OPEN

Differentially private density estimation with skew-normal mixtures model

Weisan Wu

The protection of private data is a hot research issue in the era of big data. Differential privacy is a strong privacy guarantees in data analysis. In this paper, we propose DP-MSNM, a parametric density estimation algorithm using multivariate skew-normal mixtures (MSNM) model to differential privacy. MSNM can solve the asymmetric problem of data sets, and it is could approximate any distribution through expectation-maximization (EM) algorithm. In this model, we add two extra steps on the estimated parameters in the M step of each iteration. The first step is adding calibrated noise to the estimated parameters based on Laplacian mechanism. The second step is post-processes those noisy parameters to ensure their intrinsic characteristics based on the theory of vector normalize and positive semi definition matrix. Extensive experiments using both real data sets evaluate the performance of DP-MSNM, and demonstrate that the proposed method outperforms DPGMM.

With the rapid developments of big data technology, data privacy protection has become a important issue. Differential privacy is an attack-resistant model with strict theoretical proof and good quantitative representation. Differential private density estimation has been done on mixtures model. The most of them are considered in the Gaussian mixtures model (GMM)^{1,2}, but the data can presents skewness or heavy tailed behavior. Nearly, the vulnerability of pre-trained models to the attack is not uniform when the training data itself is skewed, Trues et al.³ shows that risk from some attacks is routinely increased when models use skewed training data. This risk may be caused by the reduction of the accuracy of the model density estimation on assumed data normality.

To the best of our knowledge no existing work has concern in the skew-normal mixtures model with differential privacy. In skew-normal mixtures model, all data are assumed to be drawn by randomly sampling from one of component of the skew-normal mixture model, and the learning task is to estimate the proportion parameters, location parameters, scale parameters and skewness parameters of each component distribution in model. In particular, we need to consider some problem of skew-normal mixtures model, where (1) skew-normal distribution compose the skew-normal mixtures model, and (2) we not clear the sample belongs to which component, and (3) we just only have perturbed data, and (4) we hope to get estimate values of all parameters.

Addressing on these problems, we propose DP-MSNM, which appends two extra steps (noise adding step and post-processing step) to the general method of density estimation with MSNM in each iteration like. The noise adding step is adding classical Laplace noise to the original estimated parameters in each iteration to achieve differential privacy. The post-processing step will through eigen decomposition on scale parameter to prevent its not positive semi-definite and use normalize method to ensure sum the weight parameter as to 1.

The remainder of the paper is organized as follows. Section “Preliminaries” introduces preliminaries. Given DP-MSNM details in Section “Differentially private MSNM”. Section “Experimental setup” evaluates the experimental performance. Finally, Section “Conclusions” concludes the work and discusses future works.

Preliminaries

Let D be a data set that contains n records $x_1, \dots, x_n$ with p attributes, and $x_i = (x_{i1}, \dots, x_{ip})$. There $\|x\|_p$, $\|A\|_p$ are denote the L_p norms of vector x and matrix A .

Differential privacy. The privacy definition is due to Dwork et al.⁴. The inform definition is given follows.

Definition 1 (*Differential privacy*) A randomized algorithm mechanism $\mathcal{M}$ provides ϵ -differential privacy, if for all Output O of $\mathcal{M}$ and for any two neighbor data sets D and D' , we have

¹Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China. ²School of Mathematics and Statistics, Baicheng Normal University, Baicheng 137000, China. email: wuws009@outlook.com

$$P(\mathcal{M}(D) = O) \leq e^\varepsilon P(\mathcal{M}(D') = O)$$

We called ε as privacy budget, the smaller ε means the higher level of privacy provides. Where we adding calibrated random noise to the output to achieved privacy. We use Laplacian mechanism and L_1 -sensitivity to output perturbation.

Definition 2 (L_1 -sensitivity) The L_1 sensitivity of a query function $\mathcal{M}$ as the maximum change in the L_1 -norm of the query output on the neighbor data sets D and D' . That is ,

$$S(\mathcal{M}) = \max_{d(D,D')=1} \|\mathcal{M}(D) - \mathcal{M}(D')\|_1$$

where we use Hamming distance to measure the distance between two data sets D and D' , and its distance is expressed as $d(D, D') = \{i : x_i \neq y_i, x_i \in D, y_i \in D'\}$. We denote the noisy vector or matrix are follows Laplace distribution $Lap(0, \frac{S(\mathcal{M})}{\varepsilon})$, where $\frac{S(\mathcal{M})}{\varepsilon}$ is scale parameter. Given L_1 sensitivity $S(\mathcal{M})$ and the privacy budget ε , this mechanism can ensure differential privacy.

