

Editing, Imputation, and Synthesis: A Public Use File for the Census of Manufactures

Hang Kim

NISS / Duke University

2015 Affiliates Annual Meeting, Miami, FL
Sunday, March 15

- ✧ Joint work with Jerry Reiter, Alan Karr, and Larry Cox
- ✧ Research supported by NSF [SES-11-31897]

United States
**Census
2000**



**AMERICAN
COMMUNITY
SURVEY**
U.S. CENSUS BUREAU



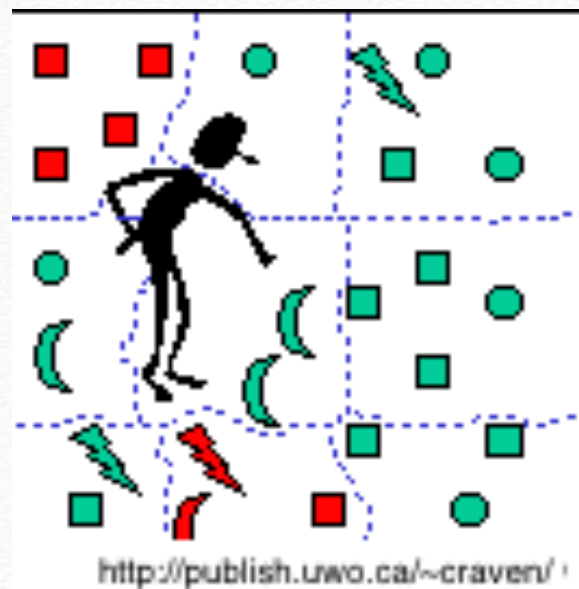
Integrated Public Use Microdata Series
census microdata for social and economic research

Background

Disseminate Public Use File

Production of Public Use File

Survey Design



Survey Data Collection



Data Processing



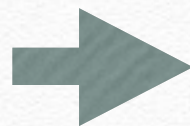
Publication

Current Practice: Data Processing

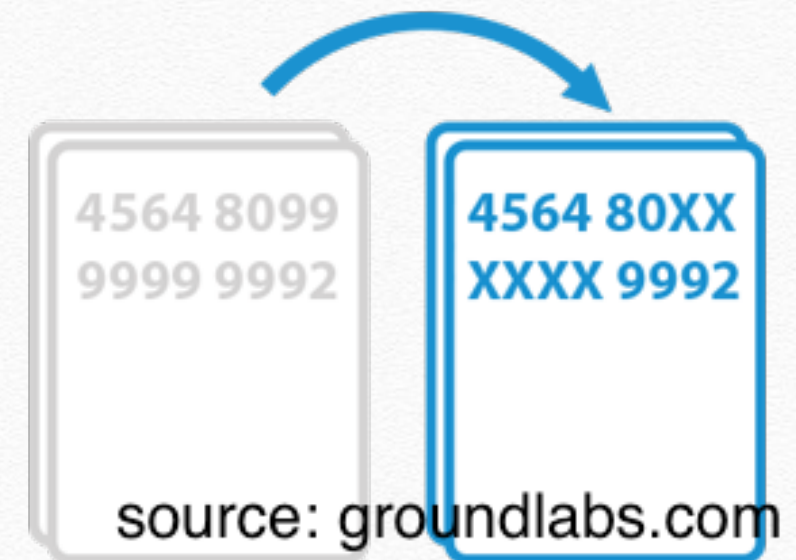
Initial Check / Recontact



Imputation



Data Editing



Disclosure Control/
Data Masking

Edit Rules

- **Logical constraints** which are satisfied by reported records to be considered plausible and consistent
 - find unacceptable errors in survey data
 - e.g. pregnant male, \$2M of avg. salary
 - specify space of reasonably imputed values
- Common edit rules for continuous values
 1. Range restriction e.g. *total emp. > 0*
 2. Ratio edit e.g. *total salary / total emp. < \$1M*
 3. Balance edit
 - e.g. *total emp. = production workers + other emp.*

