# Unsupervised Federated Multi-task Learning for Heterogeneous Tasks

## Aligning mixture components across distributed tasks

Yang Feng

Department of Biostatistics, New York University

IMSI Workshop on "New Horizons on Model Transportability and Data Integration"

# Acknowledgements

**Collaborators**

Ye Tian (Yale)

Yizheng Wang (HKUST)

Lingchong Liu (HKUST)

Haolei Weng (Michigan State)

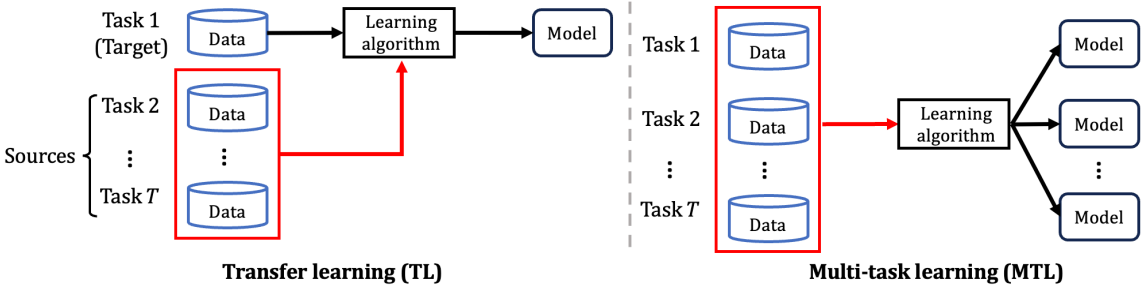
Lucy Xia (HKUST)

**Funding**

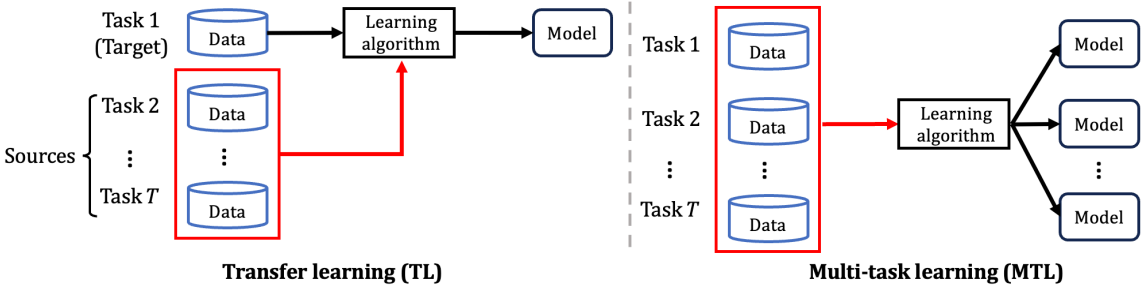
This research is supported by NSF Grant DMS-2324489.

# Outline

# Knowledge transfer



# Knowledge transfer




- Borrowing information across tasks can reduce variance and improve generalization.
- In federated settings, the gain depends on *which* tasks are truly related.
- **Goal:** adaptively transfer knowledge effectively when tasks are similar, partially similar, or not comparable at all.

# Unsupervised Federated Learning

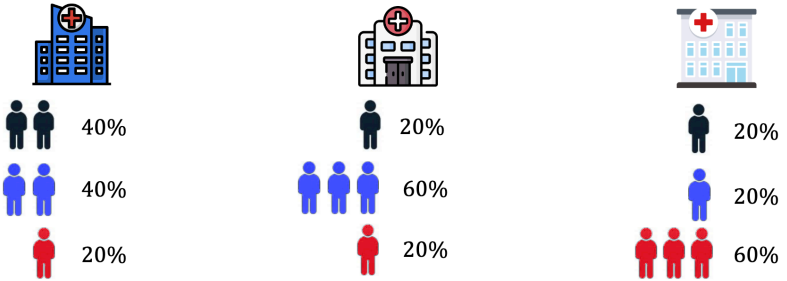
- Many multi-site scientific studies lack reliable labels or cannot harmonize labels across sites.

# Unsupervised Federated Learning

- Many multi-site scientific studies lack reliable labels or cannot harmonize labels across sites.
- Even when labels exist, collecting them at scale is expensive and often infeasible in privacy-sensitive domains.

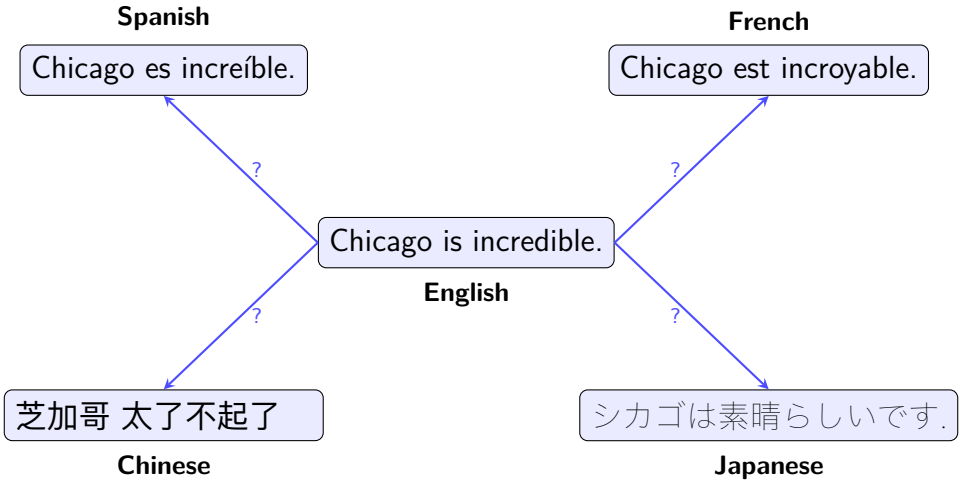
# Unsupervised Federated Learning

- Many multi-site scientific studies lack reliable labels or cannot harmonize labels across sites.
- Even when labels exist, collecting them at scale is expensive and often infeasible in privacy-sensitive domains.
- Unsupervised learning lets us pool signal for tasks such as patient phenotyping [e.g., Nakamaru et al., 2023, Jacquemyn et al., 2024]



