

Individualized Group Learning

Chencheng Cai

Department of Mathematics and Statistics
Washington State University

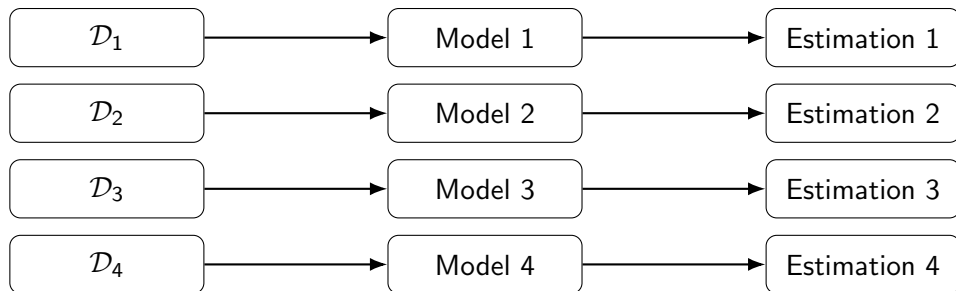
January 30, 2025

Table of Contents

- 1 Introduction
- 2 Individualized Group Learning
- 3 Theoretical Results
- 4 Simulations and Real Examples
- 5 Summary

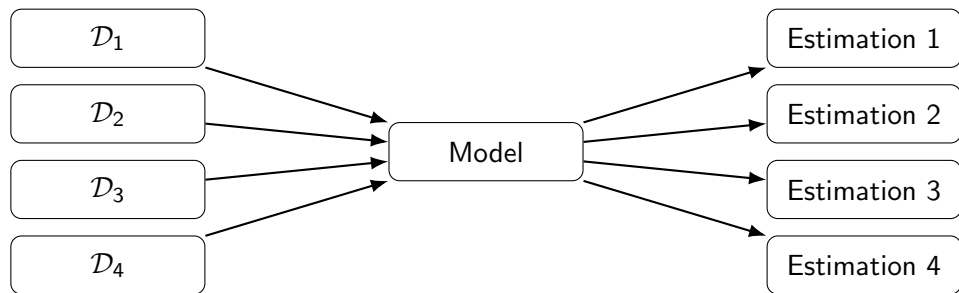
- Individualized models are in high demand:
 - ▶ personalized medicine
 - ▶ the market of one
 - ▶ targeted advertisement
 - ▶ small data models
- The dilemma:
 - ▶ Large data model provides accurate and efficient estimations for the whole population.
 - ★ targeting the common law of the population
 - ★ inadequate for under-represented sub-populations
 - ▶ A Small data model trained on a fully personalized dataset provides a completely individualized model.
 - ★ Usually vulnerable due to limited data size.
 - ★ Privacy issue.

Individual Model



- \mathcal{D}_i , $i = 1, 2, 3, 4$: datasets for four individuals.
- Model i for $i = 1, 2, 3, 4$: same model family with different parameters.
- Estimation i for $i = 1, 2, 3, 4$: estimated parameters.

Population Model



- Fit one large model with all datasets.
- Estimation i , $i = 1, 2, 3, 4$, are the same for all individuals.

- **Me:** "Let's discuss in the domain of statistics."
- **ChatGPT:** "Absolutely, statistics is a fascinating field! ..."
- **Me:** "What is Bayesian inference?"
- **ChatGPT:** "Bayesian inference is a statistical approach for updating beliefs or making predictions about unknown parameters or events based on evidence or data. It's named after the Reverend Thomas Bayes, an 18th-century mathematician, and is rooted in Bayesian probability theory...."

ChatGPT knows Bayesian inference — a widely recognized methodology.

- **Me:** "What is fusion learning?"
- **ChatGPT:** "Fusion learning, also known as ensemble learning, is a machine learning technique where multiple models (or learners) are combined to improve the overall performance of the system. The idea behind fusion learning is that by combining the predictions of multiple models, we can often achieve better accuracy, robustness, and generalization compared to using a single model alone. There are several approaches to fusion learning, including: ... "

**ChatGPT recognizes fusion learning as ensemble learning,
a very close but different concept.**

- **Me:** "What is individualized group learning?"
- **ChatGPT:** "" Individualized group learning" sounds like a combination of two concepts: individualized learning and group learning.

...

Combining these two concepts, "individualized group learning" could refer to a learning approach that integrates elements of both individualized learning and group learning."

