




Statistical Inference with Local Optima

Yen-Chi Chen


To cite this article: Yen-Chi Chen (2022): Statistical Inference with Local Optima, Journal of the American Statistical Association, DOI: [10.1080/01621459.2021.2023550](https://doi.org/10.1080/01621459.2021.2023550)



To link to this article: <https://doi.org/10.1080/01621459.2021.2023550>

 View supplementary material 

 Published online: 28 Feb 2022.

 Submit your article to this journal 

 Article views: 329

 View related articles 

 View Crossmark data 



Statistical Inference with Local Optima

Yen-Chi Chen 

Department of Statistics, University of Washington, Seattle, WA

ABSTRACT

We study the statistical properties of an estimator derived by applying a gradient ascent method with multiple initializations to a multi-modal likelihood function. We derive the population quantity that is the target of this estimator and study the properties of confidence intervals (CIs) constructed from asymptotic normality and the bootstrap approach. In particular, we analyze the coverage deficiency due to finite number of random initializations. We also investigate the CIs by inverting the likelihood ratio test, the score test, and the Wald test, and we show that the resulting CIs may be very different. We propose a two-sample test procedure even when the maximum likelihood estimator is intractable. In addition, we analyze the performance of the EM algorithm under random initializations and derive the coverage of a CI with a finite number of initializations. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received September 2020
Accepted December 2021

KEYWORDS

EM algorithm; Gradient descent; Maximum likelihood estimation; Nonconvex; Two-sample test

1. Introduction

Many statistical analyses involve finding the maximum of an objective function. The maximum likelihood estimator (MLE) is the maximum of the likelihood function. In variational inference (Blei, Kucukelbir, and McAuliffe 2017), the variational estimator is constructed by maximizing the evidence lower bound. In regression analysis, we estimate the parameter by minimizing the loss function, which is equivalent to maximizing the negative loss function. In nonparametric mode hunting (Parzen 1962; Romano 1988a, 1988b), the parameter of interest is the location of the density global mode; therefore, we are finding the point that maximizes the density function.


Each of the above analyses works well when the objective function is concave. However, when the objective function is nonconcave and has many local maxima, finding the (global) maximum can be challenging and even computationally intractable. Moreover, because the computed estimator may not be the actual MLE, the resulting confidence set may not have the nominal coverage.

In this paper, we focus on the analysis of the MLE of a multi-modal likelihood function. Our analysis can also be applied to the examples mentioned before and other types of M-estimators (Van der Vaart 1998). Maximizing a multi-modal likelihood function is a common scenario encountered while we fit a mixture model (Redner and Walker 1984; Titterton, Smith, and Makov 1985). Figure 1 plots the log-likelihood function of fitting a 2-Gaussian mixture model to data generated from a 3-Gaussian mixture model, in which the orange color indicates the regions of parameter space with high likelihood values. There are two local maxima, denoted by the blue and green crosses. The blue maximum is the global maximum. To find the

maximum of a multi-modal likelihood function, we often apply a gradient ascent method such as the EM algorithm (Redner and Walker 1984; Titterton, Smith, and Makov 1985) with an appropriate initial guess of the MLE. The right panel of Figure 1 shows the result of applying a gradient ascent algorithm to a few initial points. Each black dot is an initial guess of the MLE, and the corresponding black curve indicates the gradient ascent path starting from this initial point to a nearby local maximum. Although it is ensured that a gradient ascent method does not decrease the likelihood value (when the step size is sufficiently small), it may converge to a local maximum or a critical point rather than the global maximum. For instance, in the right panel of Figure 1, three initial point converges to the green cross, which is not the global maximum. To resolve this issue, we often randomly initialize the starting point (initial guess) many times and then choose the convergent point with the highest likelihood value as the final estimator (McLachlan and Peel 2004; Jin et al. 2016). However, as we have not explored the entire parameter space, it is hard to determine whether the final estimator is indeed the MLE. Although the theory of MLEs suggests that the MLE is a \sqrt{n} -consistent estimator of the population maximum (population MLE) under appropriate conditions (Titterton, Smith, and Makov 1985), our estimator may not be a \sqrt{n} -consistent estimator because it is generally not the MLE.¹ The CI constructed from the estimator inherits the same problem; if our estimator is not the MLE, it is unclear what population quantity the resulting CI is covering.

The goal of this article is to analyze the statistical properties of this estimator. Note that we do not provide a solution to resolve the problem causing by multiple local maxima; instead, we attempt to analyze how the local maxima affect the performance

CONTACT Yen-Chi Chen  yenchic@uw.edu  Department of Statistics, University of Washington, Seattle, WA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

¹In fact, for a mixture model, the convergence rate could be slower than \sqrt{n} if the number of mixture k is not fixed; see, for example, Li and Barron (1999) and Genovese and Wasserman (2000).

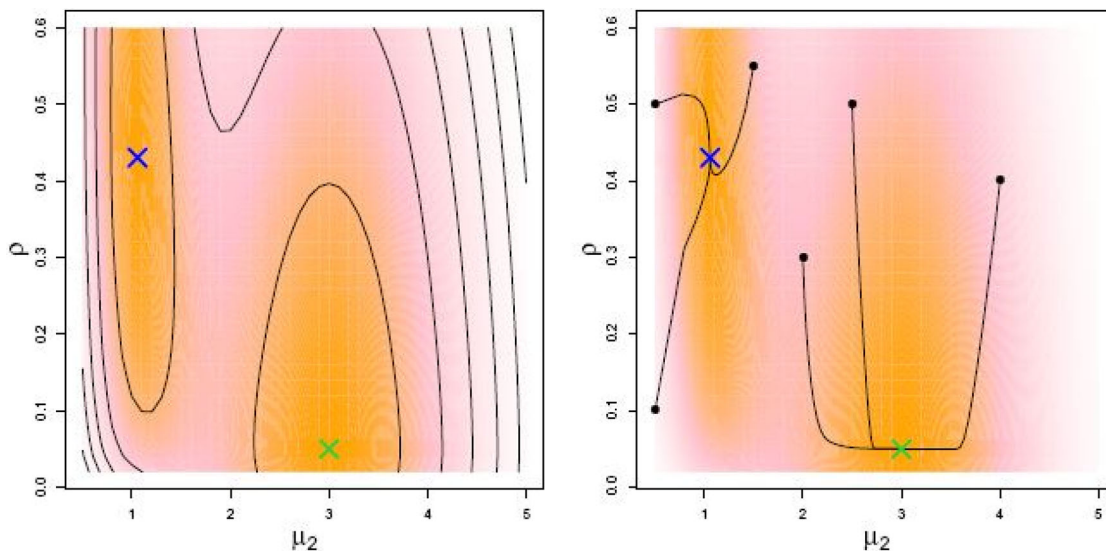


Figure 1. Log-likelihood function of fitting a 2-Gaussian mixture model to a data that is generated from a 3-Gaussian mixture model. The true distribution has a density function: $p_0(x) = 0.5\phi(x; 0, 0.2^2) + 0.45\phi(x; 0.75, 0.2^2) + 0.05\phi(x; 3, 0.2^2)$, where $\phi(x; \mu, \sigma^2)$ is the density of a Gaussian with center μ and variance σ^2 . We fit a 2-Gaussian mixture with the center of the first Gaussian being set to 0 and the variance of both Gaussians being 0.2^2 . The parameters of interest are the center of the second Gaussian μ_2 and the proportion of the second Gaussian ρ . Namely, the log-likelihood function is $L(\mu_2, \rho) = \mathbb{E}(\log((1 - \rho)\phi(X; 0, 0.2^2) + \rho\phi(X; \mu_2, 0.2^2)))$, where X has a pdf p_0 . Left: contour plot of the log-likelihood function $L(\mu_2, \rho)$. Regions with orange color are where the log-likelihood function has a high value. The two local maxima are denoted by the blue and green crosses. Right: the trajectories of the gradient ascent method with multiple initial points. Each solid black dot is an initial point and the curve attached to it indicates the trajectory of the gradient ascent method starting from that initial point.

of the estimator and the validity of a related statistical procedure. Although this estimator is not the MLE, it is commonly used in practice. To understand what population quantity this estimator is estimating, we study the behavior of estimators obtained from applying a gradient ascent algorithm to a likelihood function that has multiple local maxima. We investigate the underlying population quantity being estimated and analyze the properties of resulting CIs. Specifically, our main contributions are summarized as follows.

Main Contributions.

1. We derive the population quantity being estimated by the MLE when the likelihood function has multiple local maxima (Theorems 1 and 2).
2. We analyze the population quantity that a normal CI covers and study its coverage (Theorem 3).
3. We discuss how to use the bootstrap method to construct a meaningful CI and derive its coverage (Theorem 4).
4. We show that the CIs from inverting the likelihood ratio test, score test, and Wald test can be different (Section 3.3 and Figure 4).
5. We also discuss how to perform a two-sample test maintaining Type I error when the MLE is intractable (Section 3.4).
6. We analyze the probability that the EM algorithm recovers the actual MLE (Section 4) and study the coverage of its normal CI (Theorem 6).
7. We apply our developed framework to investigate the old faithful dataset (Section 5).

Related Work. The analysis of MLE under a multi-modal likelihood function has been analyzed for decades; see, for example, Redner (1981), Redner and Walker (1984), Sundberg (1974), and Titterington, Smith, and Makov (1985). In the multi-modal case, finding the MLE is often accomplished by applying a gradi-

ent ascent method such as the EM-algorithm (Dempster, Laird, and Rubin 1977; Wu 1983; Titterington, Smith, and Makov 1985) with random initializations. The analysis of initializations and convergence of the gradient ascent method can be found in Lee et al. (2016), Panageas and Piliouras (2017), Jin et al. (2016), and Balakrishnan, Wainwright, and Yu (2017). In our analysis, we use the Morse theory (Milnor 1963; Morse 1930; Banyaga and Hurtubise 2013) to analyze the behavior of a gradient ascent algorithm. The analysis using the Morse theory is related to the work of Chazal et al. (2017), Mei, Bai, and Montanari (2018), and Chen et al. (2017).

Outline. We begin with providing the necessary background in Section 2. Then, we discuss how to perform statistical inference with local optima in Section 3. We extend our analysis to EM algorithm in Section 4. We provide data analysis in Section 5. Finally, we discuss issues and opportunities for further work in Section 6. In the supplementary material, we include a simulation study to investigate the effect of initialization (Section A) and a generalization to mode hunting problem (Section E) and all technical assumptions and proofs (Section G).

2. Background

In the first few sections, we will focus on an estimator that attempts to maximize the likelihood function. Let $X_1, \dots, X_n \sim P_0$ be a random sample. For simplicity, we assume that each X_i is continuous. In parametric estimation, we impose a model on the underlying population distribution function $P(\cdot; \theta)$. This gives a parameterized probability density function $p(\cdot; \theta)$. The MLE estimates the parameter using

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \hat{L}_n(\theta) = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^n \log p(X_i; \theta),$$

which can be viewed as an estimator of the population MLE:

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} L(\theta) = \underset{\theta}{\operatorname{argmax}} \mathbb{E}(\log p(X_1; \theta)).$$

When the likelihood function has multiple local modes (maxima), the MLE does not in general have a closed form; therefore, we need a numerical method to find it. A common approach is to apply a gradient ascent algorithm to the likelihood function with a randomly chosen initial point. To simplify our analysis, we study a continuous-time gradient ascent flow of the likelihood function (this is like conducting a gradient ascent with an infinitely small step size). When the likelihood function has multiple local maxima, the algorithm may converge to a local maximum rather than to the global maximum. As a result, we need to repeat the above procedure several times with different initial values and choose the convergent point with the highest likelihood value.