Three Related Research Topics

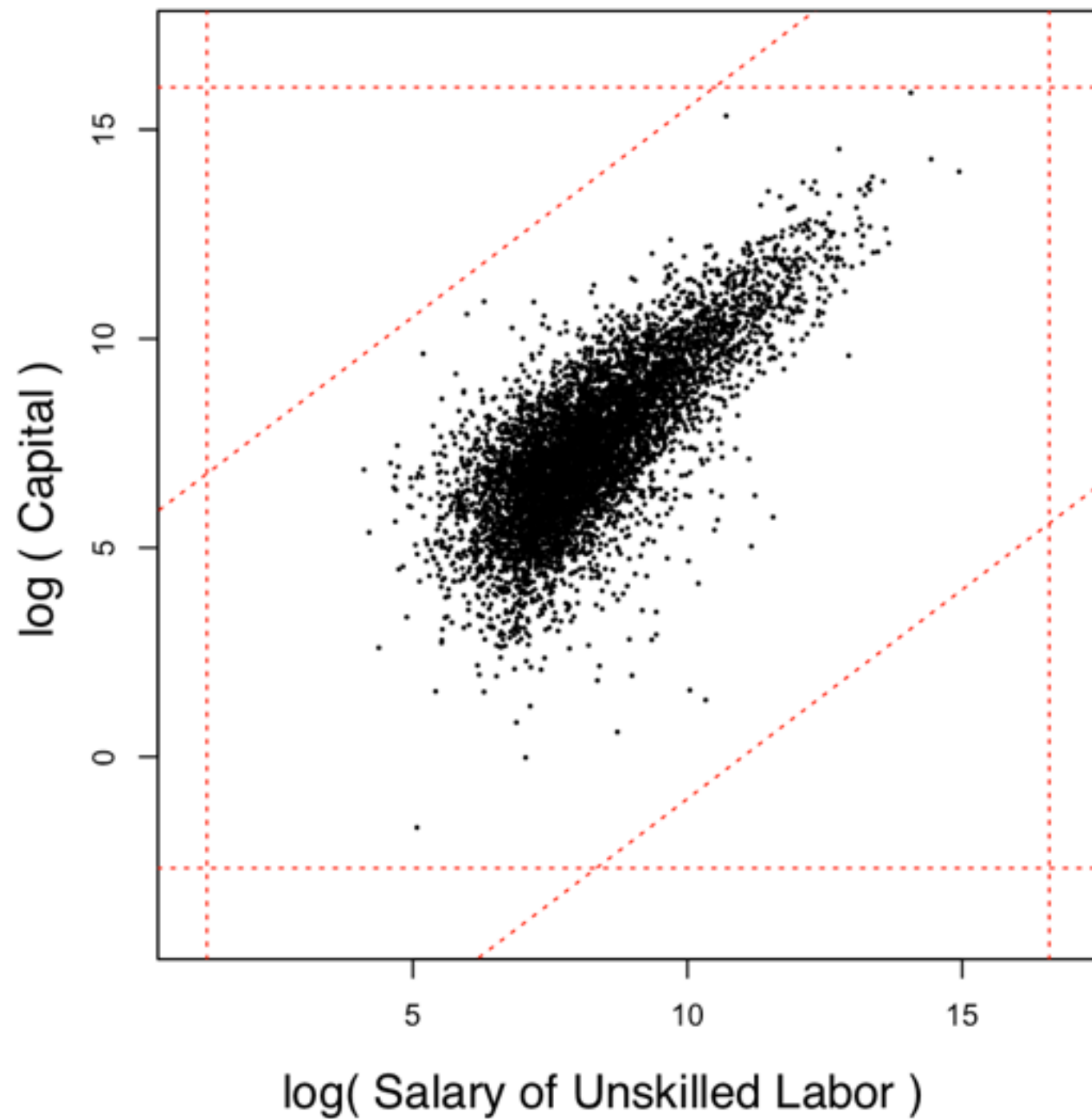
1. **Imputation** under linear constraints
 - Technical Report No. 182, NISS
 - Journal of Business and Economic Statistics, 2015, Vol 32
2. Simultaneous **data editing** and imputation
 - Technical Report No. 189, NISS
3. **Synthetic** microdata for the U.S. Census of Manufactures
 - work in progress