ChatGPT does not know individualized group learning at all.

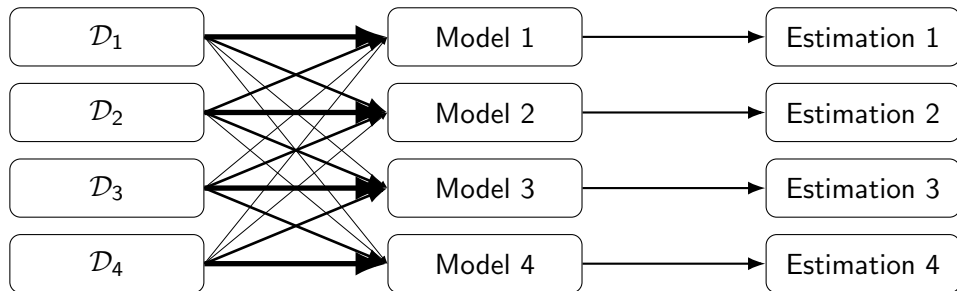
- Individualized Group Learning is an under-represented topic on the internet.

- **Me:** "What is individualized group learning?"
- **Gemini:** "The term "individualized group learning" can have two different meanings:
 1. In education: It refers to ...
 2. In statistics and machine learning: It's a more technical concept known as "iGroup". This approach focuses on creating a customized group for each individual based on their data. ..."

Gemini correctly recognizes individualized group learning.

- Use a pre-trained large language model (population model) for linguistic correctness.
- Use Google search results (individual model) for a specified answer to the topic.

Individualized Group Learning (*i*Group)



- Model i , $i = 1, 2, 3, 4$, are fusion models that use all the data but target individual i .
- Each model uses the datasets with different weights (illustrated by the line width).
- Model i is fitted in a clique of individual i :
 - ▶ clique: similar/relevant individuals
 - ▶ clique size controlled by the weights.

Table of Contents

1 Introduction

2 Individualized Group Learning

3 Theoretical Results

4 Simulations and Real Examples

5 Summary

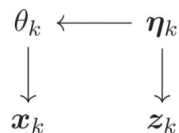
Problem Setup

$$\begin{aligned}\theta_k &\sim \pi(\cdot), \\ \mathbf{x}_k | \theta_k &\sim p(\cdot; \theta_k),\end{aligned}$$

\mathbf{x} model

$$\begin{aligned}\theta_k &= g(\boldsymbol{\eta}_k), \\ \mathbf{z}_k | \boldsymbol{\eta}_k &\sim q(\cdot; \boldsymbol{\eta}_k).\end{aligned}$$

\mathbf{z} model



diagram

- Individuals $k = 0, 1, \dots, K$. Individual 0 is the target.
- \mathbf{x} -model: $\pi(\cdot)$: unknown population distribution. $p(\cdot; \theta)$ known distribution family.
- \mathbf{z} -model: $g(\cdot)$ unknown continuous function. $q(\cdot; \boldsymbol{\eta})$ unknown distribution family.
- Observed: $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k=0}^K$
- Goal: estimate θ_0 .

Aggregated Estimation

- $\hat{\theta}_k$: individual point estimator.
- $M_k(\theta) := M_k(\theta, \mathbf{x}_k)$: individual objective function

$$\hat{\theta}_k = \arg \min_{\theta} M_k(\theta)$$

- We consider the following two pooled estimators for θ_0 .

(aggregating estimators)
$$\hat{\theta}_0^{(c)} = \frac{\sum_{k=0}^K \hat{\theta}_k w(k; 0)}{\sum_{k=0}^K w(k; 0)}$$

(aggregating objective functions)
$$\tilde{\theta}_0^{(c)} = \arg \min_{\theta} \sum_{k=0}^K M_k(\theta, \mathbf{x}_k) w(k; 0)$$

where $w(k; 0)$ is some weight function measure the similarity between individual k and 0.

