A Data Synthesizer for Legally Anonymized Location Data

Cameron Bale
August 5, 2021

Abstract

Recent research has shown the need for anonymization methods that incorporate the legal criteria defining non-personal or anonymized data. This paper proposes a method for generating legally anonymized synthetic location data with a binary mark. This synthetic data can be released and stored in place of the original data, reducing the privacy risk for individuals related to the data, and reducing the cost for companies to comply with privacy laws. To test this method, we generate synthetic legally anonymized datasets containing the locations and disease outcome for South Korean COVID-19 patients. We show that the synthetic datasets preserve the spatial relationships in the original data while meeting legal anonymization criteria.

Acknowledgements

I would like to thank Jordan Fischer, Matthew Schneider, and Steven Weber, as the inspiration and legal anonymization criteria used for this project were taken from our prior work [5]. Previous versions of this work were submitted as projects for OPR 992, OPR 922, and CS 613, and I would like to thank Hande Benson for her helpful feedback based on the OPR project submissions.

1 Introduction

While location data is useful in many contexts such as location based advertising [25], disease mapping [26], and contact tracing [20], there are growing concerns over the privacy (or lack thereof) of this type of data [35],[11]. Anonymization methods can improve the privacy of location data, e.g., [22], but the enactment of privacy legislation, e.g., the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), raises questions on how mathematical and legal privacy interact. In general under privacy law, data controllers can either implement privacy and security measures for personal data (failing to do so can result in significant fines), or convert personal data to non-personal data, a process known as anonymization [15],[14]. Recent privacy legislation [14],[38] and related research [5] highlight the need for robust anonymization methods that produce data that meet the legal and mathematical criteria defining non-personal data.

To help meet this need, we propose an anonymization method for location data with a binary mark. More specifically, we propose a privacy preserving synthetic data method built on a Gaussian Mixture Model (GMM), which approximates the distribution of individuals' locations, and a novel anonymization algorithm for the mark. The method produces synthetic data sets that accurately capture the spatial relationships of the locations in the original data. Importantly, the method produces synthetic data in which singling out, linkability, and inference (as they are legally defined) [5] are prevented, i.e., the synthetic data is legally anonymous. This method will enable data controllers to release highly useful anonymized location data with a binary mark. We show in our test that the proposed method is directly applicable to data collected during the recent COVID-19 pandemic.

This paper proceeds as follows: Section 2 contains an overview of legal data privacy, methods for generating synthetic location data, previous privacy research involving GMMs, and methods used to obtain the parameters of GMMs. Section 3 outlines the two methods utilized in this paper to solve for the parameters of our GMM, and the algorithm used to assign marks that meet the criteria defining anonymized data. Section 4 gives a brief overview of the South Korean COVID-19 location data used to test the proposed anonymization method, and Section 5 illustrates the results of this test. Section 6 concludes with a discussion of the test results and future research topics.

2 Literature Review

2.1 Legal Data Privacy

The interested reader should consult [5] for a detailed review of privacy legislature, legal privacy terms and their definitions. These authors provide three mathematical criteria used to evaluate whether an anonymization method produces data that is legally anonymized. These criteria are designed to prevent singling out (isolating an individual in a larger group based on the data), linkability (linking together two records concerning the same individual), and inference (deducing, with significant probability, unknown information about an individual) [14]. These three terms were found to be exhaustive in defining legally anonymized data. A data controller may use an anonymization method to alter personal data so that it meets the criteria proposed in [5], thereby avoiding the need to comply with strict, and potentially expensive, privacy and security measures for personal data. For example, under the GDPR, data controllers are required to track the flow of personal data throughout their organization [2]. Data controllers are not required to track the flow of legally anonymized data, and this data is not subject to the same restrictions as personal data when being transferred across national borders (e.g., from the EU to the US) [17]. While we do not provide exact estimates of the savings to companies from legally anonymizing data, estimates of the cost for a company to comply with the CCPA range from $50,000 to over $5,000,000 depending on the size of the company [6]. Publicly available legal anonymization methods that reduce compliance costs have the potential to be very valuable to data controllers. Next, we discuss a subset of anonymization methods, synthetic data methods, and their application to location data.

2.2 Synthetic Location Data Methods

Synthetic data methods are designed to generate partially or fully synthetic data which preserves utility and reduces privacy risk compared to the original data. In practice, a statistical model (e.g., [31]) is fit to the original data, or the parameters of the underlying probability distribution of the data are estimated [21]. These are then used to generate the synthetic data.

Synthetic location data should preserve the spatial dependence and statistical properties of the original locations. Prior research has produced methods for generating synthetic data with precise geographical points and their associated marks [28], and larger geographic regions such as counties [27]. While consideration is given to disclosure risks in these scenarios [28],[27],[21], the criteria defining legally anonymized data are not considered, and synthetic location data produced by previous methods may not be legally anonymized.

2.3 Gaussian Mixture Models and Privacy

Gaussian mixture models are used for density estimation and classification tasks. One strength of GMMs is that they can accurately estimate the density of multi-modal distributions. This is our main motivation for using a GMM, since Releasing the values of GMM parameters may compromise the privacy of the data the model was fit to. Central and local differentially private GMMs have been created to alleviate this concern. In general, privacy protection based on central differential privacy involves a data controller, who possesses personal data, using a differentially private mechanism to release private statistics about the personal data. In the case of GMMs, a data controller estimates the joint density of personal data using a differentially private GMM and releases this model to the public, who can use the model to learn about the underlying data [39]. On the other hand, local differential privacy involves implementing protection (noise addition) at the point of data collection so that data controllers are never in possession of sensitive data. In this case, the authors in [9] propose a variant of a GMM designed to estimate the joint density of the original data based on the noisy protected data. Note that any calculation performed on the output of a differentially private mechanism is also differentially private. This is one of the main strengths of differential privacy. However, differential privacy is known to severely reduce data utility, especially for detailed microdata [30], which motivates the use of synthetic data methods for anonymizing data. For example, the United States Census Bureau has announced that it will move toward using a statistical model, rather than differential privacy based methods, to release synthetic microdata from the American Community Survey [1].

GMMs have not been implemented in the context of data synthesis, perhaps because these models are only appropriate for numeric data, and many datasets consist of numeric and categorical attributes. An

alternative option is to use a Gaussian copula model. A differentially private Gaussian copula model has been used to generate synthetic mixed variable datasets that provide highly accurate answers to one-way and two-way positive conjunction queries [4]. However, the utility of this data is limited to the queries it is designed to answer. And, a major assumption underlying a copula model is that the joint density of data samples is unimodal, which is often not the case. For these reasons, we designed a synthetic data method that consists of two pieces. For numeric data, we take advantage of the ability of GMMs to estimate multimodal densities, and for categorical data, we use a novel anonymization algorithm that incorporates legal anonymization criteria.