To study the behavior of a gradient ascent flow, we define the following quantities. Given an initial point θ^\dagger , let $\widehat{\gamma}_{\theta^\dagger} : \mathbb{R} \rightarrow \Theta$ be a gradient flow such that

$$\widehat{\gamma}_{\theta^\dagger}(0) = \theta^\dagger, \quad \widehat{\gamma}'_{\theta^\dagger}(t) = \nabla \widehat{L}_n(\gamma_{\theta^\dagger}(t)).$$

Namely, the flow $\widehat{\gamma}_{\theta^\dagger}$ starts at θ^\dagger and moves according to the gradient direction of \widehat{L}_n . The stationary point $\widehat{\gamma}_{\theta^\dagger}(\infty) = \lim_{t \rightarrow \infty} \widehat{\gamma}_{\theta^\dagger}(t)$ is the destination of the gradient flow starting at θ^\dagger . Different starting points lead to flows that may end at different points.

Because our initial points are chosen randomly, we view these initial points $\theta_1^\dagger, \dots, \theta_M^\dagger$ as iid draws from a distribution $\widehat{\Pi}_n(\cdot)$ (see, e.g., McLachlan and Peel 2004, chap. 2.12.2) that may depend on the original data X_1, \dots, X_n . The number M denotes the number of the initializations. Later we will assume that $\widehat{\Pi}_n$ converges to a fixed distribution Π when the sample size increases to infinity. Note that different initialization methods lead to a different distribution of $\widehat{\Pi}_n$. As an example, in the Gaussian mixture model, we often draw random points from the observed data as the initial centers of each mixture component. In this case, $\widehat{\Pi}_n$ can be viewed as the empirical distribution function.

By applying the gradient ascent to each of the M initial parameters, we obtain a collection of stationary points $\widehat{\gamma}_{\theta_1^\dagger}(\infty), \dots, \widehat{\gamma}_{\theta_M^\dagger}(\infty)$. The estimator is the one that maximizes the likelihood function so it can be written as

$$\widehat{\theta}_{n,M} = \operatorname{argmax}_{\widehat{\gamma}_{\theta_\ell^\dagger}(\infty)} \left\{ \widehat{L}_n \left(\widehat{\gamma}_{\theta_\ell^\dagger}(\infty) \right) : \ell = 1, \dots, M \right\}. \quad (1)$$

In practice, we often treat $\widehat{\theta}_{n,M}$ as $\widehat{\theta}_{\text{MLEs}}$ and use it to make inferences about the underlying population. However, unless the likelihood function is concave, there is no guarantee that $\widehat{\theta}_{n,M} = \widehat{\theta}_{\text{MLE}}$. Thus, our inferences and conclusions, which were based on treating $\widehat{\theta}_{n,M}$ as the MLE, could be problematic.

2.1. Population-Level Analysis

To better understand the inferences we make when treating $\widehat{\theta}_{n,M}$ as $\widehat{\theta}_{\text{MLE}}$, we start with a population level analysis over $\widehat{\theta}_{n,M}$. The population version of the gradient flow $\widehat{\gamma}_{\theta^\dagger}$ starting at θ^\dagger is a gradient flow $\gamma_{\theta^\dagger}(t)$ such that

$$\gamma_{\theta^\dagger}(0) = \theta^\dagger, \quad \gamma'_{\theta^\dagger}(t) = \nabla L(\gamma_{\theta^\dagger}(t)).$$

Algorithm 1 Gradient ascent with random initialization

1. Choose θ^\dagger randomly from a distribution $\widehat{\Pi}_n$.
2. Starting with θ^\dagger , apply a gradient ascent algorithm to \widehat{L}_n until it converges. Let $\widehat{\gamma}_{\theta^\dagger}(\infty)$ be the stationary point.
3. Repeat the above two steps M times, leading to

$$\widehat{\gamma}_{\theta_1^\dagger}(\infty), \dots, \widehat{\gamma}_{\theta_M^\dagger}(\infty).$$

4. Compute the corresponding log-likelihood value of each of them:

$$\widehat{L}_n \left(\widehat{\gamma}_{\theta_1^\dagger}(\infty) \right), \dots, \widehat{L}_n \left(\widehat{\gamma}_{\theta_M^\dagger}(\infty) \right).$$

5. Choose the final estimator as

$$\widehat{\theta}_{n,M} = \operatorname{argmax}_{\widehat{\gamma}_{\theta_\ell^\dagger}(\infty)} \left\{ \widehat{L}_n \left(\widehat{\gamma}_{\theta_\ell^\dagger}(\infty) \right) : \ell = 1, \dots, M \right\}.$$

The destination of this gradient flow, $\gamma_{\theta^\dagger}(\infty) = \lim_{t \rightarrow \infty} \gamma_{\theta^\dagger}(t)$, is one of the critical points of $L(\theta)$.

For a critical point m of L , we define the basin of attraction of m as the collection of initial points where the gradient flow converges to m :

$$\mathcal{A}(m) = \{ \theta \in \Theta : \gamma_\theta(\infty) = m \}.$$

Namely, $\mathcal{A}(m)$ is the region where the (population) gradient ascent flow converges to critical point m .

Throughout this article, we assume that L is a Morse function and L has a continuous second derivatives. That is, critical points of L are nondegenerate (well-separated). By the stable manifold theorem (e.g., Theorem 4.15 of Banyaga and Hurtubise 2013), $\mathcal{A}(m)$ is a k -dimensional manifold, where k is the number of negative eigenvalues of $H(m)$, the Hessian matrix of $L(\cdot)$ evaluated at m . Thus, the Lebesgue measure of $\mathcal{A}(m)$ is nonzero only when m is a local maximum. Because of this fact, we restrict our attention to local maxima and ignore other types of critical points; a randomly chosen initial point has probability zero of falling within the basin of attraction of a critical point that is not a local maximum when $\widehat{\Pi}_n$ is continuous. Note that a similar argument also appears in Lee et al. (2016) and Panageas and Piliouras (2017). Let \mathcal{C} be the collection of local maxima with

$$\mathcal{C} = \{ m_1, \dots, m_K \},$$

$$L(m_1) \geq L(m_2) \geq \dots \geq L(m_K),$$

where K is the number of local maxima. The population MLE is $m_1 = \theta_{\text{MLE}}$.

Figure 2 provides an illustration of the critical points and the basin of attraction. The left panel displays the contour plot of a log-likelihood function. The three solid black dots are the local maxima (m_1, m_2 , and m_3), the three crosses are the critical points, and the empty box indicates a local minimum. In the middle panel, we display gradient flows from some starting points. The right panel shows the corresponding basins of attraction ($\mathcal{A}(m_1), \mathcal{A}(m_2)$, and $\mathcal{A}(m_3)$). Each color patch is a basin of attraction of a local maximum.

For the ℓ th local maximum, we define the probability

$$q_\ell^\pi = \Pi(\mathcal{A}(m_\ell)) = \int_{\mathcal{A}(m_\ell)} d\Pi(\theta), \quad (2)$$

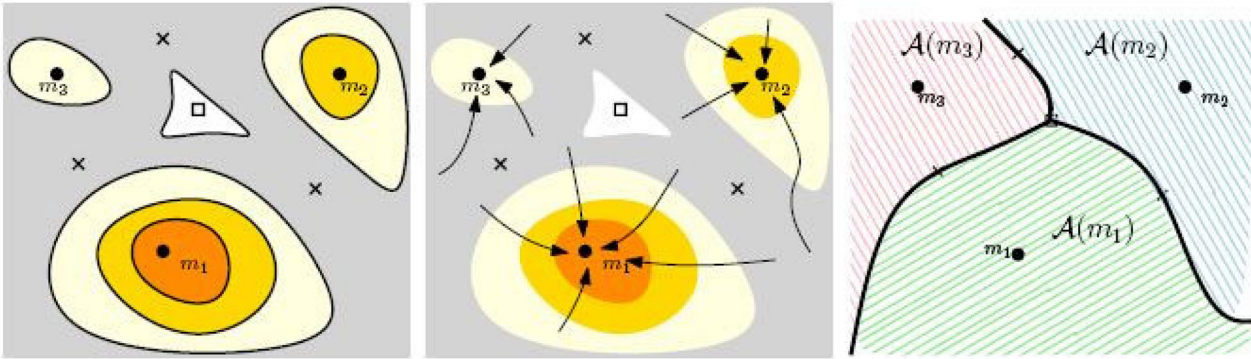


Figure 2. An illustration of critical points and basin of attractions. Left: the colored contours show the level sets of the log-likelihood function. The three sold black dots are the locations of local maxima (m_1, m_2 , and m_3); the crosses are the locations of saddle points; and the empty box indicates the location of a local minimum. Middle: gradient flows with different starting points. Each arrow indicates the gradient flow starting from an initial point that ends at a local maximum. Right: basins of attractions of basins of attraction of local maxima. Note that by the Morse theory, saddle points and local minima will be on the boundary of basins of attraction of local maxima.

where Π is the population version of $\widehat{\Pi}_n$ (i.e., $\widehat{\Pi}_n$ converges to Π in the sense of assumption (A4) in the Section D.1, supplementary material). q_ℓ^π is the probability that the initialization method chooses an initial point within the basin of attraction of m_ℓ . Namely, q_ℓ^π is the chance that the gradient ascent flow converges to m_ℓ from a random initial point. Note that we add a superscript π to q_ℓ^π to emphasize the fact that this quantity depends on how we choose the initialization approach. Varying the initialization method leads to different probabilities q_ℓ^π .

We define a ‘‘cumulative’’ probability of the top N local maxima as $Q_N^\pi = \sum_{\ell=1}^N q_\ell^\pi$, where q_ℓ^π is defined in Equation (2). The quantity Q_N^π plays a key role in our analysis because it is the probability of seeing one of the top N local maxima after applying the gradient ascent method with a single initialization. Note that $q_1^\pi = Q_1^\pi$ is the probability of selecting an initial parameter value within the basin of attraction of the MLE, which is also the probability of obtaining the MLE with only one initialization. Later we will give a bound on the number of initializations we need in order to obtain the MLE with a high probability (Proposition 1).

Because the estimator $\widehat{\theta}_{n,M}$ is constructed by M initializations, we introduce a population version of it. Let $\theta_1^\dagger, \dots, \theta_M^\dagger \sim \Pi$ be the initial points and let $\gamma_{\theta_1^\dagger}(\infty), \dots, \gamma_{\theta_M^\dagger}(\infty)$ be the corresponding destinations. The quantity

$$\bar{\theta}_M = \operatorname{argmax}_{\gamma_{\theta_\ell^\dagger}(\infty)} \left\{ L \left(\gamma_{\theta_\ell^\dagger}(\infty) \right) : \ell = 1, \dots, M \right\}.$$

is the population analog of $\widehat{\theta}_{n,M}$.