Topic I

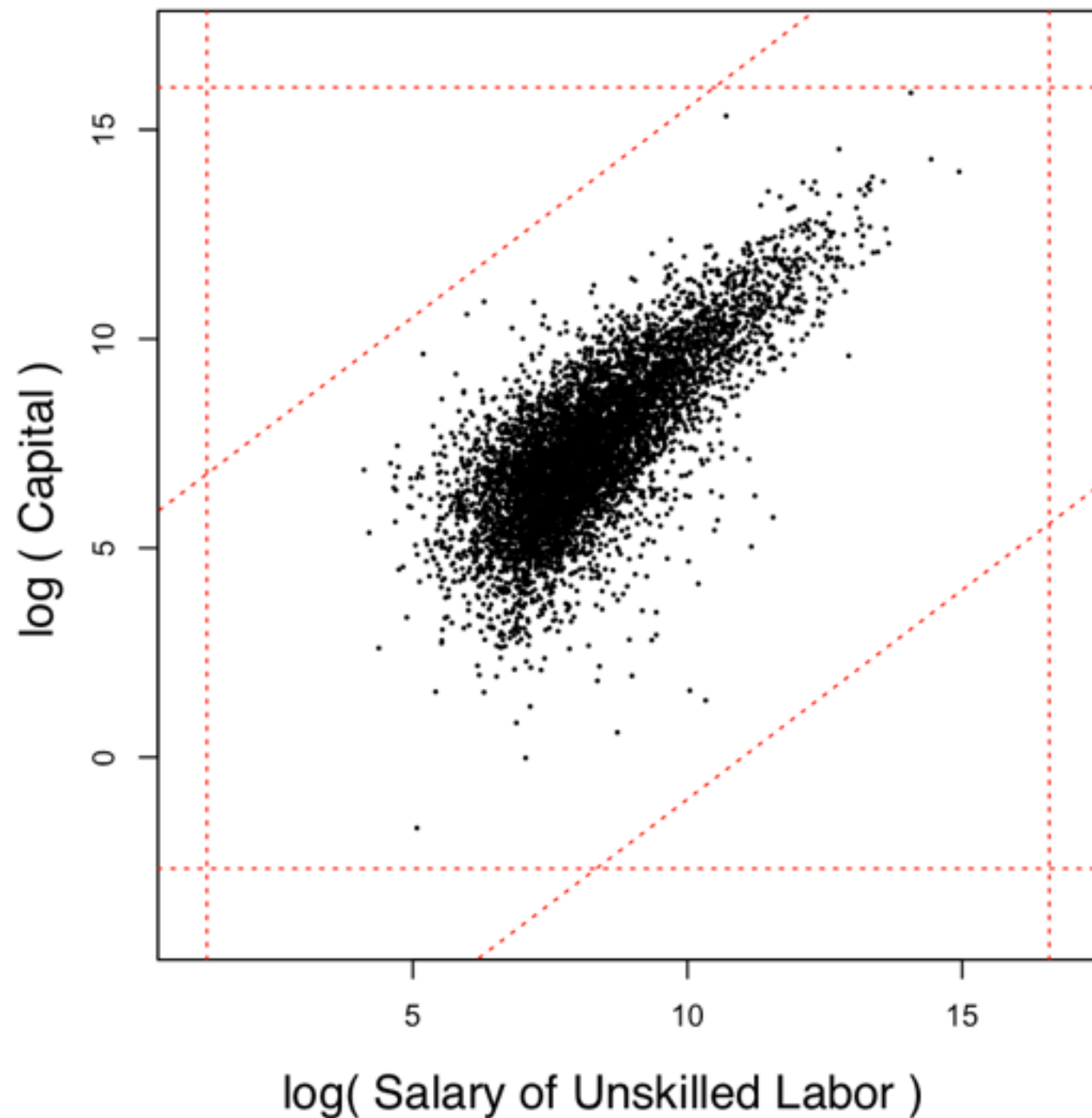
Imputation under linear
constraints



Example: Colombian Manufacturing Survey



Example: Colombian Manufacturing Survey



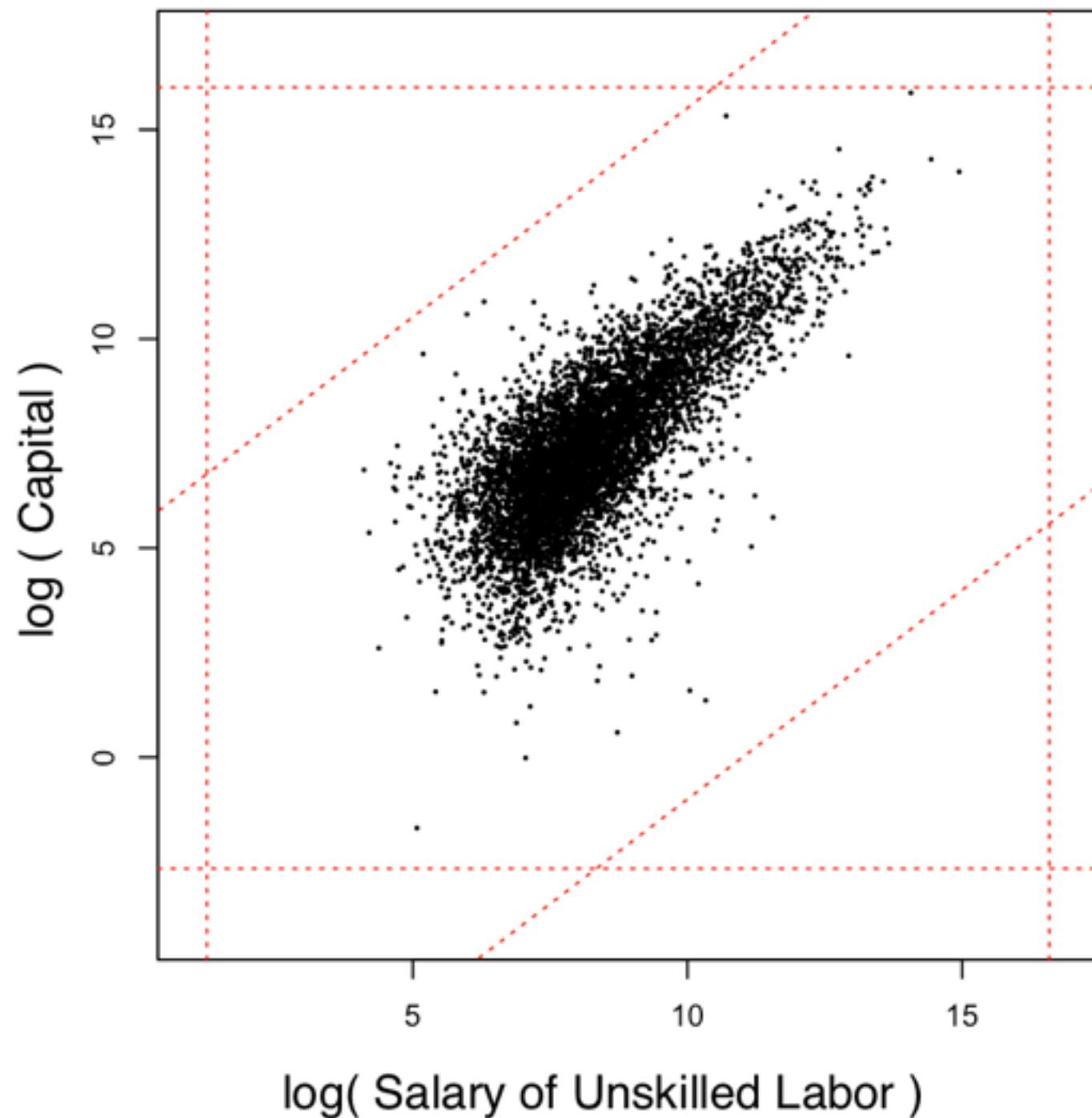
Similar to U.S. Census of Manufactures

Example: Colombian Manufacturing Survey