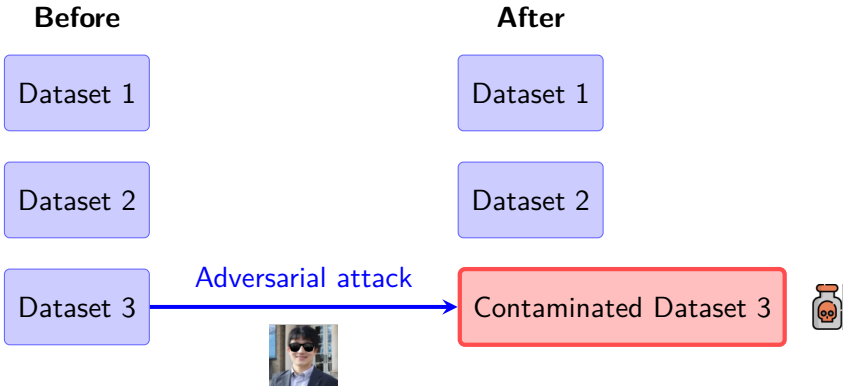

- Constraint: raw data cannot be shared, so the method must work through local computation and limited communication.

# Challenge 1: Unknown Similarity

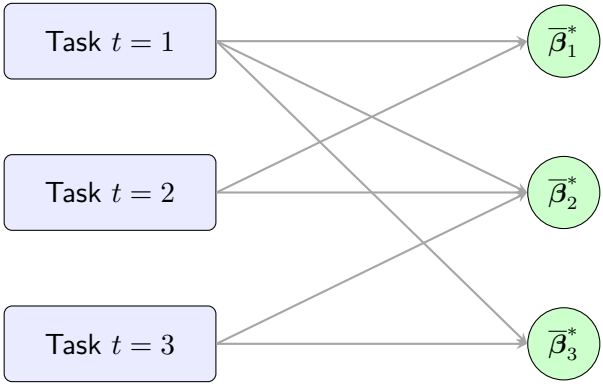



The true similarities are **unknown a priori**; the method must learn how much to share.

# Challenge 2: Data Contamination

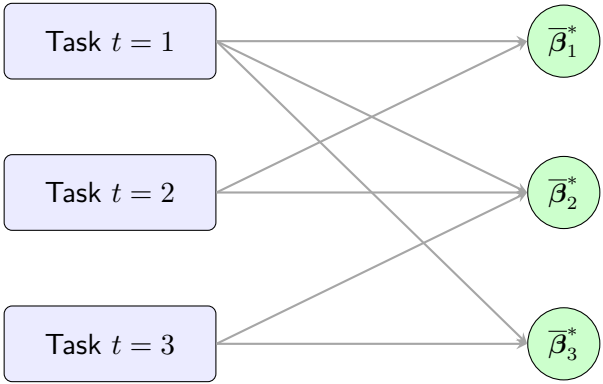


# Challenge 3: Heterogeneous Components




- Tasks share *overlapping but non-identical* component sets
- Each task observes only a subset of global components
- Correct alignment requires identifying both **matches** and **missing components**

# Challenge 3: Heterogeneous Components




- Tasks share *overlapping but non-identical* component sets
- Each task observes only a subset of global components
- Correct alignment requires identifying both **matches** and **missing components**
- **Today:** how to address them in **unsupervised federated learning**

# Why Heterogeneous Components Are Inevitable

- **Classical assumption:** every task shares the same latent mixture components and only the weights differ.
- **Reality:** each site or subject often sees only a subset of latent populations.
- **Examples:**
  - Hospitals may never observe rare diseases.
  - Human subjects perform only subsets of activities.
  - Cell types can be present only in certain tissues or conditions.
- **Consequence:** forcing all tasks to share all components leads to wrong matching, negative transfer, and unstable EM updates.

# This Talk in One Slide

**Problem**

- Federated mixture learning
- Tasks have missing components
- Standard FL over-shares

**Key Idea**

- Estimate each task locally
- Align components across tasks
- Borrow strength adaptively

**Takeaway**

- Better robustness to heterogeneity
- Competitive when tasks are homogeneous
- Works without sharing raw data

**Roadmap**

1. Homogeneous federated EM as the baseline problem.
2. Why missing components break that baseline.
3. AdaFedEM-H: alignment + adaptive regularization.

# Baseline: Homogeneous Federated Mixture Learning

- Start with the simpler baseline where every task has the same $K$ mixture components.
- For task $t \in S \subseteq [T]$, $\sum_{k=1}^{K} w_k^{(t)*} = 1$:

$$\boldsymbol{x}_i^{(t)} \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k^{(t)*} \cdot p(\cdot|\boldsymbol{\beta}_k^{(t)*}), \quad i = 1:n,$$

$w_k^{(t)*}$: mixture proportions; $p(\cdot|\boldsymbol{\beta}_k^{(t)*})$: density functions.