Due to the fact that $\bar{\theta}_M$ is constructed by M initializations, it may not be the population MLE θ_{MLE} . However, it is still the best among all these M candidates so it should be one of the top local maxima in terms of the likelihood value. Let $C_N = \{m_1, \dots, m_N\}$ be the top N local maxima, where $N \leq K$ and $L(m_1) \geq \dots \geq L(m_K)$. By simple algebra, we have

$$P(\bar{\theta}_M \in C_N) = 1 - P(\bar{\theta}_M \notin C_N) = 1 - (1 - Q_N^\pi)^M.$$

Given any fixed number N , such a probability converges to 1 as $M \rightarrow \infty$ when Π covers the basin of attraction of every local maximum. Therefore, we can pick $N = 1$ and choose M sufficiently large to ensure that we obtain the MLE with

an overwhelming probability. However, when M is finite, the chance of obtaining the population MLE could be slim.

To acknowledge the effect from the initializing M times, we introduce a new quantity called the *precision level*, denoted as $\delta > 0$. Given a precision level δ , we define an integer

$$N_{M,\delta}^\pi = \min\{N : (1 - Q_N^\pi)^M \leq \delta\}$$

that can be interpreted as: with a probability of at least $1 - \delta$, $\bar{\theta}_M$ is among the top $N_{M,\delta}^\pi$ local maxima. We further define

$$C_{M,\delta}^\pi = C_{N_{M,\delta}^\pi}^\pi, \quad (3)$$

which satisfies

$$P(\bar{\theta}_M \in C_{M,\delta}^\pi) \geq 1 - \delta.$$

Namely, with a probability of at least $1 - \delta$, $\bar{\theta}_M$ recovers one element of $C_{M,\delta}^\pi$. We often want δ to be small because later we will show that common CIs have an asymptotic coverage $1 - \alpha - \delta$ containing an element of $C_{M,\delta}^\pi$ (Section 3.1). If we want to control the Type I error to be, say 5%, we may want to choose $\alpha = 2.5\%$ and $\delta = 2.5\%$.

Example 1 (Modal regression). To illustrate the idea, consider a regression problem where we observe $(X_1, Y_1), \dots, (X_n, Y_n)$ and the goal is to fit a linear model of the conditional mode of Y given X . This problem is called linear modal regression (Yao and Li 2014; Chen 2018; Feng et al. 2020). Suppose that our data is generated by the following mixture model (intercept is 0):

$$Y = \theta X + \epsilon, \quad X \sim \text{Uni}[0, 1], \quad \epsilon \sim N(0, 0.05^2).$$

However, θ is also random such that $P(\theta = 0) = 0.6, P(\theta = 1) = 0.2, P(\theta = 2) = P(\theta = 3) = 0.1$. Namely, it is a mixture of four component regression problem and the left panel of Figure 3 shows a scatterplot of the data. It is known that (Chen 2018) if we are looking for the conditional (global) mode of Y given X , we can maximize the following objective function:

$$\ell_n(\theta) = \frac{1}{nh} \sum_{i=1}^n K \left(\frac{Y_i - \theta X_i}{h} \right).$$

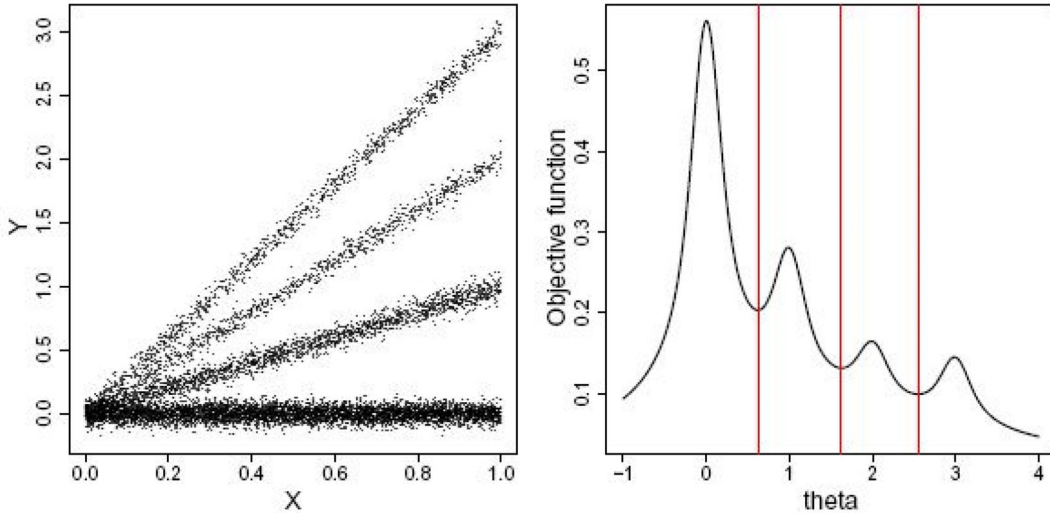


Figure 3. A modal regression method to the mixture regression problem. Left: we display a data generated by a 4-mixture regression model. Right: the objective function of the linear modal regression as a function of the parameter θ . The three red vertical lines display the boundary of basins of attraction of the four local modes.

The right panel plots $\ell_n(\theta)$ when we choose $h = 0.1$ and K to be the Gaussian kernel. The four local modes correspond to the four mixture components. The global mode is $\theta_{MLE} = m_1 = 0$, which corresponds to the component with the highest proportion. There are three other local modes $m_2 = 1, m_3 = 2, m_4 = 3$. The three vertical lines in the right panel indicate the boundaries of basins of attraction of different modes; they correspond to $\theta = 0.62, 1.62, 2.55$. So the basin corresponding to the global mode is $\mathcal{A}(m_1) = (-\infty, 0.62)$.

Suppose that we randomly initialize the starting point within $[-1, 4]$ (i.e., $\Pi \sim \text{Uni}[-1, 4]$), then $q_1^\pi = \Pi((-\infty, 0.62)) = 1.62 \times 0.2 = 0.324$ and $q_2^\pi = 0.2, q_3^\pi = 0.186, q_4^\pi = 0.29$. Therefore, we only have around 32% chance of getting the actual maximizer if we only initialize it once. Suppose that $\delta = 1\%$ and we randomly initialize the program four times, we obtain $N_{4,0.01}^\pi = 3$ so $C_{4,0.01}^\pi = \{m_1, m_2, m_3\}$. To ensure $C_{M,0.01}^\pi = \{m_1\}$, we need at least $M = 12$ random initializations (this corresponds to $(1 - 0.324)^{12} \approx 0.0091 < \delta = 0.01$).

The above example is a particularly simple case (one-dimensional so we can clearly see the landscape of the objective function), so we can work out the minimal number M to guarantee a probability of at least $1 - \delta$ of finding the global mode (i.e., $C_{M,\delta}^\pi = \{\theta_{MLE}\}$). In Section 3.5, we propose a practical rule of choosing M based on our judgment of the problem. In what follows, we provide a theoretical upper bound of the minimal number using the curvature around the global mode. Let ∇_v denotes the directional derivative with respect to $v \in \mathbb{S}_d$, where $\mathbb{S}_d = \{v \in \mathbb{R}^d : \|v\| = 1\}$ is the collection of all unit vectors in d dimensions. When either M or δ increase, the set $C_{M,\delta}^\pi$ may shrink. Under smoothness conditions of L , a sufficiently large M ensures $C_{M,\delta}^\pi = \{\theta_{MLE}\}$ as described in the following proposition.

Proposition 1. Assume that Θ is a compact parameter space. Assume all eigenvalues of $H(\theta_{MLE}) = \nabla \nabla L(\theta_{MLE})$ are less than $-\lambda_0$ for some positive constant λ_0 and

$$\sup_{\theta \in \Theta} \sup_{v_1, v_2, v_3 \in \mathbb{S}_d} |\nabla_{v_1} \nabla_{v_2} \nabla_{v_3} L(\theta)| < c_3.$$

Moreover, assume that θ_{MLE} is unique within Θ . Then for every $\delta > 0$, $P(C_{M,\delta}^\pi = \{\theta_{MLE}\}) \geq 1 - \delta$ when $M \geq \frac{\log \delta}{\log(1 - \Pi(B(\theta_{MLE}, \frac{\lambda_0}{c_3})))}$, where $B(\theta, r) = \{x \in \Theta : \|x - \theta\|_2 \leq r\}$ is the ball centered at θ with a radius r .

Proposition 1 describes a desirable scenario: when M is sufficiently large, with a probability of at least $1 - \delta$ the set $C_{M,\delta}^\pi$ contains only the MLE.

If the uniqueness assumption is violated, that is, there are multiple parameters attaining the maximum value of the likelihood function, then the set $C_{M,\delta}^\pi$ converges to the collection of all these maxima in probability when $M \rightarrow \infty$. One common scenario in which we encounter this situation is in mixture models where permuting some parameters results in the same model (this is known as the label switching problem in Titterton, Smith, and Makov 1985).

2.2. Sample-Level Analysis

In this section, we show that $\hat{\theta}_{n,M}$ converges to an element of $C_{M,\delta}^\pi$ with a probability at least $1 - \delta$. We first introduce some generic assumptions. We define the projection distance $d(x, A) = \inf_{y \in A} \|x - y\|$ for any point x and any set A . For a set A and a scalar r , define $A \oplus r = \{x : d(x, A) \leq r\}$.

We start with a useful lemma which states that with a high probability (a probability tending to 1 as the sample size increases), the local maxima of \hat{L}_n and the local maxima of L have a one-to-one correspondence. Denote the collection of local maxima of \hat{L}_n as $\hat{\mathcal{C}} = \{\hat{m}_1, \dots, \hat{m}_{\hat{K}}\}$ such that $\hat{L}_n(\hat{m}_1) \geq \dots \geq \hat{L}_n(\hat{m}_{\hat{K}})$, where \hat{K} is the number of local maxima of \hat{L}_n . Note that by definition, $\hat{m}_1 = \hat{\theta}_{MLE}$. Let

$$\epsilon_{1,n} = \sup_{\theta \in \Theta} \left\| \nabla \hat{L}_n(\theta) - \nabla L(\theta) \right\|_{\max},$$

$$\epsilon_{2,n} = \sup_{\theta \in \Theta} \left\| \nabla \nabla \hat{L}_n(\theta) - \nabla \nabla L(\theta) \right\|_{\max},$$

be the bounds on gradient and Hessian.

Lemma 1. Assume (A1) and (A3) in Section D.1, supplementary material. Then there exists a constant C_0 such that when $\epsilon_{1,n}, \epsilon_{2,n} < C_0$, $\widehat{K} = K$ and for every $\ell = 1, \dots, K$,

$$\|\widehat{m}_\ell - m_\ell\| < \min \left\{ \min_{j \neq \ell} \|\widehat{m}_\ell - m_j\|, \min_{j \neq \ell} \|\widehat{m}_j - m_\ell\| \right\}.$$

This result appears in many places in the literature so we will omit the proof in this exposition. Interested readers are encouraged to consult Chazal et al. (2017), Mei, Bai, and Montanari (2018), and Chen et al. (2017). Note that in most cases, we in fact have a stronger result than what Lemma 1 suggests—not only is there a one-to-one correspondence between a pair of estimated and population local maxima, but also between pairs of other types of critical points.