Data 1977-1991 have been used for researchers

Example: Colombian Manufacturing Survey



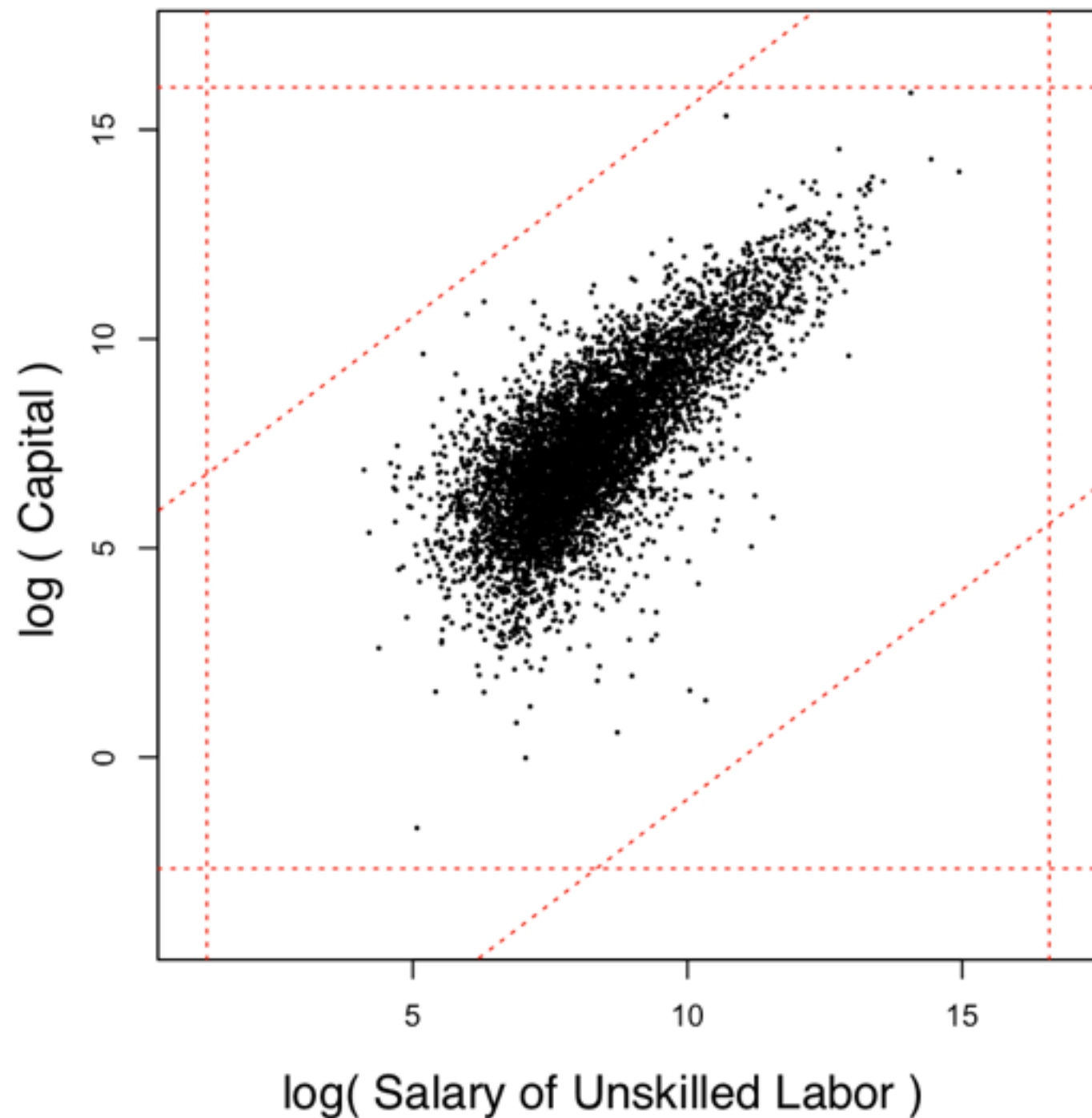
All variables are continuous

Example: Colombian Manufacturing Survey



Complex feasible region given edit rules

Example: Colombian Manufacturing Survey