# Baseline: Homogeneous Federated Mixture Learning

- Start with the simpler baseline where every task has the same $K$ mixture components.
- For task $t \in S \subseteq [T]$, $\sum_{k=1}^{K} w_k^{(t)*} = 1$:

$$\boldsymbol{x}_i^{(t)} \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k^{(t)*} \cdot p(\cdot|\boldsymbol{\beta}_k^{(t)*}), \quad i = 1:n,$$

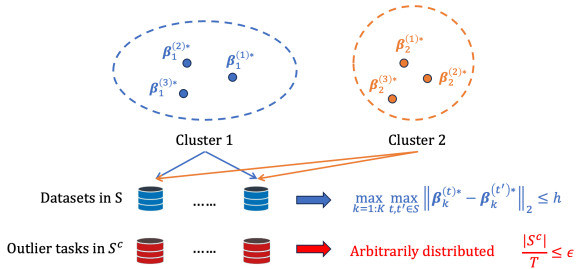
  $w_k^{(t)*}$: mixture proportions; $p(\cdot|\boldsymbol{\beta}_k^{(t)*})$: density functions.
- **Goal:** estimate $\{w_k^{(t)*}, \boldsymbol{\beta}_k^{(t)*}\}_{k=1}^{K}$ for $t \in S$ while sharing only intermediate updates.

# Baseline: Homogeneous Federated Mixture Learning

- Start with the simpler baseline where every task has the same $K$ mixture components.
- For task $t \in S \subseteq [T]$, $\sum_{k=1}^{K} w_k^{(t)*} = 1$:

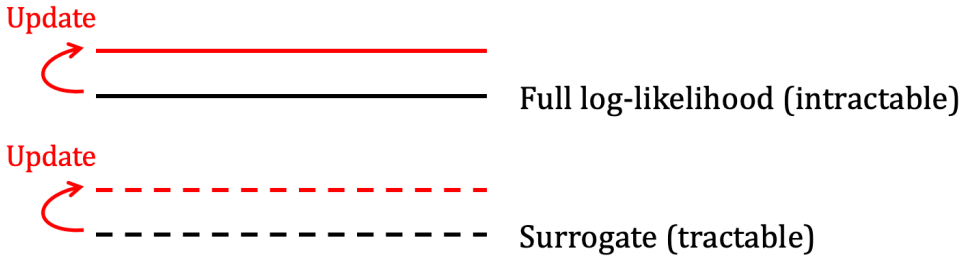
$$\boldsymbol{x}_i^{(t)} \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k^{(t)*} \cdot p(\cdot|\boldsymbol{\beta}_k^{(t)*}), \quad i = 1:n,$$

  $w_k^{(t)*}$: mixture proportions; $p(\cdot|\boldsymbol{\beta}_k^{(t)*})$: density functions.
- **Goal:** estimate $\{w_k^{(t)*}, \boldsymbol{\beta}_k^{(t)*}\}_{k=1}^{K}$ for $t \in S$ while sharing only intermediate updates.



# EM Algorithm

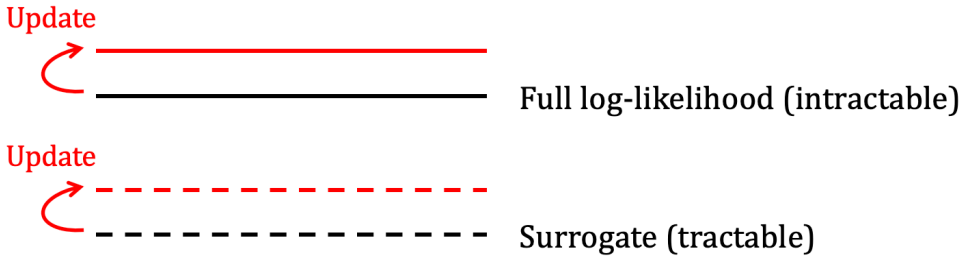
- EM and Gradient-EM are standard workhorses for mixture estimation [e.g., Wu, 1983, Redner and Walker, 1984, Balakrishnan et al., 2017].




[BWY17] Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis.
[MNB+21] Marfoq, O. et al. (2021). Federated multi-task learning under a mixture of distributions.
[DFMR21] Dieuleveut, A. et al. (2021). Federated-EM with heterogeneity mitigation and variance reduction.

# EM Algorithm

- EM and Gradient-EM are standard workhorses for mixture estimation [e.g., Wu, 1983, Redner and Walker, 1984, Balakrishnan et al., 2017].




- Federated extensions of EM already exist [Marfoq et al., 2021, Dieuleveut et al., 2021, Wu et al., 2023], but they typically assume:
  - the same component structure across tasks,
  - limited finite-sample understanding.

[BWY17] Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis.
[MNB+21] Marfoq, O. et al. (2021). Federated multi-task learning under a mixture of distributions.
[DFMR21] Dieuleveut, A. et al. (2021). Federated-EM with heterogeneity mitigation and variance reduction.

# Federated Gradient-EM: High-Level Template

- **Step 1: Local E/M update**
  - each task computes soft assignments and parameter updates from its own data
- **Step 2: Server aggregation**
  - the server shrinks task-specific parameters toward a shared center
- **Step 3: Repeat**
  - iterate until estimates stabilize

[SO11] She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression.