Aggregated Estimation

(aggregating estimators)

$$\hat{\theta}_0^{(c)} = \frac{\sum_{k=0}^K \hat{\theta}_k w(k; 0)}{\sum_{k=0}^K w(k; 0)}$$

(aggregating objective functions)

$$\tilde{\theta}_0^{(c)} = \arg \min_{\theta} \sum_{k=0}^K M_k(\theta, \mathbf{x}_k) w(k; 0)$$

- Example:
 - ▶ $\hat{\theta}_k$ is the least squares estimator and $M_k(\theta, \mathbf{x}_k)$ is the sum of squared error.
 - ▶ $\hat{\theta}_k$ is the maximum likelihood estimator and $M_k(\theta, \mathbf{x}_k)$ is the negative log-likelihood function.
- When $w(k; 0) \equiv 1$, the combined estimators corresponds to a population model.
- When $w(k; 0) = \mathbf{1}_{\{k=0\}}$, the combined estimators corresponds to the individual model.
- **The design of the weight function $w(k, 0)$ is crucial for the combined estimators.**

Weight Function

We would like to construct the weight function as follows:

$$w(k; 0) = w_1(\mathbf{z}_k, \mathbf{z}_0) w_2(\hat{\theta}_k, \hat{\theta}_0 \mid \mathbf{z}_0, \mathbf{z}_k)$$

- $w_1(\mathbf{z}_k, \mathbf{z}_0)$ measures the similarity between \mathbf{z}_k and \mathbf{z}_0 . A common choice is

$$w_1(\mathbf{z}_k, \mathbf{z}_0) = \mathcal{K}_1 \left(\frac{\|\mathbf{z}_k - \mathbf{z}_0\|}{b_1} \right)$$

for some kernel function \mathcal{K}_1 and bandwidth b_1 .

- $w_2(\hat{\theta}_k, \hat{\theta}_0 \mid \mathbf{z}_0, \mathbf{z}_k)$ measures the similarity between individual estimators. The choice for $w_2(\hat{\theta}_k, \hat{\theta}_0 \mid \mathbf{z}_0, \mathbf{z}_k)$ is not trivial!

Weight Function

An intuitive but **incorrect** choice for $w_2(\hat{\theta}_k, \hat{\theta}_0 | \mathbf{z}_0, \mathbf{z}_k)$ is

$$w_2(\hat{\theta}_k, \hat{\theta}_0 | \mathbf{z}_0, \mathbf{z}_k) = \mathcal{K}_2 \left(\frac{\|\hat{\theta}_k - \hat{\theta}_0\|}{b_2} \right)$$

for some kernel function \mathcal{K}_2 and bandwidth b_2 .

Example: when $K \rightarrow \infty$ and $b_2 \rightarrow 0$, $\hat{\theta}_0^{(c)} \rightarrow \hat{\theta}_0$.



Weight Function

We propose to use a novel weight function:

$$w_2(\hat{\theta}_k, \hat{\theta}_0 | \mathbf{z}_0, \mathbf{z}_k) = \frac{\int p(\hat{\theta}_k | \theta)p(\hat{\theta}_0 | \theta)p(\theta | \mathbf{z}_0)d\theta}{p(\hat{\theta}_k | \mathbf{z}_k)p(\hat{\theta}_0 | \mathbf{z}_0)}$$

- Interpretation 1:

$\int p(\hat{\theta}_k | \theta)p(\hat{\theta}_0 | \theta)p(\theta | \mathbf{z}_0)d\theta$ measures the joint density for $(\hat{\theta}_k, \hat{\theta}_0)$ if **they were generated independently from identical individuals**.

- Interpretation 2:

$$\frac{\int p(\hat{\theta}_k | \theta)p(\hat{\theta}_0 | \theta)p(\theta | \mathbf{z}_0)d\theta}{p(\hat{\theta}_0 | \mathbf{z}_0)} = \int p(\hat{\theta}_k | \theta)p(\hat{\theta}_k | \hat{\theta}_0, \mathbf{z}_0)d\theta$$

It gives the probability of $\hat{\theta}_k$ in the posterior predictive distribution of individual 0. Or, the probability of getting $\hat{\theta}_k$ **if the individual 0 generated a new and independent copy of $\hat{\theta}_0$** .

Computation of the Weight Function

$$w_2(\hat{\theta}_k, \hat{\theta}_0 | \mathbf{z}_0, \mathbf{z}_k) = \frac{\int p(\hat{\theta}_k | \theta)p(\hat{\theta}_0 | \theta)p(\theta | \mathbf{z}_0)d\theta}{p(\hat{\theta}_k | \mathbf{z}_k)p(\hat{\theta}_0 | \mathbf{z}_0)}$$

- The two conditional probabilities can be estimated from the joint dataset $\{(\hat{\theta}_k, \mathbf{z}_k)\}_{k=0}^K$.