The following theorem provides a bound on the distance from $\widehat{\theta}_{n,M}$ to $\mathcal{C}_{M,\delta}^\pi$.

Theorem 1. Assume (A1–A4) in Section D.1, supplementary material and let $\epsilon_{3,n}, \epsilon_{4,n}$ be the two bounds in Assumption (A4) in Section D.1, supplementary material. Let C_0 be the constant in Lemma 1 and

$$\xi_n = P(\epsilon_{1,n} \geq C_0 \text{ or } \epsilon_{2,n} \geq C_0).$$

When $\epsilon_{1,n}, \dots, \epsilon_{4,n}$ are nonrandom, with a probability of at least $1 - \delta - \xi_n + O(\epsilon_{1,n} + \epsilon_{3,n} + \epsilon_{4,n})$,

$$d(\widehat{\theta}_{n,M}, \mathcal{C}_{M,\delta}^\pi) = O(\epsilon_{1,n}).$$

Moreover, when $\epsilon_{1,n}, \dots, \epsilon_{4,n}$ are random and there exists $\eta_{n,j}(t)$ such that

$$P(\epsilon_{j,n} > t) \leq \eta_{n,j}(t) \quad \text{for } j = 1, \dots, 4,$$

then for any sequence $t_n \rightarrow 0$, with a probability of at least $1 - \delta - \xi_n + O(t_n) - \sum_{j=1,3,4} \eta_{n,j}(t_n)$,

$$d(\widehat{\theta}_{n,M}, \mathcal{C}_{M,\delta}^\pi) = O_P(\epsilon_{1,n}).$$

In the first claim ($\epsilon_{1,n}, \dots, \epsilon_{4,n}$ are nonrandom), the probability comes from the randomness of initializations. In the second claim ($\epsilon_{1,n}, \dots, \epsilon_{4,n}$ are random), the probability statement accounts for both the randomness of initializations and $\epsilon_{1,n}, \dots, \epsilon_{4,n}$.

In many applications, the probability ξ_n is very small because that statement is true when both $\epsilon_{1,n}, \epsilon_{2,n}$ are less than a fixed threshold (see Lemma 16 of Chazal et al. 2017). Further, the chance that these two quantities are less than a fixed number has a probability of $1 - e^{-C \cdot a_n}$ for some $a_n \rightarrow \infty$ as $n \rightarrow \infty$ and $C > 0$ so often ξ_n can be ignored.

When we made further assumptions of the likelihood function to obtain a \sqrt{n} rate (assumptions (A3L) and (A4L) in Section D.1, supplementary material), we have the following result on a concrete rate.

Theorem 2. Assume (A1), (A2) (A3L), and (A4L) in Section D.1, supplementary material. Then when $n \rightarrow \infty$, with a probability of at least $1 - \delta - O\left(\sqrt{\frac{\log n}{n}}\right)$,

$$d(\widehat{\theta}_{n,M}, \mathcal{C}_{M,\delta}^\pi) = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Theorem 2 bounds the distance from the estimator to an element of $\mathcal{C}_{M,\delta}^\pi$ when M initializations are used. Note that if M is sufficiently large (so Proposition 1 holds), then we can claim that $\|\widehat{\theta}_{n,M} - \theta_{\text{MLE}}\| = O_P\left(\frac{1}{\sqrt{n}}\right)$ with a probability around $1 - \delta$.

3. Statistical Inference

In this section we study the procedure of making inferences when the likelihood function has multiple maxima.

To simplify the problem of constructing CIs, we focus on constructing CIs of $\tau_{\text{MLE}} = \tau(\theta_{\text{MLE}})$, where $\tau : \Theta \rightarrow \mathbb{R}$ is a known function. We estimate τ_{MLE} using $\widehat{\tau}_{\text{MLE}} = \tau(\widehat{\theta}_{n,M})$. Recall that $\mathcal{C}_{M,\delta}^\pi$ from Equation (3) is the top local modes that we can discover with a precision level $1 - \delta$ and M initializations and Π initialization method. Moreover, we define

$$\tau(\mathcal{C}_{M,\delta}^\pi) = \{\tau(\theta) : \theta \in \mathcal{C}_{M,\delta}^\pi\}.$$

The set $\tau(\mathcal{C}_{M,\delta}^\pi)$ will be the population quantity that the CIs are covering.

3.1. Normal Confidence Interval

A naive approach to constructing a CI is to estimate the variance of $\tau(\widehat{\theta}_{n,M})$ and invert it into a CI. Such CIs are in one-dimensional space and are based on the asymptotic normality of the MLE (Redner and Walker 1984):

$$\sqrt{n}(\widehat{\theta}_{\text{MLE}} - \theta_{\text{MLE}}) \xrightarrow{D} N(0, \sigma^2),$$

for some $\sigma^2 > 0$. In practice, we only have access to $\widehat{\theta}_{n,M}$, not $\widehat{\theta}_{\text{MLE}}$, so we replace $\widehat{\theta}_{\text{MLE}}$ by $\widehat{\theta}_{n,M}$ and construct a CI using the normality. This is perhaps the most common approach to the construction of a CI and the representation of the error of estimation (see, McLachlan and Peel 2004, chap. 2.15 and 2.16 for examples of mixture models). However, we will show that when the likelihood function has multiple local maxima, this CI undercovers for τ_{MLE} and has $1 - \alpha - \delta$ coverage for an element in $\tau(\mathcal{C}_{M,\delta}^\pi)$.

To fully describe the construction of this normal CI, we begin with an analysis of the asymptotic covariance of the MLE. Let $S(\theta) = \nabla L(\theta)$ be the score function and $H(\theta) = \nabla S(\theta)$ be the Hessian matrix of the log-likelihood function. Moreover, let $S(\theta|X_i) = \nabla \log p(X_i; \theta)$ and $H(\theta|X_i) = \nabla S(\theta|X_i)$. The MLE $\widehat{\theta}_{\text{MLE}}$ has an asymptotic covariance matrix

$$\text{cov}(\widehat{\theta}_{\text{MLE}}) = H(\theta_{\text{MLE}})^{-1} \mathbb{E}(S(\theta_{\text{MLE}}|X_1)S(\theta_{\text{MLE}}|X_1)^T) H(\theta_{\text{MLE}})^{-1} + o(1).$$

Note that under regularity conditions,

$$H(\theta_{\text{MLE}}) = -\mathbb{E}(S(\theta_{\text{MLE}}|X_1)S(\theta_{\text{MLE}}|X_1)^T) = -I(\theta_{\text{MLE}})$$

is the Fisher's information matrix, which further implies $\text{cov}(\widehat{\theta}_{\text{MLE}}) = I^{-1}(\theta_{\text{MLE}})$. However, when the model is misspecified, $H(\theta_{\text{MLE}}) \neq I(\theta_{\text{MLE}}) = \mathbb{E}(S(\theta_{\text{MLE}}|X_1)S(\theta_{\text{MLE}}|X_1)^T)$, and in this case, we cannot use the information matrix to construct a normal CI.

Using the delta method (Van der Vaart 1998; Wasserman 2006), the variance of $\tau(\widehat{\theta}_{\text{MLE}})$ is

$$\text{var}(\tau(\widehat{\theta}_{\text{MLE}})) = g_\tau^T(\theta_{\text{MLE}}) \text{cov}(\widehat{\theta}_{\text{MLE}}) g_\tau(\theta_{\text{MLE}}),$$

where $g_\tau(\theta) = \nabla \tau(\theta)$.

Thus, given an estimator $\hat{\theta}_{n,M}$, we can estimate the covariance matrix using

$$\begin{aligned} \widehat{\text{Cov}}(\hat{\theta}_{n,M}) &= \hat{H}_n(\hat{\theta}_{n,M})^{-1} \left(\frac{1}{n} \sum_{i=1}^n S(\hat{\theta}_{n,M}|X_i) S(\hat{\theta}_{n,M}|X_i)^T \right) \hat{H}_n(\hat{\theta}_{n,M}), \\ \hat{H}_n(\theta) &= \frac{1}{n} \sum_{i=1}^n H(\theta|X_i). \end{aligned} \quad (4)$$

And the CI is

$$C_{n,\alpha} = \left\{ t : \sqrt{n} \left| \frac{t - \tau(\hat{\theta}_{n,M})}{g_\tau^T(\hat{\theta}_{n,M}) \widehat{\text{Cov}}(\hat{\theta}_{n,M}) g_\tau(\hat{\theta}_{n,M})} \right| \leq z_{1-\alpha/2} \right\}, \quad (5)$$

where z_α is the α quantile of a standard normal distribution. Note that under suitable assumptions, one can also use Fisher's information matrix or the empirical information matrix to replace $\widehat{\text{Cov}}(\hat{\theta}_{n,M})$.

However, because $\hat{\theta}_{n,M}$ is never guaranteed to be the MLE, $C_{n,\alpha}$ may not contain the population MLE with the right coverage. In what follows, we show that $C_{n,\alpha}$ has an asymptotic $1 - \alpha - \delta$ coverage of covering an element of $\tau(C_{M,\delta}^\pi)$ and $1 - \alpha - (1 - q_1^\pi)^M$ coverage for covering the MLE, where $q_1^\pi = \Pi(\mathcal{A}(\theta_{\text{MLE}}))$ is defined in Equation (2).

Theorem 3. Assume (A1), (A2), (A3L), (A4L), (A5), and (T) in Sections D.1 and D.2, supplementary material. Then $P(C_{n,\alpha} \cap \tau(C_{M,\delta}^\pi) \neq \emptyset) \geq 1 - \alpha - \delta - O\left(\sqrt{\frac{\log n}{n}}\right)$. Thus, by choosing $\delta = (1 - q_1^\pi)^M$, we have

$$P(\tau_{\text{MLE}} \in C_{n,\alpha}) \geq 1 - \alpha - (1 - q_1^\pi)^M - O\left(\sqrt{\frac{\log n}{n}}\right).$$

The population quantity covered by the normal CI is given by the fact that $C_{n,\alpha}$ has an asymptotic $1 - \alpha - \delta$ coverage for an element of $\tau(C_{M,\delta}^\pi)$. The quantities α and δ play similar roles in terms of coverage but they have different meanings. The quantity α is the conventional confidence level, which aims to control the fluctuation of the estimator. On the other hand, δ is the precision level that corrects for the multiple local optima.

When M is sufficiently large (greater than the bound given in Proposition 1), Proposition 1 guarantees that we asymptotically have at least $1 - \alpha - \delta$ coverage of the population MLE. Equivalently, when δ is sufficiently small ($\delta \leq (1 - q_1^\pi)^M \Rightarrow C_{M,\delta}^\pi = \{\theta_{\text{MLE}}\}$), the first assertion implies the second assertion: $C_{n,\alpha}$ has a coverage of $1 - \alpha - (1 - q_1^\pi)^M$ of containing τ_{MLE} .

3.2. Bootstrap

The bootstrap method (Efron 1982, 1979) is a common approach for constructing a CI. While there are many variants of bootstrap, we focus on the empirical bootstrap with the percentile approach.

When applying a bootstrap approach to an estimator that requires multiple initializations (such as our estimator or the estimator from an EM-algorithm), there is always a question: *How should we choose the initial point for each bootstrap sample?*

Should we rerun the initialization several times to pick the highest value for each bootstrap sample?