Density estimation with MSNM. In the multivariate skew-normal mixture models (MSNM), we assumed sample x_i is i.i.d. with a density $f_{MSN}(x_i)$ that can be written as a linear composition of skew-normal distributions⁵ in the form

$$\begin{aligned} f_{MSNM}(x; \Theta) &= \sum_{k=1}^G \pi_k f_{MSN}(x; \theta_k) \\ &= \sum_{k=1}^G \pi_k \{2\phi^d(x; \xi_k, \Sigma_k, \alpha_k) \Phi(\alpha_k^T \sigma^{-1}(x - \xi_k))\}, \end{aligned} \quad (1)$$

where $x = (x_1, \dots, x_n)$, and $x_i, i = 1, \dots, n$ is a p -dimensional vector. $\Theta = (\theta_1, \dots, \theta_G)$, $\theta_k = (\xi_k, \Sigma_k, \alpha_k)$, $k = 1, \dots, G$. where $\pi_k \geq 0$ is mixture proportion and satisfied that $\sum_{k=1}^G \pi_k = 1$. $\xi_k, \Sigma_k, \alpha_k$ are parameter in the k -th component of the mixture models, and α_k is the p -dimensional skewness parameter, Σ_k is a $p \times p$ positive semi-definite (PSD) correlation matrix, ξ_k is the p -dimensional local parameter. $\phi_d(x; \xi, \Sigma)$ is the density of a normal distribution $N(\xi, \Sigma)$, $\Phi(\cdot)$ is the cumulate distribution function of standard normal. The function $f_{MSN}(x; \xi, \Sigma, \alpha)$ can be present by

$$f_{MSN}(x; \xi, \Sigma, \alpha) = 2\phi_d(x; \xi, \Sigma, \alpha) \Phi(\alpha^T \sigma^{-1}(x - \xi)) \quad (2)$$

where $\sigma = (\Sigma \odot I_p)^{\frac{1}{2}}$, $\odot$ denote Hadamard product. If a random variable X has density function like Eq. (2), we called the variable X is a Multivariate skew-normal distribution, and denote $X \sim SN_p(\xi, \Sigma, \alpha)$ like as Contreras-Reyes and Cortés, 2016⁶.

Recalling the variables $X_0 \sim SN_p(0, \bar{\Sigma}, \alpha)$ and $X \sim SN_p(\xi, \Sigma, \alpha)$, we can obtain an additional representation for multivariate skew-normal distribution from the results of Azzalini et al.⁵ as follows:

$$X_0 = D_\delta Z_0 + \delta |Z_1|, \quad (3)$$

where $\delta = (\delta_1, \delta_2, \dots, \delta_p)^T$ is a vector with the elements in $(-1, 1)$, and

$$D_\delta = (I_p - \text{diag}(\delta)^2)^{1/2},$$

U_0 and U_1 are the normal variables of dimension p and 1, the joint distribution is

$$\begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} \sim N_{p+1} \left(0, \begin{pmatrix} \bar{\Omega} & 0 \\ 0 & 1 \end{pmatrix} \right),$$

$\bar{\Psi}$ is a full-rank correlation matrix.

By the transformation $X = \xi + \sigma X_0$, we have

$$Y = \xi + \sigma X_0 = \xi + \sigma D_\delta Z_0 + \sigma \delta |Z_1|.$$

Let $|Z_1| = \tau$. Given τ , we can obtain the condition representation, that is,

$$X|\tau \sim N_p(\xi + \sigma \delta \tau, \sigma D_\delta \bar{\Omega} D_\delta \sigma). \quad (4)$$

For a p -dimensional SN distribution, the parameters have the following relationships by simple algebraic work (See⁵):

$$\begin{aligned} \omega &= D_\delta^{-1} \delta; \\ \bar{\Sigma} &= D_\delta (\bar{\Omega} + \omega \omega^T) D_\delta; \\ \alpha &= (1 + \omega^T \bar{\Omega}^{-1} \omega)^{-1/2} D_\delta^{-1} \bar{\Omega}^{-1} \omega; \\ \delta &= (1 + \alpha^T \bar{\Sigma} \alpha)^{-1/2} \bar{\Sigma} \alpha. \end{aligned} \quad (5)$$

Furthermore, we have

$$\tilde{\Sigma} = D_{\delta} \tilde{\Omega} D_{\delta} + D_{\delta} \omega \omega^T D_{\delta} = D_{\delta} \tilde{\Omega} D_{\delta} + \delta \delta^T.$$

Thus, let $\eta = \sigma \delta$, by the relation between Σ and $\tilde{\Sigma}$, we have

$$\Sigma = \sigma \tilde{\Sigma} \sigma = \sigma D_{\delta} \tilde{\Omega} D_{\delta} \sigma + \sigma \delta \delta^T \sigma = \sigma D_{\delta} \tilde{\Omega} D_{\delta} \sigma + \eta \eta^T.$$

Therefore

$$X|\tau \sim N_p(\xi + \sigma \delta \tau, \Sigma - \eta \eta^T).$$

Let Z_{ik} be the membership indicator variable such that it equals 1 when x_i is from the k -th component of the MSNM model, and equals 0 otherwise. Consider the complete data $(X, Z) = \{X_i, Z_i\}_{i=1}^n$, where the latent component-indicators vector $Z_i = (Z_{i1}, \dots, Z_{iG})$ follows a multi-nomial distribution with the trial and cell probabilities of $\pi_1, \dots, \pi_G$. Let us write it as $Z_i \sim M(1; \pi_1, \dots, \pi_G)$. Based on the component indicators, for each X_i , $i = 1, \dots, n$, the hierarchical representation for MSNMs can be given by Cabral et al.⁷

$$\begin{aligned} X_i|\tau_i, Z_{ik}=1 &\sim N_p(\xi_k + \sigma_k \delta_k \tau_i, \Sigma_k - \eta_k \eta_k^T), \\ \tau_i|Z_{ik}=1 &\sim TN_{[0,+\infty)}(0, 1), \\ Z_i &\sim Multi(1; \pi_1, \dots, \pi_G), \end{aligned} \quad (6)$$