2.4 Methods for Obtaining GMM Parameters

The largest barrier to obtaining the parameters of a GMM is the constraint that covariance parameters must be positive definite. Optimization methods such as Newton, quasi-Newton, and scoring methods can be used to obtain mixture parameters, but come with significant downsides in terms of computation cost and complexity [29].

The parameters of Gaussian Mixture Models (GMM) are typically found using the expectation maximization (EM) algorithm [8]. EM has significant benefits such as automatically satisfying probabilistic constraints, exhibiting monotonic convergence without the need to set a learning rate, and requiring low computational overhead [40]. While the EM algorithm exhibits slow convergence toward optimal parameter values in problems where mixture components are poorly separated, it still produces rapid increases in likelihood [40],[29]. In problems where mixture components are poorly separated, other methods are likely to perform poorly as well [40].

The semi-definite programming literature describes techniques for solving problems with a nature similar to a GMM - maximizing an objective with a positive definite matrix constraint. One general strategy for handling this constraint is to factor the positive definite matrix. Such factorizations leverage problem-specific properties such as sparsity or the ability to use a low-rank approximation of the matrix of interest [23]. These properties are not afforded to GMMs, since covariance matrices cannot be singular and may not be sparse. Still, some semi-definite programming techniques have been tested in solving for mixture model parameters and on related statistical problems.

One factorization proposed in [36] enables semi-definite programming problems to be expressed as standard convex optimization problems. This strategy was found to be inferior to an 'EM-like' fixed point algorithm in terms of speed when solving a problem similar to a mixture model. This was due to the computational expense of each iteration [33].

The authors in [19] tested a conjugate gradient method in solving for the parameters of a GMM in its original form and with a Cholesky decomposition of the covariance matrices. The EM algorithm was significantly faster in most tests. However, these authors also proposed a reformulated version of the GMM that lends itself to Riemannian optimization. This reformulation, paired with manifold stochastic gradient descent, was shown to outperform the EM algorithm by achieving better log-likelihood values and faster convergence to an optimum in some cases.

The reformulated GMM solved via Riemannian optimization has already been applied in forecasting wind power [16] and Markov decision process problems [37]. Due to this recent success, this paper tests both EM and Riemannian optimization in solving for the parameters of our GMM.

3 Methodology

Consider a dataset $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ containing N samples of D dimension. These samples are labeled $[N] = \{1, \ldots, N\}$ and indexed $n \in [N]$. Using a GMM with K mixture components, $p(\mathbf{x})$ is estimated as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1}$$

Here, $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a multivariate Gaussian distribution (the kth mixture component), π_k is the weight of the kth mixture component, and $\sum_{k=1}^{K} \pi_k = 1$ so that $p(\mathbf{x})$ is a proper probability density.

3.1 GMM and EM

Assuming the samples $\mathbf{x}_n$, $n \in [N]$ are i.i.d., the parameters $\mathbf{\Theta} = \{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \pi_k\}_{k=1}^K$ that maximize the log-likelihood of observing $\mathbf{X}$ are obtained by solving the following problem:

$$\begin{aligned} \max_{\{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \pi_k\}_{k=1}^K} \quad & \sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \\ \text{s.t.} \quad & \sum_{k=1}^K \pi_k = 1 \\ & \pi_k \geq 0, \quad \forall k \end{aligned} \tag{2}$$

See [7] for a thorough explanation and derivation of the EM algorithm updates for GMMs. Let $r_{n,k}$ denote the probability that $\mathbf{x}_n$ was generated by the kth mixture component:

$$r_{n,k} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j)}. \tag{3}$$

The mean, covariance, and mixture weight parameters are updated according to the following rules:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{R_k} \sum_{n=1}^N r_{n,k} \mathbf{x}_n \tag{4}$$

$$\mathbf{\Sigma}_k^{new} = \frac{1}{R_k} \sum_{n=1}^N r_{n,k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \tag{5}$$

$$\pi_k^{new} = \frac{R_k}{N} \tag{6}$$

where $R_k = \sum_{n=1}^N r_{n,k}$. Given initial values for $\mathbf{\Theta}$, the EM algorithm alternates between calculating the responsibilities $r_{n,k}$ and updating $\mathbf{\Theta}$ until convergence. This will require computing the inverse of the covariance parameters, which may be singular if the mean of a component is identical to some $\mathbf{x}_n$ [7]. In practice, this is handled by adding a small number to the diagonal of the covariance parameters [32].

3.2 Reformulated GMM and Riemannian Optimization

The interested reader should consult [18] for detailed explanations and derivations of the GMM reformulated for Riemannian optimization. A brief overview and essential equations are presented here. The log-likelihood presented in (2) must be reformulated to be geodesically convex to enable Riemannian optimization methods to compete with EM in solving for GMM parameters. First, augment all samples $\mathbf{x}_n$ with an extra dimension:

$$\mathbf{y}_n^{\mathrm{T}} = [\mathbf{x}_n^{\mathrm{T}} \quad 1] \tag{7}$$

Next, consider the following maximization problem, which consists of a single Gaussian, using the augmented samples $\mathbf{y}_n$:

$$\max_{\mathbf{S}} \quad \sum_{n=1}^N \log q_{\mathcal{N}}(\mathbf{y}_n; \mathbf{S}) \tag{8}$$

where $q_{\mathcal{N}}(\mathbf{y}_n; \mathbf{S}) = 2\pi \exp(\frac{1}{2}) \mathcal{N}(\mathbf{y}_n; \mathbf{0}, \mathbf{S})$. Let $\mathbf{S}^*$ denote the maximum of (8). Solving for $\mathbf{S}^*$ yields the optimal solution to (2) when $K = 1$:

$$\mathbf{S}^* = \begin{bmatrix} \mathbf{\Sigma}^* + \boldsymbol{\mu}^* \boldsymbol{\mu}^{*\mathrm{T}} & \boldsymbol{\mu}^* \\ \boldsymbol{\mu}^{*\mathrm{T}} & 1 \end{bmatrix} \tag{9}$$

Overall, the original GMM problem is replaced with a reparameterized log-likelihood consisting of geodesically convex components with local maxima that agree with those of (2). Let $\omega_k = \log \frac{\pi_k}{\pi_K}$, where $\omega_K = 0$.

Using this change of variables to incorporate the contraints in (2) produces the following optimization problem on the product manifold $(\Pi_{k=1}^{K}\mathbb{P}^{D+1}) \times \mathbb{R}^{K-1}$, where $\mathbb{P}^{D+1}$ is the manifold of $(D+1)\times(D+1)$ positive definite matrices:

$$\max_{\{\mathbf{S}_k\}_{k=1}^{K},\{\omega_k\}_{k=1}^{K-1}} \quad \sum_{n=1}^{N}\log\sum_{k=1}^{K}\frac{\exp(\omega_k)}{\sum_{j=1}^{K}\exp(\omega_j)}q_{\mathcal{N}}(\mathbf{y}_n;\mathbf{S}_k) \tag{10}$$

In this paper, both problems are solved - (2) and (10) - for various values of K. The values of K that minimize the Bayesian Information Criterion of the models are selected [24]. Based on these values of K, we select the solver that produces a model that generates the synthetic data with spatial properties most similar to those of the original data.