Based on the following arguments, we recommend using *the estimator of the original sample, $\hat{\theta}_{n,M}$, as the initial point for every bootstrap sample*. The purpose of using the bootstrap is to approximate the distribution of the estimator $\hat{\theta}_{n,M}$. In the M-estimator theory (Van der Vaart 1998), we know that the variation of $\hat{\theta}_{n,M}$ is caused by the randomness of the function $\hat{L}_n(\theta)$ around $\hat{\theta}_{n,M}$. Thus, to make sure the bootstrap approximates such randomness, we need to ensure that the bootstrap estimator $\hat{\theta}_{n,M}^*$ is around $\hat{\theta}_{n,M}$ so the distribution of $\hat{\theta}_{n,M}^* - \hat{\theta}_{n,M}$ approximates the distribution of $\hat{\theta}_{n,M} - \theta^\dagger$ for some $\theta^\dagger \in \mathcal{C}$. By Lemma 1, we know that there is a local maximum $\hat{\theta}^*$ of the bootstrap log-likelihood function that is close to $\hat{\theta}_{n,M}$. Therefore, we need that the initial point to which we apply the gradient ascent method in the bootstrap sample is within the basin of attraction of $\hat{\theta}^*$ asymptotically (with a probability tending to 1 when the sample size $n \rightarrow \infty$). Because $\hat{\theta}_{n,M}$ is close to $\hat{\theta}^*$, $\hat{\theta}_{n,M}$ will be within the basin of attraction of $\hat{\theta}^*$ asymptotically; as a result, $\hat{\theta}_{n,M}$ is a good initial point for the bootstrap sample.

Moreover, using the same initial point in every bootstrap sample avoids the problem of label switching (Redner and Walker 1984). Label switching occurs when the distribution function is the same after permuting some parameters. For instance, in a Gaussian mixture model with equal variance and proportion, permuting the location parameters leads to the same model. When we use the same initial point in every bootstrap sample, we alleviate this problem.

Now we describe the formal bootstrap procedure. Let X_1^*, \dots, X_n^* be a bootstrap sample. We first calculate the bootstrap log-likelihood function \hat{L}_n^* . Next, we start a gradient ascent flow from the initial point $\hat{\theta}_{n,M}$. The gradient ascent flow leads to a new local maximum, denoted as $\hat{\theta}_{n,M}^*$. By evaluating the function $\tau(\cdot)$ at this new local maximum, we obtain a bootstrap estimate of the parameter of interest, $\tau(\hat{\theta}_{n,M}^*)$. We repeat the above procedure many times and construct a CI using the upper and lower $\alpha/2$ quantile of the distribution of $\tau(\hat{\theta}_{n,M}^*)$. Namely, let

$$\hat{\omega}_{1-\alpha} = \hat{G}^{-1}(1 - \alpha), \quad \hat{G}_\omega(s) = P(\tau(\hat{\theta}_{n,M}^*) \leq s | X_1, \dots, X_n).$$

The CI is $\hat{C}_{n,\alpha}^* = [\hat{\omega}_{\alpha/2}, \hat{\omega}_{1-\alpha/2}]$. Algorithm 2 outlines the procedure of this bootstrap approach.

A benefit of this CI is that $\hat{C}_{n,\alpha}^*$ does not require any knowledge about the variance of $\hat{\theta}_{n,\alpha}$. When the variance is complicated or does not have a closed form, being able to construct a CI without knowledge of the variance makes this approach particularly appealing.

Theorem 4. Assume (A1), (A2), (A3L), (A4L), (A5), and (T) in Sections D.1 and D.2, supplementary material. Let $\hat{C}_{n,\alpha}^*$ be defined as the above. Then

$$P(\hat{C}_{n,\alpha}^* \cap \tau(C_{M,\delta}^\pi) \neq \emptyset) \geq 1 - \alpha - \delta - O\left(\sqrt{\frac{\log n}{n}}\right).$$

Therefore, by choosing $\delta = (1 - q_1^\pi)^M$, where q_1^π is defined in Theorem 3, we conclude that

$$P(\tau_{\text{MLE}} \in \hat{C}_{n,\alpha}^*) \geq 1 - \alpha - (1 - q_1^\pi)^M - O\left(\sqrt{\frac{\log n}{n}}\right).$$

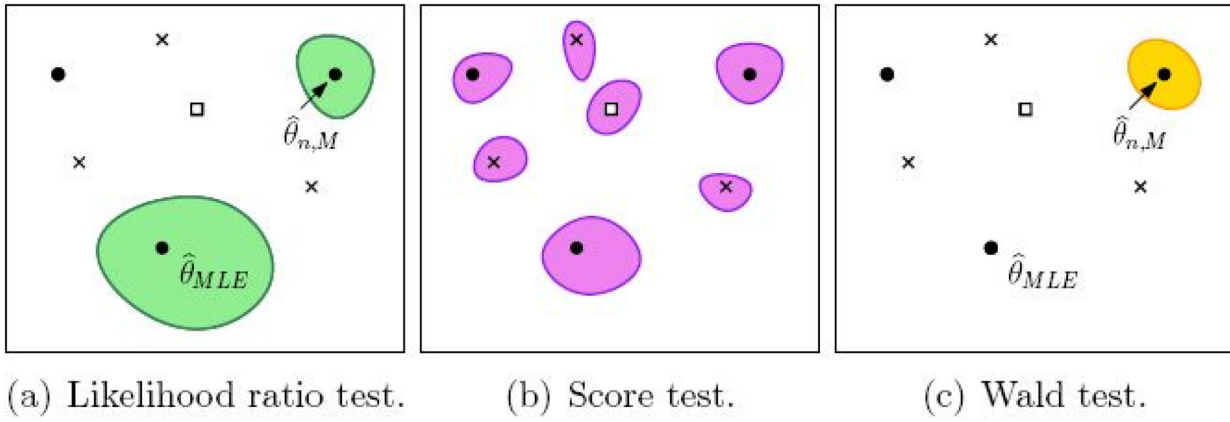


Figure 4. Illustration of CIs from inverting a test. The black dots are local maxima of the estimated likelihood function. The black crosses are saddle points. The empty box is a local minimum. Assume that the estimator we compute, $\hat{\theta}_{n,M}$, is the local maximum in the top-right corner and the actual MLE is the local maximum at the bottom. Left: the CI (green areas) from the likelihood ratio test. This CI contains not only the regions around our estimator but also regions around the actual MLE. Middle: the CI (purple areas) from the score test. This CI contains regions around each critical point because the gradient around every critical point is close to 0. Right: The CI (yellow area) from the Wald test. This CI will be an ellipsoid around the estimator $\hat{\theta}_{n,M}$. Note that this figure is only for the purpose of illustration; it was not created from a real dataset.

Algorithm 2 Percentile bootstrap method

1. Let $\hat{\theta}_{n,M}$ be the output from [Algorithm 1](#).
2. Generate a bootstrap sample and let \hat{L}_n^* denote the bootstrap log-likelihood function.
3. Use $\hat{\theta}_{n,M}$ as the initial point, apply the gradient ascent algorithm to \hat{L}_n^* until it converges. Let $\hat{\theta}_{n,M}^*$ be the convergent.
4. Repeat Step 2 and 3 B times, leading to $\hat{\theta}_{n,M}^{*(1)}, \dots, \hat{\theta}_{n,M}^{*(B)}$. Let $\tau(\hat{\theta}_{n,M}^{*(1)}), \dots, \tau(\hat{\theta}_{n,M}^{*(B)})$ be the corresponding value of the parameter of interest.
5. Compute the quantile

$$\hat{\omega}_{1-\alpha} = \hat{G}_\omega^{-1}(1 - \alpha), \quad \hat{G}_\omega(s) = \frac{1}{B} \sum_{\ell=1}^B I(\tau(\hat{\theta}_{n,M}^{*(\ell)}) \leq s).$$

6. Form the CI as $\hat{C}_{n,\alpha}^* = [\hat{\omega}_{\alpha/2}, \hat{\omega}_{1-\alpha/2}]$.
-

The conclusions of [Theorem 4](#) are similar to those of [Theorem 3](#): under appropriate conditions, with a (asymptotic) coverage $1 - \alpha - \delta$, the CI covers an element of $\tau(\mathcal{C}_{M,\delta}^\pi)$, and with a coverage $1 - \alpha - (1 - q_1^\pi)^M$, the CI covers τ_{MLE} .

Remark 1. There are many other variants of bootstrap approaches and [Algorithm 2](#) describes only a simple one. A common alternative to the method so far presented is bootstrapping the pivotal quantity (also known as the studentized pivotal approach in [Wasserman 2006](#) and the percentile- t approach in [Hall 2013](#)). In certain scenarios, bootstrapping a pivotal quantity leads to a CI with a higher order accuracy (namely, the coverage will be $1 - \alpha - O(\frac{1}{n})$). However, we may not have such a property because the bottleneck of the coverage error $O\left(\sqrt{\frac{\log n}{n}}\right)$ comes from the uncertainty of the basins of attraction. Such uncertainty may not be reduced when using the pivotal approach.

3.3. Confidence Intervals by Inverting a Test

In this section, we introduce three CIs of τ_{MLE} created by inverting hypothesis tests. We consider three famous tests: the likelihood ratio test, the score test, and the Wald test. Although the three tests are asymptotically equivalent in a regular setting (when the likelihood function is concave and smooth), they lead to very different CIs when the likelihood function has multiple local maxima. [Figure 4](#) provides an example illustrating these three CIs of a multi-modal likelihood function.

Because it is easier to invert a test for a CI of θ_{MLE} , we focus on describing the procedure of constructing a CI of θ_{MLE} in this section. With a $1 - \alpha$ CI of θ_{MLE} , say $\hat{\Theta}_{n,\alpha}$, one can easily invert it into a $1 - \alpha$ CI of τ_{MLE} by

$$\hat{\mathcal{T}}_{n,\alpha} = \{\tau(\theta) : \theta \in \hat{\Theta}_{n,\alpha}\}. \quad (6)$$

3.3.1. Likelihood Ratio Test

One classical approach to inverting a test to a CI is to use the likelihood ratio test ([Owen 1990](#)). Such a CI is also called a likelihood region in [Kim and Lindsay \(2011\)](#).

Under appropriate conditions, the likelihood ratio test implies

$$2n(\hat{L}_n(\hat{\theta}_{MLE}) - \hat{L}_n(\theta_{MLE})) \xrightarrow{d} \chi_d^2,$$

where χ_k^2 is a χ^2 distribution with k degrees of freedom and d is the dimension of the parameter. This motivates a $1 - \alpha$ CI of θ_{MLE} of the form

$$\Theta_{n,\alpha}^0 = \{\theta : 2n(\hat{L}_n(\hat{\theta}_{MLE}) - \hat{L}_n(\theta)) \leq \zeta_{d,1-\alpha}\},$$

where $\zeta_{d,1-\alpha}$ is the $1 - \alpha$ quantile of χ_d^2 .

