MCMC for Bayesian nonparametric mixture modeling under differential privacy

Beraha Mario
Department of Economics and Statistics
University of Torino, Torino, Italy
and
Stefano Favaro
Department of Economics and Statistics
University of Torino and Collegio Carlo Alberto, Torino, Italy
and
Vinayak Rao
Department of Statistics
Purdue University, West Lafayette, IN 47907

October 15, 2023

Abstract

Estimating the probability density of a population while preserving the privacy of individuals in that population is an important and challenging problem that has received considerable attention in recent years. While the previous literature focused on frequentist approaches, in this paper, we propose a Bayesian nonparametric mixture model under differential privacy (DP) and present two Markov chain Monte Carlo (MCMC) algorithms for posterior inference. One is a marginal approach, resembling Neal's algorithm 5 with a pseudo-marginal Metropolis-Hastings move, and the other is a conditional approach. Although our focus is primarily on local DP, we show that our MCMC algorithms can be easily extended to deal with global differential privacy mechanisms. Moreover, for certain classes of mechanisms and mixture kernels, we show how standard algorithms can be employed, resulting in substantial efficiency gains. Our approach is general and applicable to any mixture model and privacy mechanism. In several simulations and a real case study, we discuss the performance of our algorithms and evaluate different privacy mechanisms proposed in the frequentist literature.

Keywords: Dirichlet Process, Data augmentation, Pseudo-marginal MCMC

1 Introduction

Mixture models offer a natural and computationally convenient framework for density estimation and clustering (Frühwirth-Schnatter et al., 2019). Bayesian nonparametric (BNP) mixture models, introduced in the seminal work of Lo (1984), assume that each observation belongs to one of a potentially infinite number of groups or components, each modeled with parametric density function $f(\cdot \mid \phi)$ for some parameter ϕ. Formally, $n \geq 1$ observations are modeled as a random sample $Y = (Y_1, \ldots, Y_n)$ distributed as $Y_i \mid P \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \theta) P(\mathrm{d}\theta)$, with the mixing distribution P being an almost surely discrete random probability measure, i.e. $P(\cdot) = \sum_{h \geq 1} w_h \delta_{\phi_h}(\cdot)$, with the w_h's providing the relative prevalences of the mixture components labelled by the ϕ_h's. Mixture models are increasingly popular in applied contexts such as sociology (Land, 2001; Collins and Lanza, 2009), healthcare analysis (Singh and Ladusingh, 2010), multi-omics data (Wang et al., 2020), and many other contexts involving sensitive data (Schlattmann, 2009). These settings raise the need to balance learning the mixing distribution P with protecting individual's privacy.

In this paper, we consider the statistical problem of fitting a nonparametric Dirichlet process mixture model when one has access only to a differentially-private (DP) (Dwork, 2006) database Z, obtained from the original confidential database Y via some publicly known privatization mechanism. We contrast our problem with so-called algorithmic approaches (Dwork and Roth, 2014), whose goal is to propose randomized algorithms that preserve individual privacy. Algorithmic approaches for (Gaussian) mixture models under DP have been proposed starting from Nissim et al. (2007); see also Kamath et al. (2019) and the references therein. In the Bayesian framework, Dimitrakakis et al. (2017), Savitsky et al. (2022), and Hu et al. (2022) studied the privacy properties of (pseudo) posteriors and established conditions under which samples from the posterior or from the posterior predictive distribution yields a perturbed database Z with valid privacy guarantees. Bernstein and Sheldon (2018, 2019) focussed on likelihoods belonging to exponential family and linear regression, and showed that by perturbing sufficient statistics appropriately, and using a suitable asymptotic approximation, the posterior distribution given Z is easy to compute.

When it comes to estimation given privatized data, in the frequentist literature, the focus is usually on estimators for quantities of interest, or statistical tests based on Z, while also quantifying the loss of efficiency due to privacy in a minimax framework (see Wasserman and Zhou, 2010; Duchi et al., 2018, and the references therein). Then, one usually seeks for an optimal pair of estimator and perturbation mechanism that match the minimax rate. Closer to our work Karwa et al. (2015) and Ju et al. (2022) treat the Y_i's as missing variables and develop general variational Bayes and MCMC algorithms for posterior inference. One of our proposed samplers, the conditional sampler, is an instance of the algorithm of Ju et al. (2022), though our other sampler, the marginal sampler, takes a pseudo-marginal MCMC approach, allowing us to better trade off computation and mixing by introducing additional auxiliary variables.

1.1 Our contributions

In this paper, we investigate posterior inference in BNP mixture models under differential privacy. We start by focusing on the notion of *local DP* (Dwork, 2006), where the sanitized $Z = (Z_1, \ldots, Z_n)$ is obtained by perturbing each Y_i individually. Here, if the original dataset Y is modeled as a realization of a BNP mixture model, then, after marginalizing out the Y_i's, the Z_i's again follow a nonparametric mixture model, whose kernel in general intractable. We propose two MCMC sampling schemes for posterior density estimation and clustering given Z. The first is a *marginal* algorithm, where we marginalize out the infinite-dimensional measure P, and can be considered a version of Algorithm 5 in Neal (2000) for private data. Here, our approach has the flavor of the pseudo-marginal MCMC approach (Andrieu and Roberts, 2009), with the latent Y_i's introduced back into the MCMC state as auxiliary variables to deal with intractable Metropolis-Hastings probabilities. The second algorithm is a *conditional* algorithm, and it can be considered as a private version of the slice-sampler of Kalli et al. (2011). In this setting, the Y_i's are part of the state of the MCMC and we adopt the same blocked-Gibbs sampling strategy as Ju et al. (2022), alternately sampling P given the Y_i's and the Y_i's given P and the Z_i's. For specific local DP channels (including the Gaussian mechanism), we show how one can combine the privacy mechanism with a carefully chosen mixture kernel and base measure, allowing us to analytically marginalize with respect to the Y_i's and obtain a conjugate mixture model of the Z_i's. Then, the posterior distribution can be sampled using efficient algorithms such as Algorithm 3 in Neal (2000) or the split-merge algorithm in Jain and Neal (2004). We also extend our framework to the setting of *global DP*, where the sanitized database $W = (W_1, \dots, W_k)$ is obtained by perturbing the sample $Y = (Y_1, \dots, Y_n)$ jointly, and show that the two algorithms proposed can be adapted in a straightforward way to this setting.

For concreteness, we consider BNP mixture models based of a Dirichlet process prior for P (Ferguson, 1973; Lo, 1984), such that the random probabilities w_h's follows the stick-breaking constructions (Sethuraman, 1994). However, the ideas developed in this paper can be easily extended to more general prior distributions, under which marginal and conditional MCMC sampling schemes for BNP mixture models have been developed in the non-private setting (Griffin and Walker, 2011; Barrios et al., 2013; Favaro and Teh, 2013; Favaro and Walker, 2013; Lomeli et al., 2017; Miller and Harrison, 2018; Argiento and De Iorio, 2022). See Wade (2023) for a recent review. Our MCMC sampling schemes do not depend on a specific choice of the DP channel used to produce the sanitized database Z, and we consider popular channels such as adding Laplace and Gaussian noise, wavelet-based perturbation (Duchi et al., 2018; Butucea et al., 2020), and sampling from a perturbed histograms (Wasserman and Zhou, 2010). We find that for moderate sample sizes, the wavelet-based perturbation method and the approach of sampling from a perturbed histograms are sensitive to the hyper-parameters, and often lead to poor density estimates. Instead, the approach of adding noise tend produce sensible posterior inferences.

The paper is structured as follows. In Section 2 we present an overview of DP, focussing on some definitions of local DP and DP channels. Section 3 contains the main results of the paper, introducing the private Neal 5 algorithm, the private slice-sampling and some variations thereof. In Section 4 we present a simulation study comparing the proposed algorithms over different choices of the DP channel, and Section 6 contains an application to real data on blood donors at the Milano Department of the Associazione Volontari Italiani del Sangue (AVIS). In Section 7 we discuss the limitations of the proposed approach, and present some further research directions. Our algorithms have been implemented in `C++` as part of the `BayesMix` library (Beraha et al., 2022) and interested users can easily extend our examples to different prior distributions, mixture kernels and/or privacy mechanisms.

2 Local Differential Privacy

We present a brief overview of DP, focusing on *local* and *non-interactive* DP, arguably the most popular notion of privacy in the statistical framework of density estimation (Butucea et al., 2020; Kroll, 2021; Butucea et al., 2022; Sart, 2023). For any $n \geq 1$, denote by $Y_1, \ldots, Y_n$ the confidential observations, defined on a probability space $(\Omega, \mathcal{A}, \Pr)$, and taking values in a general measurable (Polish) space $(\mathbb{Y}, \mathcal{Y})$. Under local DP, the sample $Y = (Y_1, \ldots, Y_n)$ is randomly perturbed to produce a *sanitized* database $Z = (Z_1, \ldots, Z_n) \in \mathbb{Z}^n$, with the perturbation being encoded by a collection of *channels* $Q_1, \ldots, Q_n$, where $Q_i : \mathbb{Z} \times \mathbb{Y} \to \mathbb{R}_+$ is such that, for any $y \in \mathbb{Y}$, $Q_i(\cdot \mid y) := Q_i(\cdot, y)$ is a conditional distribution with density function over $\mathbb{Z}$, $i = 1, \ldots, n$. The *sanitized* database $Z \sim \otimes_{i=1}^n Q_i(\cdot \mid Y_i)$ is released for public use. Note that local DP does not require a trusted data holder even for collecting the data, since each individual can perturb their datum Y_i independently of all the other individuals. Depending on the definition of local DP, the channels Q_i's need to satisfy some properties. We recall the most popular definitions below.

In her seminal work, Dwork (2006) introduced the framework of ε-DP, where

$$\sup_{S \in \mathcal{Z}} \frac{Q_i(S \mid Y_i = y)}{Q_i(S \mid Y_i = y')} \leq e^{\varepsilon} \quad \text{for any } y, y' \in \mathbb{Y}. \tag{1}$$

The parameter ϵ controls the *privacy loss* that an individual can suffer when analyzing their data. See Bun and Steinke (2016) for further discussions. As shown in Wasserman and Zhou (2010), ε-DP also implies an upper bound on the power of the statistical test aimed at identifying the presence of a particular individual in the unobserved original sample. Among several relaxations of ε-DP, the most popular is (ε, δ)-DP, requiring:

$$Q_i(S \mid Y_i = y) \leq e^{\varepsilon} Q_i(S \mid Y_i = y') + \delta, \quad \text{for all measurable } S, \text{ and for any } y, y' \in \mathbb{Y}. \tag{2}$$

According to this definition, DP holds *most of the times*, with the probability that individuals suffer a privacy loss greater than ε bounded by δ. Another relaxation, referred to as concentrated DP (Dwork and Rothblum, 2016; Bun and Steinke, 2016), has gained significant interest, to the extent of being adopted by the US Census Bureau (Garfinkel, 2022). We say that a channel Q_i is ρ zero concentrated DP (ρ-zCDP) if $D_\alpha\left(Q_i(\cdot \mid y) || Q_i(\cdot \mid y')\right) \le \rho\alpha \quad \text{for any } y, y' \in \mathbb{Y},$ where D_α is the α-Rényi divergence. We refer to Dwork and Rothblum (2016) and Bun and Steinke (2016) for additional details.

