbol{\mu}_k^2/\epsilon_i^2$, and $\sigma_3^2 \geq 2\log(1.25/\delta_i)\Delta\Sigma_k^2/\epsilon_i^2$, where $0 < \epsilon_i < 1$.

These results are obtained by using the following results in [12]: Proposition 1.4. If $\mathcal{M}$ satisfies ϵ_i-DP, then $\mathcal{M}$ satisfies $\frac{1}{2}\epsilon_i^2$-zCDP; Proposition 1.6. Gaussian mechanism satisfies $\Delta^2/(2\sigma^2)$-zCDP, where Δ is a sensitivity; Lemma 1.7. If two mechanisms satisfy ρ_1-zCDP and ρ_2-zCDP, respectively, then their composition satisfies $(\rho_1 + \rho_2)$-zCDP; and Proposition 1.3. If $\mathcal{M}$ provides ρ-zCDP, then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho\log(1/\delta)}, \delta)$-DP for any $\rho > 0$.

Moments Accountant composition (MA). For using MA, as a first step, we identify the form of privacy loss random variable and its λ-th moment in each mechanism we use. For ϵ_i-DP Laplace mechanism $\mathcal{M}_i^L$ outputting $f(\mathcal{D})$ and $x \sim \text{Lap}(0, \frac{\Delta f}{\epsilon_i})$, $L^{(o)}$ at $o = f(\mathcal{D}) + x$ has the following form:

$$L^{(o)} = \begin{cases} \epsilon_i, \text{ if } x < 0, \text{ w.p. } \frac{1}{2} \\ -\epsilon_i, \text{ if } x > \Delta f, \text{ w.p. } \frac{1}{2}\mathrm{e}^{-\epsilon_i} \\ -\frac{\epsilon_i}{\Delta f}(2x - \Delta f), \text{ if } 0 \leq x \leq \Delta f, \text{ w.p.} \frac{1}{2}(1 - \mathrm{e}^{-\epsilon_i}). \end{cases}$$

Following the definition in Eq (2), the λ-th moment is given by

$$\alpha_{\mathcal{L}} = \log\left[\frac{\lambda+1}{2\lambda+1}e^{\lambda\epsilon_i} + \frac{\lambda}{2\lambda+1}e^{(-\epsilon_i(\lambda+1))}\right]. \quad (19)$$

For (ϵ_i, δ_i)-DP Gaussian mechanism $\mathcal{M}_i^G$ with noise magnitude σ and $x \sim \mathcal{N}(0, \sigma^2)$, $L^{(o)}$ at $o = f(\mathcal{D}) + x$ is $L^{(o)} = \left(\frac{\Delta f}{\sigma}\right)\left(\frac{x}{\sigma}\right) + \frac{1}{2}\left(\frac{\Delta f}{\sigma}\right)^2$. The λ-th moment is then

$$\alpha_{\mathcal{G}} = (\lambda^2 + \lambda)\frac{(\Delta f)^2}{2\sigma^2}. \quad (20)$$

Note that multi-dimensional Laplace/Gaussian mechanisms also have the same form of the λ-th moment as the scalar version. See the Supplementary material for the derivation.

For achieving (ϵ, δ)-DP, the tail bound is given by $\delta = \min_\lambda \exp\left[J(K+1)\alpha_{\mathcal{L}} + JK\alpha_{\mathcal{G}} - \lambda\epsilon\right]$ under scenario LLG; and $\delta = \min_\lambda \exp\left[J(2K+1)\alpha_{\mathcal{G}} - \lambda\epsilon\right]$ under scenario GGG. Under each case, we calculate ϵ_i satisfying the tail bound with the fixed budget (ϵ, δ). Algorithm 1 summarizes our method.

6 DPEM for Factor Analysis

Under FA, the conditional distributions over observed variables $\mathbf{x}_i$ are assumed to be Gaussian, $p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|W\mathbf{z}_i, \Psi)$, and the prior over latent variables $\mathbf{z}_i$ is also assumed to be Gaussian: $p(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|0, I)$.