3.3 Criteria Defining Anonymized Location Data

Recall that singling out, linkability, and inference as they are legally and mathematically defined [5] must be prevented for data to be legally anonymized. In theory, linkability will automatically be prevented in synthetic data with continuous variables since it will be impossible to create linkages on those variables with outside data. Note that the criteria (11) and (13) were originally developed in [5]. We present them here for ease of reference.

Consider a synthetic data set with $\check{N}$ observations containing locations drawn from $p(\mathbf{x})$ and marks assigned using the algorithm which will be shown later. Denote this synthetic dataset $\check{\mathbf{X}} = [\check{\mathbf{x}}_1,\ldots,\check{\mathbf{x}}_{\check{N}}]^{\mathrm{T}}$ which consists of synthetic samples $\check{\mathbf{x}}_n = [\check{x}_{n,1}\ \ \check{x}_{n,2}\ \ \check{s}_n]^{\mathrm{T}}$. Here, $\check{x}_{n,1}$ and $\check{x}_{n,2}$ are the location measurements for the nth sample, and $\check{s}_n$ is the mark. Let $\bar{\mathbf{X}} = [\bar{\mathbf{x}}_1,\ldots,\bar{\mathbf{x}}_I]^{\mathrm{T}}$ represent the matrix of unique samples in the dataset $\check{\mathbf{X}}$, where $I \leq \check{N}$. These samples are labeled $[I] = \{1,\ldots,I\}$ and indexed $i \in [I]$. A counting measure $\mu(\bar{\mathbf{x}}_i) = \#\{n \in [\check{N}] : \check{\mathbf{x}}_n = \bar{\mathbf{x}}_i\}$ indicates the number of times a sample $\bar{\mathbf{x}}_i$ occurs in the dataset $\check{\mathbf{X}}$. Define a vector $\bar{\mathbf{z}} = [\bar{z}_1,\ldots,\bar{z}_I] \in \mathbb{N}^I$ with $\bar{z}_i = \mu(\bar{\mathbf{x}}_i)$ and $\sum_{i=1}^{I}\bar{z}_i = \check{N}$. This vector contains the number of times each $\bar{\mathbf{x}}_i$ occurs in $\check{\mathbf{X}}$. Singling out is prevented in $\check{\mathbf{X}}$ when,

$$\inf \bar{\mathbf{z}} \geq 2. \tag{11}$$

When (11) is met, then linkability is also prevented.

Preventing inference in $\check{\mathbf{X}}$ requires protecting against the most likely means a third party may employ to re-identify individuals [5], [14]. We consider a variant of the 'Type S Risk' discussed in [28], a type of inferential disclosure risk [21]. This risk is designed to measure the ability of an adversary to infer the mark of an individual based on their spatial location(s) and the samples in $\check{\mathbf{X}}$.

Suppose an adversary knows some individual was at the location in an original data sample $\mathbf{x}_n$. That is, the adversary knows the values $\mathbf{x}_n^{\ell} = [x_{n,1}\ x_{n,2}]$. Since the synthetic data will not contain any samples located *exactly* at this location, the adversary will infer s_n, the mark of the individual, based on the marks of synthetic samples that are spatially 'close' to $\mathbf{x}_n^{\ell}$. Let s^* denote the true value of s_n. Define $p(s^*|\delta,\check{\mathbf{X}},\mathbf{x}_n^{\ell})$, an estimate of the probability that an adversary will correctly infer s^*, as

$$p(s^*|\delta,\check{\mathbf{X}},\mathbf{x}_n^{\ell}) = \frac{\sum_{\check{\mathbf{x}}_j^{\ell}\in\mathscr{N}(\mathbf{x}_n^{\ell})} I(\check{s}_j = s^*)}{|\mathscr{N}(\mathbf{x}_n^{\ell})|}, \tag{12}$$

where $\mathscr{N}(\mathbf{x}_n^{\ell}) = \{j \in [\check{N}] : ||\check{\mathbf{x}}_j^{\ell} - \mathbf{x}_n^{\ell}|| \leq \delta\}$. That is, $\mathscr{N}(\mathbf{x}_n^{\ell})$ is the set of synthetic data samples with locations that are within a δ-neighborhood of $\mathbf{x}_n^{\ell}$, and $p(s^*|\delta,\check{\mathbf{X}},\mathbf{x}_n^{\ell})$ is the proportion of synthetic samples with locations in $\mathscr{N}(\mathbf{x}_n^{\ell})$ with a mark equal to s^*. The criteria defining the prevention of inference follows:

$$p(s^*|\delta,\check{\mathbf{X}},\mathbf{x}_n^{\ell}) - p(s^*|\mathbf{x}_n^{\ell}) \leq c,\ \forall n \in [N]. \tag{13}$$

Here, $p(s^*|\mathbf{x}_n^{\ell})$ denotes a baseline, or prior probability of an adversary inferring the mark s^* for $\mathbf{x}_n$. This quantity will be unknown to a data controller, and should be estimated based on an assessment of available external information that would inform the prior beliefs of an adversary. Likewise, the quantity c is chosen as deemed appropriate by the data controller.

Synthetic datasets in which singling out, linkability, and inference are prevented are created by ensuring that (11) and (13) are met. In [5], these criteria were prevented using methods based on generalization and

deleting individuals' data for which the criteria were not met. In a sense, the method in this paper works in the opposite way - instead of deleting data for which the criteria are violated, additional locations are generated and marks are assigned in such a way that the criteria are satisfied for every $\mathbf{x}_n$.

This is accomplished using a novel anonymization algorithm, which performs the following steps. N synthetic locations are initially sampled from $p(\mathbf{x})$. For each $\mathbf{x}_n$ with a positive mark, one randomly chosen synthetic location in $\mathscr{N}(\mathbf{x}_n^\ell)$ is assigned a positive mark. The left-hand side of (13) is then calculated for each $\mathbf{x}_n$ in the original data. If (13) does not hold for some $\mathbf{x}_n$, additional synthetic locations with a mark $\check{s} = 1 - s_n$ are generated from the kth mixture component corresponding to the largest $r_{n,k}$ for $\mathbf{x}_n$ until (13) is met. Then, each unique synthetic point is duplicated once so that (11) is met.

To generate legally anonymous $\check{\mathbf{X}}$, the algorithm takes the original data $\mathbf{X}$, the prior probabilities $\mathbf{p}$ associated with each value of the mark, the desired number of mixture components K, the neighborhood radius δ, and the value c. In our application, we found that preventing inference in $\check{\mathbf{X}}$ was usually achieved after adding a few hundred additional synthetic samples.