In practice, we do not know the actual MLE $\hat{\theta}_{MLE}$ and have only the estimator $\hat{\theta}_{n,M}$. Therefore, we replace $\hat{\theta}_{MLE}$ by $\hat{\theta}_{n,M}$, leading to a CI

$$\hat{\Theta}_{n,\alpha} = \{\theta : 2n(\hat{L}_n(\hat{\theta}_{n,M}) - \hat{L}_n(\theta)) \leq \zeta_{d,1-\alpha}\}. \quad (7)$$

The CI $\widehat{\Theta}_{n,\alpha}$ has asymptotic $1 - \alpha$ coverage for θ_{MLE} , regardless of whether or not $\widehat{\theta}_{n,M}$ equals to $\widehat{\theta}_{\text{MLE}}$ because $\widehat{L}_n(\widehat{\theta}_{n,M}) \leq \widehat{L}_n(\widehat{\theta}_{\text{MLE}})$ implies $\widehat{\Theta}_{n,\alpha} \supset \Theta_{n,\alpha}^0$. Because the set $\Theta_{n,\alpha}^0$ is a CI with asymptotic $1 - \alpha$ coverage of θ_{MLE} , $\widehat{\Theta}_{n,\alpha}$ also enjoys this property. Thus, even when we only have a small number of initializations, the CI in Equation (7) has the asymptotic (in terms of sample size) coverage.

The CI $\widehat{\Theta}_{n,\alpha}$ can be used to carry out a hypothesis test. Consider testing the null hypothesis

$$H_0 : \tau_{\text{MLE}} = \tau_0 \subset \mathbb{R}. \tag{8}$$

We can simply check if the set $\tau(\widehat{\Theta}_{n,\alpha})$ and τ_0 intersects or not to decide if we can reject the null hypothesis. This controls the Type I error asymptotically.

Although $\widehat{\Theta}_{n,\alpha}$ is valid regardless of the number M , it is often very conservative and computationally intractable. When $\widehat{\theta}_{n,M}$ is not $\widehat{\theta}_{\text{MLE}}$, the set $\widehat{\Theta}_{n,\alpha}$ is often nonconcave and composed of many disjoint regions, each of which corresponds to a local mode of \widehat{L}_n with a likelihood value greater than $\widehat{\theta}_{n,M}$. See the left panel of Figure 4 for an illustration. Moreover, we do not know the exact locations of other regions because they correspond to the local modes whose basins of attraction contain no initial points when we apply the gradient ascent method.

Remark 2. Although the number M does not affect the coverage of $\widehat{\Theta}_{n,\alpha}$, it does affect the size of $\widehat{\Theta}_{n,\alpha}$. The higher log-likelihood value of the estimator $\widehat{\theta}_{n,M}$, the smaller $\widehat{\Theta}_{n,\alpha}$. This is because the CI includes all parameters whose likelihood values are greater than or equal to $\widehat{L}_n(\widehat{\theta}_{n,M}) - \frac{1}{2n}\zeta_{d,1-\alpha}$. Thus, increasing M does improve the CI, but not in the sense of coverage. This is a distinct feature compared to the bootstrap or normal CIs.

3.3.2. Score Test

In addition to the likelihood ratio test, one may invert the score test (Rao 1948) to obtain a CI. The score test is based on the following observation: when $\theta = \theta_{\text{MLE}}$ and the likelihood function is smooth,

$$n \cdot \nabla \widehat{L}_n(\theta)^T \widehat{I}_n(\theta)^{-1} \nabla \widehat{L}_n(\theta) \xrightarrow{d} \chi_d^2, \tag{9}$$

where $\widehat{I}(\theta) = \frac{1}{n} \sum_{i=1}^n S(\theta|X_i)S(\theta|X_i)^T$ is the observed Fisher's information matrix. Thus, we can construct a CI of θ_{MLE} via

$$\{\theta : n \cdot \nabla \widehat{L}_n(\theta)^T \widehat{I}_n(\theta)^{-1} \nabla \widehat{L}_n(\theta) \leq \zeta_{d,1-\alpha}\}$$

and then use it to construct a CI of τ_{MLE} as Equation (6).

Although this CI is an asymptotically valid $1 - \alpha$ CI, it tends to be very large because θ_{MLE} is not the only case in which Equation (9) holds—all critical points, including local minima and saddle points, of $L(\cdot)$ satisfy this equation. Thus, this CI is the collection of regions around critical points, and as such it tends to be a complicated set of large total size. The middle panel of Figure 4 illustrates a CI from the score test. In terms of testing Equation (8), we can use this CI or use the score test because the CI has the right coverage asymptotically.

3.3.3. Wald Test

Another common approach to finding CIs is inverting the Wald test (Wald 1943). It relies on the following fact:

$$n \cdot (\widehat{\theta}_{\text{MLE}} - \theta_{\text{MLE}})^T \widehat{\text{cov}}(\widehat{\theta}_{\text{MLE}})^{-1} (\widehat{\theta}_{\text{MLE}} - \theta_{\text{MLE}}) \xrightarrow{d} \chi_d^2.$$

By the above property, a CI of θ_{MLE} is

$$\{\theta : n \cdot (\widehat{\theta}_{\text{MLE}} - \theta)^T \widehat{\text{cov}}(\widehat{\theta}_{\text{MLE}})^{-1} (\widehat{\theta}_{\text{MLE}} - \theta) \leq \zeta_{d,1-\alpha}\},$$

where $\widehat{\text{cov}}$ is defined in Equation (4).

Because we do not have $\widehat{\theta}_{\text{MLE}}$ but only $\widehat{\theta}_{n,M}$, we use

$$\{\theta : n \cdot (\widehat{\theta}_{n,M} - \theta)^T \widehat{\text{cov}}_n(\widehat{\theta}_{n,M})^{-1} (\widehat{\theta}_{n,M} - \theta) \leq \zeta_{d,1-\alpha}\}$$

as the CI. By construction, this CI is an ellipsoid; see the right panel of Figure 4 for an illustration.

The CI that results from this inversion will be asymptotically the same as the normal CI so it has the same coverage property. Namely, the CI has asymptotic $1 - \alpha - \delta$ coverage for covering one element of $C_{M,\delta}^\pi$ and $1 - \alpha - (1 - q_1^\pi)^M$ coverage for containing θ_{MLE} .

Note that unlike the two previous CIs constructed from inverting tests that can be applied to testing the null hypothesis in Equation (8), this CI may not control Type I error because it does not have the asymptotic coverage of covering θ_{MLE} . The same issue also occurs in the normal CI $C_{n,\alpha}$ and the bootstrap CI $C_{n,\alpha}^*$.

Compared to the other two tests, the Wald test leads to a CI that can be represented easily—it is an ellipsoid around $\widehat{\theta}_{n,M}$. If we make further use of Equation (6) to construct a CI of τ_{MLE} , the result is an interval centered at the estimator $\tau(\widehat{\theta}_{n,M})$ so the CI can be succinctly expressed as the estimator plus and minus the standard error.

3.4. Two-Sample Test

We now explain how to do a two-sample test using a multimodal likelihood function. In a two-sample test, we observe two sets of data $X_1, \dots, X_n \sim P_X$ and $Y_1, \dots, Y_m \sim P_Y$ and we would like to test if the two datasets are from the same distribution. That is, the null hypothesis being tested is

$$H_0 : P_X = P_Y. \tag{10}$$

A common way of testing (10) is to fit a parametric model $P(\cdot; \theta)$ to both samples and then compare the fitted parameters. An advantage of this approach is that we can interpret the results based on the likelihood model. When rejecting H_0 , we not only know that H_0 is not feasible, but also are able to describe the degree of difference between the two datasets by comparing their corresponding parameters.

Let $L_X(\theta) = \mathbb{E} \log p(X_1; \theta)$ and $L_Y(\theta) = \mathbb{E} \log p(Y_1; \theta)$ be the likelihood functions from the two populations. The null hypothesis in Equation (10) implies

$$H_0 : L_X = L_Y. \tag{11}$$

Because this equality is derived from Equation (10), rejecting the null hypothesis in Equation (11) implies that the null hypothesis in Equation (10) should be also rejected.

A naive idea of how to test Equation (11) is to compute the MLEs in both samples and then compare the MLEs to determine the significance. This method implicitly assumes that we can compute the actual MLEs. Indeed, the H_0 in Equation (11) implies that the two MLEs should be the same so we can directly test the locations of MLEs. However, when L_X or L_Y is multimodal, our estimators could be local maxima rather than the

Algorithm 3 Two-sample test without the MLE

1. Pool both samples together, to form a joint sample $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$.
2. Fit the log-likelihood model to this joint sample and apply the gradient ascent algorithm with a random initialization to find a local maximum.
3. Iterate the above procedure M times and then select from among the local maxima that has the highest log-likelihood value. Denoted this local maximum by $\hat{\theta}_{\text{opt}}$.
4. Now for each of the two samples, fit the likelihood function and apply the gradient ascent algorithm with initial point being $\hat{\theta}_{\text{opt}}$. Let $\hat{\theta}_X$ and $\hat{\theta}_Y$ denote the destination of each of the two samples, respectively.
5. Compare $\hat{\theta}_X$ and $\hat{\theta}_Y$ using conventional two-sample test techniques.

MLEs. Thus, the two estimators may be very different even if H_0 (in Equation (11)) is true because the estimators happen to be different local maxima.

To ensure that the two estimators converge to the same destination when the null hypothesis H_0 is true, the two estimators must be estimating the same local maximum. A simple way is to choose the same initial point in both samples. We therefore, recommend the procedure in [Algorithm 3](#).

In the first step, we combine the data from the two samples because under H_0 , combining them gives us the largest sample from the population. The second and the third steps are the same as the algorithm described in [Algorithm 1](#) to the pooled sample. The resulting estimator should be an estimator with a high likelihood value by [Theorem 2](#). Under H_0 , this estimator should also have a high value in terms of L_X and L_Y . Moreover, because $\hat{\theta}_{\text{opt}}$ is a local maximum of the pooled likelihood function and H_0 implies that L_X , L_Y , and pooled likelihood function are all the same, $\hat{\theta}_{\text{opt}}$ should be close to both the local maximum of L_X and the local maximum of L_Y that correspond to the same local maximum of the underlying population likelihood function. Thus, both $\hat{\theta}_X$ and $\hat{\theta}_Y$ are close to the same local maximum of the underlying population likelihood function, so a comparison between them would control the Type I error (asymptotically).

We do not specify how to compare $\hat{\theta}_X$ and $\hat{\theta}_Y$ because there are many ways to perform this comparison. For instance, we can compare them by constructing their CIs and determining if the two CIs intersect. Or we can do a permutation test where the test statistics are some particular distance between them, for example, $T_1 = \|\hat{\theta}_X - \hat{\theta}_Y\|$.

3.5. A Practical Procedure of Choosing M

As is shown in the above analysis, the choice of M plays a key role in the coverage of a confidence set. Here we propose a practical procedure to choose M based on the analyst's judgment about q_1^π .

We first pick a the precision level δ . A simple rule is to choose $\delta = 1\%$ when the significance level $\alpha = 5\%$ or 10% . Then we hypothesize a threshold q^* such that we believe that $q_1^\pi \geq q^*$.

Namely, we assume that the chance of initializing in the MLEs basin of attraction is no smaller than q^* .

Under this threshold, to ensure the coverage deficiency is less than δ , we need

$$(1 - q^*)^M \leq \delta \Rightarrow M \geq \frac{\log \delta}{\log(1 - q^*)} = M^*(\delta, q^*).$$

When δ and q^* are given, the number of initializations needed is $M = M^*(\delta, q^*)$.