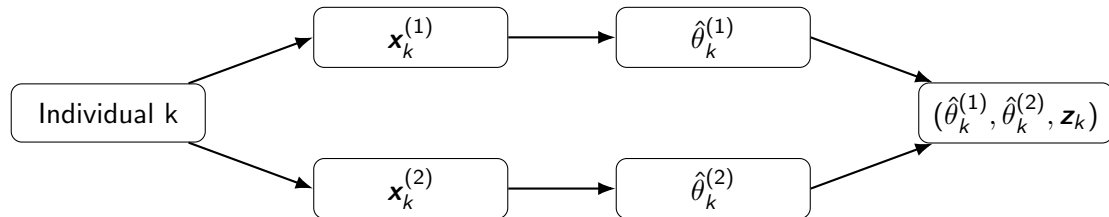
$$\hat{p}(\hat{\theta}_k | \mathbf{z}_k) = \frac{\sum_{j=0}^K \mathcal{K}_1(\|\mathbf{z}_k - \mathbf{z}_j\|/b_1)\mathcal{K}_2(\|\hat{\theta}_k - \hat{\theta}_j\|/b_2)}{\sum_{j=0}^K \mathcal{K}_1(\|\mathbf{z}_k - \mathbf{z}_j\|/b_1)}$$

- **If we have independent copies** $\hat{\theta}_k^{(1)}$ and $\hat{\theta}_k^{(2)}$ from each individual k . The numerator is:

$$\frac{\sum_{j=0}^K \mathcal{K}_1(\|\mathbf{z}_0 - \mathbf{z}_j\|/b_1)\mathcal{K}_2(\|\hat{\theta}_j^{(1)} - \hat{\theta}_0\|/b_2)\mathcal{K}_3(\|\hat{\theta}_j^{(2)} - \hat{\theta}_k\|/b_3)}{\sum_{j=0}^K \mathcal{K}_1(\|\mathbf{z}_0 - \mathbf{z}_j\|/b_1)}$$

Computation of the Weight Function

We need independent copies of \mathbf{x}_k from each individual k .



- However, this is impossible because we only observe one copy of \mathbf{x}_k for each k .
- We approximate the procedure by letting $\mathbf{x}_k^{(1)}$ and $\mathbf{x}_k^{(2)}$ be two **bootstrap samples** of \mathbf{x}_k .
- Bootstrap requires a minimum sample size ~ 10 and $\hat{\theta}_k^{(1)}$ and $\hat{\theta}_k^{(2)}$ are weakly dependent.

Individualized Group Learning (iGroup) Workflow

1. Observe $(\mathbf{x}_k, \mathbf{z}_k)$ for individual k , $k = 0, 1, \dots, K$, where individual 0 is the target.
2. For each \mathbf{x}_k , obtain estimator $\hat{\theta}_k$ and obtain $\hat{\theta}_k^{(1)}$ and $\hat{\theta}_k^{(2)}$ from two bootstrap samples.
3. Calculate the weight function

$$w(k; 0) = \mathcal{K}_1 \left(\frac{\|\mathbf{z}_k - \mathbf{z}_0\|}{b_1} \right) \frac{\int p(\hat{\theta}_k | \theta) p(\hat{\theta}_0 | \theta) p(\theta | \mathbf{z}_0) d\theta}{p(\hat{\theta}_k | \mathbf{z}_k) p(\hat{\theta}_0 | \mathbf{z}_0)}$$

where the second part is estimated nonparametrically from $\{(\hat{\theta}_k^{(1)}, \hat{\theta}_k^{(2)}, \mathbf{z}_k)\}_{k=0}^K$.

4. Construct the combined estimator for θ_0 in either of the following two ways.

$$\hat{\theta}_0^{(c)} = \frac{\sum_{k=0}^K \hat{\theta}_k w(k; 0)}{\sum_{k=0}^K w(k; 0)} \quad \text{or} \quad \tilde{\theta}_0^{(c)} = \arg \min_{\theta} \sum_{k=0}^K M_k(\theta, \mathbf{x}_k) w(k; 0)$$

- Note that the iGroup estimator can be constructed based solely on \mathbf{x}_k or \mathbf{z}_k .