where $TN_{[0,+\infty)}(0, 1)$ denotes the half normal distribution. According to the hierarchical representation (6), ignoring the added constants and denoting

$$\Lambda_k = \Sigma_k - \eta_k \eta_k^T \quad (7)$$

the complete data log-likelihood function is

$$\begin{aligned} p\ell_n^c(\Theta) &= \sum_{i=1}^n \sum_{k=1}^G Z_{ij} \left\{ \log(\pi_k) - \frac{1}{2} \log |\Lambda_k| - \frac{1}{2} (x_i - \xi_k)^T \Lambda_k^{-1} (x_i - \xi_k) \right. \\ &\quad \left. + (x_i - \xi_k)^T \Lambda_k^{-1} \eta_k \tau_i - \frac{1}{2} \eta_k^T \Lambda_k^{-1} \eta_k \tau_i^2 - \frac{1}{2} \tau_i^2 \right\} \end{aligned}$$

According to the Bayesian theorem, we have

$$\tau_i|X_i = x_i, Z_{ik} = 1 \sim TN_{[0,+\infty)}(\mu_{\tau_{ik}}, \sigma_{\tau_{ik}}^2),$$

where

$$\mu_{\tau_{ik}} = \frac{(x_i - \xi_k)^T \Lambda_k^{-1} \eta_k}{1 + \eta_k^T \Lambda_k^{-1} \eta_k}, \quad \sigma_{\tau_{ik}}^2 = \frac{1}{1 + \eta_k^T \Lambda_k^{-1} \eta_k}.$$

Thus, for the current parameters $\Theta^{(t)}$, let

$$\mu_{\tau_{ik}}^{(t)} = \frac{(x_i - \xi_k^{(t)})^T (\Lambda_k^{(t)})^{-1} \eta_k^{(t)}}{1 + \eta_k^{(t)T} (\Lambda_k^{(t)})^{-1} \eta_k^{(t)}}, \quad \sigma_{\tau_{ik}}^{2(t)} = \frac{1}{1 + (\eta_k^{(t)})^T (\Lambda_k^{(t)})^{-1} \eta_k^{(t)}}.$$

The EM algorithm then proceeds as follows: *E-step*: Let us compute the conditional expectations as

$$\begin{aligned} r_{0ik}^{(t)} &= \frac{\pi_k^{(t)} f_{MSN}(x_i; \xi_k^{(t)}, \Sigma_k^{(t)}, \alpha_k^{(t)})}{\sum_{j=1}^G \pi_j^{(t)} f_{MSN}(x_i; \xi_j^{(t)}, \Sigma_j^{(t)}, \alpha_j^{(t)})}, \\ r_{1ik}^{(t)} &= E(\tau_i|X_i = x_i, Z_{ik} = 1, \Theta^{(t)}) = \mu_{\tau_{ik}}^{(t)} + \sigma_{\tau_{ik}}^{(t)} \Delta_{ik}^{(t)}, \\ r_{2ik}^{(t)} &= E(\tau_i^2|X_i = x_i, Z_{ik} = 1, \Theta^{(t)}) = \mu_{\tau_{ik}}^{2(t)} + \sigma_{\tau_{ik}}^{2(t)} + \mu_{\tau_{ik}}^{(t)} \sigma_{\tau_{ik}}^{(t)} \Delta_{ik}^{(t)}, \end{aligned}$$

where $\Delta_{ik}^{(t)} = \phi\left(\frac{\mu_{\tau_{ik}}^{(t)}}{\sigma_{\tau_{ik}}^{(t)}}\right) / \Phi\left(\frac{\mu_{\tau_{ik}}^{(t)}}{\sigma_{\tau_{ik}}^{(t)}}\right)$. Thus, we get $E(Z_{ik} \tau_i|X_i = x_i, \Theta^{(t)}) = r_{0ik}^{(t)} r_{1ik}^{(t)}$, and

$$E(Z_{ik} \tau_i^2|X_i = x_i, \Theta^{(t)}) = r_{0ik}^{(t)} r_{2ik}^{(t)}.$$

Therefore, the Q function can be written as

$$\begin{aligned} Q(\Theta, \Theta^{(t)}) &= \sum_{i=1}^n \sum_{k=1}^G r_{0ik}^{(t)} \left\{ \log(\pi_k) - \frac{1}{2} \log |\Lambda_k| - \frac{1}{2} (x_i - \xi_k)^T \Lambda_k^{-1} (x_i - \xi_k) \right. \\ &\quad \left. + (x_i - \xi_k)^T \Lambda_k^{-1} \eta_k r_{1ik}^{(t)} - \frac{1}{2} (1 + \eta_k^T \Lambda_k^{-1} \eta_k) r_{2ik}^{(t)} \right\}. \end{aligned}$$

M-Step: Let us maximize $Q(\Theta, \Theta^{(t)})$ with respect to Θ under the restriction with $\sum_{k=1}^G \pi_k = 1$.