Under $\delta = 0.01(1\%)$, the above threshold becomes $M^*(0.01, q^*) \approx \frac{4.6}{|\log(1 - q^*)|}$. In the extreme case where $q^* \approx 0$ (i.e., the chance of obtaining the actual MLE is very small), we further have $|\log(1 - q^*)| \approx q^*$, so the above threshold becomes

$$M^*(0.01, q^*) \approx 4.6/q^*. \quad (12)$$

The above threshold provides an easy-to-use reference rule of choosing M . For instance, suppose we believe that the chance of getting the MLE is no smaller than $0.1\%(10^{-3})$, then we need at least $M \geq 4.6 \times 10^3 = 4600$ initializations to ensure the coverage deficiency is less than 1% . The choice of q^* should be determined by the analyst's judgment about the problem.

In practice, when the dimension is large, q^* is often small (see Section A and Table 1 in supplementary material), so we need a large number of initializations to control this uncertainty. However, the threshold in Equation (12) is independent of the dimension as long as q^* is fixed. This implies that if we can design a method (such as using a strongly convex penalty) such that $q^* = q_d^* \rightarrow q_0^*$ when $d \rightarrow \infty$ and q_0^* is not a tiny number, the bound $M^*(\delta, q^*)$ can be small. For instance, if $q_0^* = 0.1$ and $\delta = 0.01$, we only need about 46 initializations even if the dimension d is large.

While the above analysis assumes q^* to be fixed and nonrandom, this analysis can be applied to the case where $q^* = q_n^*$ is random and its distribution depends on n as well. As long as we have a concentration bound such that $P(q_n^* > q_{\dagger, \delta}^*) > 1 - \delta/2$ for some fixed quantity $q_{\dagger, \delta}^*$, we can plug $q_{\dagger, \delta}^*$ into (12) and use the revised bound $M^*(\delta/2, q_{\dagger, \delta}^*)$ as the minimal number of initializations needed (we use $\delta/2$ to account for the randomness of q_n^*).

4. EM-Algorithm

In this section, we use the above framework to analyze estimators obtained from the EM-algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Peel 2004; McLachlan and Krishnan 2007). For simplicity, we consider a latent variable model, assuming that our observations are iid random variables X_1, \dots, X_n from some unknown distribution and each individual has a latent variable Z . Namely, our dataset consists of pairs $(X_1, Z_1), \dots, (X_n, Z_n)$ that are iid from an unknown distribution function P_0 but the Z_1, \dots, Z_n are unobserved.

With the latent variable, we assume that the density of (X, Z) forms a parametric model $p_\theta(X, Z)$, where $\theta \in \Theta$ is the underlying parameter. We define

$$L(\theta|X, Z) = \log p_\theta(X, Z),$$

$$L(\theta|X) = \mathbb{E}(L(\theta|X, Z)|X),$$

$$L(\theta) = \mathbb{E}(L(\theta|X)) = \mathbb{E}(L(\theta|X, Z)).$$

The function $L(\theta)$ is the population log-likelihood function and its sample estimator is $\widehat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n L(\theta|X_i)$. Under this model, the population MLE and sample MLE are

$$\theta_{\text{MLE}} = \operatorname{argmax}_{\theta \in \Theta} L(\theta), \quad \widehat{\theta}_{\text{MLE}} = \operatorname{argmax}_{\theta \in \Theta} \widehat{L}_n(\theta).$$

To describe the EM-algorithm, we follow the notations of Balakrishnan, Wainwright, and Yu (2017) and define $Q(\theta|\theta') = \mathbb{E}(Q(\theta|\theta', X))$ and $\widehat{Q}_n(\theta|\theta') = \frac{1}{n} \sum_{i=1}^n Q(\theta|\theta', X_i)$, where

$$Q(\theta|\theta', X) = \int p_{\theta'}(z|X)L(\theta|X, z)dz.$$

Given an initial parameter $\theta^{(0)}$, the population EM-algorithm updates it by

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta \in \Theta} Q(\theta|\theta^{(t)})$$

for $t = 0, 1, 2, 3, \dots$. When applied to data, the sample EM-algorithm uses the following update

$$\widehat{\theta}^{(t+1)} = \operatorname{argmax}_{\theta \in \Theta} \widehat{Q}_n(\theta|\theta^{(t)}).$$

It is known that under smoothness conditions and good initializations (Titterton, Smith, and Makov 1985; McLachlan and Peel 2004; McLachlan and Krishnan 2007), the stationary point (also called the destination) satisfying the following conditions:

$$\theta^{(\infty)} = \lim_{t \rightarrow \infty} \theta^{(t)} = \theta_{\text{MLE}}, \quad \widehat{\theta}^{(\infty)} = \lim_{t \rightarrow \infty} \widehat{\theta}^{(t)} = \widehat{\theta}_{\text{MLE}}.$$

Namely, the EM algorithm leads to the actual MLE.

If the initial point $\theta^{(0)}$ is not well-chosen, the EM-algorithm can converge to a local maximum or a saddle point instead of the MLE (Wu 1983). Therefore, the EM-algorithm is often applied to multiple initial points and the stationary point with the highest likelihood value is used as the final estimator. In this case, the estimator can be viewed as the one generated by the method described in Algorithm 1 with the gradient ascent method being replaced by the EM-algorithm. Let $\widehat{\theta}_{n,M}^{\text{EM}}$ be the stationary point with the highest likelihood value after M initializations from $\widehat{\Pi}_n$. Note that we ignore the algorithmic error by assuming that for each initial point, we run the EM-algorithm until it converges (the algorithmic error of EM-algorithm has been studied in Balakrishnan, Wainwright, and Yu 2017). By viewing the initialization procedure as choosing the starting points from a distribution $\widehat{\Pi}_n$, we fall into the same framework as described in Section 2. As a result, the set of top local modes with $1 - \delta$ precision level and M initializations, $C_{M,\delta}^{\pi}$, is well-defined.

Although we can attest that $C_{M,\delta}^{\pi}$ is well-defined, it is unclear how to analyze the stability of the basin of attraction of the EM-algorithm, so we cannot develop a theoretical guarantee for inferring $C_{M,\delta}^{\pi}$ as we had in Theorem 2. However, we are at least able to determine that a ball centered at θ_{MLE} with a sufficiently small radius will be within the basin of attraction of the MLE when the function $Q(\theta|\theta')$ is sufficiently smooth (Balakrishnan, Wainwright, and Yu 2017). We use this fact to bound the estimator $\widehat{\theta}_{n,M}^{\text{EM}}$ and the population MLE θ_{MLE} .

Theorem 5. Assume (A3L) and (EM1–4) in Sections D.1 and D.3, supplementary material. Define

$$q_{\text{EM}} = \frac{1}{2} \cdot \Pi \left(B \left(\theta_{\text{MLE}}, \frac{r_0}{3} \right) \right).$$

Then when $n \rightarrow \infty$, there exist positive numbers c_1 and c_2 such that

$$P \left(\widehat{\theta}_{n,M}^{\text{EM}} = \widehat{\theta}_{\text{MLE}} \in \mathcal{A}^{\text{EM}}(\theta_{\text{MLE}}) \right) \geq 1 - (1 - q_{\text{EM}})^M - \eta_n(q_{\text{EM}}) - c_1 e^{-c_2 n},$$

where $\eta_n(t)$ is a concentration bound in (EM4) that is often in the form of $\eta_n(t) = A_1 e^{-A_2 n t^2}$ for some fixed constant $A_1, A_2 > 0$.

Theorem 5 shows that the EM-algorithm recovers the MLE with a probability of at least $1 - (1 - q_{\text{EM}})^M - \eta_n(q_{\text{EM}}) - c_1 e^{-c_2 n}$. Note that both $\eta_n(q_{\text{EM}})$ and $c_1 e^{-c_2 n}$ converge to 0 when $n \rightarrow \infty$. Thus, we have a bound on the number of initializations M needed to ensure that we have a good chance of obtaining the MLE $\widehat{\theta}_{\text{MLE}}$.

In addition, Theorem 5 allows us to bound the coverage of a normal CI (Section 3.1) with the estimator computed from the EM algorithm. Let $C_{n,\alpha}^{\text{EM}}$ be the normal CI by replacing $\widehat{\theta}_{n,M}$ by $\widehat{\theta}_{n,M}^{\text{EM}}$ in Equation (5). Namely,

$$C_{n,\alpha}^{\text{EM}} = \left\{ t : \sqrt{n} \left| \frac{t - \tau(\widehat{\theta}_{n,M}^{\text{EM}})}{g_{\tau}^T(\widehat{\theta}_{n,M}^{\text{EM}}) \widehat{\text{cov}}(\widehat{\theta}_{n,M}^{\text{EM}}) g_{\tau}(\widehat{\theta}_{n,M}^{\text{EM}})} \right| \leq z_{1-\alpha/2} \right\},$$

where $g_{\tau}(\theta) = \nabla \tau(\theta)$ and $\widehat{\text{cov}}(\theta)$ is the estimated covariance matrix from Equation (4) (see Section 3.1 for more details).

Theorem 6. Assume (EM1–5), (A3,5), and (T) in Sections D.1, D.2 and D.3, supplementary material. Let q_{EM} be the quantity defined in Theorem 5. Then when $n \rightarrow \infty$,

$$P \left(\tau_{\text{MLE}} \in C_{n,\alpha}^{\text{EM}} \right) \geq 1 - \alpha - (1 - q_{\text{EM}})^M - \eta_n(q_{\text{EM}}) - O \left(\sqrt{\frac{\log n}{n}} \right).$$

Theorem 6 can be proved using Theorems 3 and 5, so we omit the proof in this presentation.

Theorem 6 shows that while we can use the asymptotic normality to construct a CI, we may not have the nominal coverage. If one wants an asymptotic $1 - \alpha$ CI, we can use $C_{n,\alpha/2}^{\text{EM}}$ with $M \geq \frac{\log(\alpha/2)}{\log(1-q_{\text{EM}})}$ because $M \geq \frac{\log(\alpha/2)}{\log(1-q_{\text{EM}})}$ implies $(1 - q_{\text{EM}})^M \leq \alpha/2$ so the coverage of $C_{n,\alpha/2}^{\text{EM}}$ is at least $1 - \alpha/2 - \alpha/2 - O \left(\sqrt{\frac{\log n}{n}} \right)$.

The CI from the bootstrap approach also works and the coverage is similar—the coverage is decreased by $(1 - q_{\text{EM}})^M$. One can also invert a testing procedure to a CI as described in Section 3.3; the behaviors of the three CIs are similar to the ones in Section 3.3—the likelihood ratio test gives a CI that is asymptotically valid regardless of M ; the score test gives an asymptotically valid CI but tends to be very large; and the Wald test gives a CI whose asymptotic coverage is $1 - \alpha - (1 - q_{\text{EM}})^M$.