Algorithm 1 DP-EM under MoG using MA

```
Require: Dataset $\mathcal{D}$, per-iteration budget $(\epsilon_i, \delta_i)$ calculated by MA or zCDP composition
Ensure: $(\epsilon, \delta)$-DP parameters $\tilde{\boldsymbol{\theta}}$
  Iterate until convergence (J iterations):
  Compute parameters by plugging in the responsibilities given in Eq (10).
  Noise up $\boldsymbol{\pi}$ by Eq (13), $\boldsymbol{\mu}$ by Eq (15), and $\Sigma$ by Eq (18).
```

In this case, the complete-data likelihood is proportional to $p(X, Z) \propto \exp(\phi(\boldsymbol{\theta})^\top T(X, Z))$, where $\phi(\boldsymbol{\theta})$ is a vectorized version of the concatenated matrix $[W^\top\Psi^{-1}, \Psi^{-1}, -\frac{1}{2}G^{-1}]$, $G^{-1} = I + W^\top\Psi^{-1}W$, and where the sufficient statistics are also a vectorized version of a concatenated matrix $T(X, Z) = [\sum_{i=1}^N \mathbf{x}_i\mathbf{z}_i{}^\top, \sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i{}^\top, \sum_{i=1}^N \mathbf{z}_i\mathbf{z}_i{}^\top]$.

Due to conjugacy the posterior over $\mathbf{z}_i$ is also Gaussian, where the first and second moments are given by $\bar{\mathbf{z}}_i = GW^\top\Psi^{-1}\mathbf{x}_i$ and $\langle\mathbf{z}_i\mathbf{z}_i{}^\top\rangle = G + \bar{\mathbf{z}}_i\bar{\mathbf{z}}_i{}^\top$. The expected sufficient statistics become a function of the data second moment matrix, denoted by $\Lambda := \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i{}^\top$,

$$\begin{aligned} &\langle T(X, Z)\rangle_{q(Z)} \\ &= N\left[\Lambda\Psi^{-1}WG^\top,\ \Lambda,\ G + GW^\top\Psi^{-1}\Lambda\Psi^{-1}WG^\top\right]. \end{aligned}$$

For privacy-preserving EM, we perturb Λ by Analyze Gauss [17], resulting in a perturbed matrix $\tilde{\Lambda}$, which we use when updating the parameters by

$$W^{new} = \left[\tilde{\Lambda}\Psi^{-1}WG^\top\right]\left[G + GW^\top\Psi^{-1}\tilde{\Lambda}\Psi^{-1}WG^\top\right]^{-1},$$

$$\Psi^{new} = \text{diag}\left[\tilde{\Lambda} - WGW^\top\Psi^{-1}\tilde{\Lambda}\right].$$

until convergence, at no extra privacy cost. Therefore, unlike MoGs, FA only requires perturbing the data second moment matrix *once* for privacy preservation. The EM iterations are then *post-processing* steps which are free from cumulative differential privacy loss.

7 Experiments

We used four real-world datasets to test our algorithm. In all datasets, we preprocessed the data such that the input vectors had maximum norm 1.

Stroke dataset was used in [18] for predicting the occurrence of a stroke within a year after an atrial fibrillation diagnosis. We used 100 principal components ($d = 100$) of 4,096 raw features (conditions and

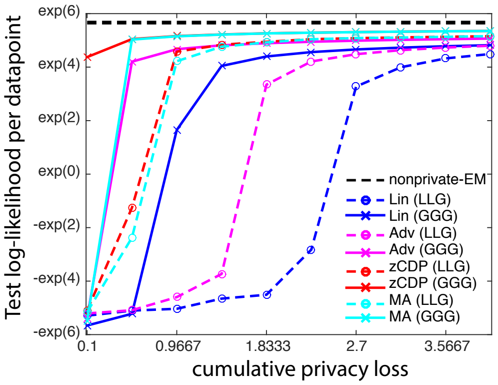

Figure 2: Stroke dataset. Test log-likelihood per data point as a function of cumulative privacy loss after 20 EM iterations. We fit the data with MoG using the conventional EM first (in black dotted line). We then ran the private EM algorithm with a different per-iteration privacy budget resulting from different composition methods, in order to achieve (ϵ, δ)-DP EM parameters, where ϵ varies from 0.1 to 4 and δ is fixed to 10^{-4}. We fixed $\delta_i = 10^{-6}$ when using Gaussian mechanisms.

