Minimax Optimal Rate for Parameter Estimation in Multivariate Deviated Model

NeurIPS 2023

Dat Do[†,⋆], Huy Nguyen[‡,⋆], Khai Nguyen[‡], Nhat Ho[‡]

University of Michigan at Ann Arbor[†]; University of Texas at Austin[‡]

Goals

In this work, we aim to study the parameter estimation rate of the *Multivariate Deviated Model*:

$$p_G(x) = (1 - \lambda)h_0(x) + \lambda f(x|\mu, \Sigma), \quad (1)$$

where

- h_0 is a known density, f is a known family of densities.
- $\lambda \in (0, 1), \mu \in \mathbb{R}^{d_1}, \Sigma \in \mathbb{R}^{d_2 \times d_2}$ are parameters to be estimated.

Motivation

$$p_G(x) = (1-\lambda)h_0(x) + \lambda f(x|\mu, \Sigma),$$

- **Hypothesis testing:** The null hypothesis h_0 and the alternative is p_G. Applications in microarray data analysis.
- **Contaminated model:** h_0 is previously known data distribution, and we want to estimate the contaminated part
- **Domain adaptation:** h_0 is a pre-trained large model estimated from a domain, and f is a low-rank adaptation part to be estimated for a new domain.

Setup, Goals, and Challenges

- Observe n i.i.d. data from

$$p_G(x) = (1-\lambda^*)h_0(x) + \lambda^* f(x|\mu^*, \Sigma^*),$$

  and we get the MLE $\widehat{G}_n = (\widehat{\lambda}_n, \widehat{\mu}_n, \widehat{\Sigma}_n) = \arg\max \sum_{i=1}^{n} \log p_G(x_i)$.
- We want to obtain the optimal uniform rate

$$(\widehat{\lambda}_n, \widehat{\mu}_n, \widehat{\Sigma}_n) \to (\lambda^*, \mu^*, \Sigma^*).$$

- Challenges:
  1. When $\lambda^* \approx 0$, it is harder to estimate (μ^*, Σ^*) (singularity);
  2. In the setting $h_0 = f(x|\mu_0, \Sigma_0)$, it is harder to estimate λ^* when $(\mu^*, \Sigma^*) \approx (\mu_0, \Sigma_0)$ (identifiability)

Uniform rate of convergence

Suppose there is a Machine Learning model $(f_\theta)_{\theta\in\Theta}$

- Data is generated from f_{θ^*} (θ^*: true parameter);
- We obtain an estimator $\widehat{\theta}_n$ from n i.i.d. data.
- How many data to obtain $\epsilon-$error of the estimator? (i.e., $\|\widehat{\theta}_n - \theta^*\| \leq \epsilon$)

Rate of convergence: $\left\|\widehat{\theta}_n - \theta^*\right\| \lesssim C_{\theta^*} \times rate(n)$

Uniform rate of convergence: $\left\|\widehat{\theta}_n - \theta^*\right\| \lesssim C \times rate(n)$, where C does not depend on θ^*.

Main result 1: Distinguishable setting

Theorem 1

Suppose h_0 is linearly independent with $f(\cdot|\mu, \Sigma)$ and its derivatives, for all (μ, Σ). Then,

$$\sup_{G_*} \mathbb{E}_{p_{G*}} \left(\lambda^* \|(\widehat{\mu}_n, \widehat{\Sigma}_n) - (\mu^*, \Sigma^*)\| \right) \lesssim \frac{\log(n)}{\sqrt{n}},$$

$$\sup_{G_*} \mathbb{E}_{p_{G*}} \left(|\widehat{\lambda}_n - \lambda^*| \right) \lesssim \frac{\log(n)}{\sqrt{n}},$$

and this is also the minimax rate.

Main result 2: Non-distinguishable and Strongly identifiable setting

Theorem 2

Suppose $h_0(\cdot) = f(\cdot|\mu_0, \Sigma_0)$*, and the family of densities f with its derivatives up to second-order are linearly independent. Then,*

$$\sup_{G_*} \mathbb{E}_{p_{G*}} \left(\lambda^* \|(\mu^*, \Sigma^*) - (\mu_0, \Sigma_0)\| \|(\widehat{\mu}_n, \widehat{\Sigma}_n) - (\mu^*, \Sigma^*)\| \right) \lesssim \frac{\log(n)}{\sqrt{n}},$$

$$\sup_{G_*} \mathbb{E}_{p_{G*}} \left(\|(\mu^*, \Sigma^*) - (\mu_0, \Sigma_0)\|^2 |\widehat{\lambda}_n - \lambda^*| \right) \lesssim \frac{\log(n)}{\sqrt{n}}.$$

and this is also the minimax rate.