2.1 Local DP: channels

We present some popular choices of the local DP perturbation channel Q we will consider. The most natural one adds independent and identically distributed zero-mean noise:

$$Z_i = Y_i + \gamma_i, \qquad \gamma_i \overset{\text{iid}}{\sim} f_\gamma.$$

To match any of the notions of DP discussed above, we must require that the Y_i's have a bounded support. In particular, consider unidimensional data on $\mathbb{Y}$, and denote by $\Delta = \sup_{y,y'} |y - y'|$ the diameter of $\mathbb{Y}$. If $\gamma_i \sim \mathcal{L}(0, \Delta/\varepsilon)$, with $\mathcal{L}$ the Laplace distribution, then Q_i satisfies ε-DP, and hence (ε, δ)-DP for any choice of δ, as well as $\frac{1}{2}\varepsilon^2$-zCDP (Bun and Steinke, 2016, Propositon 1.4). The use of the Gaussian distribution is also popular, that is $\gamma_i \sim \mathcal{N}(0, \eta^2)$, with $\mathcal{N}$ being the Gaussian distribution. The choice of the Gaussian distribution is known to satisfy (ε, δ)-DP (Dwork and Roth, 2014; Balle and Wang, 2018), for suitable values of ε and δ, and ρ-zCDP, though it does not satisfy ε-DP.

Other DP mechanisms, not based on adding random noise, have been proposed in the literature. The minimax framework of Duchi et al. (2018) seeks the optimal pair of perturbation and estimator for a given problem, thus making the optimal estimator and perturbation mechanism related (Duchi et al., 2018; Butucea et al., 2020, 2022; Kroll, 2021; Sart, 2023). We review the wavelet-based perturbation mechanism of Butucea et al. (2020) in greater detail in Appendix A. Later, we will focus on one particular example of the this based on the Haar basis, whereby $Z_i = (Z_{i,j,k})$ for $j = 0, \dots, J$ and $k \le 2^j$ is distributed as

$$Z_{i,j,k} \mid Y_i \sim \mathcal{L}(\psi_{j,k}(Y_i), s), \tag{3}$$

where $\psi_{j,k}(x) = 2^{j/2}\left(\mathrm{I}_{[0,1/2)}(2^j x - k) - \mathrm{I}_{[1/2,1)}(2^j x - k)\right)$, $s = \frac{12}{\varepsilon}\frac{\sqrt{2}}{\sqrt{2}-1}2^{J/2}$. Here J is a parameter that trades-off the amount of information released and the amount of noise added. This perturbation guarantees ε-DP (Butucea et al., 2020, Equation 3.1).

3 BNP mixture modeling under local DP

We assume a BNP mixture model for the confidential Y_i's. Taking into account the local DP channels Q_i's, we write the following hierarchical model for sanitized observations:

$$\begin{aligned} Z_i \mid Y_i &\overset{\text{ind}}{\sim} Q_i(\cdot \mid Y_i), && i = 1, \dots, n \\ Y_i \mid \theta_i &\overset{\text{ind}}{\sim} f(\cdot \mid \theta_i), \quad \theta_i \mid P \overset{\text{iid}}{\sim} P, && i = 1, \dots, n \\ P &\sim \mathscr{Q}. \end{aligned} \tag{4}$$

Here, $f(\cdot \mid \theta)$ is a probability density kernel indexed by a parameter $\theta \in \Theta$, and P is an almost surely discrete random probability measure over Θ, with its law $\mathscr{Q}$ being the Dirichlet process with concentration parameter α and base measure G_0 (Ferguson, 1973). Thus, $P = \sum_{h \geq 1} w_h \delta_{\phi_h}$ where $(\phi_h)_{h \geq 1}$ is a collection of i.i.d. random variables from G_0 and $(w_h)_{h \geq 1}$ is a sequence of positive random weights summing to one, obtained, for instance, via the so-called stick breaking process: $w_h = \nu_h \prod_{j<h}(1 - \nu_h)$, where $\nu_h \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$.

We propose two MCMC algorithms for posterior sampling from this model. The first is a marginal MCMC sampler that is an instance of the so-called pseudo-marginal MCMC, while the second is a conditional MCMC sampler based on missing data imputation.

3.1 Marginal MCMC sampling: private Neal 5 algorithm

Starting with (4), we operate two marginalizations: first we integrate out P by relying on the well-known generalized Pólya urn characterization of the Dirichlet measure (Ferguson, 1973), and then marginalize out the Y_i's, To this end, letting $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_k^*)$ be the unique values in $\theta_1, \dots, \theta_n$ and denoting by $n_h = \sum_{i=1}^n \mathrm{I}[\theta_i = \theta_h^*]$, we obtain

$$Z_i \mid \theta_i \overset{\text{ind}}{\sim} g(\cdot \mid \theta_i) = \int_{\mathbb{Y}} Q_i(\cdot \mid y) f(y \mid \theta_i) \mathrm{d}y, \qquad i = 1, \dots, n \tag{5}$$

$$\Pr(\boldsymbol{\theta} \in \mathrm{d}\boldsymbol{\theta}) = \frac{\alpha^k}{(\alpha)_{(k)}} \prod_{h=1}^k (n_h - 1)! G_0(\mathrm{d}\theta_h^*). \tag{6}$$

The main challenge in sampling from the conditional distribution (5) is that $g(\cdot \mid \theta)$ is intractable, in the sense that it cannot be evaluated analytically. Beraha and Corradin (2021) propose a general strategy to deal with intractable kernels in the context of BNP mixtures, proposing the use of approximate Bayesian computation. However, this might be inefficient for large datasets. Instead, we develop an exact pseudomarginal MCMC sampler (Andrieu and Roberts, 2009) using an unbiased estimator $\hat{g}$ for the mixture kernel.

From (5), it is clear that the conditional distribution of θ_i is

$$\theta_i \mid \theta_{-i}, Z \sim b\pi(\theta_i \mid \theta_{-i}) g(Z_i \mid \theta_i), \tag{7}$$

where θ_{-i} is $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, b is a normalizing constant, and $\pi(\theta_i \mid \theta_{-i}) := \sum_{j=1}^{n-1} \frac{1}{\alpha+n-1} \delta_{\theta_j}(\cdot) + \frac{\alpha}{\alpha+n-1} G_0(\cdot)$ is the conditional prior of θ_i given θ_{-i}. As in Algorithm 5 of Neal (2000), we can sample from the conditional distribution (7) by a Metropolis-Hastings step, where the proposal is $\theta_i' \sim \pi(\cdot \mid \theta_{-i})$, and the acceptance rate is

$$\min\left\{1; \frac{\pi(\theta_i \mid \theta_{-i}) g(Z_i \mid \theta_i)}{\pi(\theta_i' \mid \theta_{-i}) g(Z_i \mid \theta_i')} \frac{\pi(\theta_i' \mid \theta_{-i})}{\pi(\theta_i \mid \theta_{-i})}\right\} = \min\left\{1; \frac{g(Z_i \mid \theta_i)}{g(Z_i \mid \theta_i')}\right\}.$$

Since $g(\cdot \mid \theta)$ cannot be evaluated analytically, we employ a pseudo-marginal MCMC step that only requires an unbiased estimator for $g(\cdot \mid \theta)$. Assuming without loss of generality that Q_i has a density q_i with respect to some dominating measure, a natural estimator is

$$\hat{g}_i := \hat{g}(Z_i \mid \theta_i) = \frac{1}{m} \sum_{j=1}^{m} q_i(Z_i \mid \tilde{Y}_{i,j}), \qquad \tilde{Y}_{i,j} \sim f(\cdot \mid \theta_i). \tag{8}$$

The state space of the resulting sampler consists then of $(\theta_i, \hat{g}_i)$. Each iteration, and for each i, the sampler proposes new values $(\theta_i', \hat{g}_i')$ from equations (7) and (8) and accepts them with probability $\min\{1; \hat{g}_i'/\hat{g}_i\}$. Following Algorithm 2 in Neal (2000), rather that storing the θ_i's, one can also just store their unique values θ_h^* and cluster allocations c_i, such that $\theta_i = \theta_h^*$ if and only if $c_i = h$. In this setting, it is possible to additionally update the θ^* directly to improve mixing. If this is needed, we must include the $\tilde{Y}_{i,j}$ in our state space (rather than just $\hat{g}$), with $p(\theta_h^* \mid \cdots) \propto \prod_{i:c_i=h} f(\{\tilde{Y}_{i,j}\} \mid \theta_h^*) G_0(\mathrm{d}\theta_h^*)$. Algorithm 1 details the pseudocode of our marginal sampling.

For the setting of ε-DP, we can easily extend the results in Ju et al. (2022) to our pseudo-marginal move obtaining lower bound for the acceptance rate:

Proposition 1. *If the channel q_i satisfies ε-DP as in* (1) *then* $\min\{1; \hat{g}_i'/\hat{g}_i\} \geq e^{-\varepsilon}$

Proof. The proof follows by observing that for any pair $(\tilde{Y}'_{i,j}, \tilde{Y}_{i,j})$, ε-DP imples $q_i(Z_i \mid \tilde{Y}'_{i,j}) > e^{-\varepsilon} q_i(Z_i \mid \tilde{Y}_{i,j})$, so that $\sum_{j=1}^m (q_i(Z_i \mid \tilde{Y}'_{i,j}) - e^{-\varepsilon} q_i(Z_i \mid \tilde{Y}_{i,j})) > 0$. Then, noting that $\frac{\hat{g}'_i}{\hat{g}_i} = \frac{\sum_{j=1}^m q_i(Z_i \mid \tilde{Y}'_{i,j})}{\sum_{j=1}^m q_i(Z_i \mid \tilde{Y}_{i,j})}$, we immediately have our result. □

In practice, we observe much larger acceptance rates than this theoretical lower-bound. Moreover, we also empirically observe that for (ε, δ)-DP channels and ρ-ZCDP channels the acceptance rates are satisfactory, e.g., larger than 10%.

Algorithm 1. Private Neal 5 algorithm

Input: Sanitized data $Z_1, \dots, Z_n$.
Initialize: $c_1, \dots, c_n$, the unique values $\theta^*_1, \dots, \theta^*_k$, and the Y_i's
for *each MCMC iteration* **do**
  for $i = 1, \dots, n$ **do**
    Sample $c'_i \in \{1, \dots, k+1\}$ with $\Pr(c'_i = h) \propto \begin{cases} n_h^{-i} & \text{if } 1 \le h \le k \\ \alpha & \text{if } h = k+1 \end{cases}$,
    where n_h^{-i} is n_h computed when observation i is removed from the state.
    if $c'_i = k+1$ **then**
      Sample $\theta^*_{k+1} \sim G_0$
    Sample $\{\tilde{Y}'_{i,j}\} \sim f(\cdot \mid \theta^*_{c'_i})$.
    With prob. $\min\{1; \hat{g}'_i / \hat{g}_i\}$ set $(c_i, \{\tilde{Y}_{i,j}\}) = (c'_i, \{\tilde{Y}'_{i,j}\})$, else revert all changes.
  for $h = 1, \dots, k$ **do**
    Sample θ^*_h from the density $p(\theta^*_h \mid \cdots) \propto \prod_{i: c_i = h} f(\{\tilde{Y}_{i,j}\} \mid \theta^*_h) G_0(\mathrm{d}\theta^*_h)$
end

3.2 Conditional MCMC sampling: private slice-sampling

The development of a private version of the slice-sampling of Kalli et al. (2011) is straightforward. From the model (4) with a Dirichlet measure $\mathscr{Q}$, we write the joint distribution

$$\Pr(Z \in \mathrm{d}z, Y \in \mathrm{d}y, \theta \in \mathrm{d}\theta, P \in \mathrm{d}\tilde{p}) = \left\{ \prod_{i=1}^{n} q_i(z_i \mid y_i)\mathrm{d}z_i\, f(y_i \mid \theta_i)\mathrm{d}y_i\, \tilde{p}(\mathrm{d}\theta_i) \right\} \mathscr{Q}(\mathrm{d}\tilde{p}), \quad (9)$$

from which all the full-conditional distributions for the random variables involved can be derived. Note that (P, θ) is independent of Z given Y. Therefore, with respect to the original slice-sampling of Kalli et al. (2011), we must only additionally sample the Y_i's from their full-conditional distributions. From (9) the full-conditional density of Y_i is

$$f_{Y_i}(y \mid \mathrm{rest}) \propto q_i(z_i \mid y) f(y \mid \theta_i),$$

which can be easily sampled by a standard Metropolis-Hastings step. Following Ju et al. (2022), we used $f(\cdot \mid \theta_i)$ as the proposal proposal distribution, giving in the same mixing guarantee of Proposition 1 for ε-DP, and resulting in good empirical performance. Having sampled the Y_i's, the updates of P and θ proceed as in Kalli et al. (2011), disregarding the Z_i's. Other conditional algorithms like Papaspiliopoulos and Roberts (2008), Ishwaran and James (2001), Canale et al. (2022), or Arbel and Prünster (2017) may also be considered.