Table of Contents

1 Introduction

2 Individualized Group Learning

3 Theoretical Results

4 Simulations and Real Examples

5 Summary

Assumptions

Assumption (Dense Assumption)

There exists a constant $d \geq 1$ such that for all $i = 1, \dots, K$, $|\mathcal{C}_0(\epsilon)| \asymp K\epsilon^d$ in probability when $K \rightarrow \infty, \epsilon \rightarrow 0$, where $\mathcal{C}_0(\epsilon) = \{k : \|\theta_k - \theta_0\| < \epsilon, k = 0, \dots, K\}$.

Assumption (Smooth Parameter)

There exists a positive constant κ such that, for all $\theta, \theta' \in \Omega_\theta$,

$$\sup_{\mathbf{x}} |p(\mathbf{x} | \theta) - p(\mathbf{x} | \theta')| \leq \kappa \|\theta - \theta'\|.$$

Assumption (Kernel Assumption)

The kernel function \mathcal{K} used in w_1 satisfies

$$\mathcal{K}(\cdot) \geq 0, \quad \int \mathcal{K}(u) du < \infty, \quad \lim_{u \rightarrow \infty} u\mathcal{K}(u) \rightarrow 0.$$

In addition, when $K \rightarrow \infty, b \rightarrow 0$ and $b^d K \rightarrow \infty$.

Target Estimator

Let δ_0 be any estimator for θ_0 using the data $\{(\mathbf{x}_k, \mathbf{z}_k)\}_{k=0}^K$. Both $\hat{\theta}_0^{(c)}$ and $\tilde{\theta}_0^{(c)}$ are such estimators.

Proposition (Risk Decomposition under Squared Loss)

Suppose θ_0 has a prior distribution $\pi(\cdot)$. Then we have

$$\mathbb{E}[(\delta_0 - \theta_0)^2] = \mathbb{E}[(\delta_0 - \mathbb{E}_\pi[\theta \mid \mathbf{z}_0, \mathbf{x}_0])^2] + \mathbb{E}[(\mathbb{E}_\pi[\theta \mid \mathbf{z}_0, \mathbf{x}_0] - \theta_0)^2]$$

The minimum mean squared error is obtained when $\delta_0 = \mathbb{E}_\pi[\theta \mid \mathbf{z}_0, \mathbf{x}_0]$, the posterior mean. We call

$$\Theta_0(\mathbf{x}_0, \mathbf{z}_0; \ell_2) := \mathbb{E}_\pi[\theta \mid \mathbf{z}_0, \mathbf{x}_0]$$

the **target estimator**, where $\pi(\cdot)$ takes the population distribution of θ .

Target Estimator

For general loss function $L(\hat{\theta}, \theta)$, we define the **target estimator** by

$$\Theta_0(\mathbf{x}_0, \mathbf{z}_0; L) := \arg \min_{\delta} \mathbb{E}_{\pi} [L(\delta, \theta) \mid \mathbf{x}_0, \mathbf{z}_0]$$

And a similar risk decomposition:

Proposition (Risk Decomposition for General Loss)

If $L(\hat{\theta}, \theta)$ is second-order partially differentiable with respect to $\hat{\theta}$, then

$$\mathbb{E}[L(\delta_0, \theta_0)] = \frac{1}{2} \mathbb{E}[L''(\Theta_0, \theta_0)(\delta_0 - \Theta_0)^2] + \mathbb{E}[L(\Theta_0, \theta_0)] + o((\delta_0 - \Theta_0)^2)$$

where $\Theta_0 = \Theta_0(\mathbf{x}_0, \mathbf{z}_0; L)$ is the target estimator.

Target Estimator

- The target estimators

$$\Theta_0(\mathbf{x}_0, \mathbf{z}_0; \ell_2) = \mathbb{E}_\pi[\theta \mid \mathbf{z}_0, \mathbf{x}_0]$$

$$\Theta_0(\mathbf{x}_0, \mathbf{z}_0; L) = \arg \min_{\delta} \mathbb{E}_\pi[L(\delta, \theta) \mid \mathbf{x}_0, \mathbf{z}_0]$$

are the **best estimators** we can construct based on the observed data.

- They are **Bayes estimators** under the empirical prior $\pi(\cdot)$, which is unknown.
- Instead of converging to θ_0 , we hope our iGroup estimators converge to target estimators

$$\hat{\theta}_0^{(c)} \rightarrow \Theta_0 \quad \text{and} \quad \tilde{\theta}_0^{(c)} \rightarrow \Theta_0$$

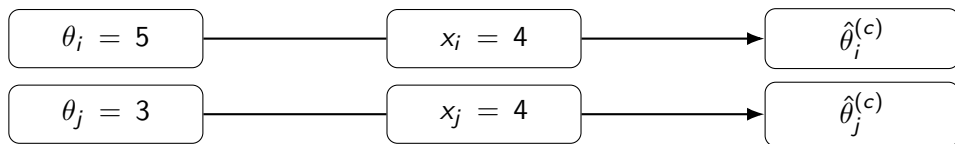
under certain loss $L(\cdot, \cdot)$.