# Federated Gradient-EM

- **Local update:** For task $t \in [T]$:
  - **E-step:** Surrogate $Q^{(t)}(\boldsymbol{\beta}, \boldsymbol{w}) = n^{-1} \sum_{i,k} \mathbb{P}(z^{(t)} = k | \boldsymbol{x}_i^{(t)}; \widehat{\boldsymbol{w}}^{(t)\mathsf{old}}, \widehat{\boldsymbol{\beta}}^{(t)\mathsf{old}})$
    $\times \big[ \log p(\boldsymbol{x}_i^{(t)} | z^{(t)} = k; \boldsymbol{\beta}) + \log w_k \big]$
  - **M-step:** $\star$ $\widehat{w}_k^{(t)} \leftarrow n^{-1} \sum_{i=1}^{n} \mathbb{P}(z^{(t)} = k | \boldsymbol{x}_i^{(t)}; \widehat{\boldsymbol{w}}^{(t)\mathsf{old}}, \widehat{\boldsymbol{\beta}}^{(t)\mathsf{old}})$
    $\star$ $\widetilde{\boldsymbol{\beta}}_k^{(t)} \leftarrow \widehat{\boldsymbol{\beta}}_k^{(t)\mathsf{old}} + \eta \cdot \nabla_{\boldsymbol{\beta}_k} Q^{(t)}(\boldsymbol{\beta}, \boldsymbol{w})|_{\widehat{\boldsymbol{\beta}}^{(t)\mathsf{old}}, \widehat{\boldsymbol{w}}^{(t)\mathsf{old}}}$

[SO11] She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression.

# Federated Gradient-EM

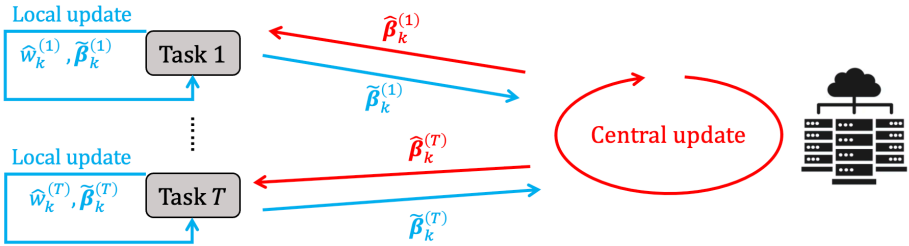
- **Local update:** For task $t \in [T]$:
  - **E-step:** Surrogate $Q^{(t)}(\boldsymbol{\beta}, \boldsymbol{w}) = n^{-1} \sum_{i,k} \mathbb{P}(z^{(t)} = k | \boldsymbol{x}_i^{(t)}; \widehat{\boldsymbol{w}}^{(t)\mathsf{old}}, \widehat{\boldsymbol{\beta}}^{(t)\mathsf{old}})$
    $\times \big[ \log p(\boldsymbol{x}_i^{(t)} | z^{(t)} = k; \boldsymbol{\beta}) + \log w_k \big]$
  - **M-step:** $\star\ \widehat{w}_k^{(t)} \leftarrow n^{-1} \sum_{i=1}^n \mathbb{P}(z^{(t)} = k | \boldsymbol{x}_i^{(t)}; \widehat{\boldsymbol{w}}^{(t)\mathsf{old}}, \widehat{\boldsymbol{\beta}}^{(t)\mathsf{old}})$
    $\star\ \widetilde{\boldsymbol{\beta}}_k^{(t)} \leftarrow \widehat{\boldsymbol{\beta}}_k^{(t)\mathsf{old}} + \eta \cdot \nabla_{\boldsymbol{\beta}_k} Q^{(t)}(\boldsymbol{\beta}, \boldsymbol{w})|_{\widehat{\boldsymbol{\beta}}^{(t)\mathsf{old}}, \widehat{\boldsymbol{w}}^{(t)\mathsf{old}}}$
- **Central update:** For each cluster $k \in [K]$:

$$\{\widehat{\boldsymbol{\beta}}_k^{(t)}\}_{t=1}^T, \bar{\boldsymbol{\beta}}_k \leftarrow \underset{\{\boldsymbol{\beta}^{(t)}\}, \bar{\boldsymbol{\beta}}}{\arg\min} \big\{ \tfrac{1}{2} \textstyle\sum_{t=1}^T (\|\boldsymbol{\beta}^{(t)} - \widetilde{\boldsymbol{\beta}}_k^{(t)}\|_2^2 + \frac{\lambda}{\sqrt{n}} \|\boldsymbol{\beta}^{(t)} - \bar{\boldsymbol{\beta}}\|_2) \big\}$$

Only component parameters reach the server (never raw data); the shared center $\bar{\boldsymbol{\beta}}_k$ is a robust **coordinate-wise median**, tolerating up to $\epsilon < 1/3$ outlier tasks.

[SO11] She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression.