Algorithm 1 Sketch of Generating Anonymized $\check{\mathbf{X}}$

$\check{\mathbf{X}} \leftarrow$ sample N locations from $p(\mathbf{x})$
for $\mathbf{x}_n,\ n \in \{1, \ldots, N\}$ **do**
  Find $\mathscr{N}(\mathbf{x}_n^\ell)$
  while $|\mathscr{N}(\mathbf{x}_n^\ell)| == 0$ **do**
    $\tilde{k} \leftarrow \text{argmax}\ r_{n,k},\ k = 1, \ldots, K$
    Sample $\check{\mathbf{x}}^\ell$ from $\mathcal{N}_{\tilde{k}}$, append to $\check{\mathbf{X}}$
  end while
  if $s_n == 1$ **then**
    Randomly select some $\check{\mathbf{x}}_j^\ell \in \mathscr{N}(\mathbf{x}_n^\ell)$
    $\check{s}_j \leftarrow 1$
  end if
end for
for $\mathbf{x}_n,\ n \in \{1, \ldots, N\}$ **do**
  $\mathrm{i}_n \leftarrow p(s^*|\delta, \check{\mathbf{X}}, \mathbf{x}_n^\ell) - p(s^*|\mathbf{x}_n^\ell)$
end for
while $\mathrm{i}_n >$ c for some $\mathbf{x}_n$ **do**
  for all $\mathbf{x}_n$ where $\mathrm{i}_n >$ c **do**
    $\tilde{k} \leftarrow \text{argmax}\ r_{n,k},\ k = 1, \ldots, K$
    while $\mathrm{i}_n >$ c **do**
      Sample $\check{\mathbf{x}}_j^\ell$ from $\mathcal{N}_{\tilde{k}}$, append to $\check{\mathbf{X}}$
      $\check{s}_j \leftarrow 1 - s_n$
      $\mathrm{i}_n \leftarrow p(s^*|\delta, \check{\mathbf{X}}, \mathbf{x}_n^\ell) - p(s^*|\mathbf{x}_n^\ell)$
    end while
  end for
  for $\mathbf{x}_n,\ n \in \{1, \ldots, N\}$ **do**
    $\mathrm{i}_n \leftarrow p(s^*|\delta, \check{\mathbf{X}}, \mathbf{x}_n^\ell) - p(s^*|\mathbf{x}_n^\ell)$
  end for
end while
Compute $\bar{\mathbf{X}}$ from $\check{\mathbf{X}}$
for $\bar{\mathbf{x}}_i,\ i \in \{1, \ldots, I\}$ **do**
  Compute $\mu(\bar{\mathbf{x}}_i)$
  if $\mu(\bar{\mathbf{x}}_i) == 1$ **then**
    Append $\bar{\mathbf{x}}_i$ to $\check{\mathbf{X}}$
  end if
end for

4 Data and Testing

The dataset [10] used for testing was collected by the Korean Centers for Disease Control and Prevention (KCDC) and was eventually taken down from its online source due to privacy concerns. This data contains GPS coordinates for 1,472 individuals who were diagnosed with COVID-19 in South Korea. Anywhere from 1 to 45 locations were tracked for each individual between January 20, 2020 and June 01, 2020. Approximately 63% of the locations are unique. Prior to analysis, we project latitude and longitude measurements to meters using the appropriate coordinate reference system, and then standardize the data [12].

Location data is often accompanied by attributes (marks) of the individuals' whose locations are recorded. This creates a high privacy risk since locations and marks can be cross-referenced or combined with outside information to identify individuals. While the South Korean COVID-19 patient data contains multiple marks such as age, gender, and disease status (whether the individual passed away from COVID-19), we focus on protecting disease status. Only 0.67% of the locations in the data correspond to individuals who passed away, where $s_n = 1$ indicates that the individual passed away, and $s_n = 0$ indicates that the individual survived.

5 Results

5.1 Testing Overview

The GMMs discussed previously are fit only to individuals' locations. Once the best method (EM or Riemannian optimization) for obtaining the parameters is found (based on BIC and L function values), multiple sets of synthetic locations are generated from the GMM obtained using the selected method. The anonymization algorithm is then used to assign marks to the locations such that the anonymization criteria are met, and the resulting data is analyzed to determine whether it maintains the spatial properties of the original data.

5.2 EM and Original GMM

The EM algorithm from Scikit Learn [13] is used to solve for the parameters of (2). To avoid covariance singularities, 1e−10 is added to the diagonal of each covariance matrix. The algorithm is initialized using mixture centers produced from k-means++ [3]. The problem in (2) is solved five times for each value of $K = 2, \ldots, 1000$, and the lowest BIC value from those solutions is kept. These values are plotted in Figure 1. $K = 844$ is chosen as this produced the lowest BIC overall.

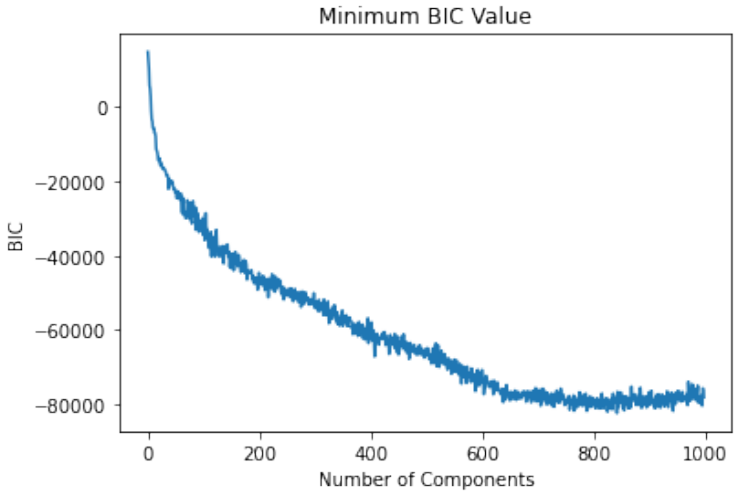


Figure 1: *EM and Original GMM Formulation*: Minimum BIC value from five runs for each $K = 2, \ldots, 1000$.