1. Update $\pi_k^{(t)}$ by $\pi_k^{(t+1)} = n^{-1} \sum_{i=1}^n r_{0ik}^{(t)}$.
2. Update $\xi_k^{(t)}$ by

$$\xi_k^{(t+1)} = \frac{\sum_{i=1}^n r_{0ik}^{(t)} x_i - \eta_k^{(t)} \sum_{i=1}^n r_{0ik}^{(t)} r_{1ik}^{(t)}}{\sum_{i=1}^n r_{0ik}^{(t)}}. \quad (8)$$

3. Update $\eta_k^{(t)}$ by

$$\eta_k^{(t+1)} = \frac{\sum_{i=1}^n r_{0ik}^{(t)} (x_i - \xi_k^{(t+1)}) r_{1ik}^{(t)}}{\sum_{i=1}^n r_{0ik}^{(t)}}. \quad (9)$$

4. Update $\Lambda_k^{(t)}$ by

$$\Lambda_k^{(t+1)} = \frac{1}{\sum_{i=1}^n r_{0ik}^{(t)}} \left\{ \sum_{i=1}^n r_{0ik}^{(t)} (x_i - \xi_k^{(t+1)}) (x_i - \xi_k^{(t+1)})^T - 2\eta_k^{(t+1)} \sum_{i=1}^n r_{0ik}^{(t)} r_{1ik}^{(t)} (x_i - \xi_k^{(t+1)}) + \eta_k^{(t+1)} \sum_{i=1}^n r_{0ik}^{(t)} r_{2ik}^{(t)} (\eta_k^{(t+1)})^T \right\}. \quad (10)$$

Differentially private MSNM

In this section, we present DP-MSNM method which is a differentially private density estimation algorithm with MSNM.

Main idea. Let G denotes the component order of the MSNM. The main idea of DP-MSNM is add to two extra steps after getting the original estimated parameters in the M-step of each iteration. The original estimated parameters include mixture proportion (π), local parameter vector (ξ_k), scale parameter matrix (Σ_k) and skewness parameter vector (α_k) of each Skew-Normal distribution, where $\pi = (\pi_1, \dots, \pi_G)$ and $k \in \{1, \dots, G\}$. First, we need to get noise of each parameter by L_1 -sensitivity and allocated privacy budget, and add these noises to the original estimated parameters. We get noisy parameters $\tilde{\pi}$ and $\tilde{\xi}_k, \tilde{\Lambda}_k, \tilde{\eta}_k$. The second step is post-processes $\tilde{\pi}$ and $\tilde{\xi}_k, \tilde{\Lambda}_k, \tilde{\eta}_k$, since the noise added will break some intrinsic characteristics of weight parameter and component parameters, this step will output $\hat{\pi}$ and $\hat{\xi}_k, \hat{\Lambda}_k, \hat{\eta}_k$.

Noise adding step. In this part, we will analyze the L_1 sensitivities of $\pi_k, \xi_k, \Lambda_k, \eta_k$, and we will add calibrated noise to the original estimated parameters. We suppose that $D = \{x_1, \dots, x_n\}$ and $D' = \{x_1, \dots, x_{n-1}\}$ are two neighbor data sets, where D' has the same $n-1$ records as D , but not have the n -th record. Let $N_k = \sum_{i=1}^n r_{0ik}$, we also assume the data set is well-separated⁸ and $N_k \geq \frac{n}{2G}, k = 1, \dots, G$. We use differentially private k-means algorithm⁹ to get the centers of clusters, and get η_k from skew parameter α_k through Mardia measure method¹⁰. We denote $S(\pi), S(\xi), S(\Lambda), S(\eta)$ as the L_1 -sensitivities of $\pi, \xi_k, \Lambda_k, \eta_k$ in each iteration. According to Eqs. (8)–(10), we give $S(\pi), S(\xi), S(\Lambda), S(\eta)$ in Lemmas 1, 2, 3 and 4.

Lemma 1 In each iteration, the L_1 -sensitivity of π is $S(\pi) = \frac{G}{n}$.

The proof process of Lemma 1 is identical as Wu et al.¹. In order to make the proof of subsequent lemmas clear, we still write it down as follows.

Proof Since $\pi = (\pi_1, \dots, \pi_G)$, let $\pi(D)$ and $\pi(D')$ denote the weights of data sets D and D' . Then the k -th elements of $\pi(D) - \pi(D')$ is $\frac{\sum_{i=1}^n r_{0ik}}{n} - \frac{\sum_{i=1}^{n-1} r_{0ik}}{n-1} = \frac{\sum_{i=1}^{n-1} (r_{0nk} - r_{0ik})}{n(n-1)}$. Since $|r_{0nk} - r_{0ik}| < 1$, we have $\|\pi(D) - \pi(D')\|_1 = \sum_{k=1}^G \left| \frac{\sum_{i=1}^{n-1} (r_{0nk} - r_{0ik})}{n(n-1)} \right| \leq \sum_{k=1}^G \frac{\sum_{i=1}^{n-1} |r_{0nk} - r_{0ik}|}{n(n-1)} \leq \frac{G}{n}$. $\square$

Lemma 2 In each iteration, if $\|x_i\|_1 \leq R, N_k \geq \frac{n}{2G}$ and $\|\eta_k\|_1 \leq R_3$, the L_1 -sensitivity of ξ_k is $S(\xi) = \frac{4G(R-R_3)}{n}$.