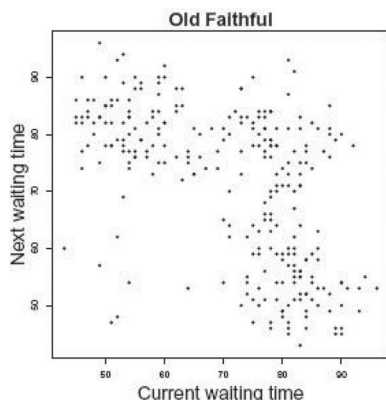
Remark 3. The probability bound in [Theorem 5](#) is a conservative lower bound because the basin of attraction $\mathcal{A}^{\text{EM}}(\theta_{\text{MLE}})$ can be much larger than the ball $B(\theta_{\text{MLE}}, r_0/3)$. To improve the bound on the coverage, we need to know the stability of this basin because the basin of attraction of the sample MLE, $\hat{\mathcal{A}}^{\text{EM}}(\hat{\theta}_{\text{MLE}}) = \{\hat{\theta}^{(0)} : \hat{\theta}^{(\infty)} = \hat{\theta}_{\text{MLE}}\}$, can be different from $\mathcal{A}^{\text{EM}}(\theta_{\text{MLE}})$ and the probability that the EM-algorithm recovers $\hat{\theta}_{\text{MLE}}$ is $\hat{\Pi}(\hat{\mathcal{A}}^{\text{EM}}(\hat{\theta}_{\text{MLE}}))$, not $\Pi(\mathcal{A}^{\text{EM}}(\theta_{\text{MLE}}))$. Therefore, we need to know the asymptotic behavior of $\hat{\Pi}(\hat{\mathcal{A}}^{\text{EM}}(\hat{\theta}_{\text{MLE}}))$ to improve the results in [Theorem 5](#). Intuitively, we expect that the set $\hat{\mathcal{A}}^{\text{EM}}(\hat{\theta}_{\text{MLE}})$ converges to $\mathcal{A}^{\text{EM}}(\theta_{\text{MLE}})$ under some set metrics. However, to our knowledge, such convergence has not yet been established, so we cannot improve the bound in [Theorem 5](#).

5. Real Data: Old Faithful Data

To illustrate the prevalence of the local modes in mixture models, we consider old faithful data that can be obtained by the object `faithful` in R. It is a dataset consisting of $n = 272$ observations of the eruption and waiting time of Old Faithful geyser in Yellowstone National Park. Here we consider two variables: the current waiting time and the next waiting time.

Left panel of [Figure 5](#) shows the scatterplot of the data. We see that clearly there are three major bumps in the data. Thus, we fit a 3-Gaussian mixture model with the package `mixtools` and use the default method for initialization and draw $M = 1000$ initializations. While it seems that the 3-Gaussian mixture should be clear, it turns out that we have more than 20 local modes! This is caused by the outliers in the bottom-right corner of the left panel so the covariance matrices have multiple local modes. The right panel of [Figure 5](#) shows the chance of obtaining one of the 8 local modes corresponding to the top likelihood values.

In this case, the chance of obtaining the MLE is 21%. Suppose we want to reach the precision level $\delta = 1\%$, the number of initialization M has to satisfy $(1 - 0.21)^M \leq 0.01 \Rightarrow M \geq 19.53$. Thus, we need at least $M = 20$ initializations to achieve such precision. Note that the approximation method in [Equation \(12\)](#) leads to $M^*(0.01, 0.21) \approx 4.6/0.21 = 21.9$, which suggests that we need at least $M = 22$ initializations to control the precision level to be 1%.



From the analysis of [Section 3.5](#), the number of initialization needed is $M^*(\delta, q^*) = \frac{\log \delta}{\log(1-q^*)}$, which is logarithmic in the precision level δ . Thus, we can improve the precision without drastically increasing M as long as q^* is not small. To see this, in the old faithful data, if we want to improve precision level from $\delta = 1\%$ to $\delta = 0.1\%$, we only need $M \geq \frac{\log 0.001}{\log(0.79)} = 29.3$, so we only need $M = 30$ initializations. There is no much cost to improve the precision level in this case.

6. Discussion

In this article, we analyzed the performance of an estimator derived from applying a gradient ascent method with multiple initializations. We study the asymptotic theory of such estimator and investigate the properties of the corresponding CIs. In what follows, we discuss possible extensions and future directions.

6.1. Applications and Extensions

6.1.1. Reproducibility

Because the initializations are random, it is nontrivial to “reproduce” the result. Even we use the same dataset and the same estimating procedure, we may not obtain the same estimator. Both the number of initializations M and the initialization method $\hat{\Pi}_n$ affect the realization of the estimator. We should provide details on how we initialized the starting points and how many times the initialization was applied to fully describe how we obtained our results. Unlike a conventional statistical analysis, the statistical model and data alone are not enough for reproducing the results.

Even when all the information above is provided, we still may not obtain an estimator with the same numerical value because of random initializations. A remedy is to report the distribution of the log-likelihood values corresponding to the local maxima discovered from every initialization. If the result is reproducible, other research teams would be able to recover a similar distribution when rerunning the same program. In this case, checking the reproducibility becomes a two-sample test problem as follows. Suppose that another research team obtains N log-likelihood values. If the result is reproducible, then the

Likelihood	Proportion
-2038.44	12.5%
-2036.92	14.4%
-2036.72	6.2%
-2035.79	0.7%
-2035.65	2%
-2033.35	7.6%
-2033.23	1.6%
-2029.62 (MLE)	21%

Figure 5. The old faithful data. Left: the scatterplot of the current waiting time versus the next waiting time. Right: the result of applying EM algorithm to the old faithful data. There are more than 20 local modes and here we only display the results of eight local modes corresponding to the top likelihood values. The proportion indicates the chance of obtaining that local mode from a random initialization (default method in `mixtools`).

new N values and the original M values reported in the literature should be from the same distribution. Thus,

H_0 : The result is reproducible \Leftrightarrow

H_0 : The two samples are from the same distribution.

We can then apply a two-sample test to see if the result is indeed reproducible.

6.1.2. Comparing Initialization Approaches

Our analysis provides two new ways of comparing different initialization approaches. As discussed in Section F, supplementary material, when M is fixed, the only way to reduce the size of $C_{M,\delta}^\pi$ or the coverage loss, $(1 - q_1^\pi)^M$, is to choose a better initialization method ($\hat{\Pi}_n$ and Π). Ideally, we would like to put as much probability mass in the basin of attraction of the actual MLE as possible so that we have a high chance of finding MLE with a small number of M . When M and δ are both fixed, a better initialization approach would have either a smaller set $C_{M,\delta}^\pi$ or a higher value of $q_1^\pi = \Pi(\mathcal{A}(\theta_{\text{MLE}}))$. The simulation study in Section A.1 is based on this idea in comparing three initialization methods.

Supplementary Materials

The online supplementary material contains additional simulations, technical assumptions, proofs of the theorems, and analysis on related topics.

Funding

Supported by NSF grant DMS-1952781, DMS-2112907, DMS-2141808, and NIH grant U24-AG072122.

ORCID

Yen-Chi Chen  <http://orcid.org/0000-0002-4485-306X>

References

- Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017), "Statistical Guarantees for the EM Algorithm: From Population to Sample-based Analysis," *The Annals of Statistics*, 45, 77–120. [2,11]
- Banyaga, A., and Hurtubise, D. (2013), *Lectures on Morse homology* (Vol. 29), Dordrecht: Springer. [2,3]
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [1]
- Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., Rinaldo, A., and Wasserman, L. (2017), "Robust Topological Inference: Distance to a Measure and Kernel Distance," *The Journal of Machine Learning Research*, 18, 5845–5884. [2,6]
- Chen, Y.-C. (2018), "Modal Regression Using Kernel Density Estimation: A Review," *Wiley Interdisciplinary Reviews: Computational Statistics*, 10, e1431. [4]
- Chen, Y.-C., Genovese, C. R., Wasserman, L. (2017), "Statistical Inference Using the Morse-Smale Complex," *Electronic Journal of Statistics*, 11, 1390–1433. [2,6]
- Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38. [2,10]
- Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [7]
- (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia, PA: SIAM. [7]
- Feng, Y., Fan, J., Suykens, J. A. (2020), "A Statistical Learning Approach to Modal Regression," *Journal of Machine Learning Research*, 21, 1–35. [4]
- Genovese, C. R., and Wasserman, L. (2000), "Rates of Convergence for the Gaussian Mixture Sieve," *The Annals of Statistics*, 28, 1105–1127. [1]
- Hall, P. (2013), *The Bootstrap and Edgeworth Expansion*, New York: Springer. [8]
- Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and Jordan, M. I. (2016), "Local Maxima in the Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences," in *Advances in Neural Information Processing Systems*, pp. 4116–4124. [1,2]
- Kim, D., and Lindsay, B. G. (2011), "Comparing Wald and Likelihood Regions Applied to Locally Identifiable Mixture Models," in *Mixtures: Estimation and Applications*, eds. K. L. Mengersen, C. P. Robert and D. M. Titterton, pp. 77–100, New York: Wiley. [8]
- Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016), "Gradient Descent Only Converges to Minimizers," in *Conference on Learning Theory*, pp. 1246–1257. [2,3]
- Li, J. Q., and Barron, A. R. (1999), "Mixture Density Estimation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pp. 279–285. Cambridge, MA: MIT Press. [1]
- McLachlan, G., and Krishnan, T. (2007), *The EM Algorithm and Extensions* (Vol. 382), New York: Wiley. [10,11]
- McLachlan, G., and Peel, D. (2004), *Finite Mixture Models*, New York: Wiley. [1,3,6,10,11]
- Mei, S., Bai, Y., and Montanari, A. (2018), "The Landscape of Empirical Risk for Nonconvex Losses," *The Annals of Statistics*, 46, 2747–2774. [2,6]
- Milnor, J. W. (1963), *Morse Theory* (Vol. 51), Princeton, NJ: Princeton University Press. [2]
- Morse, M. (1930), "The Foundations of a Theory of the Calculus of Variations in the Large in m -Space (Second Paper)," *Transactions of the American Mathematical Society*, 32, 599–631. [2]
- Owen, A. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120. [8]
- Panageas, I., and Piliouras, G. (2017), "Gradient Descent Only Converges to Minimizers: Non-isolated Critical Points and Invariant Regions," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. [2,3]
- Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065–1076. [1]
- Rao, C. R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 44, pp. 50–57. Cambridge, UK: Cambridge University Press. [9]
- Redner, R. (1981), "Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions," *The Annals of Statistics*, 9, 225–228. [2]
- Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239. [1,2,6,7]
- Romano, J. P. (1988a), "Bootstrapping the Mode," *Annals of the Institute of Statistical Mathematics*, 40, 565–586. [1]
- (1988b), "On Weak Convergence and Optimality of Kernel Density Estimates of the Mode," *The Annals of Statistics*, 16, 629–647. [1]
- Sundberg, R. (1974), "Maximum Likelihood Theory for Incomplete Data from an Exponential Family," *Scandinavian Journal of Statistics*, 1, 49–58. [2]
- Titterton, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley. [1,2,5,11]
- Van der Vaart, A. W. (1998), *Asymptotic Statistics* (Vol. 3), Cambridge, UK: Cambridge University Press. [1,6,7]
- Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54, 426–482. [9]
- Wasserman, L. (2006), *All of Nonparametric Statistics*, New York: Springer. [6,8]
- Wu, C. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103. [2,11]
- Yao, W., and Li, L. (2014), "A New Regression Model: Modal Linear Regression," *Scandinavian Journal of Statistics*, 41, 656–671. [4]