# FedGMMAT: Federated Generalized Linear Mixed Model Association Tests

Wentao Li
BMI 6313
*The School of Biomedical Informatics |The University of Texas Health Science Center at Houston*

## Acknowledgment

This research is supported by Dr. Xiaoqian Jiang, Dr. Arif Harmanci and Dr. Han Chen at UTHealth.

***Scan the QR code for the recorded presentation***

## Introduction

The generalized Linear Mixed Model Association Test (GMMAT) is a well-known approach for genome-wide analysis (Chen, H, et al., 2016). Due to the privacy awareness in Protected Health Information (PHI), and genetic information is restrictively protected by sharing according to Health Insurance Portability and Accountability Act (HIPAA), GMMAT performance will be limited.

Federated Learning (Li, T, et al., 2020) is a privacy-preserving machine learning technique that promises collaborative training on isolated datasets without actually sharing the raw data. Motivated by this concept, we proposed a GMMAT algorithm under Federated Learning settings (Figure 1).

## Step 1: Federated Learning on the null model

a) **Broadcast model parameters:** The central server initiates the process of null model learning by broadcasting the parameters to clients.

b) **Upload local information:** Each federated client will calculate and send local information (i.e. Gradient and Hessian matrix) to the central server.

c) **Update model parameters:** The central server will gather the model information from clients and update the model parameters with the aggregated information. Then broadcasts the up-to-date model parameters to all clients.

d) **Broadcast model status:** After the update on model parameters, the central server will determine whether the model is converged or not by the log-likelihood score. If the difference from the last iteration is within 10^(-6), then stop the iteration. If not, goes to step a).

The communicated information during Step 1 is listed in Table 2.

## Step 2: Federated Learning on the mixed-effects model

a) **Broadcast model parameters:** The central server will set the parameters from the null model as the initial fix-effect coefficients, and the variance of targets as the initial mixed-effect hyper-parameters. Then send them to the federated clients.

b) **Upload local information:** Once the federated client receives the parameters, computes and uploads the local information.

c) **Update model parameters:** The central server aggregates then updates mixed-effect hyper-parameters, fixed-effect coefficient, and mixed-effect coefficient. And broadcast the up-to-date parameters to all federated clients.

d) **Broadcast model status:** After the update on model parameters, the central server will determine whether the model is converged with the difference of parameters from the last iteration. If it has not converged, go to step b).

The communicated information during Step 2 is listed in Table 2.

## Step 3: Federated Learning on the Score test

Once the parameter space is optimized under the objective, each federated node can collaboratively compute the score test with the mixed-effects model on their local genotype data. First, the central server will broadcast the projection matrix to each federated node. Then the score test statistics can be collected and aggregated from federated nodes.

The communicated information during Step 3 is listed in Table 2.

## Tables

Table 1. **Coefficients of GMMAT and FedGMMAT**

| | fixed-effect | | | mixed-effect |
|---|---|---|---|---|
| | Intercept | age | sex | |
| GMMAT | 0.4721 | -0.0068 | -0.0864 | 0.3377 |
| fedGMMAT | 0.4721 | -0.0068 | -0.0864 | 0.3377 |

Table 2. **Communication information**

| | Client to Server | Data size | Server to Client | Data size |
|---|---|---|---|---|
| Null model | $enc(\sum_{j=1}^{k} \tilde{\ell}_j^{(t)})$ | $p$-dim vector | $\boldsymbol{\alpha}^{(t)}$ | $p$-dim vector |
| | $enc(\sum_{j=1}^{k} \mathbf{H}_j^{(t)})$ | $p \times p$ matrix | $pk$ | scalar |
| Mixed model | $enc(\sum_{j=1}^{k} \boldsymbol{\Xi}_j)$ | $p \times p$ matrix | $\boldsymbol{\alpha}^{(t)}$ | $p$-dim vector |
| | $enc(\sum_{j=1}^{k} \boldsymbol{\Omega}_j)$ | $n \times p$ matrix | $\boldsymbol{b}^{(t)}$ | 2-dim vector |
| | $enc(\tilde{\mathbf{Y}})$ | $n$-dim vector | $\hat{\tau}$ | scalar |
| | $enc(\sum_{j=1}^{k} \boldsymbol{\Psi}_j)$ | $n$-dim vector | $pk$ | scalar |
| Score test | $enc(\sum_{j=1}^{k} T_j)$ | $nSNPs$-dim vector | $pk$ | scalar |
| | | | $P_{jj}$ | $n_j \times n_j$ matrix |

Intermediate information with data sizes being communicated during the Federated Learning training

## Diagram

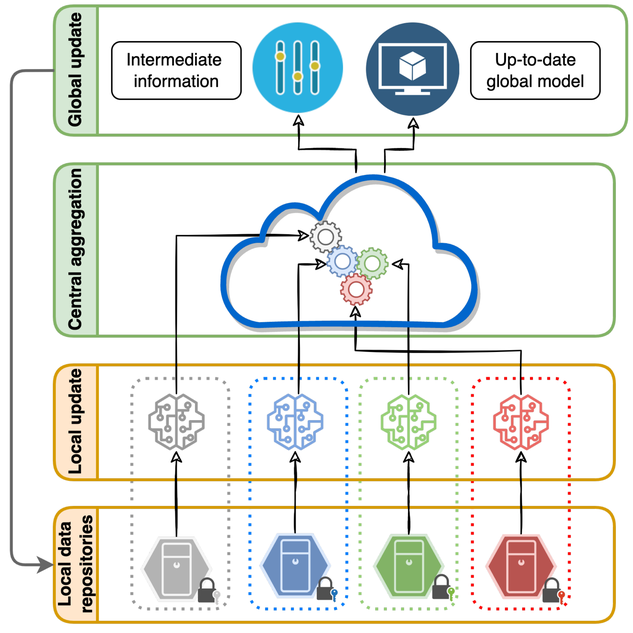


Figure 1. **Flowchart of FedGMMAT.** In the FedGMMAT framework, each local data repository will maintain its unique dataset locally and gather intermediate model information from Global updates. Then each local data repository will compute a local model with their protected data, and transfer part of the local model information to the central server for global aggregation.

## Conclusion

In this study, we show the lossless performance of fedGMMAT by using the same synthetic datasets in the R package `GMMAT'. The synthetic dataset includes 400 samples and 100 SNPs. And the model will fit a GLMM with age and sex as covariates and disease status as the outcome. The coefficients of fixed-effects and mixed-effects of GMMAT and fedGMMAT are shown in Table 1.

we are well aware that there are some limitations in current work, including the missing protection layer when transmitting the model information, lacking asynchronous learning ability, and stress on communication bandwidth. Thus, our future work will focus on refining the current method by adding secure multi-computation protections, embedding tree-based asynchronous communication framework, and using lightweight communication protocols.

## References

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., ... & Lin, X. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653-666.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, *37*(3), 50-60.

***Please contact via email: wentao.li@uth.tmc.edu***