Proof Let $\xi(D)$ and $\xi(D')$ denote the local parameter of data sets D and D' . The k -th elements of $\xi(D) - \xi(D')$ is

$$\begin{aligned} & \frac{\sum_{i=1}^n r_{0ik} x_i - \eta_k \sum_{i=1}^n r_{0ik} r_{1ik}}{\sum_{i=1}^n r_{0ik}} - \frac{\sum_{i=1}^{n-1} r_{0ik} x_i - \eta_k \sum_{i=1}^{n-1} r_{0ik} r_{1ik}}{\sum_{i=1}^{n-1} r_{0ik}} \\ &= \frac{r_{0nk} \sum_{i=1}^{n-1} r_{0ik} (x_n - x_i) - \eta_k r_{0nk} \sum_{i=1}^n r_{0ik} (r_{1nk} - r_{1ik})}{\sum_{i=1}^n r_{0ik} \sum_{i=1}^{n-1} r_{0ik}} \end{aligned}$$

Because of $\|x_i\|_1 \leq R, N_k \geq \frac{n}{2G}, \|\eta_k\|_1 \leq R_3$, we have

$$\begin{aligned}\|\xi(D) - \xi(D')\|_1 &\leq \frac{2G}{n} \frac{|Rr_{0nk} \sum_{i=1}^{n-1} r_{0ik} - \eta_k r_{0nk} \sum_{i=1}^{n-1} r_{0ik}|}{\sum_{i=1}^{n-1} r_{0ik}} \\ &\leq \frac{2G}{n} (|Rr_{0nk} - \eta_k r_{0nk}|) \\ &\leq \frac{4G(R - R_3)}{n}.\end{aligned}$$

Lemma 3 In each iteration, if $\|x_i\|_1 \leq R, N_k \geq \frac{n}{2G}$ and $\|r_{1ik}\|_1 \leq R_1, \|r_{2ik}\|_1 \leq R_2$, the L_1 -sensitivity of η_k is $S(\eta) = \frac{2GRR_1}{nR_2}$.

Proof let $\eta(D)$ and $\eta(D')$ denote the parameter in the Eq. (9) D and D' . The k -th elements of $\eta(D) - \eta(D')$ is

$$\begin{aligned}&\frac{\sum_{i=1}^n r_{0ik}(x_i - \xi_k)r_{1ik}}{\sum_{i=1}^n r_{0ik}r_{2ik}} - \frac{\sum_{i=1}^{n-1} r_{0ik}(x_i - \xi_k)r_{1ik}}{\sum_{i=1}^{n-1} r_{0ik}r_{2ik}} \\ &= \frac{r_{0ik} \sum_{i=1}^{n-1} r_{0ik}[r_{2ik}r_{1ik}(x_n - \xi_k) - r_{2ik}r_{1ik}(x_i - \xi_k)]}{\sum_{i=1}^n r_{0ik}r_{2ik} \sum_{i=1}^{n-1} r_{0ik}r_{2ik}}\end{aligned}$$

Then

$$\|\eta(D) - \eta(D')\|_1 \leq \frac{2GRR_1}{nR_2}.$$

□

Lemma 4 In each iteration, if $\|x_i\|_1 \leq R, N_k \geq \frac{n}{2G}, \|r_{1ik}\|_1 \leq R_1, \|r_{2ik}\|_1 \leq R_2, \|\eta_k\|_1 \leq R_3$, the L_1 -sensitivity of Λ_k is $S(\Lambda) = \frac{2G(R^2 + 4RR_1R_3 + 4R_2R_3^2)}{n}$.

Proof let $\Lambda(D)$ and $\Lambda(D')$ denote the parameter in the Eq. (10) D and D' . The k -th element of vector $\Lambda(D) - \Lambda(D')$ is

$$\begin{aligned}&\frac{\sum_{i=1}^n r_{0ik}(x_i - \xi_k)(x_i - \xi_k)^T - 2\eta_k \sum_{i=1}^n r_{0ik}r_{1ik}(x_i - \xi_k)^T + \eta_k \sum_{i=1}^n r_{0ik}r_{2ik}\eta_k^T}{\sum_{i=1}^n r_{0ik}} \\ &- \frac{\sum_{i=1}^{n-1} r_{0ik}(x_i - \xi_k)(x_i - \xi_k)^T - 2\eta_k \sum_{i=1}^{n-1} r_{0ik}r_{1ik}(x_i - \xi_k)^T + \eta_k \sum_{i=1}^{n-1} r_{0ik}r_{2ik}\eta_k^T}{\sum_{i=1}^{n-1} r_{0ik}} \\ &= \frac{1}{\sum_{i=1}^n r_{0ik} \sum_{i=1}^{n-1} r_{0ik}} \left\{ r_{0nk} \sum_{i=1}^{n-1} [(x_n - \xi_k)(x_n - \xi_k)^T - (x_i - \xi_k)(x_i - \xi_k)^T] \right. \\ &\quad \left. - 2\eta_k r_{0nk} \sum_{i=1}^{n-1} [r_{1nk}(x_n - \xi_k)^T - r_{1ik}(x_i - \xi_k)^T] + \eta_k \sum_{i=1}^{n-1} [r_{0nk}r_{2nk}\eta_k^T - r_{0ik}r_{2ik}\eta_k^T] \right\}\end{aligned}$$

Then

$$\|\Lambda(D) - \Lambda(D')\|_1 \leq \frac{2G(R^2 + 4RR_1R_3 + 4R_2R_3^2)}{n}.$$