# Federated Gradient-EM



# Existing Theory: Error Decomposition

**Theorem 1 (Estimation error, [Tian et al., 2024])**

$$\max_{t\in S}\max_{k\in[K]}|\widehat{w}_k^{(t)}-w_k^{(t)*}|\vee\|\widehat{\boldsymbol{\beta}}_k^{(t)}-\boldsymbol{\beta}_k^{(t)*}\|_2$$

$$\lesssim_{\mathbb{P}} \underbrace{R_1}_{\text{iterative error}} + \underbrace{R_2}_{\text{aggregation rate}} + \underbrace{R_3}_{\text{heterogeneous mixture proportions}} + \underbrace{R_4}_{\text{task similarity}} + \underbrace{R_5}_{\text{outliers}}$$

# Existing Theory: Error Decomposition

**Theorem 1 (Estimation error, [Tian et al., 2024])**

$$\max_{t\in S}\max_{k\in[K]} |\widehat{w}_k^{(t)} - w_k^{(t)*}| \vee \|\widehat{\boldsymbol{\beta}}_k^{(t)} - \boldsymbol{\beta}_k^{(t)*}\|_2$$

$$\lesssim_{\mathbb{P}} \underbrace{R_1}_{\text{iterative error}} + \underbrace{R_2}_{\text{aggregation rate}} + \underbrace{R_3}_{\text{heterogeneous mixture proportions}} + \underbrace{R_4}_{\text{task similarity}} + \underbrace{R_5}_{\text{outliers}}$$

For Gaussian Mixture Models or Mixture of GLMs:

- $R_1 = \delta^r$, $r$: # of iterations, $\delta \in (0,1)$
- $R_2 = \sqrt{\frac{d}{nT}}$
- $R_3 = \sqrt{\frac{1}{n}}$
- $R_4 = \min\left\{h, \sqrt{\frac{d}{n}}\right\}$
- $R_5 = \epsilon\sqrt{\frac{d}{n}}$

# Existing Theory: Error Decomposition

**Theorem 1 (Estimation error, [Tian et al., 2024])**

$$\max_{t\in S}\max_{k\in[K]} |\widehat{w}_k^{(t)} - w_k^{(t)*}| \vee \|\widehat{\boldsymbol{\beta}}_k^{(t)} - \boldsymbol{\beta}_k^{(t)*}\|_2$$

$$\lesssim_{\mathbb{P}} \underbrace{R_1}_{\text{iterative error}} + \underbrace{R_2}_{\text{aggregation rate}} + \underbrace{R_3}_{\text{heterogeneous mixture proportions}} + \underbrace{R_4}_{\text{task similarity}} + \underbrace{R_5}_{\text{outliers}}$$

For Gaussian Mixture Models or Mixture of GLMs:

- $R_1 = \delta^r$, $r$: # of iterations, $\delta \in (0,1)$
- $R_2 = \sqrt{\frac{d}{nT}}$
- $R_3 = \sqrt{\frac{1}{n}}$
- $R_4 = \min\left\{h, \sqrt{\frac{d}{n}}\right\}$
- $R_5 = \epsilon\sqrt{\frac{d}{n}}$

When $r \to \infty, T \to \infty, h \ll \sqrt{\frac{d}{n}}, \epsilon \ll 1$:

Better than single-task learning $\sqrt{\frac{d}{n}}$

# What the Existing Theory Says

**Interpretation of the error bound**

- optimization error decreases with more EM iterations
- aggregation improves with more tasks through a $\sqrt{nT}$ effect
- performance still depends on task similarity and contamination

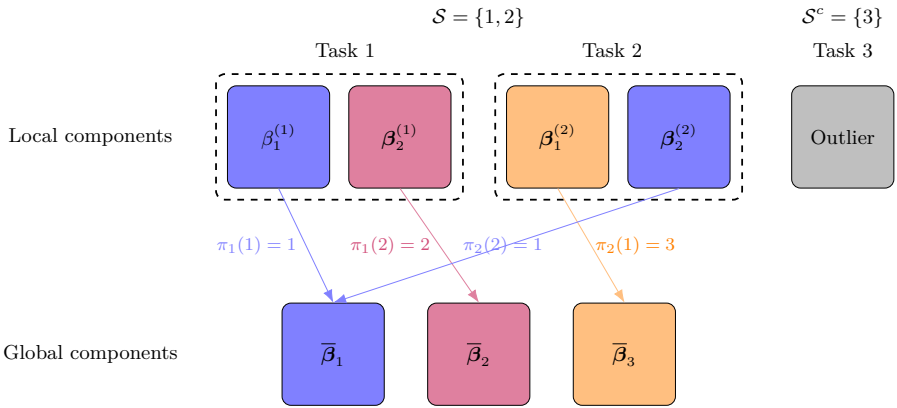
**What is still missing**

- the theory assumes a fixed component label across tasks
- it does *not* allow a task to simply lack one of the global components

[TWF24] Tian, Y., Weng, H., & **Feng, Y.** (2024). Towards the theory of unsupervised federated learning: Non-asymptotic analysis of federated EM algorithms. *ICML 2024*.

# Outline

# A Simple Failure Mode




- Tasks can share some clusters but not others.
- If the server forces a common label set, it can average incompatible components.

# What Changes in the Heterogeneous Setting

**Key modeling shift**

Each task can have its own number of local components $K_t$, and those local components only approximately align with a larger collection of global components.

- **Local view:** estimate the latent structure for each task.
- **Global view:** identify which local components correspond across tasks.
- **New requirement:** detect both *matches* and *true absences*.