3.3 Algorithms for tractable marginalized kernels

Whenever the marginalized kernel $g_i(z \mid \theta) = \int_{\mathbb{Y}} q_i(z \mid y) f(y \mid \theta)\mathrm{d}y$ has a tractable analytic form, we can integrate out the Y_i's and sample from the posterior distribution of the (marginal) model (5), or its equivalent formulation when P is not marginalized out, by using standard algorithms. This is the case, for instance, of a Gaussian mixture model for the Y_i's coupled with the Gaussian DP mechanism, or the case of the Binomial DP mechanism in Chen et al. (2022) paired with a mixture of Beta kernels. We refer to Dimitrakakis et al. (2017) for other examples.

As an example, we consider the case of Gaussian mixtures with a Gaussian DP mechanism for univariate data. After marginalizing out Y_i's, we have that the kernel g is

$$g(\cdot \mid \theta_i) = \int_{\mathbb{R} \times \mathbb{R}_+} \mathcal{N}(\cdot \mid y, \eta^2) \mathcal{N}(y \mid \mu_i, \sigma_i^2)\mathrm{d}y = \mathcal{N}(\cdot \mid \mu_i, \eta^2 + \sigma_i^2).$$

Accordingly, the Bayesian model for the Z_i's, with the Y_i's marginalized out, is

$$\begin{aligned} Z_i \mid \mu_i, \sigma_i^2 &\overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \eta^2 + \sigma_i^2) \\ \mu_i, \sigma_i^2 \mid P &\overset{\text{iid}}{\sim} P, \quad P \sim DP(\alpha G_0). \end{aligned} \quad (10)$$

The posterior distribution under (10) can be easily sampled. Note that standard choices for G_0, such as the Normal-Inverse-Gamma distribution, are not conjugate to $g(\cdot \mid \cdot)$, and one must resort to, e.g., Algorithm 8 in Neal (2000). However, by means of a careful choice of the prior distribution, we show it is possible to marginalize out the unique values (μ_h^*, τ_h^*), obtaining an MCMC algorithm that samples only the cluster allocations. Setting $\tau_i^2 = \eta^2 + \sigma_i^2$, which is a random variable supported on $[\eta^2, +\infty)$, we have

$$\begin{aligned} Z_i \mid \mu_i, \tau_i^2 &\stackrel{\text{ind}}{\sim} \mathcal{N}(Y_i, \tau_i^2) \\ \mu_i, \tau_i^2 \mid P &\stackrel{\text{iid}}{\sim} P, \quad P \sim DP(\alpha G_0). \end{aligned} \tag{11}$$

By choosing $G_0(\mathrm{d}\mu, \mathrm{d}\tau^2) = \mathcal{N}(\mu \mid \mu_0, \tau^2/\lambda) IG(\tau^2 \mid a, b) \mathrm{I}_{[\eta^2, +\infty)}(\tau^2) \mathrm{d}\mu \, \mathrm{d}\tau^2$ we obtain a conjugate model. Now (11) can be sampled by means of Algorithm 3 in Neal (2000), or by means of the split and merge algorithm in Jain and Neal (2004), which usually leads to a better mixing. Critical to this derivation is (i) the choice of an appropriate perturbation channel that allows for analytic convolution with respect to the mixture kernel, and (ii) the choice of G_0 that takes into account the mixture kernel as well as the privacy constraint. The main difference with the standard Gaussian mixture with a Normal-Inverse-Gamma centering measure is that the marginal distribution of data inside a cluster is not a Student t distribution. Nonetheless its density function can be computed by Bayes' formula.

4 Numerical Illustrations

We present two simulation studies to assess the performance of our marginal and conditional sampling schemes across different privacy mechanisms, privacy levels and sample sizes. To evaluate samplers, we monitor the effective sample size (ESS) of the number of clusters, i.e. the number of unique values across the c_i's. We note that when varying the privacy level ε, the posterior distribution is itself changing, and it makes little sense to compare effective sample sizes across different posteriors. In this case, we monitor the acceptance rate of the update of the Y_i's, allowing us to evaluate how tight the bound in Proposition 1 is. We note this is not directly comparable across samplers, since for the marginal sampling scheme the Y_i's are updated together with the c_i's, while for the conditional sampling scheme these updates are done separately. Unless otherwise specified, chains are ran for 10,000 iterations, discarding the first 5,000 as burn-in steps, and we report the median over 50 independent repetitions of the acceptance rates and ESS.

4.1 Private Neal 5 algorithm versus private slice-sampling

We simulate $n = 50, 200, 500, 1000$ data points from a mixture of three equally weighted truncated Gaussian distributions on the interval $I = [-10, 10]$, with means $-5, 0, 5$ and variance equal to one. We sanitize data through the Laplace mechanism, for $\varepsilon = 1.0, 2, 5, 10, 50$. We model the confidential data with a location-scale mixture of Gaussians, i.e. $\theta_i = (\mu_i, \sigma_i^2)$ and $f(y \mid \theta_i)$ is the Gaussian density function with mean μ_i and variance σ_i^2. The base measure $G_0(\mu, \sigma^2)$ is assumed to be the Normal-Inverse-Gamma distribution $\mathcal{N}(\mu \mid 0, 10\sigma^2) \times IG(\sigma^2 \mid 3, 3)$. Note that despite the confidential data being compactly supported, for simplicity, we assume a mixture model with support on the whole set $\mathbb{R}$.

We compare three proposed algorithms, the private Neal 5 algorithm with $m = 1$ auxiliary variables, with $m = 5$ auxiliary variables, and the private slice-sampler. Figure 1 shows the posterior density estimate for the Y_i's when $n = 200$, that is the posterior expectation of $\int f(\cdot \mid \theta) P(\mathrm{d}\theta)$. See Figure 9 in the Appendix for other values of n. We note how for strong privacy levels, such as $\varepsilon = 1, 2$, the density estimate is poor, even with $n = 1,000$ observations. For looser privacy levels, such as $\varepsilon = 50$, even $n = 50$ observations are sufficient for accurate density estimates. As a general observation, both the private Neal 5 algorithm and the private slice-sampling agree on the density estimate, with some slight differences when $n = 1000$ and $\varepsilon = 5, 10$ when mixing is more challenging. Moreover, we see the marginal sampler performs slightly better when $m = 5$, both in terms of mixing and density estimate. However this increases the computational cost by almost a factor m. We arrive at similar conclusions when we focus on clustering performance, as measured by the Adjusted Rand Index (ARI) between the true and estimated clustering structure of the Y_i's (see Figure 10 in the appendix).

Figure 2 shows the median acceptance rate and ESS across the 50 replicates. As expected, the acceptance rates decrease with ε, but are still significantly larger than the

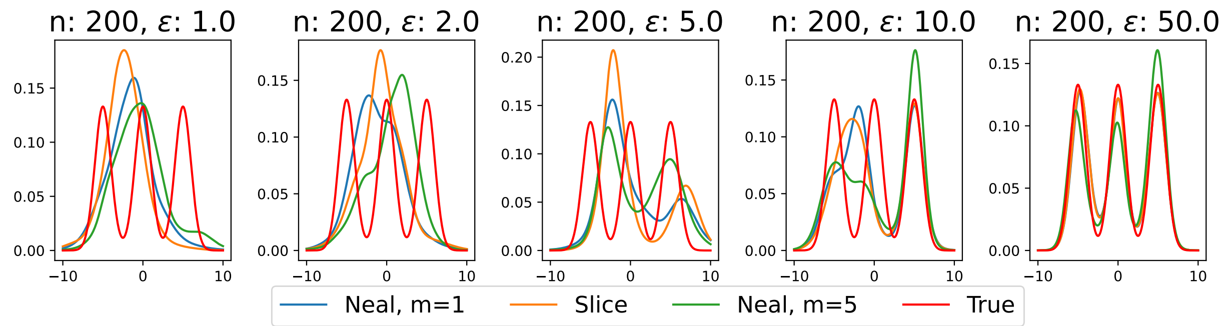

Figure 1: Density estimate for different privacy levels ε when $n = 200$ (Section 4.1).

theoretical lower bound $\exp(-\varepsilon)$ in Proposition 1. For the private Neal 5 algorithm, the ESS for the number of clusters also decreases with ε, since the update of the cluster allocations is intimately tied to the update of the private observation. On the other hand, for the private slice-sampling, the ESS increases with ε, likely due to the decoupling of the sampling of the private data from the remaining of the parameters in the MCMC sampling. We also see that for lower values of ε, the private Neal 5 algorithm yields higher ESSs than the private slice-sampling. The runtimes are essentially unchanged when varying ε, and comparable for Neal 5 algorithm and the slice-sampler for all settings except when $n = 1000$, when the slice-sampler is approximately ten times faster. The slice-sampler could be further accelerated by using parallel computations, since all the updates are easily parallelizable, in contrast to Algorithm 1 which is inherently sequential. All things being equal, we recommend the marginal sampler for high-privacy settings, and the conditional sampler when only small amounts of noise are added.

4.2 Laplace mechanism versus wavelet-based mechanism

We simulate $n = 250$ data points from a mixture of three equally weighted Beta distributions with parameters $(5, 50)$, $(50, 50)$ and $(50, 5)$, respectively. We assume a Dirichlet process mixture of Beta kernels with parameters $a > 0, b > 0$ for the Y_i's, i.e. $f(\cdot \mid \theta) = \text{Beta}(\cdot \mid a, b)$, and set the base measure as $G_0(\mathrm{d}a, \mathrm{d}b) = \text{Gamma}(\mathrm{d}a \mid 2, 2)\text{Gamma}(\mathrm{d}b \mid 2, 2)$. See Appendix B for further details. Below, we report posterior inference obtained using the private Neal 5 algorithm, though similar results hold for the private slice-sampler.

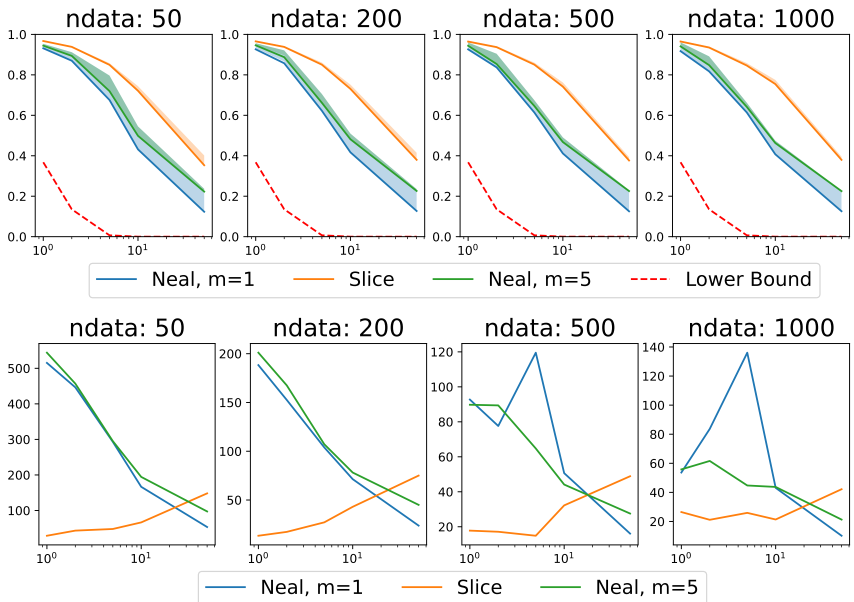

Figure 2: Top: acceptance rate as a function of the privacy level (the dashed line is the theoretical lower bound). Bottom: effective sample size for the number of clusters.