In each iteration, let $\varepsilon_\pi, \varepsilon_\xi, \varepsilon_\Lambda, \varepsilon_\alpha$ present the privacy budget allocated to $\pi, \xi_k, \Lambda_k, \alpha_k$ respectively, where $k = 1, \dots, G$. So, we know the privacy budget in each iteration is $\varepsilon_\pi + G(\varepsilon_\xi + \varepsilon_\Lambda + \varepsilon_\alpha)$ from composite theorem. Next we can get the noisy parameters $\tilde{\pi}, \tilde{\xi}_k, \tilde{\Lambda}_k, \tilde{\alpha}_k$ through Lemmas 1, 2, 3 and 4 and the Laplace mechanism, it was given in follows:

$$\begin{aligned}\tilde{\pi}_k &= \pi_k + e_{\pi_k}^{1 \times 1}, \\ \tilde{\xi}_k &= \xi_k + e_{\xi_k}^{p \times 1}, \\ \tilde{\Lambda}_k &= \Lambda_k + e_{\Lambda_k}^{p \times p}, \\ \tilde{\eta}_k &= \eta_k + e_{\eta_k}^{p \times 1}.\end{aligned}$$

where $e_{\pi_k} \sim \text{Lap}(0, \frac{S(\pi)}{\varepsilon_\pi}), e_{\xi_k} \sim \text{Lap}(0, \frac{S(\xi)}{\varepsilon_\xi}), e_{\Lambda_k} \sim \text{Lap}(0, \frac{S(\Lambda)}{\varepsilon_\Lambda}), e_{\eta_k} \sim \text{Lap}(0, \frac{S(\eta)}{\varepsilon_\eta})$ denote noise respect to $\pi_k, \xi_k, \Lambda_k, \eta_k$.

Post-processing step. The noise parameters $\tilde{\pi}, \tilde{\xi}_k, \tilde{\Lambda}_k, \tilde{\eta}_k$ which obtained in the Noise adding step may not satisfy some of the basic properties of the original parameters. Firstly, $\tilde{\pi}$ not satisfied $\pi_k \geq 0$ and $\sum_{k=1}^n = 1$

generally. Secondly, $\bar{\Lambda}_k$ would not be a positive semi-definite matrix after adding noisy. If the noisy covariance matrix $\bar{\Lambda}_k$ is not PSD matrix, the other parameters, for instance $\hat{\xi}_k, \hat{\alpha}_k$ can not be calculated exactly. We need to solve these problems as follow.

First, we use the method of normalized to the $\bar{\pi}$. We let $\bar{\pi}_k = \frac{\bar{\pi}_k - \min\{\bar{\pi}\}}{\max\{\bar{\pi}\} - \min\{\bar{\pi}\}}$, and make them sum to one by $\hat{\pi}_k = \frac{\bar{\pi}_k + \zeta}{\sum_{k=1}^G \bar{\pi}_k + G\zeta}$. There ζ is a small number, it can be smooth $\bar{\pi}$

Secondly, we can use theory of Higham¹¹ to ensure the covariance matrix $\bar{\Sigma}_k$ to be PSD through post-processing $\bar{\Lambda}_k$. Like Wu et al.¹, we use the follow equation

$$\hat{\Lambda}_k = \bar{\Lambda}_k + \varrho(\bar{\Lambda}_k)I, \quad (11)$$

where $\varrho(\bar{\Lambda}_k) = \min\{r \geq 0 : \bar{\Lambda}_k + rI \geq 0\}$, ≥ 0 means that all the eigenvalues are not less than 0, I is the identity matrix. We use $\hat{\Lambda}_k$ to get $\hat{\xi}_k$, and next we will get $\hat{\eta}_k, \hat{\Sigma}_k$ and $\hat{\alpha}_k$ by Eqs. (5) and (7). The parameters $\hat{\xi}_k, \hat{\Sigma}_k, \hat{\alpha}_k$ will be join into the next iteration calculation.

Although adding two extra steps on original parameters, we can know the parameters $\hat{\pi}, \hat{\xi}, \hat{\alpha}, \hat{\Sigma}$ also satisfied ε -differential privacy¹².

DP-MSNM. From what discussed above, we give the DP-MSNM algorithm in Algorithm 2, where lines 1–2 allocate the privacy budget and initialize the parameters, lines 3 calculates the L_1 -sensitivities of all parameters, lines 5–6 execute the normal E-step and M-step, lines 7–12 execute the noise adding step, lines 13–20 execute the post-processing step. Given total privacy budget ε , Wu et al.¹ shows that under the maximum iteration T , Algorithm 2 satisfies ε -differential privacy. for $k = 1, \dots, G$, we allocate the privacy budget $\varepsilon_\Lambda = \frac{0.6}{T*G}\varepsilon$ to Λ_k , $\varepsilon_\xi = \frac{0.13}{T*G}\varepsilon$ and $\varepsilon_\eta = \frac{0.13}{T*G}\varepsilon$ to ξ_k and η_k respectively, because of the higher ε_Λ can reduce the amount of noise added to the Λ_k .