The model produced by EM is assessed by comparing the spatial properties of the original locations and the synthetic locations drawn from the GMM. We use the K and L functions to assess these spatial

properties. The K function [28] is the expected number of points within a distance h of an arbitrary point. An estimate of this function can be obtained using the following:

$$\hat{K}(h) = \frac{|D|}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} I(||\mathbf{x}_n - \mathbf{x}_m|| \leq h)/N, \text{ for } m \neq n. \tag{14}$$

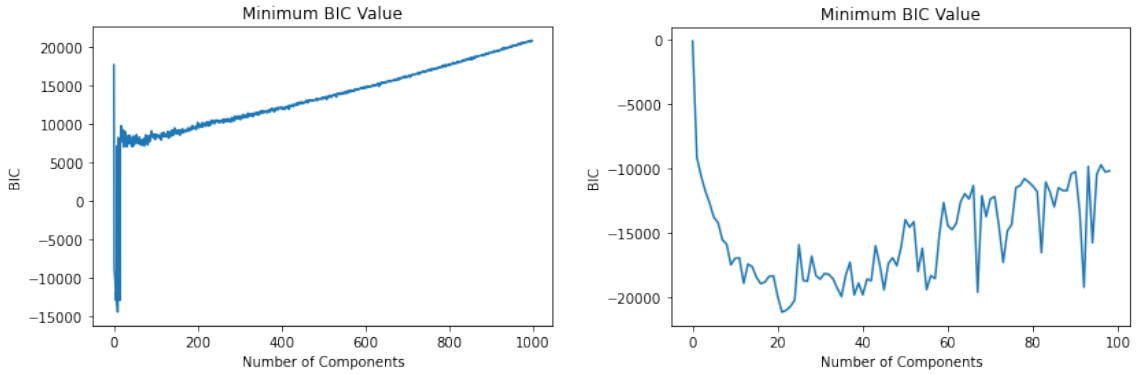
Here, $I(.)$ is the indicator function which is equal to one whenever points $\mathbf{x}_n$ and $\mathbf{x}_m$ are within distance h of each other, and $|D|$ is the area of the spatial domain. The L function can be estimated using the following:

$$\hat{L}(h) = \sqrt{\hat{K}(h)/\pi} - h, \tag{15}$$

where positive values of this estimate indicate spatial clustering. Similar to [28], 50 synthetic datasets are generated from the EM GMM with $K = 844$ components. The L function values for the original and synthetic datasets are computed for a range of h values. We include the mean and 95% empirical confidence interval estimated from the L function values calculated from the synthetic datasets. These results are shown in Figure 3.

5.3 Riemannian Optimization and Reformulated GMM

The python module Pymanopt [34] is used to solve for the parameters of (10). Both steepest descent (gradient descent) (SD) and conjugate gradient (CG) solvers were used to solve the GMM reformulation without the covariance singularity and mixture weight penalizers described in [18]. The optimal numbers of components for the SD and CG variants were found to be $K = 23$ and $K = 11$, respectively. The SD solver achieved lower BIC values than the CG solver, however, both solvers began producing models with increasing BIC values much earlier than the EM variant.



(a) Conjugate gradient (CG) solver.

(b) Steepest descent (gradient descent) (SD) solver.

Figure 2: Minimum BIC value from five runs for each $K = 2, \ldots, 1000$ for the conjugate gradient (CG) solver, and each $K = 2, \ldots, 100$ for the steepest descent (SD) solver. We only obtain solutions for $K = 2, \ldots, 100$ for SD since the performance decreases past $K = 23$.

Again, we generate 50 synthetic datasets consisting of locations draws from the optimal CG and SD GMMs and calculate the values of the K and L functions as described in Section 5.2. These results are also shown in Figure 3.

5.4 Assessing Preservation of Spatial Dependencies

The L function values computed from the synthetic locations drawn from the EM GMM are very similar to the values from the original data. Patterns of spatial clustering and the expected number of points near a given point are well maintained in this synthetic data. The L function estimates from the synthetic locations

drawn from the SD GMM appear closer to the original estimates than those from the CG GMM. However, the L function values from the EM GMM are still the closest to those from the original data. Overall, the model fit and L function values from synthetic datasets drawn from the reformulated GMM are underwhelming. Except for when $h = 0$, the L function is consistently overestimated in locations drawn from the SD GMM, and consistently underestimated in locations drawn from the CG GMM (Figure 3).

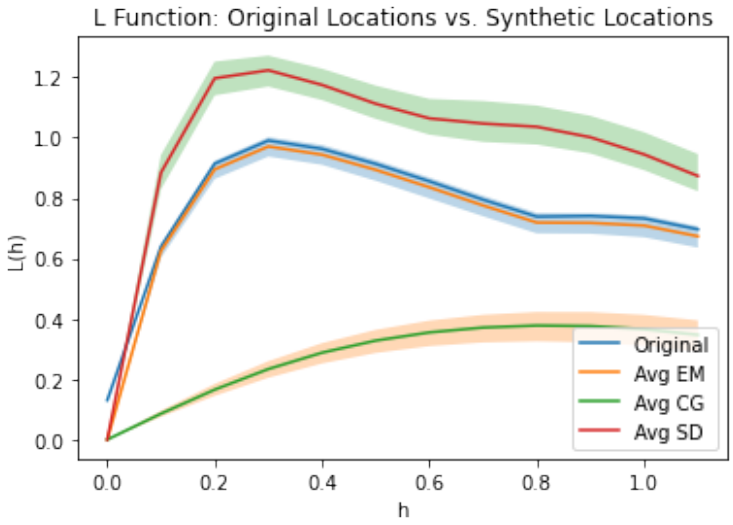
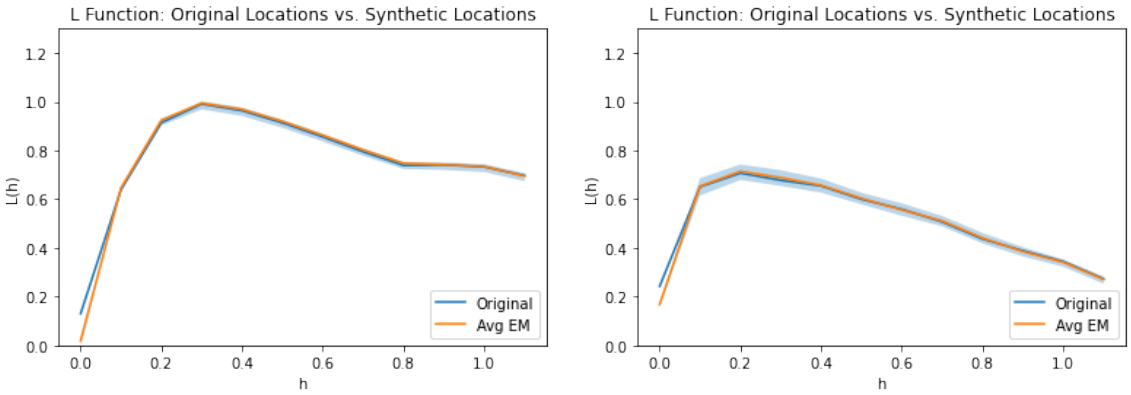


Figure 3: *All Solvers*: L function estimates for values of h from 0 to 1.2. The blue denotes the estimates from the original data. The orange, green, and red lines denote the average values across 50 synthetic location datasets generated from the EM, CG, and SD GMMs, respectively. The shaded areas are the 95% empirical confidence intervals on the L function estimates from the synthetic datasets.