We consider privacy levels $\varepsilon = 2, 10, 50, 100, 250$ and, when adopting the wavelet-based mechanism, we consider $J = 2, 4, 6$. Figure 3 reports the posterior estimates of the density function, and Figure 12 in the Appendix the corresponding acceptance rates. We see that when $J = 2$ the density estimate is rather poor: for $\varepsilon = 2, 10, 50$ the estimated density does not capture the three modes and pushes all the mass to the boundaries of the interval, when $\varepsilon = 50$ it captures the mode in 1/2 but still fails to properly capture the boundaries. When $J = 6$, the acceptance rates are rather low when $\varepsilon \geq 100$. The density estimates are poor as well: for all values of ε the estimated density is very different from the true one. When $J = 4$ instead, the estimated density matches the true one for $\varepsilon \geq 50$. Similarly, when using the Laplace mechanism, the estimated densities match the true one for $\varepsilon \geq 50$.

4.3 Private Neal 5 versus non-private (marginal) algorithms

Consider the BNP mixture model (4) where $Q_i(\cdot \mid Y_i) = \mathcal{N}(\cdot \mid Y_i, \eta^2)$. Following Dwork and Roth (2014) we set set $\eta = \sqrt{2 \log \frac{1.25}{\delta}} \frac{\Delta}{\varepsilon}$, but other choices can be considered to ensure (ε, δ)-DP (Balle and Wang, 2018; Zhao et al., 2019). We consider the same data generating process as in Section 4.1 and let $\delta = \{0.01, 0.1, 0.25\}$, $\varepsilon = \{5, 10, 25, 50\}$. For smaller values of ε, the variance η^2 is extremely large (e.g., when $\varepsilon = 0.5, \delta = 0.01$, $\eta^2 > 3800$) thus making density estimation essentially impossible. We simulate $n = 250$ observations and

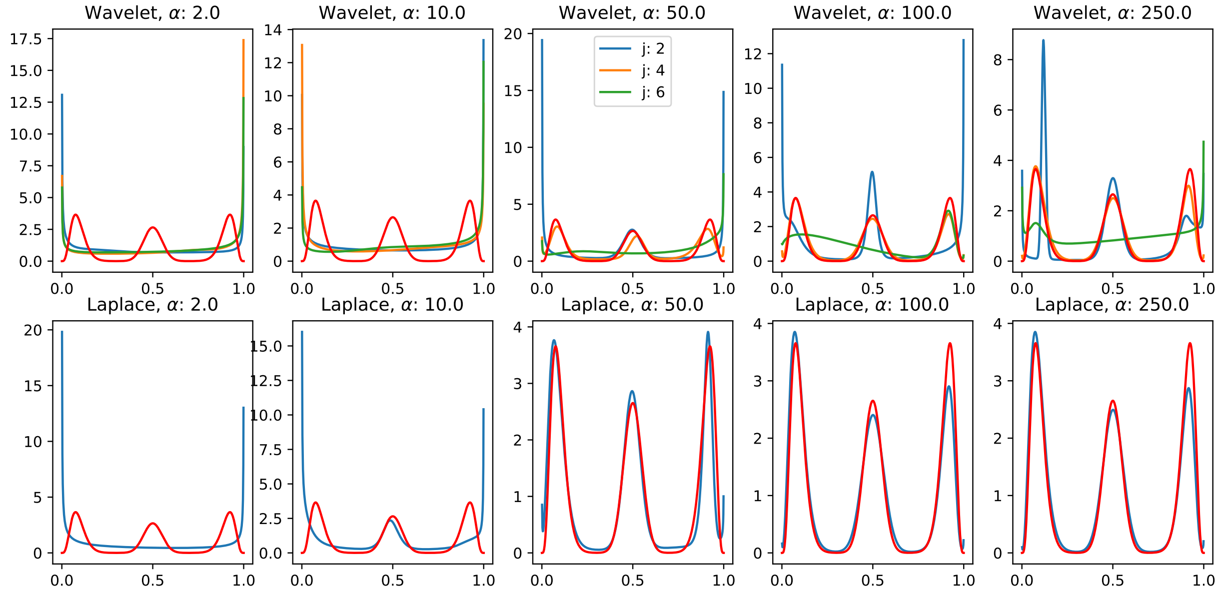

Figure 3: Density estimate for the simulated example in Section 4.2. Top row: wavelet privacy mechanism. Bottom row: Laplace privacy mechanism.

compute posterior inference under model (11) using either Neal's algorithm 2 or algorithm 3, which marginalizes w.r.t. the unique values (μ_h^*, τ_h^*), and under model (5) using the private Neal 5 algorithm with $m = 1$.

Figure 4 shows the posterior density estimates for $\delta = 0.1$, see Figure 13 in the Appendix for the other cases. For the privacy level $\varepsilon = 50$, the three marginal algorithms produce essentially identical density estimates. Instead, when $\varepsilon = 25$ and $\delta \leq 0.1$, the private Neal 5 algorithm estimates a unimodal density, while Neal 2 and Neal 3 algorithms estimate a bimodal density closer to the true data generating process. For smaller values of ε all the algorithms produces rather poor posterior estimates. Figure 5 instead shows the effective sample sizes of the number of clusters. Neal 2 and Neal 3 algorithms outperform the private Neal 5 algorithm, and Neal 3 algorithm is always better than Neal 2 algorithm. Such a behaviour was somehow expected, since these algorithms marginalize over the Y_i's and Neal 3 algorithm further marginalizes over the unique values. Moreover, while the effective sample size for the private Neal 5 algorithm decreases with the privacy level ε, as already observed in Section 4.1, this does not happen for Neal 2 and Neal 3 algorithms.

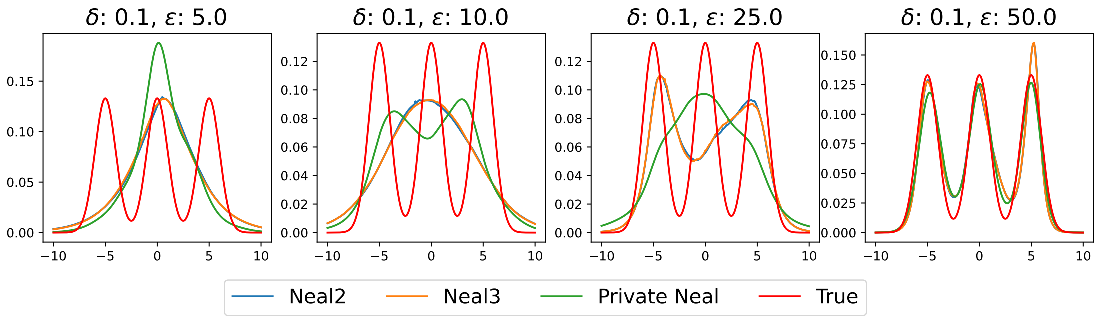

Figure 4: Density estimate for the example in Section 4.3 for different ε and $\delta = 0.1$.

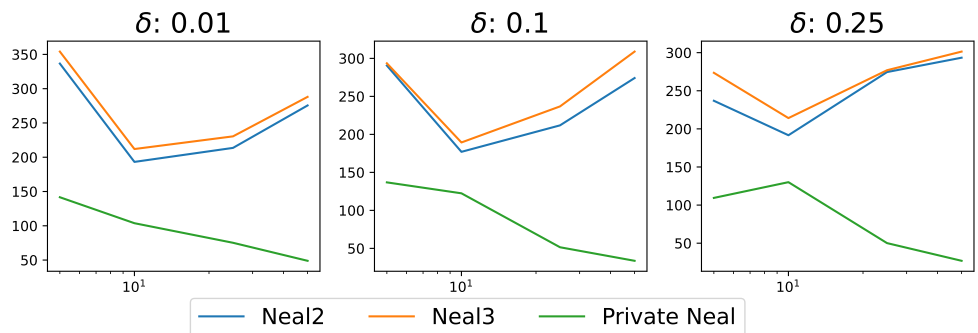

Figure 5: Effective sample size of the number of clusters for the example in Section 4.3

5 BNP mixture modeling under global DP

In this section, we extend of our framework to the setting of global DP, whereby the privacy mechanism is not applied observation-wise but to the whole dataset $Y = (Y_1, \dots, Y_n)$. Y is now stored with a trusted data holder, who then releases the sanitized $W \sim Q(\cdot \mid Y)$ where $Q(\cdot \mid \cdot)$ is the privacy channel. Under ε-DP, Q must satisfy

$$\sup_A \sup_{y,y' : d_H(y,y')=1} \frac{Q(W \in A \mid Y = y)}{Q(W \in A \mid Y = y')} \leq e^{\varepsilon}$$

where the first supremum is over all measurable sets and d_H is the Hamming distance, i.e. $d(y, y') = |\{i : y_i \neq y'_i\}|$. Extensions to (ε, δ)-DP and ρ-zCDP are straightforward.

A few global privacy mechanisms for density estimation have been proposed in Wasserman and Zhou (2010). We consider the "sampling from a smoothed histogram" mechanism, for $Y_i \in [0, 1]$ which releases an i.i.d. sample $W = (W_1, \dots, W_k) \in \mathbb{Y}^k$ from the density

$$f_{m,\delta}(x) := (1 - \delta) \sum_{j=1}^{m} \frac{c_j}{n} \mathrm{I}_{B_j}(x) + \delta, \tag{12}$$

where $\{B_1, \dots, B_m\}$ is a partition of $[0,1]$, $c_j = \sum_{i=1}^n I(Y_i \in B_j)$, and n, m, δ and ε satisfy $k \log\left(\frac{1-\delta}{\delta}\frac{m}{n} - 1\right) \leq \varepsilon$. Other alternatives include directly releasing a perturbed histogram as proposed in Dwork (2006), or the projection estimator in Lalanne et al. (2023). Importantly, these perturbation mechanisms yield a channel $Q(\cdot \mid Y)$ whose density is easily computable. On the other hand, the "sampling from a perturbed mechanism" or the mechanism based on the truncated perturbed Fourier expansion proposed in Wasserman and Zhou (2010) yield to channels $Q(\cdot \mid Y)$ with intractable densities which makes posterior inference cumbersome.

Under the global DP setting for density estimation, the BNP mixture model is essentially identical to (4), where we replace the first line with $W \mid Y \sim Q(\cdot \mid Y)$. A significant difference between the local DP setting and global DP setting is that global DP does not typically allow sensible inferences on the latent clustering, since W does not preserve information about individual observations.

5.1 Marginal and conditional MCMC sampling

We consider the joint distribution of data and parameters under global DP setting:

$$\Pr(W \in \mathrm{d}w, Y \in \mathrm{d}y, \theta \in \mathrm{d}\theta, P \in \mathrm{d}\tilde{p}) = Q(\mathrm{d}w \mid Y) \left\{ \prod_{i=1}^n f(y_i \mid \theta_i) \mathrm{d}y_i \, \tilde{p}(\mathrm{d}\theta_i) \right\} \mathscr{Q}(\mathrm{d}\tilde{p}). \quad (13)$$

It is straightforward to extend the private slice-sampling to the global DP setting. The only difference from the local DP setting is that the conditional distribution of Y_i is

$$f_{Y_i}(y \mid \mathrm{rest}) \propto Q(\mathrm{d}w \mid Y) f(y_i \mid \theta_i),$$

from which we can sample by a standard Metropolis-Hastings step with $f(\cdot \mid \theta_i)$ as proposal, similarly to Section 3.2. Depending on the choice of the channel Q, this might become inefficient since evaluating the full-conditional density might scale as $\mathcal{O}(nk)$. However, for the mechanism we consider, updating Y_i changes at most one of the counts c_j in (12) and thus we do not need to re-compute the histogram $f_{m,\delta}$ at every step.