Algorithm 1 Post-Processing of $\hat{\Lambda}_k$

Input: $\bar{\Lambda}_k$

Output: $\hat{\Lambda}_k$

- 1: $E \leftarrow \text{Eigen Decomposition}(\bar{\Lambda}_k)$; //Obtain eigenvalues
 - 2: $\bar{\lambda}_{\min} \leftarrow \min(E)$; //Find the minimum eigenvalues
 - 3: $\delta(\bar{\Lambda}_k) \leftarrow \max\{0, -\bar{\lambda}_{\min}\}$
 - 4: $\hat{\Lambda}_k = \bar{\Lambda}_k + \delta(\bar{\Lambda}_k)I$;
 - 5: **return** $\hat{\Lambda}_k$
-

Algorithm 2 Differentially private density estimation with DP-SMSN

Input: $D, \varepsilon, T, G, n, R, R_1, R_2, R_3$; // ε is the total privacy budget

Output: $\Theta = (\pi, \xi_1, \dots, \xi_G, \Sigma_1, \dots, \Sigma_G, \alpha_1, \dots, \alpha_G)$

- 1: $\varepsilon_I = 0.1\varepsilon, \varepsilon_\pi = \frac{0.04}{T}\varepsilon, \varepsilon_\xi = \frac{0.13}{T*G}\varepsilon, \varepsilon_\Lambda = \frac{0.6}{T*G}\varepsilon, \varepsilon_\eta = \frac{0.13}{T*G}\varepsilon$; //Allocate privacy budget
 - 2: $\text{Init}(\Theta, \varepsilon_I)$; //Initialize parameters using PrivGene⁹
 - 3: $S(\pi), S(\xi), S(\Lambda), S(\eta) \leftarrow \text{CalSense}(R, G, n, R_1, R_2, R_3)$; //Lemma 1, 2, 3, 4
 - 4: **for** $j = 1$ to T **do**
 - 5: $\Gamma \leftarrow \text{doEStep}(\Theta)$;
 - 6: $\Theta \leftarrow \text{doMStep}$;
 - 7: $\bar{\pi} \leftarrow \text{NoiseAdding}(\varepsilon_\pi, \pi, S(\pi))$;
 - 8: **for** $k = 1$ to G **do**
 - 9: $\bar{\xi}_k \leftarrow \text{NoiseAdding}(\varepsilon_\xi, \xi_k, S(\xi))$;
 - 10: $\bar{\Lambda}_k \leftarrow \text{NoiseAdding}(\varepsilon_\Lambda, \Lambda_k, S(\Lambda))$;
 - 11: $\bar{\eta}_k \leftarrow \text{NoiseAdding}(\varepsilon_\eta, \eta_k, S(\eta))$;
 - 12: **end for**
 - 13: $\hat{\pi} \leftarrow \text{Post-Process}(\bar{\pi})$; //Normalize $\bar{\pi}$
 - 14: **for** $k = 1$ to G **do**
 - 15: $\hat{\Lambda}_k \leftarrow \text{PPS}(\bar{\Lambda}_k)$; //Call Algorithm 1
 - 16: $\hat{\xi}_k \leftarrow \hat{\Lambda}_k$;
 - 17: $\hat{\eta}_k \leftarrow \hat{\Lambda}_k, \hat{\xi}_k$;
 - 18: $\hat{\Sigma}_k \leftarrow \hat{\Lambda}_k, \hat{\eta}_k$; // Through Eq(7)
 - 19: $\hat{\alpha}_k \leftarrow \hat{\Sigma}_k, \hat{\eta}_k$; // Through Eq(5)
 - 20: **end for**
 - 21: $\Theta \leftarrow \hat{\pi}, \hat{\xi}_1, \dots, \hat{\xi}_G, \hat{\Sigma}_1, \dots, \hat{\Sigma}_G, \hat{\alpha}_1, \dots, \hat{\alpha}_G$; //Update Θ
 - 22: **end for**
 - 23: **return** Θ ;
-

	n	d	K	B	T
3DRoad	434,847	3	6	2	30
AIS	202	2	2	2	30

Table 1. Description of data sets. **Figure 1.** Effect of total privacy budget on AIS data set.

Experimental evaluation

Our experimental implemented in R version 3.6.3¹³ based on mixsmsn package⁷ to get DP-MSNM results, and all experiments are performed on a laptop with Intel Core i7-9750H and 2.60 GHz CPU and 6.00 RAM.

Experimental setup. We evaluate DP-MSNM on two real data sets, the Australian Institute of Sport data set and the 3DRoad Network data set, which from UCI Machine Learning Repository. We choose the columns 2–3 in AIS data set and the columns 2–4 in 3DRoad data set. For convenience, we'll still assume that the default values for both data sets are the same as those in Wu et al.¹. In this article, we use the complete data set, and the default values are recorded in the Table 1.

We use log-likelihood mean $\frac{1}{n} \log\{f_{MSNM}(x; \pi, \xi, \Sigma, \alpha)\}$ which is the average log-likelihood of the whole data set to evaluate the performance of DP-MSNM. A larger logarithmic likelihood mean means that the parameters estimated by the model have a better performance. We compare the density estimation effects of SAGMM, DPGMM and DP-MSNM with log-likelihood mean models for the two data sets respectively (see Figs. 1 and 2). In particular, we use AIC and BIC criteria⁶ to compare the fitting effect of DPGMM algorithm and DP-MSNM algorithm on two data sets.