# Heterogeneous Multi-Task Mixture Model

- **Model**: for each task $t \in \mathcal{S}$, the observation $\boldsymbol{x}_i^{(t)}$ is generated from a finite mixture model with $K_t$ *local* components.

$$\boldsymbol{x}_i^{(t)} \overset{i.i.d.}{\sim} \sum_{k=1}^{K_t} w_k^{(t)*} p\big(\cdot \,|\boldsymbol{\beta}_k^{(t)*}\big) \quad (t \in \mathcal{S}, i \in [n]),$$

# Heterogeneous Multi-Task Mixture Model

- **Model**: for each task $t \in \mathcal{S}$, the observation $\boldsymbol{x}_i^{(t)}$ is generated from a finite mixture model with $K_t$ *local* components.

$$\boldsymbol{x}_i^{(t)} \overset{i.i.d.}{\sim} \sum_{k=1}^{K_t} w_k^{(t)*} p\big(\cdot \,|\boldsymbol{\beta}_k^{(t)*}\big) \quad (t \in \mathcal{S}, i \in [n]),$$

- **Local Components**: $\boldsymbol{\beta}_k^{(t)*} : t \in \mathcal{S}, k \in [K_t]$
- **Global Components**: $\overline{\boldsymbol{\beta}}_r^* : r \in [K]$

# Heterogeneous Multi-Task Mixture Model

- **Model**: for each task $t \in \mathcal{S}$, the observation $\boldsymbol{x}_i^{(t)}$ is generated from a finite mixture model with $K_t$ *local* components.

$$\boldsymbol{x}_i^{(t)} \overset{i.i.d.}{\sim} \sum_{k=1}^{K_t} w_k^{(t)*} p\big(\cdot \,|\boldsymbol{\beta}_k^{(t)*}\big) \quad (t \in \mathcal{S}, i \in [n]),$$

- **Local Components**: $\boldsymbol{\beta}_k^{(t)*} : t \in \mathcal{S}, k \in [K_t]$
- **Global Components**: $\overline{\boldsymbol{\beta}}_r^* : r \in [K]$
- **Approximate alignment**: for each $t \in \mathcal{S}$, there exists a unique mapping $\pi_t(\cdot) \in \Pi_{K_t,K}$ such that

$$\max_{k \in [K_t]} \left\| \boldsymbol{\beta}_k^{(t)*} - \overline{\boldsymbol{\beta}}_{\pi_t(k)}^* \right\|_2 \leq h,$$

  where $\Pi_{K_t,K}$ is the collection of all injections from $[K_t]$ to $[K]$.
- **Interpretation**: $h$ quantifies how far a task-specific component can drift from a shared prototype.

# AdaFedEM-H: Main Idea

**Three-step workflow:**

1. **Single-Task Initialization**:
   - Fit each task locally with `GSF` (Group-Sort-Fuse) to estimate both $K_t$ and the local parameters.
2. **Component Alignment**:
   - Match components across tasks.
   - Introduce a new global component only when existing ones are poor matches.
3. **Adaptive Federated EM**:
   - Adaptive penalty $\lambda_{ada} = \lambda/\widehat{h}$: large $\widehat{h}$ (high heterogeneity) $\rightarrow$ weaker sharing; small $\widehat{h} \rightarrow$ stronger sharing.

**Global components $K$:** chosen by an elbow (`Kneedle`) method, correct in 85–99% of simulation runs.

**Practical message**

Share aggressively when tasks are similar; back off automatically when they are not.

Three challenges $\rightarrow$ three ingredients: alignment $\leftrightarrow$ heterogeneous components, adaptive $\lambda_{ada}$ $\leftrightarrow$ unknown similarity, robust median $\leftrightarrow$ contamination.

# Component Alignment via Greedy Matching

**Goal:** align *local* mixture components across tasks without pretending every task has every cluster.

**Greedy alignment strategy**

- Tasks are processed sequentially
- The first task initializes the current pool of global components
- For each later task, every local component must either
  - align to an existing global component $r \in [K]$, or
  - introduce a new global component
- The alignment $\widehat{\pi}_t \in \Pi_{K_t,K}$ is chosen adaptively

**Why this matters**

This prevents the server from averaging two components that represent genuinely different latent groups.

# Component Alignment Score

**Alignment rule for task $t$:**

$$\widehat{\pi}_t = \underset{\pi \in \Pi_{K_t,K}}{\arg\min} \left[ \sum_{r \in \mathcal{R}(\pi)} \frac{1}{|\mathcal{A}_{t-1}(r)|} \sum_{t' \in \mathcal{A}_{t-1}(r)} \left\| \boldsymbol{\beta}^{(t)*}_{\pi^{-1}(r)} - \boldsymbol{\beta}^{(t')*}_{\widehat{\pi}^{-1}_{t'}(r)} \right\|_2 + \gamma \, |\mathsf{New}(\pi)| \right]$$

- First term: closeness to previously aligned components.
- Second term: a penalty for creating a new global component.
- Net effect: new components appear only when old ones are a poor fit.

# Alignment Is Provably Correct

**Theorem 2 (Alignment correctness)**

*If each global component appears in enough tasks before the first outlier, and the initialization error and heterogeneity $h$ are small, then for a suitable penalty $\gamma$ the greedy alignment is* ***exact*** *up to a single global relabeling:*

$$\widehat{\pi}_t = \iota \circ \pi_t \quad \textit{for all } t \in S, \qquad \textit{w.p. } \geq 1 - \Delta.$$

- Tolerates a fraction $\epsilon < 1/3$ of outlier (contaminated) tasks.
- Conditions are met with high probability after randomizing the task order a few times.
- This is the foundation for the estimation guarantee next.

Wang, Y., Liu, L., **Feng, Y.**, & Xia, L. (2026). Unsupervised Federated Multi-task Learning for Heterogeneous Tasks.