We can also design a marginal MCMC sampler for global DP, analogous to what we have done for local DP. In particular, by integrating out P in (13), we obtain

$$\Pr(W \in \mathrm{d}w, Y \in \mathrm{d}y, \theta \in \mathrm{d}\theta) = Q(\mathrm{d}w \mid Y) \left\{ \prod_{i=1}^n f(y_i \mid \theta_i) \mathrm{d}y_i \right\} m(\mathrm{d}\theta),$$

where $m(\mathrm{d}\theta) = \frac{\alpha^k}{(\alpha)_{(k)}} \prod_{h=1}^{k} (n_h - 1)! G_0(\mathrm{d}\theta_h^*)$ is the marginal distribution of a sample from a Dirichlet process. Taking inspiration from Algorithm 1, here we propose to update (Y_i, θ_i) jointly. Writing $\pi(\theta_i \mid \theta_{-i})$ for the conditional prior of θ_i given θ_{-i} as in (7), we obtain

$$\Pr(Y_i \in \mathrm{d}y_i, \theta_i \in \mathrm{d}\theta_i \mid \text{rest}) \propto Q(\mathrm{d}w \mid Y) f(y_i \mid \theta_i) \mathrm{d}y_i \pi(\theta_i \mid \theta_{-i}).$$

We sample from this using a Metropolis-Hastings step with a proposal $\Pr(Y_i' \in \mathrm{d}y', \theta_i' \in \mathrm{d}\theta') = \pi(\theta' \mid \theta_{-i}) f(y' \mid \theta_i') \mathrm{d}y'$ and acceptance probability $\min\left\{1, \frac{q(W \mid Y_1,\ldots,Y_i',\ldots,Y_n)}{q(W \mid Y_1,\ldots,Y_i,\ldots,Y_n)}\right\}$. Here $q(\cdot \mid \cdot)$ is the density of the channel Q. Further considerations of marginal MCMC sampling schemes under global DP follow along the same lines discussed for local DP.

In Appendix C, we report a simulation where we compare the posterior inference under the privacy mechanism in 12 using the two MCMC algorithms described above. We obtain the same conclusions of Section 4.1, namely both algorithms present much higher acceptance rates than the theoretical lower bound. The density estimates show a behavior similar to what is observed for the wavelet-based perturbation: one carefully needs to choose the hyper-parameters involved in Q to get good estimates.

6 Real data analysis

We consider a dataset of blood donors at the Milano Department of the Associazione Volontari Italiani del Sangue (AVIS) between January 1st, 2010 and May 15th, 2016. Each data point is the log of the time between the first and second blood donation of one of $n = 100$ individuals, so that $Y_i = \log(\#\text{ number of days between first and second donation})$. Due to health concerns, individuals are not allowed to donate blood for 90 days after their donation. Moreover, the timespan in which data are collected ensures a maximum value for Y_i. Together, these imply $Y_i \in [4.49, 7.6]$, so that $\Delta = 3.11$. We refer to Argiento et al. (2022) for further analyses on the blood donors data.

We analyze this dataset under zero concentrated differential privacy, fixing a privacy loss budget $\rho = 17.8$ following the US Census Bureau standard for data referring to individuals (Garfinkel, 2022). To obtain sanitized data, we consider the Laplace mechanism (with scale $\Delta/\sqrt{2\rho}$) and the Gaussian mechanism (with variance $\Delta^2/(2\rho)$). We assume a Gaussian mixture model for the Y_i's, with G_0 the Normal-Inverse-Gamma distribution: $G_0(\mathrm{d}\mu, \mathrm{d}\sigma^2) = \mathcal{N}(\mathrm{d}\mu \mid \mu_0, 10\sigma^2) \times IG(\mathrm{d}\sigma^2 \mid a, b)$. The hyperparameters for G_0 are chosen in an empirical Bayes fashion, with $\mu = n^{-1} \sum_{i=1}^n Z_i$, and (a, b) are chosen so that $\mathbb{E}[\sigma^2]$ is equal to one half of the empirical variance of the Z_i's and $\mathrm{Var}[\sigma^2] = 0.5$. When using the Gaussian mechanism, we consider the private Neal 5 algorithm, the private slice-sampling, and the private Neal 2 and Neal 3 algorithms as discussed in Section 3.3. When using the Laplace mechanism, we consider the private Neal 5 algorithm and the private slice-sampling. We also consider a non-Bayesian alternative recently proposed in Farokhi (2020) which estimates the density function of the Y_i's by the deconvolution kernel density estimator of Stefanski and Carroll (1990). See also Delaigle and Gijbels (2004) and references therein. To compute the estimator we use the R package `decon` (Wang and Wang, 2011), which applies a bootstrap approach to select the optimal kernel bandwidth.

The density estimates are reported in Figure 6 (top). The Bayesian estimates look essentially identical across both perturbation mechanisms and for any of the proposed algorithms, reflecting the importance of explicitly modeling the confidential data as well as the privacy-mechanism. Instead, while the deconvolution kernel density estimator looks similar to the Bayesian estimate under the Gaussian mechanism, under the Laplace mechanism, it seems to be overfitting the data. We also compare the posterior of the mixture density given the privatized data and given the true non-sanitized data. Figure 6 (bottom) shows the posterior density estimates and the 95% pointwise credible bands. It is clear how the density estimates are essentially identical when considering the privatized or the original dataset. On the other hand, as one would expect, the credible bands are thinner when considering the original observations.

7 Discussion

We considered the problem of Bayesian density estimation under differential privacy, providing a "marginal" and a "conditional" algorithm for posterior inference. These algorithms are amenable to any perturbation mechanism as long as it has a tractable density. We also showed how in some cases, a careful choice of perturbation mechanism, mixture kernel, and

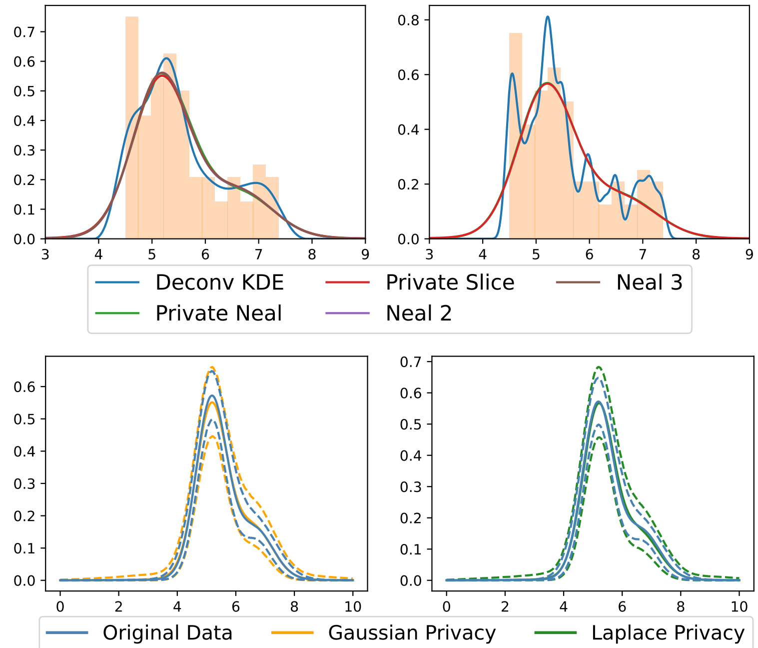

Figure 6: Density estimates for the AVIS dataset in Section 6. Left plots: Gaussian perturbation. Right plots: Laplace perturbation. The top row shows posterior point estimates and the deconvolution KDE estimator. The bottom row shows the point estimates and pointwise 95% credible bands. The histograms refer to the non-privatized data Y_i.

base measure allow for several latent parameters to be marginalized out, thus obtaining more efficient algorithms. While we have focused on Dirichlet process mixtures for univariate data in the exposition, our algorithms are general regarding both the mixing measure and the mixture kernels (including the case of multivariate observations), and using the `BayesMix` library, it is already possible to consider different modeling choices.

Our algorithms are not specific to the setting of differential privacy, and can be applied to general deconvolution problems. In particular, in the case of the Laplace mechanism we recover the setting in Rousseau and Scricciolo (2021), where theoretical guarantees on the convergence of the posterior distribution are provided. Our examples in Section 4.1 can be seen as an empirical confirmation of their theoretical work. As mentioned in Ju et al. (2022), the efficiency of our algorithms are linked to the special structure of differential privacy, and it is interesting to see if other deconvolution problems possess structure that can be exploited similarly. In the other direction, extending the analysis in Rousseau and Scricciolo (2021) to other kind of privacy mechanisms is a challenging and interesting task. Another question that is worth pursuing is to more formally study the effect that the inclusion of a privacy mechanism has on the mixing time of the MCMC algorithms.

References

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697 – 725.

Arbel, J. and I. Prünster (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing 27*, 3–17.

Argiento, R., R. Corradin, A. Guglielmi, and E. Lanzarone (2022). Clustering blood donors via mixtures of product partition models with covariates. *arXiv:2210.08297*.

Argiento, R. and M. De Iorio (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics 50*(5), 2641–2663.

Balle, B. and Y.-X. Wang (2018, 10–15 Jul). Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *The 35th International Conference on Machine Learning*.

Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). Modeling with normalized random measure mixture models. *Statistical Science 28*(3), 313 – 334.

Beraha, M. and R. Corradin (2021). Bayesian nonparametric model based clustering with intractable distributions: an ABC approach. *arXiv preprint arXiv:2112.10393*.

Beraha, M., B. Guindani, M. Gianella, and A. Guglielmi (2022). BayesMix: Bayesian mixture models in C++. *arXiv preprint arXiv:2205.08144*.

Bernstein, G. and D. R. Sheldon (2018). Differentially private bayesian inference for exponential families. *Advances in Neural Information Processing Systems 31*.

Bernstein, G. and D. R. Sheldon (2019). Differentially private Bayesian linear regression. *Advances in Neural Information Processing Systems 32*.

Bun, M. and T. Steinke (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography: 14th International Conference, TCC 2016-B, Beijing, China, 2016, Proceedings, Part I*, pp. 635–658. Springer.

Butucea, C., A. Dubois, M. Kroll, and A. Saumard (2020). Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids.

Butucea, C., A. Rohde, and L. Steinberger (2022, July). Interactive versus non-interactive locally differentially private estimation: Two elbows for the quadratic functional. arXiv:2003.04773 [math, stat].

Canale, A., R. Corradin, and B. Nipoti (2022). Importance conditional sampling for Pitman–Yor mixtures. *Statistics and Computing 32*(3), 40.

Chen, W.-N., A. Ozgur, and P. Kairouz (2022). The Poisson Binomial mechanism for unbiased federated learning with secure aggregation. In *Proceedings of the 39th International Conference on Machine Learning*.

Collins, L. M. and S. T. Lanza (2009). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. John Wiley & Sons.

Delaigle, A. and I. Gijbels (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis 45*(2), 249–267.

Dimitrakakis, C., B. Nelson, Z. Zhang, A. Mitrokotsa, and B. I. Rubinstein (2017). Differential privacy for Bayesian inference through posterior sampling. *Journal of Machine Learning Research 18*(11), 1–39.

Duchi, J. C., M. I. Jordan, and M. J. Wainwright (2018, January). Minimax Optimal Procedures for Locally Private Estimation. *Journal of the American Statistical Association 113*(521), 182–201.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pp. 1–12. Springer.