Results

Compare with DP-GMM and SAGMM. We implement SAGMM by applying the privacy preserving platform GUPT. The default values are the same as that of DP-MSNM. For the number of partitions, and the spherical covariance in SAGMM which is assumed known before the experiment. We set the number of partitions as 5. Different variances are applied and the best spherical covariance is chosen for each experiment. From Figs. 1 and 2 we observed that the performance of the three methods improve when B increases, DP-MSNM outperforms SAGMM and DPGMM. In terms of the shape of the curve in Fig. 1, our results are different from those of Wu et al.¹, mainly because we used a larger amount of data in the data set 3DRoad, and in order not to lose data information, we did not normalized each attribute of the data set.

We implement DPGMM by software R with the same environment as the DP-MSNM. From the Figs. 1 and 2 we find the DP-MSNM outperforms DPGMM.

Effect of number of skew-normal distribution (G). For AIS data set, as can be seen from Table 2, using DPGMM algorithm, the values of AIC and BIC are the minimum when $k = 3$. By using DP-MSNM algorithm, the values of AIC and BIC are the smallest when $k = 2$. For 3DRoad data set, as can be seen from Table 3, using

Figure 2. Effect of total privacy budget on 3DRoad data set.

Figure 3. Effect of G in AIS data set.

DPGMM algorithm, the values of AIC and BIC are the minimum when $k = 8$. By using DP-MSNM algorithm, the values of AIC and BIC are the smallest when $k = 7$.

These results are consistent with the results we can see from Figs. 3 and 4, because for AIS data, the log-likelihood mean of DPGMM algorithm at $k = 3$ is close to that of DP-MSNM algorithm (Fig. 3). A similar thing happened with 3DRoad data (Fig. 4).

So, we can conclude that the order of the component has great effect to our results.

Conclusions

In this paper, we proposed DP-MSNM, which is an effective and accurate privacy-preserving mechanism to estimate density. We added noise to original parameters and through post-processing step to ensure weight and covariance parameter have its intrinsic characteristics. In the future, we will extend the DP-MSNM model to high-dimensional data. We also plan to construct other Skew-family distributions to differential privacy.

Figure 4. Effect of G in 3DRoad data set.

k	2	3	4	5	6	7	8
DPGMM							
AIC	2217.618	2178.371	2184.419	2187.219	2186.657	2180.705	2188.74
BIC	2254.009	2234.611	2260.509	2283.159	2302.446	2316.344	2344.229
DP-MSNM							
AIC	2174.956	2186.836	2204.82	2205.515	2208.153	2195.302	2209.796
BIC	2224.58	2262.926	2307.377	2334.537	2363.642	2377.257	2418.217

Table 2. DP-SMSN algorithm in AIS data set. For each component $k(k = 2, \dots, 8)$, the Akaike (AIC) and Bayesian (BIC) information criteria, appear in two models.

k	2	3	4	5	6	7	8
DPGMM							
AIC	4,350,524	4,242,447	4,103,088	4,057,006	4,006,059	3,957,658	3,882,209
BIC	4,350,733	4,242,766	4,103,516	4,057,544	4,006,707	3,958,415	3,883,077
DP-MSNM							
AIC	4,194,876	4,148,930	4,139,086	3,995,200	3,941,379	3,898,330	3,904,047
BIC	4,195,150	4,149,348	4,139,646	3,995,903	3,942,225	3,899,318	3,905,178

Table 3. DP-SMSN algorithm in 3DRoad data set. For each component $k(k = 2, \dots, 8)$, the Akaike (AIC) and Bayesian (BIC) information criteria, appear in two models.

Received: 22 December 2020; Accepted: 10 May 2021
Published online: 26 May 2021

References

1. Wu, Y., Wu, Y. & Peng, H. Differentially private density estimation via gaussian mixtures model. In *IEEE/ACM 24th International Symposium* 1–6 (2016).
2. Kamath, G., Sheffet, O. & Singhal, V. Differentially private algorithms for learning mixtures of separated gaussians. In *NeurIPS* 1–62 (2019).
3. Truex, S., Liu, L. & Gursoy, M. Effects of differential privacy and data skewness on membership inference vulnerability. *arXiv.org* 1–16 (2019).
4. Dwork, C. & Aaron, R. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
5. Azzalini, A. & DallaValle, A. The multivariate skew-normal distribution. *Biometrika* **83**, 715–726 (1996).

6. Contreras-Reyes, J.E. & Cortes, D.D. Bounds on Renyi and Shannon entropies for finite mixtures of multivariate skew-normal distributions: application to swordfish (*Xiphias gladius linnaeus*). *Entropy* **18**, 382 (2016).
7. Cabral, C., Lachos, V. & Prates, M. Multivariate mixture modeling using skew-normal independent distributions. *Comput. Stat. Data Anal.* **56**, 126–142 (2012).
8. Nissim, K., Raskhodnikova, S. & Smith, A. Smooth sensitivity and sampling in private data analysis. *STOC* 75–84 (2007).
9. Zhang, J., Xiao, X. & Yang, Y. Privgene: Differentially private model fitting using genetic algorithms. In *SIGMOD* 665–676 (2013).
10. Mardia, K. V. Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530 (1970).
11. Higham, N. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **139**, 103–118 (1988).
12. Hay, M., Rastogi, V. & Miklau, G. Boosting the accuracy of differential private histograms through consistency. *PVLDB* **3**, 1021–1032 (2010).
13. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, 2020).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 11701071).

Author contributions

W.W. is the only author and wrote the manuscript text.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021