Target Estimator

Example: imagine a model with

$$\mathbb{P}[x_k = \theta_k + 1] = 1/2$$

$$\mathbb{P}[x_k = \theta_k - 1] = 1/2$$



- Since we only observe $x_i = x_j = 4$, therefore, our estimators $\hat{\theta}_i^{(c)}$ should be the same $\hat{\theta}_j^{(c)}$.
- It is impossible for $\hat{\theta}_i^{(c)} = \hat{\theta}_j^{(c)}$ to converge to both $\theta = 5$ and $\theta = 3$.

Convergence of iGroup Estimators

First, we have the convergence of the following iGroup estimator:

$$\hat{\theta}_0^{(c)} = \frac{\sum_{k=0}^K \hat{\theta}_k w(k; 0)}{\sum_{k=0}^K w(k; 0)}$$

Theorem

If $\hat{\theta}_k$ is a sufficient and unbiased estimator for θ_k for $k = 0, \dots, K$. Under all previous assumptions, when $K \rightarrow \infty$, we have

$$\hat{\theta}_0^{(c)} \longrightarrow \Theta_0(\mathbf{x}_0, \mathbf{z}_0; \ell_2) \quad \text{in probability.}$$

Sufficient means $\hat{\theta}_k$ is a sufficient statistic for θ_k .

Convergence of iGroup Estimators

Next, we have the convergence of the following iGroup estimator:

$$\tilde{\theta}_0^{(c)} = \arg \min_{\theta} \sum_{k=0}^K M_k(\theta, \mathbf{x}_k) w(k; 0)$$

Theorem

If for any given $\hat{\theta}$, the objective function $M(\theta, \hat{\theta})$ as a function of θ is convex and second-order differentiable, and if loss L satisfies

$$\int M(\theta, \hat{\theta}) p(\hat{\theta} | \theta') d\theta = L(\theta, \theta') + C(\theta'),$$

then, under all previous assumptions, when $K \rightarrow \infty$, we have

$$\tilde{\theta}_0^{(c)} \longrightarrow \Theta_0(\mathbf{x}_0, \mathbf{z}_0; L) \quad \text{in probability.}$$

Convergence of iGroup Estimators

The requirement

$$\int M(\theta, \hat{\theta}) p(\hat{\theta} | \theta') d\theta = L(\theta, \theta') + C(\theta')$$

links the objective function to a loss function.

- When $M(\theta, \hat{\theta}) = -\log p(\hat{\theta} | \theta)$ is the negative log-likelihood, $L(\theta, \theta')$ corresponds to the Kullback-Leibler divergence between $p(\hat{\theta} | \theta)$ and $p(\hat{\theta} | \theta')$.

In summary:

- The iGroup estimator, by **aggregating individual estimators**, converges to the Bayes estimator under **squared loss**.
- The iGroup estimator, by **aggregating log-likelihood functions**, converges to the Bayes estimator under **K-L divergence loss / cross-entropy loss**.

Further Results on Risk Decomposition

- When $\hat{\theta}$ is not available for all the individuals, we construct our estimator on \mathbf{z}_k and aggregate objective functions targeting $\Theta_0(\mathbf{z}_0; L)$
- When \mathbf{z}_k is not available for all the individuals, we construct our estimator solely on $\hat{\theta}_k$'s targeting $\Theta_0(\mathbf{x}_0; L)$
- For using only \mathbf{x} or \mathbf{z} , **all previous theoretical results hold.**
- But using different sets of information ($\hat{\theta}$, \mathbf{z} or both) can result in different risks.