# AdaFedEM-H: Same Rate, Harder Problem

**Theorem 3 (Estimation error, Gaussian mixtures)**

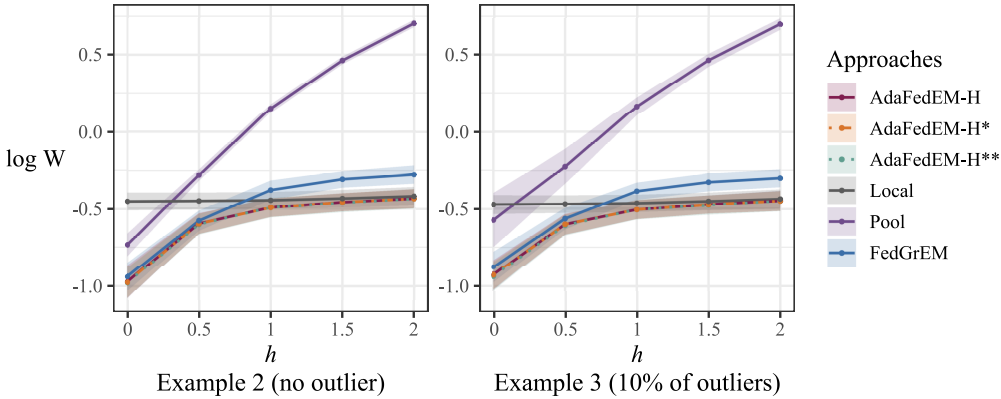
$$\max_{t\in S}\max_{k}\left|\widehat{w}_k^{(t)}-w_k^{(t)*}\right|\vee\|\widehat{\boldsymbol{\beta}}_k^{(t)}-\boldsymbol{\beta}_k^{(t)*}\|_2$$

$$\lesssim_{\mathbb{P}} \underbrace{\delta^r}_{\text{iterative}} + \underbrace{\sqrt{\frac{d}{nT}}}_{\text{aggregation}} + \underbrace{\sqrt{\frac{1}{n}}}_{\text{proportions}} + \underbrace{\min\left\{h, \sqrt{\frac{d}{n}}\right\}}_{\text{heterogeneity}} + \underbrace{\epsilon\sqrt{\frac{d}{n}}}_{\text{outliers}}$$

- **Same form as the homogeneous rate**, yet now valid with *unknown* per-task $K_t$ and *missing* components, exactly the regime the old theory excluded.
- No worse than single-task; *strictly faster* when $\epsilon \to 0$, $T \to \infty$, and $h = o(\sqrt{d/n})$.

Wang, Liu, **Feng**, & Xia (2026). $r$: # EM iterations, $\delta \in (0,1)$; $T$ tasks, $d$ dimension.

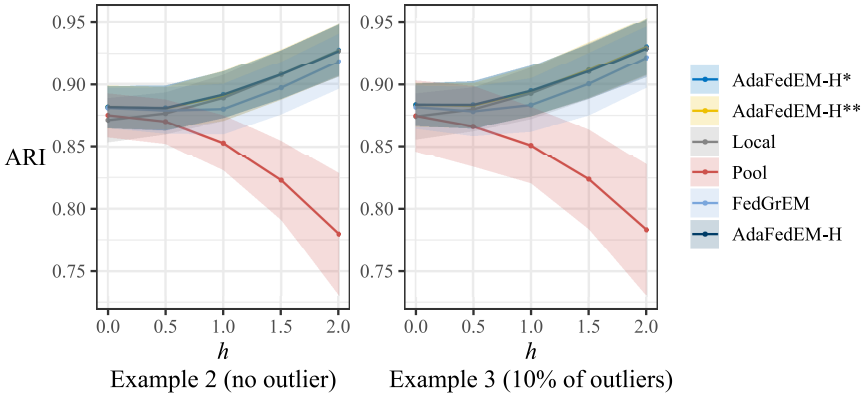
# Simulation: Wasserstein Distance ($\log W$, lower is better)




- As heterogeneity $h$ increases, component alignment becomes increasingly important.
- The revised method avoids the sharp degradation caused by a homogeneous misspecification.

`AdaFedEM-H`: estimates everything; `H*`: known # global components $K$; `H**`: also known per-task $K_t$ (oracle).

# Simulation: Clustering Accuracy (ARI, higher is better)




- Our methods (and the oracles $\texttt{H}^*$, $\texttt{H}^{**}$) stay on top across all $h$; `Pool` collapses as heterogeneity grows, with and without outliers.

# Real Data: Modeling Missing Components Helps

- Human Activity Recognition (HAR), with 30 subjects treated as tasks.
- We collapse 6 activities into 3 broader components and manually remove components to induce heterogeneity.
- Metric: adjusted Rand index (ARI); higher is better.

| | **AdaFedEM-H*** | Pooled EM | Local EM | FedGrEM | $\widehat{h}_T$ |
|---|---|---|---|---|---|
| Heterogeneous | **0.624** (0.108) | 0.486 (0.137) | 0.605 (0.108) | 0.544 (0.057) | 2.657 (0.124) |
| Homogeneous | **0.679** (0.093) | 0.513 (0.076) | 0.620 (0.101) | 0.513 (0.062) | 2.562 (0.101) |

- `AdaFedEM-H*` helps in the heterogeneous case and remains competitive in the homogeneous case.
- A larger $\widehat{h}_T$ in the heterogeneous setting signals lower similarity, so the adaptive penalty shares less, by design.