Weak identifiable setting

When $f(x|\mu,\Sigma)$ is the Gaussian distribution, we do not have the strong identifiability since $\dfrac{\partial^2 f(x|\mu,\Sigma)}{\partial\mu\partial\mu^\top} = 2\dfrac{\partial f(x|\mu,\Sigma)}{\partial\Sigma}$

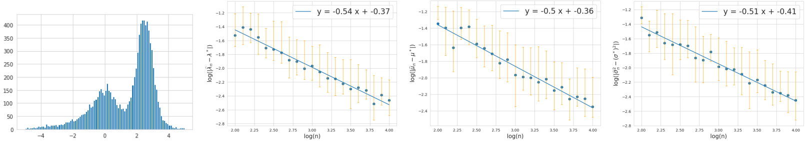
Theorem 3

$$\sup_{G_*} \mathbb{E}_{p_{G*}}\Big((\lambda^*)\left\{\|\mu^*-\mu_0\|^2+\|\Sigma^*-\Sigma_0\|\right\} \times \left\{\|\widehat{\mu}_n-\mu^*\|^2+\|\widehat{\Sigma}_n-\Sigma^*\|\right\}\Big) \lesssim \frac{\log(n)}{\sqrt{n}},$$

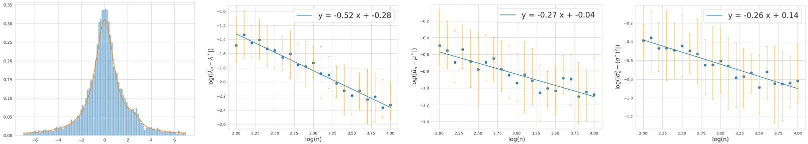
$$\sup_{G_*} \mathbb{E}_{p_{G*}}\Big(\left\{\|\mu^*-\mu_0\|^4+\|\Sigma^*-\Sigma_0\|^2\right\}|\widehat{\lambda}_n-\lambda^*|\Big) \lesssim \frac{\log(n)}{\sqrt{n}}.$$

Simulation study (1): Distinguishable setting

h_0 is a standard Cauchy distribution, and $f(\cdot|\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.



(a) Histogram (b) Rate of $\widehat{\lambda}_n$ (c) Rate of $\widehat{\mu}_n$ (d) Rate of $\widehat{\sigma}^2_n$



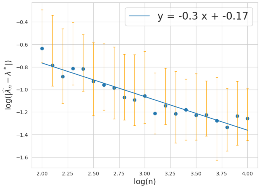
(a) Histogram (b) Rate of $\widehat{\lambda}_n$ (c) Rate of $\widehat{\mu}_n$ (d) Rate of $\widehat{\sigma}^2_n$

Figure: Case (i) $\lambda^* = 0.5$; Case (ii) $\lambda^* = 0.5/n^{1/4}$.

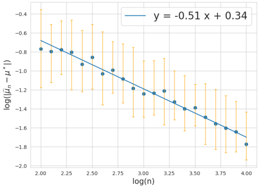
Simulation study (2): Weakly identifiable setting

Case 1: $\mu^* = \mu_0$ and $(\sigma^*)^2 \to \sigma_n^2$ in the rate $n^{-1/8}$

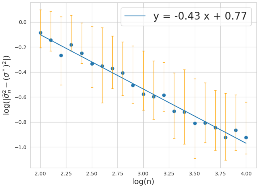
Case 2: $\sigma^* = \sigma_0$ and $\mu^* \to \mu_0$ in the rate $n^{-1/8}$.



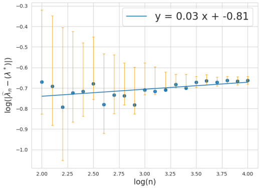
(a) Rate of $\widehat{\lambda}_n$



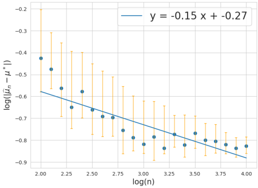
(b) Rate of $\widehat{\mu}_n$



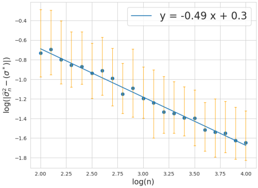
(c) Rate of $\widehat{\sigma}_n^2$



(a) Rate of $\widehat{\lambda}_n$



(b) Rate of $\widehat{\mu}_n$



(c) Rate of $\widehat{\sigma}_n^2$

Conclusions

We study the minimax rate and MLE convergence rate of the deviated model.

- Obtain the uniform rate of convergence by carefully specifying different linear independence settings between h_0 and f;
- Future direction: Uniform rate when deviating by a complex, hierarchical model or h_0 itself is a hierarchical model.