In general, the reformulated GMM does not produce synthetic locations with high utility. We explore the reason for this in Section 7.2 of the Appendix. Based on the results in sections 5.2, 5.3 and 5.4, the EM GMM with $K = 844$ components is selected and used to produce multiple sets of synthetic locations which are then anonymized using our legal anonymization algorithm.

5.5 Anonymization Algorithm

Results from this algorithm are promising. We use the EM GMM model from the previous section and our anonymization algorithm to generate 50 legally anonymized synthetic datasets, where $K = 844$, $c = 0.05$, and $\delta = 0.025$. We treat the proportion of locations in the original data that correspond to an individual who passed away from COVID-19 as an estimate of the adversary's prior probability of predicting $s_n = 1$, i.e., $p(s^*|\mathbf{x}_n^\ell) = 0.0067$ when $s_n = 1$. Similarly, $p(s^*|\mathbf{x}_n^\ell) = 1 - 0.0067 = 0.9933$ when $s_n = 0$. The anonymized data closely mirrors the L function estimates for the original data, even when restricted to only the points with positive marks (see Figures 4(a) and 4(b)).



(a) All anonymous locations.

(b) Anonymous locations with $\check{s}_n = 1$.

Figure 4: *Anonymous Locations from EM GMM and Anonymization Algorithm*: L function estimates for values of h from 0 to 1.2. In (a), the blue line denotes the estimates from the original data. In (b), the blue line denotes the estimates from the original locations with positive marks. The orange lines denote the average values across 50 anonymized synthetic datasets, for all synthetic locations (a) and synthetic locations with positive marks $\check{s}_n = 1$ (b). As before, the shaded areas denote the 95% empirical confidence intervals on the L function estimates from the synthetic datasets.

6 Discussion

This paper presented a method for legally anonymizing location data with a binary mark. The results of applying this method to South Korean COVID-19 location data are promising, as the synthetic data generated using the method were legally anonymized and matched the original data on important spatial features. Specifically, a GMM with $K = 844$ components with parameters obtained using EM produced synthetic locations that closely matched the distribution of original locations. Solving a reformulated version of the GMM using steepest descent and conjugate gradient solvers with $K = 23$ and $K = 11$ components produced much less capable models. The anonymization algorithm assigned marks that closely approximated the original distribution of marked locations, and limited the increase in probability of an adversary deducing the correct value of any individual's mark to $c = 0.05$ within a spatial radius $\delta = 0.025$. To aid in visualizing this process, we have included plots of the δ-neighborhoods of four randomly chosen points in the original data with $s_n = 1$ in section 7.1 of the Appendix.

The methodology presented in this paper is of interest to data controllers that collect or store location data. Individuals' locations with a binary mark can be anonymized, which puts the data outside the scope of privacy law. This saves data stewards some cost of complying with privacy law. For increased privacy, this methodology could be paired with legally defined privacy and security requirements; a data breach involving anonymized data will present significantly lower risk to individuals relative to a breach involving the original data. However, data controllers may be reluctant to give up the utility inherent in personal data after privacy and security measures are already implemented. Maximizing the utility of anonymized data would hopefully reduce this reluctance.

There are multiple avenues for expanding and further analyzing the method in this paper. Further testing using various values of c and δ should be performed to evaluate the generalizability of this methodology. There are other types of marks to consider as well, such as multiple binary marks, continuous, or categorical marks. All of these situations would significantly increase the complexity of applying the anonymization algorithm, and would raise the question of whether the inference criteria (13) should be applied to each mark individually, applied to marks jointly, or some combination of these.

Section 7.2 of the Appendix explores why the GMMs obtained from EM and Riemannian optimization differed so significantly in the number of optimal components and the utility of the synthetic samples they could produce. In short, the EM algorithm produced mixture components with singular covariance matrices and means equal to single data points. This also produced large likelihood values. On the other hand, the

reformulated GMM is optimized over a product manifold of positive definite matrices and does not result in singular covariance matrices. The EM algorithm is essentially overfitting to the original data. This means synthetic locations generated from the model will be extremely similar to the original locations. This is good for the utility of the synthetic data, but reduces the privacy. Synthetic samples generated from components with singular covariance matrices will be approximately equal to that original point. To address this, random noise could be added to the diagonal of singular covariance matrices, or the regularization value could be increased, e.g., from 1e−10 to 1e−5.

In the future, we will assess the effect of reducing δ on privacy and utility. In general, for a given value c, it is more difficult to meet the criteria (13) as δ decreases, especially for uncommon mark values, such as the one examined in our test. Reducing δ will improve privacy, since there is less of a chance an adversary would select a radius less than δ within which to make their inference. This highlights the fact that there is uncertainty as to what attacks may be attempted on a given anonymized dataset. We consider only one type of inferential disclosure risk in our test. Another potential disclosure risk is that an adversary may deduce the location of an individual based on their mark(s) [28]. Or, an adversary may deduce the mark for an individual based on the marks of the k-nearest synthetic neighbors to the original point. We tested our anonymized data against such an attack for $k = 1$. We found that the precision, recall, and f1 score for predicting the disease status of the original points $\mathbf{x}_n$ with $s_n = 0$ were all over 0.99. This is not surprising, as an adversary could correctly infer the disease status of over 99% of individuals in our data by predicting $s_n = 0, \forall n \in [N]$. We posit that this would not correspond to a breach of privacy for these individuals. For the predicted status of points $\mathbf{x}_n$ with $s_n = 1$, the precision, recall, and f1 score were 0.083, 0.09, and 0.087, respectively. Based on this nearest-neighbor model, the majority of an adversary's predictions of positive disease status would correspond to individuals who survived, and the majority of individuals who did pass away from the disease would be predicted to have survived. This is encouraging, and shows that our proposed method provides some protection against other inference attacks. Still, we will examine whether reducing δ will also improve protection against alternate attacks, or whether steps can be taken during the anonymization algorithm to prevent additional attacks.

Overall, any means of deducing sensitive information threaten individuals' privacy in anonymized location data, and highlight the need for careful consideration of the inference criteria (13) in the law. An arbitrary choice of the most likely means a third party may employ to re-identify individuals may leave data vulnerable to other, perhaps equally likely, re-identification strategies. A range of re-identification strategies should be considered to improve the privacy of legally anonymized data.

The utility of the anonymized datasets in this paper will be more thoroughly analyzed. For example, coefficients from spatial regression models could be compared across original and synthetic datasets. Applications to other datasets are of interest as well. The tested South Korea data was quite sparse, and the Riemannian optimization of the reformulated GMM may perform better on denser data. However, the complexity of implementing the reformulated GMM reduces the appeal of this model compared to EM. The reformulated GMM would have to offer significantly better approximations of the original locations to justify its complexity in use; this is unlikely to occur based on the results in this paper.