# Takeaways

- Missing components are a modeling issue, not just noise, in federated mixture learning.
- Standard federated EM is a useful baseline but breaks when cluster labels are not globally shared.
- AdaFedEM-H combines local estimation, component alignment, and adaptive sharing to handle that mismatch.
- The empirical results show improved robustness without sacrificing the privacy-preserving federated workflow.

# Related Papers

- Tian, Y., Weng, H., & **Feng, Y** (2024). Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms. *ICML 2024*.
- Tian, Y., Weng, H., Xia, L., & **Feng, Y.** (2025). Unsupervised Multi-task and Transfer Learning on Gaussian Mixture Models. *JASA*
- Wang, Y., Liu, L., **Feng, Y**, & Xia, L. (2026). Unsupervised Federated Multi-task Learning for Heterogeneous Tasks. Manuscript.

# References I


Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 2017.

Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Genevieve Robin. Federated expectation maximization with heterogeneity mitigation and variance reduction. In *Advances in Neural Information Processing Systems*, pages 29553–29566, 2021.

Xander Jacquemyn, Marc Bosman, Rafael Willems, et al. Unsupervised machine learning identifies distinct phenotypes in cardiac complications of pediatric patients treated with anthracyclines. *Frontiers in Cardiovascular Medicine*, 11:1–12, 2024.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. In *Advances in Neural Information Processing Systems*, pages 15434–15447, 2021.

Ryo Nakamaru, Norihiro Sato, Yasunobu Kinugasa, Masaharu Sugihara, Toshiaki Tsuji, Kotaro Nochioka, and colleagues. Phenotyping of elderly patients with heart failure focused on noncardiac conditions: A latent class analysis from a multicenter registry. *BMC Geriatrics*, 23:1–11, 2023.

Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

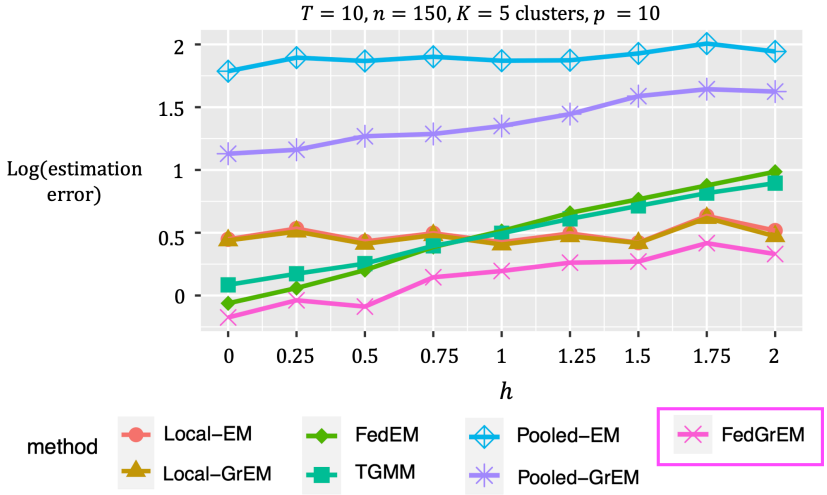
Ye Tian, Haolei Weng, and Yang Feng. Towards the theory of unsupervised federated learning: Non-asymptotic analysis of federated EM algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, pages 48226–48279. PMLR, 2024.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

# References II

Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37860–37879. PMLR, 2023.

# Simulation: Gaussian Mixture Models




Backup from the prior *homogeneous* work (Tian, Weng & **Feng**, 2024).

# Applications: Clustering of handwritten digits

- MNIST handwritten digits: $T = 100$ tasks, $K = 10$ classes, $N = 70000$ in total
  - $28 \times 28$ grayscale images
  - $80\%$ training + $20\%$ test

| $\epsilon$/Method | Local-EM | Local-GrEM | FedEM | TGMM | Pooled-EM | Pooled-GrEM | **FedGrEM** |
|---|---|---|---|---|---|---|---|
| 0% | 12.47 | 12.06 | 9.97 | 12.18 | 12.84 | 10.67 | **10.63** |
| 8% | 12.30 | 11.98 | 11.97 | 13.66 | 15.07 | 14.16 | **10.64** |
| 16% | 12.47 | 12.06 | 14.46 | 14.47 | 16.16 | 14.97 | **10.96** |
| 24% | 12.43 | 12.10 | 21.67 | 15.70 | 15.50 | 14.35 | **10.97** |

- Pen-based recognition of handwritten digits: $T = 44$, $K = 10$, $N = 10992$
  - $16$ features from the tablet
  - $80\%$ training + $20\%$ test

| $\epsilon$/Method | Local-EM | Local-GrEM | FedEM | TGMM | Pooled-EM | Pooled-GrEM | **FedGrEM** |
|---|---|---|---|---|---|---|---|
| 0% | 13.30 | 13.20 | 15.75 | 25.14 | 26.84 | 34.88 | **9.62** |
| 6.8% | 12.88 | 12.96 | 15.87 | 25.00 | 26.60 | 34.61 | **9.31** |
| 13.6% | 12.62 | 12.99 | 17.56 | 25.68 | 26.79 | 35.61 | **9.04** |
| 20.5% | 12.94 | 13.31 | 18.52 | 26.14 | 27.05 | 37.94 | **9.39** |

Backup from the prior *homogeneous* work (Tian, Weng & **Feng**, 2024).