Dwork, C. and A. Roth (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*(3–4), 211–407.

Dwork, C. and G. Rothblum (2016). Concentrated differential privacy. *arXiv:1603.01887*.

Farokhi, F. (2020). Deconvoluting kernel density estimation and regression for locally differentially private data. *Scientific Reports 10*(1), 21361.

Favaro, S. and Y. Teh (2013). MCMC for Normalized Random Measure Mixture Models. *Statistical Science 28*, 335–359.

Favaro, S. and S. Walker (2013). Slice sampling sigma-stable poisson-kingman mixture models. *Journal of Computational and Graphical Statistics 22*, 830–847.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

Frühwirth-Schnatter, S., G. Celeux, and C. P. Robert (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC.

Garfinkel, S. (2022). Differential Privacy and the 2020 US Census. *MIT Case Studies in Social and Ethical Responsibilities of Computing*.

Griffin, J. E. and S. G. Walker (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics 20*(1), 241–259.

Hu, J., M. R. Williams, and T. D. Savitsky (2022). Mechanisms for global differential privacy under Bayesian data synthesis. *arXiv preprint arXiv:2205.05003*.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association 96*(453), 161–173.

Jain, S. and R. M. Neal (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of computational and Graphical Statistics 13*(1).

Ju, N., J. A. Awan, R. Gong, and V. A. Rao (2022, June). Data augmentation MCMC for Bayesian inference from privatized data. arXiv:2206.00710 [stat].

Kalli, M., J. E. Griffin, and S. G. Walker (2011). Slice sampling mixture models. *Statistics and computing 21*, 93–105.

Kamath, G., O. Sheffet, V. Singhal, and J. Ullman (2019). Differentially private algorithms for learning mixtures of separated Gaussians. *Advances in Neural Information Processing Systems 32*.

Karwa, V., D. Kifer, and A. B. Slavković (2015). Private posterior distributions from variational approximations. *arXiv preprint arXiv:1511.07896*.

Kroll, M. (2021). On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics 15*(1), 1783 – 1813.

Lalanne, C., A. Garivier, and R. Gribonval (2023). About the cost of global privacy in density estimation. *arXiv preprint arXiv:2306.14535*.

Land, K. C. (2001). Introduction to the special issue on finite mixture models. *Sociological Methods & Research 29*(3), 275–281.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics 12*(1), 351–357.

Lomeli, M., S. Favaro, and Y. Teh (2017). A marginal sampler for sigma-Stable Poisson-Kingman mixture models. *Journal of Computational and Graphical Statistics 26*, 44–53.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics 9*(2), 249–265.

Nissim, K., S. Raskhodnikova, and A. Smith (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84.

Papaspiliopoulos, O. and G. O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika 95*(1), 169–186.

Rousseau, J. and C. Scricciolo (2021, November). Wasserstein convergence in Bayesian deconvolution models. arXiv:2111.06846.

Sart, M. (2023). Density estimation under local differential privacy and Hellinger loss. *Bernoulli 29*(3), 2318 – 2341.

Savitsky, T. D., M. R. Williams, and J. Hu (2022). Bayesian pseudo posterior mechanism under asymptotic differential privacy. *Journal of Machine Learning Research 23*, 55–1.

Schlattmann, P. (2009). *Medical applications of finite mixture models.* Springer.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica 4*.

Singh, C. H. and L. Ladusingh (2010). Inpatient length of stay: a finite mixture modeling analysis. *The European Journal of Health Economics 11*, 119–126.

Stefanski, L. A. and R. J. Carroll (1990). Deconvolving kernel density estimators. *Statistics 21*(2), 169–184.

Wade, S. (2023). Bayesian cluster analysis. *Phil. Trans. of the Royal Soc. A*.

Wang, X., Z. Sun, Y. Zhang, Z. Xu, H. Xin, H. Huang, R. H. Duerr, K. Chen, Y. Ding, and W. Chen (2020). BREM-SC: a Bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic acids research 48*(11), 5814–5824.

Wang, X.-F. and B. Wang (2011). Deconvolution estimation in measurement error models: the R package decon. *Journal of statistical software 39*(10).

Wasserman, L. and S. Zhou (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association 105*(489), 375–389.

Zhao, J., T. Wang, T. Bai, K.-Y. Lam, Z. Xu, S. Shi, X. Ren, X. Yang, Y. Liu, and H. Yu (2019). Reviewing and improving the Gaussian mechanism for differential privacy. *arXiv:1911.12060*.

A Background Material

A.1 Wavelet-Based Perturbation and Estimators

The wavelet-based perturbation mechanism of Butucea et al. (2020) considers a collection of wavelet functions

$$\psi_{j,k}(x) = \begin{cases} \varphi(x-k) & \text{if } j = -1 \\ 2^{j/2}\psi(2^j x - k) & \text{if } j \geq 0, \end{cases}$$

with φ and ψ the father and mother wavelet, respectively. We assume that the Y_i's, φ, and ψ are supported on $[0,1]$, though any other compact interval of $\mathbb{R}$ is applicable. Butucea et al. (2020) consider $Z_i = (Z_{i,j,k})$ for $j = -1, \ldots, j_1$ and $k \in \mathbb{Z}$ such that

$$Z_{ijk} \,|\, Y_i \sim \mathcal{L}(\psi_{j,k}(Y_i), s_j) \tag{14}$$

for a suitable choice of the parameter s_j. They then consider the density estimator

$$\hat{f}(x) = \sum_{j=-1}^{j_1} \sum_{k \in \mathbb{Z}} \hat{\beta}_{jk}\psi_{j,k}(x), \quad \hat{\beta}_{jk} = \sum_{i=1}^{n} Z_{i,j,k},$$

where j_1 controls the complexity of the estimator, proving that $\hat{f}(x)$ achieves the minimax optimal rate for the L_p loss over certain classes of densities. Because of the compact support of the wavelet functions, for a fixed j, $\psi_{j,k}$ is not identically zero over $[0,1]$ if and only if $k \leq 2^j$. In our work, we will consider (3) where $\psi_{j,k}$ is the Haar wavelet: $\varphi(x) = \mathrm{I}_{[0,1]}(x)$ and $\psi(x) = \mathrm{I}_{[0,1/2)}(x) - \mathrm{I}_{[1/2,1)}(x)$. Under these assumptions, ε-DP is achieved if $s_{-1} = 12/\varepsilon$ and $sj = \frac{12}{\varepsilon}\frac{\sqrt{2}}{\sqrt{2}-1}2^{j_1/2}$, see (Butucea et al., 2020, Equation 3.1). Note that $\psi_{-1,k}(x) \equiv 1$ for any $x \in [0,1]$ so this is effectively useless and we will consider only $\psi_{j,k}$ for $j \geq 0$.

B Further Details on the Experiments

B.1 Laplace versus Wavelet

In this example, the mixture kernel $f(\cdot \,|\, \theta)$, $\theta = (a,b)$ is not conjugate to the base measure G_0. The corresponding full-conditional of the cluster-specific values $\theta_h^* = (a_h^*, b_h^*)$ is

$$f(a_h^*, b_h^* \,|\, \text{rest}) \propto \left(\frac{\Gamma(a^* + b^*)}{\Gamma(a^*)\Gamma(b^*)}\right)^{n_h} \prod_{i:c_i=h} y_i^{a_h^*-1}(1-y_i)^{b_h^*-1} a_h^* b_h^* e^{-2(a_h^*+b_h^*)}.$$

To sample from the distribution above, we employ a Metropolis-adjusted Langevin algorithm step with step-size 0.05 on the unconstrained parameters $(\log a_h^*, \log b_h^*) \in \mathbb{R}^2$.

As for the initialization, when using the Laplace privacy mechanism we simply initialize the Y_i's at random on $[0, 1]$. Instead, in the case of the wavelet-based mechanism, we take advantage of the multi-resolution structure of the Haar wavelet basis. Indeed observe that $\varphi_{j,k}(x) > 0$ if and only if $x \in [2^{-j}k, 2^{-j}(k + 1/2)]$, and $\varphi_{j,k}(x) < 0$ if and only if $x \in [2^{-j}(k + 1/2), 2^{-j}(k + 1)]$. Therefore, we consider the last 2^{j_1} elements (i.e., the ones associated to the partition of the domain at the highest resolution) in Z_i and take k^* as the index associated to the maximum absolute value of such elements. Then, we initialize the Y_i's by sampling from a Uniform distribution on $[2^{-j}k, 2^{-j}(k + 1/2)]$ or on $[2^{-j}(k + 1/2), 2^{-j}(k + 1)]$ depending on whether the corresponding Z_i's were positive or negative.

C Numerical Illustrations for the Global Privacy Mechanism

We consider the privacy mechanism in (12). We generate $n = 250$ data points from a mixture of three beta distributions as in Section 4.2. We consider privacy levels $\varepsilon = 2, 10, 50, 100, 250$ and set the parameters m, k and δ following Theorem 4.3 in Wasserman and Zhou (2010), specifically $m = L\lfloor n^{1/5} + 1\rfloor$ for three choices of $L = 2, 4, 6$, $k = \lfloor n^{3/5} + 1\rfloor$ and find δ numerically as the smallest number in $(0, 1)$ satisfying

$$k \log\left(\frac{1-\delta}{\delta}\frac{m}{n} - 1\right) \leq \varepsilon.$$

The base measure G_0 is fixed as in Section 4.2 as $G_0(\mathrm{d}a, \mathrm{d}b) = \mathrm{Gamma}(\mathrm{d}a \,|\, 2, 2)\mathrm{Gamma}(\mathrm{d}b \,|\, 2, 2)$.

We sample from the posterior distribution using the two algorithms described in Section 5.1. Although not reported here, the mixing of the MCMC chain shows the same behavior as in the numerical illustration for the locally differentially private mechanism. Namely, the acceptance rates decrease with the level of privacy ε for both algorithms, while being consistently above the theoretical lower bound of $\exp(-\varepsilon)$. The effective sample sizes for the number of clusters is a decreasing function of ε for the marginal algorithm, while it is essentially constant for the conditional one. Figure 7 shows the density estimates, which are mostly in accordance for both algorithms. Similarly to the case of the wavelet-based mechanism, it is evident here that the parameters involved in the privacy mechanism must be carefully chosen, and it is not enough to simply satisfy their asymptotic growth rates. Indeed for $L = 2$ and $L = 6$ the density estimates are poor, with most of the mass being pushed to the borders and the remaining being spread uniformly through the domain. When $L = 2$ this is likely due to the number of bins m being too small, so that the resulting histogram looks uniform, when $L = 6$ the number of bins is significantly larger but δ is large as well which makes the resulting histogram rather uniform. For $L = 4$ the densities are rather satisfactory, but not as good as the ones with the locally differentially private mechanisms in Section 4.2. In all cases, the posterior assigns a sigificant mass to the boundaries of the domain. This behavior can be partially contrasted with the choice of a more concentrated base measure. See Figure 8 where we report the density estimates when $G_0(\mathrm{d}a, \mathrm{d}b) = \mathrm{Gamma}(\mathrm{d}a \mid 90, 2)\mathrm{Gamma}(\mathrm{d}b \mid 90, 2)$.