Overall, this methodology is a viable tool for creating legally anonymized location data with a binary mark. Continued research in this domain will be important to generalize and improve the robustness of the proposed methodology. Our findings show it is possible and beneficial to combine anonymization methods and legal anonymization criterion, and should be of significant interest to data controllers, statisticians, and legal practitioners alike.

References

[1] *An important census product may soon use synthetic data.* The Economist. 2021. URL: `https://www.economist.com/united-states/2021/06/24/an-important-census-product-may-soon-use-synthetic-data`.

[2] Art. 30 GDPR. *Records of Processing Activities.* 2018. URL: `https://gdpr-info.eu/art-30-gdpr/`.

[3] David Arthur and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding.* Tech. rep. 2007, pp. 1027–1035.

[4] Hassan Jameel Asghar et al. "Differentially Private Release of High-Dimensional Datasets using the Gaussian Copula". In: *Journal of Privacy and Confidentiality* 10.2 (2020). DOI: https://doi.org/10.29012/jpc.686.

[5] Cameron D. Bale et al. "Legally Anonymizing Personal Data: An Application to South Korean COVID-19 Location Data". In: *Working Paper* (2020).

[6] Berkeley Economic Advising and Research, LLC. *Standardized regulatory impact assessment: California consumer privacy act of 2018 regulations*. State of California Department of Justice Office of the Attorney General. 2019.

[7] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020. URL: https://mml-book.com.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 39 (1977), pp. 1–38. URL: https://www.jstor.org/stable/2984875.

[9] Xinrong Diao et al. "PrivGMM: Probability Density Estimation with Local Differential Privacy". In: *International Conference on Database Systems for Advanced Applications*. Springer. 2020, pp. 105–121.

[10] DS4C. *Data Science for COVID-19*. 2020. URL: https://www.kaggle.com/kimjihoo/coronavirusdataset.

[11] C. Dwork et al. "On Privacy in the Age of COVID-19". In: *Journal of Privacy and Confidentiality* 10.2 (2020). DOI: https://doi.org/10.29012/jpc.749.

[12] epsg.io. *EPSG:5179 Korea 2000 / Unified CS*. 2010. URL: https://epsg.io/5179.

[13] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: 12 (2011), pp. 2825–2830.

[14] European Data Protection Board. *Guidelines 04/2020 on the Use of Location Data and Contact Tracing Tools in the Context of the COVID-19 Outbreak*. 2020. URL: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-042020-use-location-data-and-contact-tracing%5C_en.

[15] GDPR. *Recital 26, ¶5*. 2018.

[16] F. Ge et al. "Parameter Estimation of a Gaussian Mixture Model for Wind Power Forecast Error by Riemannian L-BFGS Optimization". In: *IEEE Access* 6 (2018), pp. 38892–38899. DOI: 10.1109/ACCESS.2018.2852501.

[17] M. Hintze and K. El Emam. *Can Synthetic Data Help Organizations Respond to 'Schrems II'?* iapp. 2020. URL: https://iapp.org/news/a/can-synthetic-data-help-organizations-respond-to-schrems-ii/.

[18] Reshad Hosseini and Suvrit Sra. "An alternative to EM for Gaussian mixture models: batch and stochastic Riemannian optimization". In: *Mathematical Programming* 181 (2020), pp. 187–223. DOI: https://doi.org/10.1007/s10107-019-01381-4.

[19] Reshad Hosseini and Suvrit Sra. "Matrix Manifold Optimization for Gaussian Mixtures". In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 910–918. URL: https://proceedings.neurips.cc/paper/2015/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf.

[20] Human Rights Watch. *Mobile Location Data and COVID-19: Q&A*. 2020. URL: https://www.hrw.org/news/2020/05/13/mobile-location-data-and-covid-19-qa%5C#.

[21] A. Hundepool et al. *Statistical Disclosure Control*. Chichester: Wiley. 2012.

[22] M. Macha et al. "Perils of Location Tracking? Personalized and Interpretable Privacy Preservation in Consumer Mobile Trajectories". In: *Working Paper* (2020). URL: https://mmacha.github.io/resources/GeoTargetingPrivacy.pdf.

[23] A. Majumdar, G. Hall, and A. A. Ahmadi. "A Survey of Recent Scalability Improvements for Semidefinite Programming with Applications in Machine Learning, Control, and Robotics". In: *arXiv 1908.05209* (2019).

[24] G. J. McLachlan and S. Rathnayake. "On the number of components in a Gaussian mixture model". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.5 (2014), pp. 341–355. DOI: 10.1002/widm.1135.

[25] D. Molitor et al. "Effectiveness of Location-Based Advertising and the impact of interface design". In: *SSRN* (2020). URL: https://ssrn.com/abstract=2645281.

[26] H. Quick, L. A. Waller, and M. Casper. "Multivariate spatiotemporal modeling of age-specific stroke mortality". In: *Ann. Appl. Stat* 11 (2017), pp. 2170–2182.

[27] Harrison Quick and Lance A. Waller. "Using spatiotemporal models to generate synthetic data for public use". In: *Spatial and spatio-temporal epidemiology* 27 (2018), pp. 37–45.

[28] Harrison Quick et al. "Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography". In: *Spatial Statistics* 14 (2015), pp. 439–451.

[29] Richard A. Redner and Homer F. Walker. "Mixture Densities, Maximum Likelihood and the Em Algorithm". In: *SIAM Review* Vol. 26 (1984), pp. 195–239. URL: https://www.jstor.org/stable/2030064.

[30] Steven Ruggles et al. "Differential privacy and census data: Implications for social and economic research". In: *AEA papers and proceedings*. Vol. 109. 2019, pp. 403–08.

[31] Matthew J. Schneider et al. "A flexible method for protecting marketing data: An application to point-of-sale data". In: *Marketing Science* 37.1 (2018), pp. 153–171. DOI: https://doi.org/10.1287/mksc.2017.1064.

[32] Scikit-learn. *sklearn.mixture.GaussianMixture*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html.

[33] S. Sra and R. Hosseini. "Geometric optimisation on positive definite matrices for elliptically contoured distributions". In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 2562–2570.

[34] James Townsend, Niklas Koep, and Sebastian Weichwald. "Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differentiation". In: *Journal of Machine Learning Research* 17.137 (2016), pp. 1–5. URL: http://jmlr.org/papers/v17/16-177.html.

[35] Z. Tu et al. "A New Privacy Breach: User Trajectory Recovery From Aggregated Mobility Data". In: *IEEE/ACM Transactions on Networking* 26.3 (2018), pp. 1446–1459.

[36] R. J. Vanderbei and H. Y. Benson. "On formulating semidefinite programming problems as smooth convex nonlinear optimization problems". In: *Technical Report* (2000).

[37] Shijun Wang et al. "Riemannian Proximal Policy Optimization". In: *arXiv 2005.09195* (2020).