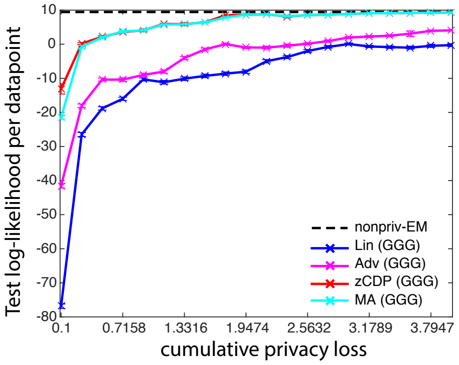

Figure 3: Life Science dataset Test log-likelihood per data point as a function of cumulative privacy loss after 10 EM iterations. We fit the data with MoG using the conventional EM first (in black dotted line). We then ran the DP-EM algorithm (GGG combination) with a different per-iteration privacy budget resulting from different composition methods, in order to achieve (ϵ, δ)-DP EM parameters, where ϵ varies from 0.1 to 4 and δ is fixed to 10^{-4}. We fixed $\delta_i = 10^{-8}$.

medicines) recorded from $50,345$ patients, by assuming that the private database was given in this form. We divided the extracted dataset into 10 different pairs of training (90%) and test sets (10%), and reported the average test log-likelihood per datapoint across the 10 independent trials in Fig. 2, setting $k = 10$.

Overall, the *GGG* scenario yielded higher test log-likelihoods than the *LLG* scenario, so we focused on this method in our experiments. We found that using the zCDP and MA compositions resulted in more accurate estimates, while also requiring less privacy budget, compared to other compositions. zCDP performed better than MA with a small privacy budget ϵ, but they both performed similarly well with a larger budget. The difference with small ϵ may be due to only searching over integer values of λ for MA, which we do for computational reasons, following [10].

Life Science dataset is from the UCI repository [19]. The dataset contains 26,733 records, consisting of 10 principal components from chemistry and biology experiments ($d = 10$). Following other approaches (e.g., [20]), we set $k = 3$. We divided the dataset into 10 different pairs of training (90%) and test sets (10%), and reported the average test log-likelihood per data point across the 10 independent trials in Fig. 3. In this experiment, we focused on scenario *GGG*. Using the zCDP and MA compositions once again resulted in more accurate estimates while requiring less privacy budget than linear and advanced compositions.

Gowalla dataset contains the social network's users' check-in locations in terms of longitude and latitude ($d = 2$). The total number of data points is 1,256,384, which we divided into 10 cross-validation sets. We then performed k-means clustering and compared our method to a differentially private k-means clustering algorithm, DPLloyd [6]. The standard Lloyd algorithm for k-means clustering first partitions the data into k clusters, with each point assigned to be in the same cluster as the nearest centroid, and then updates each centroid to be the center of the data points in the cluster. As summarized in [7], the DPLloyd adds noise to the updated centroids. Specifically, the Laplace noise is added to the number of data points assigned to each cluster as well as to the sum of each coordinate of the data points assigned to each cluster. Hence, the sensitivity becomes $d + 1$. In the original DPLloyd algorithm, due to the conventional composition theorem for DP, their noise distribution follows $\text{Lap}((d+1)J/\epsilon)$ for J iterations. We also tested the DPLloyd algorithm with zCDP compositions, which resulted in better performance in terms of normalized intra-cluster variance (NICV) across the 10 test sets. Our algorithm for k-means clustering also perturbs the centroids by adding the Laplace noise with zCDP composition, where the sensitivity of the mean locations is given in Eq (14). We set $\epsilon = 0.01$ and $\delta = 10^{-4}$ for both algorithms. As shown in Fig. 4, our method achieves smaller NICV than DPLloyd, even with a very small value of ϵ.