Further Results on Risk Decomposition

Proposition (Full Risk Decomposition)

Let $\hat{\theta}_0^{(c)}$ be any iGroup estimator by aggregating individual estimators, and let Θ_0 be its targeting estimator. Then,

$$R(\hat{\theta}_0^{(c)}) = R_{np}(\hat{\theta}_0^{(c)}) + R_{target}(\Theta_0),$$

where $R(\hat{\theta}_0^{(c)}) = \mathbb{E}[(\hat{\theta}_0^{(c)} - \theta_0)^2]$ is the overall risk, $R_{np}(\hat{\theta}_0^{(c)}) = \mathbb{E}[(\hat{\theta}_0^{(c)} - \Theta_0)^2]$ is risk premium from nonparametric estimation, and $R_{target}(\Theta_0)$ is the risk of the target estimator.

Furthermore, if we use partial information, then

$$R_{target}(\Theta_0) = R_{inf}(\Theta_0) + R_0,$$

where $R_{inf}(\Theta_0) = \mathbb{E}[(\Theta_0 - \Theta_0(\mathbf{x}_0, \mathbf{z}_0; \ell_2))^2]$ is the risk premium from information loss, and R_0 is the risk of $\Theta_0(\mathbf{x}_0, \mathbf{z}_0; \ell_2)$.

Further Results on Risk Decomposition

Risk decomposition for different iGroup models.

	iGroup Set	R_{np}	R_{target}	
			R_{inf}	R_0
Case 1	$\{z\}$	$\asymp K^{-4/(d+4)}$	> 0	same value
Case 2	$\{\hat{\theta}\}$	$\asymp K^{-1}$	> 0	
Case 3	$\{\hat{\theta}, z\}$	$\asymp K^{-4/(d+4)}$	$= 0$	

Table of Contents

- 1 Introduction
- 2 Individualized Group Learning
- 3 Theoretical Results
- 4 Simulations and Real Examples**
- 5 Summary

Short Time Series

- We simulate $K = 200$ short AR(1) time series.
- The AR(1) coefficients $\theta_1, \dots, \theta_{200}$ follow:

$$\frac{\theta_k + 1}{2} \sim \text{Beta}(4, 4), \quad k = 1, \dots, 200$$

- Each time series is generated by

$$x_{k,0} \sim \mathcal{N}\left(0, \frac{\sigma^2}{1 - \theta_k^2}\right)$$
$$x_{k,t} = \theta_k x_{k,t-1} + \epsilon_{k,t}, \quad k = 1, \dots, 200, \quad t = 1, \dots, 10,$$

where $\epsilon_{k,t} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 3$.

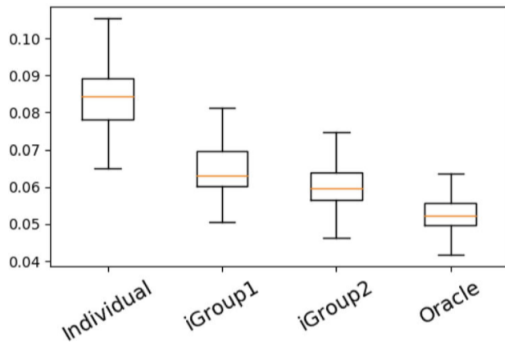
- No exogenous variable \mathbf{z}_k is involved!

- **Individual:** estimate each $\hat{\theta}_k$ from its own data (11 observations).
- **iGroup1:** aggregating log (conditional) likelihood function with weights estimated from the bootstrap samples.
- **iGroup2:** aggregating individual estimators with weights estimated from the bootstrap samples.
- **Oracle:** directly compute the posterior mean using the Beta prior.

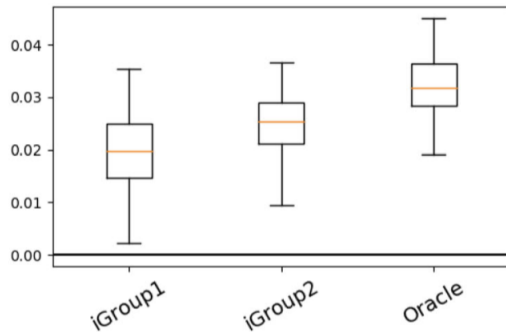
The whole simulation (including data generation) is repeated 100 times.

Short Time Series — Comparison

MSE



MSE Improvement



VaR Analysis based on Fama-French Factors

- $\alpha\%$ Value-at-Risk (VaR) of a stock is the $\alpha\%$ lower quantile of its upcoming daily return.
- $r_{t,k}$: daily return of stock k on day t .
- Exogenous variables: Fama-French three factors:

$$r_{t,k} = \alpha_{t,k} + r_f + b_{0,t,k}(MKT_t - r_f) + b_{1,t,k}SMB_t + b_{2,t,k}HML_t + \epsilon_{t,k}$$

The coefficients $b_{0,t,k}$, $b_{1,t,k}$ and $b_{2,t,k}$ are used as exogenous variables.