[38] Working Party. *Opinion 05/2014 on Anonymisation Techniques*. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[39] Y. Wu et al. "Differentially Private Density Estimation via Gaussian Mixtures Model". In: *IEEE/ACM 24th International Symposium on Quality of Service (IWQoS)* (2016), pp. 1–6. DOI: 10.1109/IWQoS.2016.7590445.

[40] L. Xu and M. I. Jordan. "On Convergence Properties of the EM Algorithm for Gaussian Mixtures". In: *Neural Comput.* 8 (1996), pp. 129–151.

7 Appendix

7.1 Delta Neighborhoods of Original Positive Points

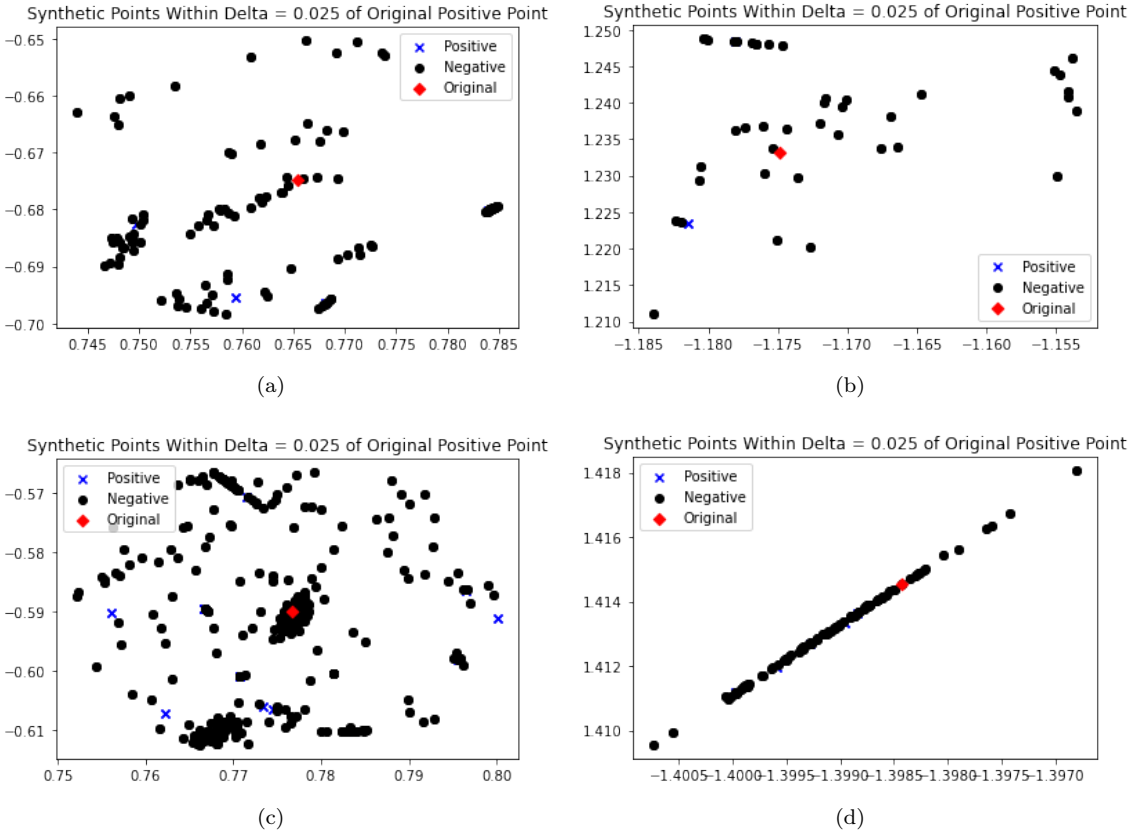


Figure 5: δ-neighborhoods for randomly chosen original positive points, marked with ♦. Positive synthetic points are marked with $\times$, negative synthetic points are marked with •.

7.2 Examining the Difference Between Riemannian and EM GMMs

The stark difference between the GMM models obtained through Riemannian optimization and the GMM model obtained through the EM algorithm warranted further exploration. We suspected that the difference in performance was due to the fact that the reformulated GMM was optimized over a product manifold of positive definite matrices. Only covariance matrices that were non-singular, or equivalently, had non-zero determinants, could be selected for the model. On the other hand, the EM algorithm requires regularization to ensure that covariance matrices remain positive definite. We hypothesized that the EM algorithm produced mixture components that were centered on a single point, resulting in a singular covariance matrix for that component, and inflating the log-likelihood for that specific point and mixture component.

To confirm our hypothesis, we examined the negative log-determinants of the covariance matrices in five models. We considered the GMM model obtained through EM with $K = 844$ components and the reformulated GMM model obtained through Riemannian optimization using the SD and CG solvers with $K \in \{23, 844\}$ and $K \in \{11, 844\}$ components, respectively. Summary statistics on these determinants can be found in Table 1. Note that the minimum(maximum) value of the negative log-determinant corresponds to the maximum(minimum) determinant value for the corresponding model.

Solver	K	**Minimum LD**	**Maximum LD**	**Median LD**	**Mean LD**
Expectation Maximization	844	16.31141	46.05170	34.71183	35.11475
Riemannian Optimization - CG	11	4.60514	4.60545	4.60519	4.60522
Riemannian Optimization - CG	844	4.60516	4.60520	4.60518	4.60518
Riemannian Optimization - SD	23	4.6087	10.4158	5.1732	5.9251
Riemannian Optimization - SD	844	4.5895	4.6845	4.6229	4.6257

Table 1: Summary statistics for the negative log-determinants of the covariance matrices of various GMMs.

Here, we see that the EM algorithm produces covariance matrices with significantly smaller determinants (10x smaller for the maximum negative log-determinant) than the models obtained via Riemannian optimization. To put these numbers in perspective, a negative log-determinant of 16.31 corresponds to a determinant equal to 8e−8, and a negative log-determinant of 4.60545 corresponds to a determinant equal to 1e−3. When the determinants of the covariance matrices of the mixture model approach zero, the likelihood in (2) will increase significantly. Furthermore, we find that 21% of the components in the EM GMM correspond to only one data point, i.e., letting $\mathcal{C}(r_{n,k}) = \#\{n \in [N] : r_{n,k} > r_{n,k'}, \forall k' \neq k\}$ denote the number of points corresponding to the kth mixture component, we find that $\mathcal{C}(r_{n,k}) = 1$ for 21% of the components in the EM GMM model. However, by virtue of optimizing over a product manifold of positive definite matrices, GMMs obtained via Riemannian optimization will not produce mixture components with singular covariance matrices, so these models cannot approximate the joint density of sparse locations as accurately as the EM GMM. In other applications, such as clustering, this behavior would be desireable, since a Riemannian optimized GMM would not overfit to a single data point in the same manner as an EM GMM.