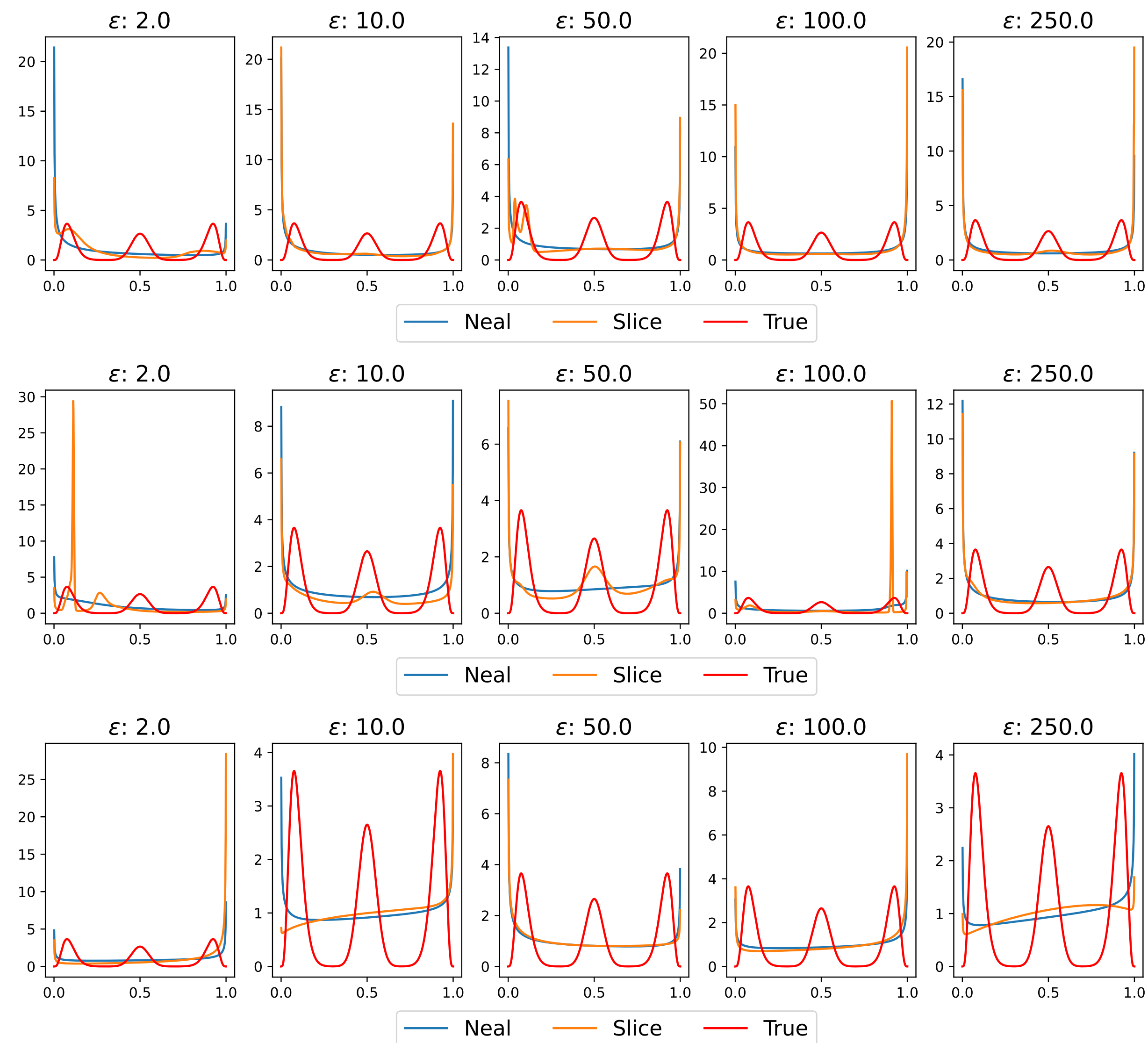

Figure 7: Density estimate for the simulated example in Section 5. From top to bottom the parameter L in $L = 2, 4, 6$

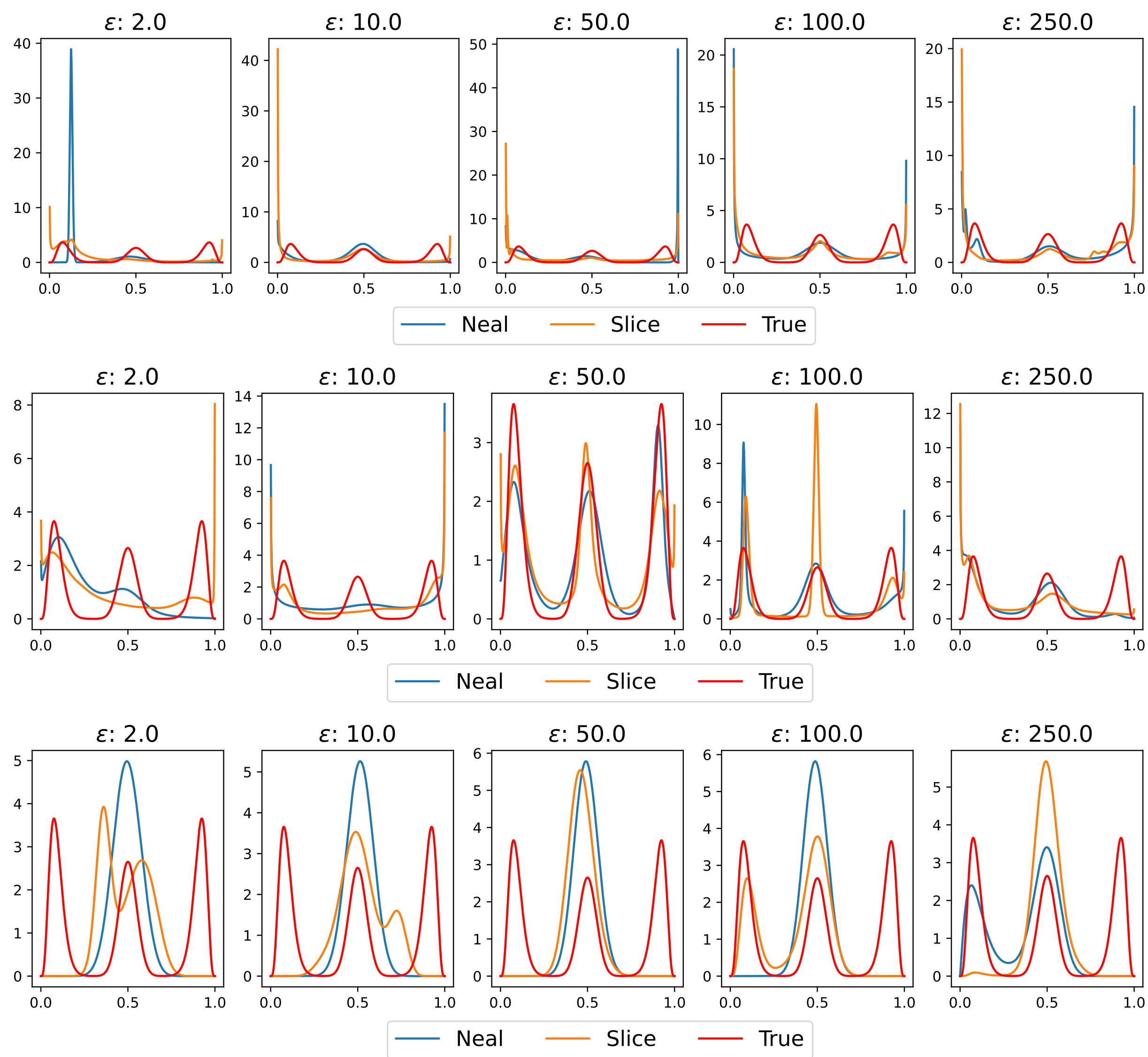

Figure 8: Density estimate for the simulated example in Section 5 when $G_0(\mathrm{d}a, \mathrm{d}b) = \mathrm{Gamma}(\mathrm{d}a \,|\, 90, 2)\mathrm{Gamma}(\mathrm{d}b \,|\, 90, 2)$. From top to bottom the parameter L in $L = 2, 4, 6$

D Futher Plots

Figure 9 shows the density estimates for one particular instance of the simulated datasets in Section 4.1.

Figure 10 shows the adjusted rand index between the true and estimated clustering for the simulation in Section 4.1

Figure 11 shows the posterior distribution for the number of clusters for the simulated example in Section 4.1.

Figure 12 shows the acceptance rate for the simulated example in Section 4.2

Figure 8 shows the density estimates for the simulated example in Section 5 when setting the base measure $G_0(\mathrm{d}a, \mathrm{d}b) = \mathrm{Gamma}(\mathrm{d}a \mid 90, 2)\mathrm{Gamma}(\mathrm{d}b \mid 90, 2)$.

Figure 13 shows the density estimates for one particular instance of the simulated datasets in Section 4.3.

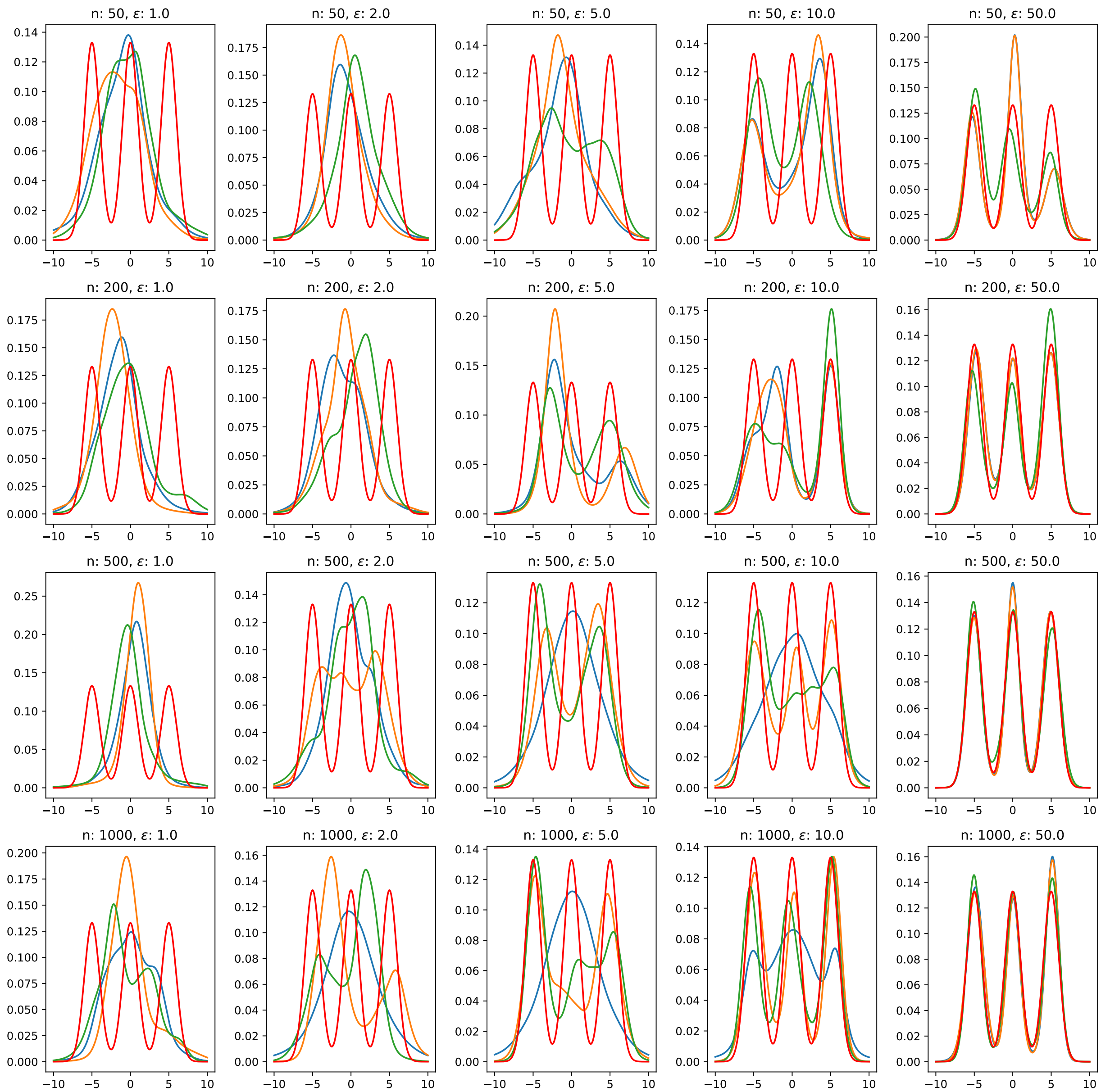

Figure 9: Density estimate for different values of the sample size n and privacy levels ε for the experiment in Section 4.1. Red line is the true data-generating density, blue and orange line are posterior estimates obtained with the Neal 5 and slice algorithms, respectively.