Olivetti Faces dataset is used to illustrate our private factor analysis method[9]. The dataset consists of ten different images for each of 40 distinct subjects

[9] We obtained the dataset from `http://scikit-learn.org/`, but the dataset is originally from *AT&T* Laboratories Cambridge.

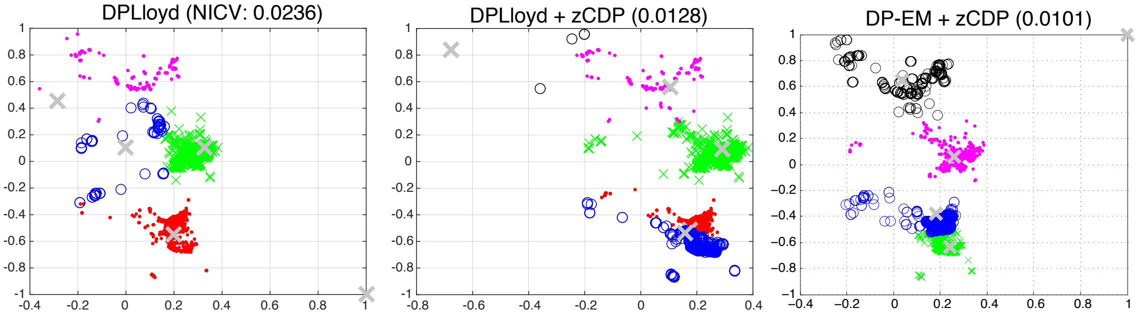

Figure 4: k-Means Clustering. Visualisation of clustering results with total privacy budget $\epsilon = 0.01$ and tolerance $\delta = 10^{-4}$. The center locations are depicted in gray cross. The numbers in parenthesis are *normalised intra-cluster variance* (NICV) values obtained by each method. **Left** The DPLloyd algorithm with linear composition performed poorly due to the relatively high level of additive noise. **Middle** DPLloyd with zCDP composition performed better than the original version. **Right** Our algorithm achieves smaller NICV than two variants of DPLloyd given the same privacy budget.

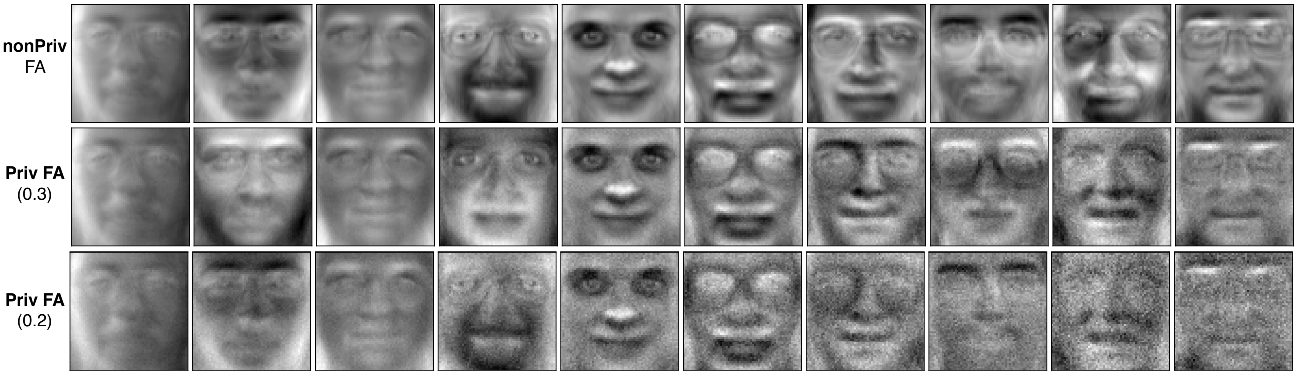

Figure 5: Private Factor Analysis. Visualisation of each column of estimated W (reshaped as 64 by 64 images). **Top**: Non-private EM. **Middle**: DP-EM with $\epsilon = 0.3$ and $\delta = 10^{-4}$. **Bottom**: DP-EM with $\epsilon = 0.2$ and $\delta = 10^{-4}$.