- We assume the dynamics of stock price change slowly such that in a look-back window of length S , the coefficients are constant.

- **Individual Estimator:**

$$\widehat{\text{VaR}}_{ind}(t, k) = Q_\alpha \left(\bigcup_{s=0}^{S-1} \{r_{t-s, k}\} \right)$$

where Q_α is the α -quantile function.

- **Market Estimator:**

$$\widehat{\text{VaR}}_{mkt}(t, k) = Q_\alpha \left(\bigcup_{k=1}^K \bigcup_{s=0}^{S-1} \{r_{t-s, k}\} \right)$$

- **Quantile Regression:** We fit a linear model of empirical quantile against the F-F coefficients. Use one-day prediction as the VaR.

VaR Analysis based on Fama-French Factors

- **iGroup Estimator:**

$$\widehat{\text{VaR}}_{igroup}(t, k) = Q_{\alpha}^{(w)} \left(\bigcup_{l=1}^K \bigcup_{s=0}^{S-1} (r_{t-s,l}, w(l; k)) \right),$$

where $Q_{\alpha}^{(w)}$ is the weighted empirical quantile.

- ▶ The weight function $w(k; l)$ can be built based on \mathbf{z} (F-F coefficients) or $\widehat{\text{VaR}}_{ind}$ or both.
- ▶ The iGroup estimator is equivalent to aggregating objective functions.

$$\widehat{\text{VaR}}_{igroup}(t, k) = \arg \min_{\theta} \sum_{l=1}^K M_k(\theta; t) w(l; k)$$

with

$$M_k(\theta; t) = \sum_{s=0}^{S-1} |r_{t-s,k} - \theta| (\alpha \mathbf{1}_{\{r_{t-s,k} > \theta\}} + (1 - \alpha) \mathbf{1}_{\{r_{t-s,k} \leq \theta\}})$$

VaR Analysis based on Fama-French Factors

- Dataset: 491 stocks from S&P 500. Predict returns for 250 trading days in 2016.
- Look-back period: $S = 100$ days.
- We measure the performance of the VaR estimators by

$$\text{RMSE} = \left(\frac{1}{491} \sum_{k=1}^{491} \left(\frac{1}{250} \sum_{t=1}^{250} \mathbf{1}_{\{r_{t,k} \leq \widehat{\text{VaR}}(t,k)\}} - 0.01 \right)^2 \right)^{1/2}$$

Method	Individual	Market	iGroup($\hat{\theta}$)	Quantile Reg	iGroup(z)	iGroup($z, \hat{\theta}$)
RMSE ($\times 10^{-3}$)	9.61	13.4	6.63	29.8	5.75	5.54

Table of Contents

- 1 Introduction
- 2 Individualized Group Learning
- 3 Theoretical Results
- 4 Simulations and Real Examples
- 5 Summary**

Comparison to Other Methods

- **Nonparametric Regression / K-Nearest Neighbor:** we have special design for similarity between $\hat{\theta}_k$ and we allow aggregating objective functions.
- **Empirical Bayes:** we do not directly estimate prior $\pi(\cdot)$.
- **Hierarchical Bayes:** we do not assume a parametric form for $\pi(\cdot)$
- **Mixed-Effects Model:** we do not assume a Gaussian prior for θ .
- **Individualized Fusion Method:** We consider $K \rightarrow \infty$ but $\text{len}(\mathbf{x}_k)$ fixed.
- **Transformer / Self-attention:**
 - ▶ Similarity: $\hat{\theta}$ weighted by a similarity measure between $\hat{\theta}$'s.
 - ▶ Difference: we aim for a Bayes estimator instead of feature extraction.

Conclusion

- We proposed a framework of individualized group learning with a novel weight function.
- The iGroup estimators converge to the best estimator one can obtain – Bayes estimator.
- The flexibility in combining individual estimators or objective functions broadens the application of the proposed method.
- Both simulation and real data examples show the superior performance of the iGroup.

References:

Individualized Group Learning. Cai, Chen, & Xie. *Journal of the American Statistical Association (JASA)*.

Individualized Inference through Fusion Learning. Cai, Chen, & Xie. *Wiley Interdisciplinary Reviews: Computational Statistics*

Thank you for listening!