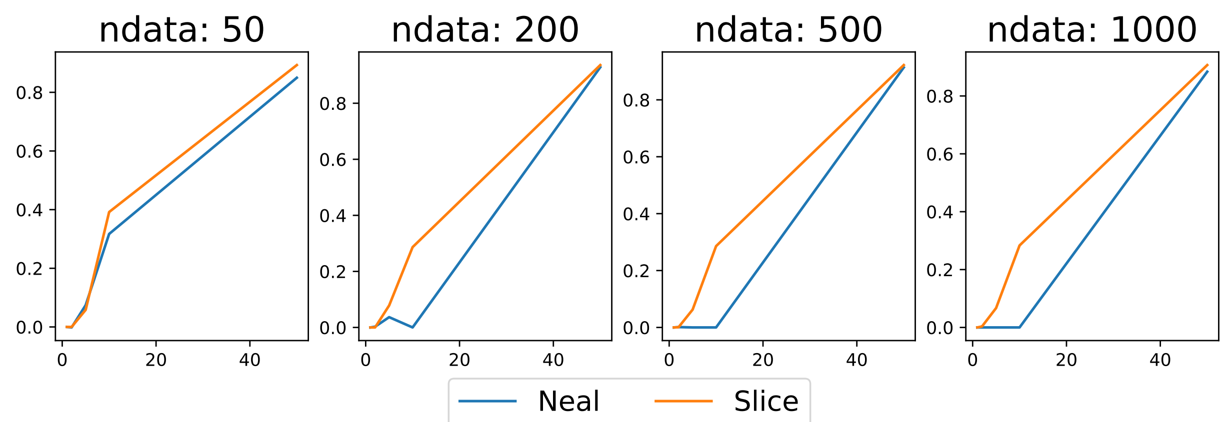

Figure 10: Adjusted Rand Index between estimated and true clustering for the experiment in Section 4.1. Figures are mediated across 50 independent experiments

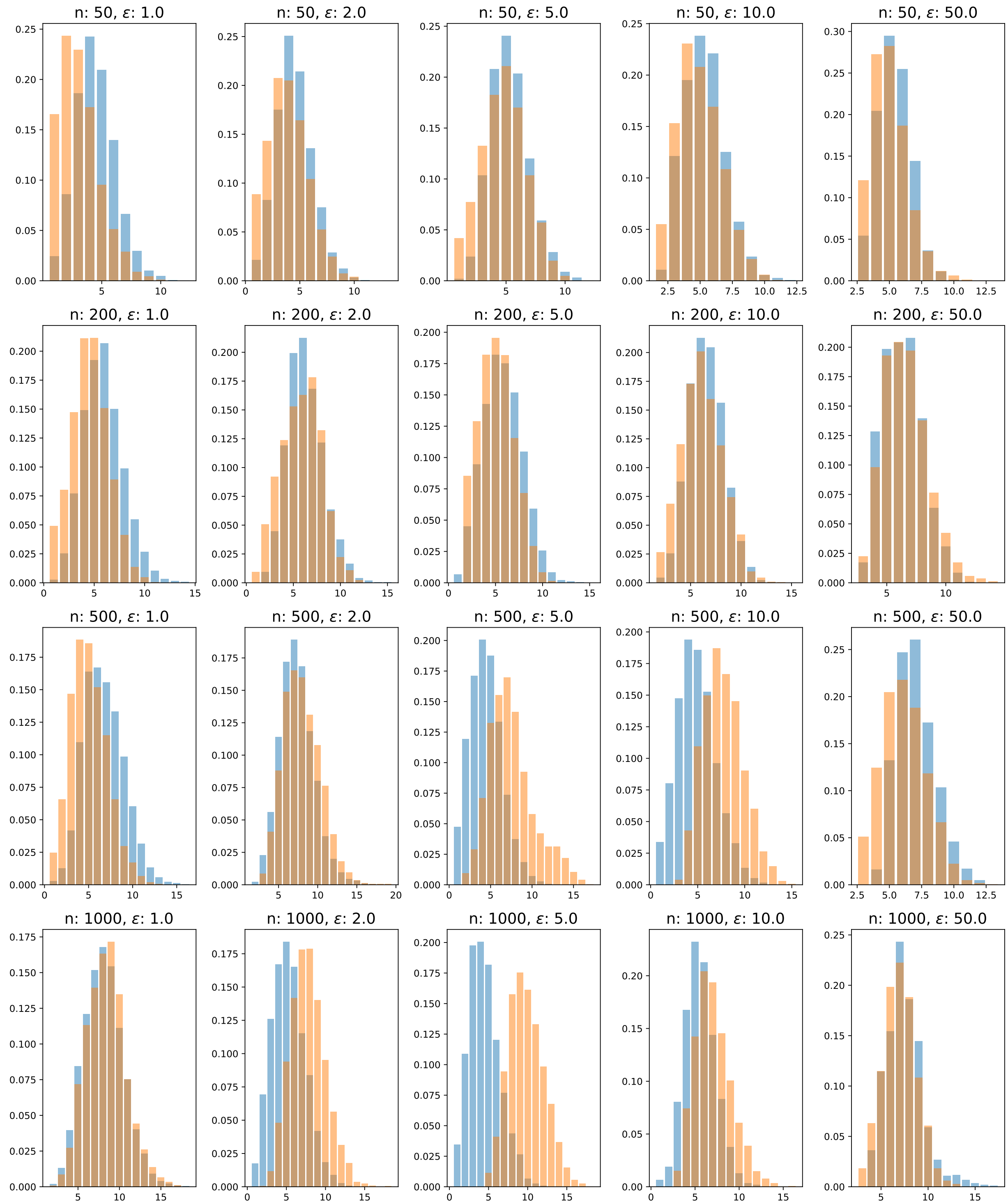

Figure 11: Posterior distribution of the number of clusters for the example in Section 4.1. The blue and orange bars correspond to Neal's and the Slice algorithms, respectively.

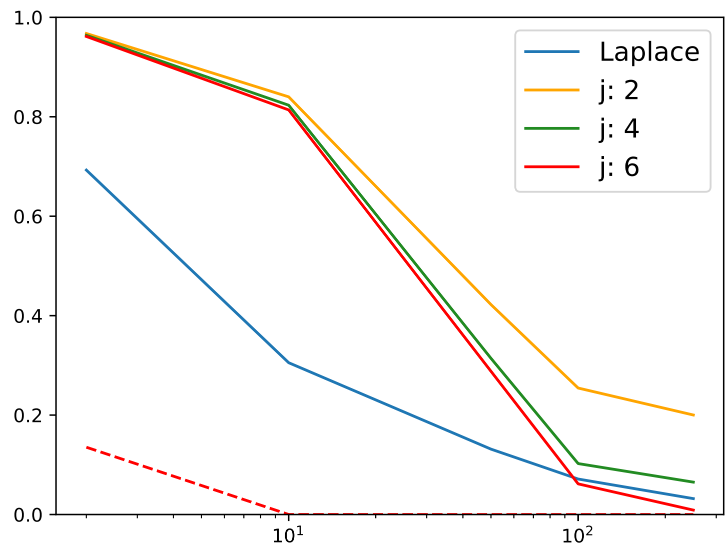

Figure 12: Acceptance rates for the simulated example in Section 4.2. The dashed line corresponds to the theoretical lower bound $\exp(-\varepsilon)$.

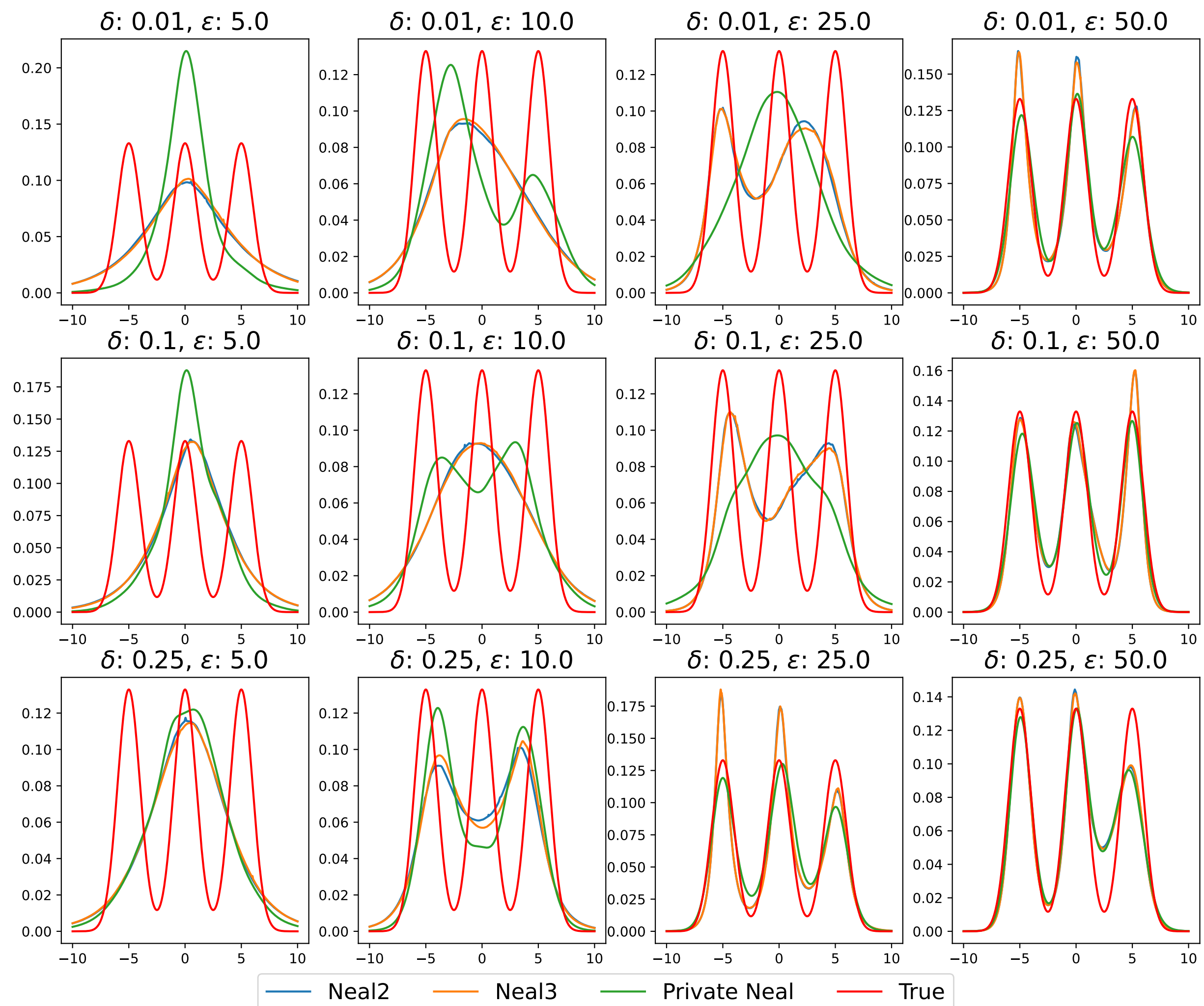

Figure 13: Density estimate for the simulated example in Section 4.3 for different values of ε and δ.