Proceedings of Machine Learning Research vol 291:1–2, 2025 38th Annual Conference on Learning Theory

Identifiability and Estimation in High-Dimensional Nonparametric Latent Structure Models

Yichen Lyu LVYC22@MAILS.TSINGHUA.EDU.CN
Pengkun Yang YANGPENGKUN@TSINGHUA.EDU.CN
Department of Statistics and Data Science, Tsinghua University[*]

Editors: Nika Haghtalab and Ankur Moitra

We study the problems of identifiability and estimation in high-dimensional nonparametric latent structure models, where the data are generated from a finite mixture of product distributions:

$$\mu = \sum_{k=1}^{m} \pi_k \left(\mu_{k1} \times \cdots \times \mu_{kd}\right). \tag{1}$$

This model captures a broad class of latent variable structures with conditionally independent coordinates and is applicable to various domains Kasahara and Shimotsu (2014); Chauveau et al. (2015).

A fundamental identifiability condition for model (1) is the linear independence of the component distributions for each variable Allman et al. (2009). This assumption underlies many algorithmic works Benaglia et al. (2009); Zheng and Wu (2020); Lu and Wang (2022), yet it fails in important cases such as Bernoulli mixtures. Recent studies have shown that high dimensionality can ensure identifiability in several special cases of model (1), including conditional i.i.d. settings and discrete mixtures Tahmasebi et al. (2018); Vandermeulen and Scott (2019); Gordon et al. (2024), despite the failure of linear independence.

To fix the gap, we introduce a new identifiability theorem that subsumes the previous results, providing a flexible criterion applicable to both continuous and discrete models. Conceptually, our result reveals that increased dimensionality and *diversity* across variables can enhance identifiability. In particular, our result implies that identifiability holds whenever the component distributions of at least $2m-1$ coordinates are pairwise distinct, thereby unifying and explaining the thresholds in Tahmasebi et al. (2018); Vandermeulen and Scott (2019); Gordon et al. (2024).

For the estimation problem, we show that under an *incoherence* condition, the component distributions μ_{kj} can be recovered from any estimate of the joint density. Statistically, we establish a perturbation theory under incoherence and near-optimal minimax rate bounds for the high-dimensional nonparametric density estimation under latent structures with smooth marginals. Contrary to the conventional curse of dimensionality, our sample complexity scales only *polynomially* with the dimension. We further propose a recovery algorithm based on classical simultaneous diagonalization methods for tensor decomposition Leurgans et al. (1993). The method provably reconstructs each component density with error proportional to the estimation error in the joint estimate.

Together, these results yield the first unified theory for identifiability and estimation in the model (1) without requiring the linear independence condition or parametric assumptions.

[*] Extended abstract. Full version appears as arXiv:2506.09165v1.

References

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A), December 2009.

Tatiana Benaglia, Didier Chauveau, and David R. Hunter. An EM-Like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, January 2009.

Didier Chauveau, David R. Hunter, and Michael Levine. Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys*, 9(none), January 2015.

Spencer L Gordon, Erik Jahn, Bijan Mazaheri, Yuval Rabani, and Leonard J Schulman. Identification of Mixtures of Discrete Product Distributions in Near-Optimal Sample and Time Complexity. *37th Annual Conference on Learning Theory*, 2024.

Hiroyuki Kasahara and Katsumi Shimotsu. Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):97–111, January 2014.

S. E. Leurgans, R. T. Ross, and R. B. Abel. A Decomposition for Three-Way Arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, October 1993.

Nan Lu and Lihong Wang. A nonparametric estimation method for the multivariate mixture models. *Journal of Statistical Computation and Simulation*, 92(17):3727–3742, November 2022.

Behrooz Tahmasebi, Seyed Abolfazl Motahari, and Mohammad Ali Maddah-Ali. On the Identifiability of Finite Mixtures of Finite Product Measures, July 2018. arXiv:1807.05444 [math, stat].

Robert A. Vandermeulen and Clayton D. Scott. An operator theoretic approach to nonparametric mixture models. *The Annals of Statistics*, 47(5):2704–2733, October 2019. Publisher: Institute of Mathematical Statistics.

Chaowen Zheng and Yichao Wu. Nonparametric Estimation of Multivariate Mixtures. *Journal of the American Statistical Association*, 115(531):1456–1471, July 2020.