($N = 400$), where each image is 64 by 64, resulting in 4096 features ($d = 4096$). Each pixel is a floating point value on the interval $[0, 1]$. Each image was treated as a datapoint, rather than each subject, though this could readily be done via *group privacy* [1]. We set the latent dimension to 10. We tested non-private EM, DP-EM with $\epsilon = 0.2$ and $\epsilon = 0.3$ (fixing $\delta = 10^{-4}$), and showed each column of the estimated loading matrix W in Fig. 5. With $\epsilon = 0.2$ (bottom) the components were noisy, but with $\epsilon = 0.3$ (middle) the FA components' faces were nearly as recognizable as for the non-private FA algorithm (top), thereby accurately recovering a set of typical faces in the dataset.

8 Conclusion

We have developed a practical algorithm that outputs accurate and privatized EM parameters based on moment perturbation under the MA and zCDP composition analyses, which effectively decrease the amount of additive noise for the same expected privacy guarantee compared to the standard analysis. We illustrated the effectiveness of our algorithm on four datasets. Based on our results, we recommend the use of zCDP composition analysis for EM, since it performed better than MA in some regimes and is easier to compute. Furthermore, we found that the GGG combination performed better than LLG under these composition methods in the context of EM, which perhaps makes sense since the zCDP and MA compositions are tailored to the Gaussian mechanism.

The private EM algorithms for the mixture of Gaussians and factor analysis models we discussed in this paper are clearly only two examples of a much broader class of models to which our private EM framework applies. Our positive empirical results with EM strongly suggest that these ideas are likely to be beneficial for privatizing many other iterative machine learning algorithms. In future work, we plan to apply this general framework to other inference methods. This fits our broader vision that *practical* privacy preserving machine learning algorithms will have an increasingly relevant role to play in our field.

Acknowledgments

MP and MW thanks Quacomm for research support. KC thanks NSF under 1253942 for research support.

References

[1] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014.

[2] Anand D Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE signal processing magazine*, 30(5):86–94, 2013.

[3] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Technical report, Department of Computer Science and Engineering, University of California, San Diego, 1994.

[4] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*, 20(1):45–57, 2001.

[5] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(8):1026–1038, 2002.

[6] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 128–138, New York, NY, USA, 2005. ACM.

[7] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means clustering. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, CODASPY '16, pages 26–37, New York, NY, USA, 2016. ACM.

[8] J. R. Foulds, J. Geumlek, M. Welling, and K. Chaudhuri. On the theory and practice of privacy-preserving Bayesian data analysis. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

[9] Adam D. Smith. Efficient, differentially private point estimators. *CoRR*, abs/0809.4794, 2008.

[10] M. Abadi, A. Chu, I. Goodfellow, H. Brendan McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. *ArXiv e-prints*, July 2016.

[11] C. Dwork and G. N. Rothblum. Concentrated differential privacy. *ArXiv e-prints*, March 2016.

[12] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *CoRR*, abs/1605.02065, 2016.

[13] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Kluwer Academic Publishers, 1998.

[14] R. P. Feynman. *Statistical Mechanics: A Set of Lectures*. Perseus, 1972.

[15] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, July 2011.

[16] Daniel Kifer, Adam Smith, Abhradeep Thakurta, Shie Mannor, Nathan Srebro, and Robert C. Williamson. Private convex empirical risk minimization and high-dimensional regression. In *In COLT*, pages 94–103, 2012.

[17] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: optimal bounds for privacy-preserving principal component analysis. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 11–20, 2014.

[18] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Department of Statistics Technical Report tr608, University of Washington, 2014.

[19] M. Lichman. UCI machine learning repository, 2013.

[20] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David E. Culler. GUPT: privacy preserving data analysis made easy. In K. Selçuk Candan, Yi Chen, Richard T. Snodgrass, Luis Gravano, and Ariel Fuxman, editors, *SIGMOD Conference*, pages 349–360. ACM, 2012.