Not easy to assume a parametric distribution

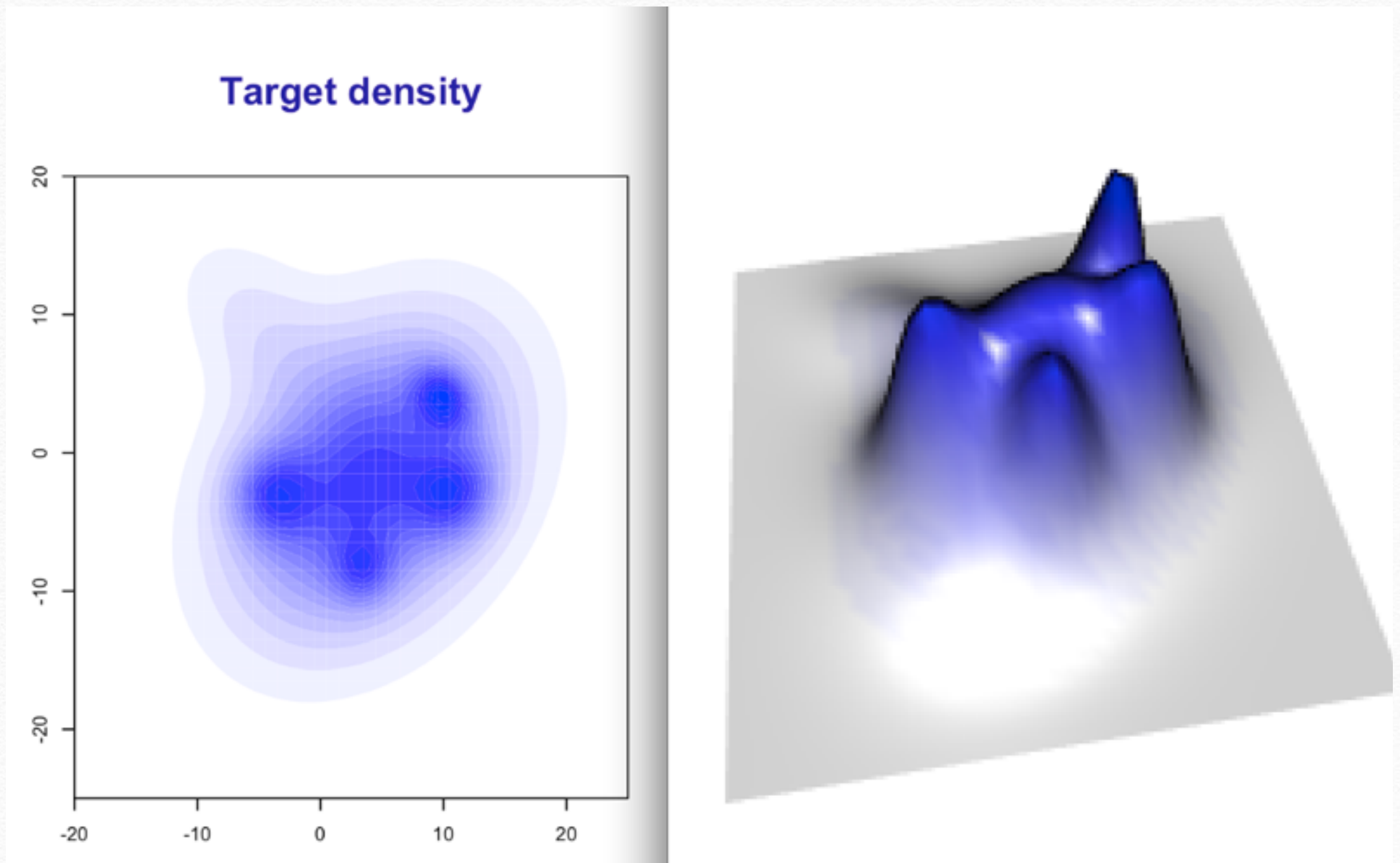
Joint Modeling Imputation (NISS report 182)

- Nonparametric Bayesian Model
 - use mixture normals with Dirichlet process (DP) priors
 - to capture complex features of data under very weak distributional assumption
 - restrict support under constraints regions
 - to guarantee that imputed values satisfy edit rules
- Multiple Imputation Approach
 - to capture uncertainty introduced by missing values and imputation process

Illustration of Mixture Normals

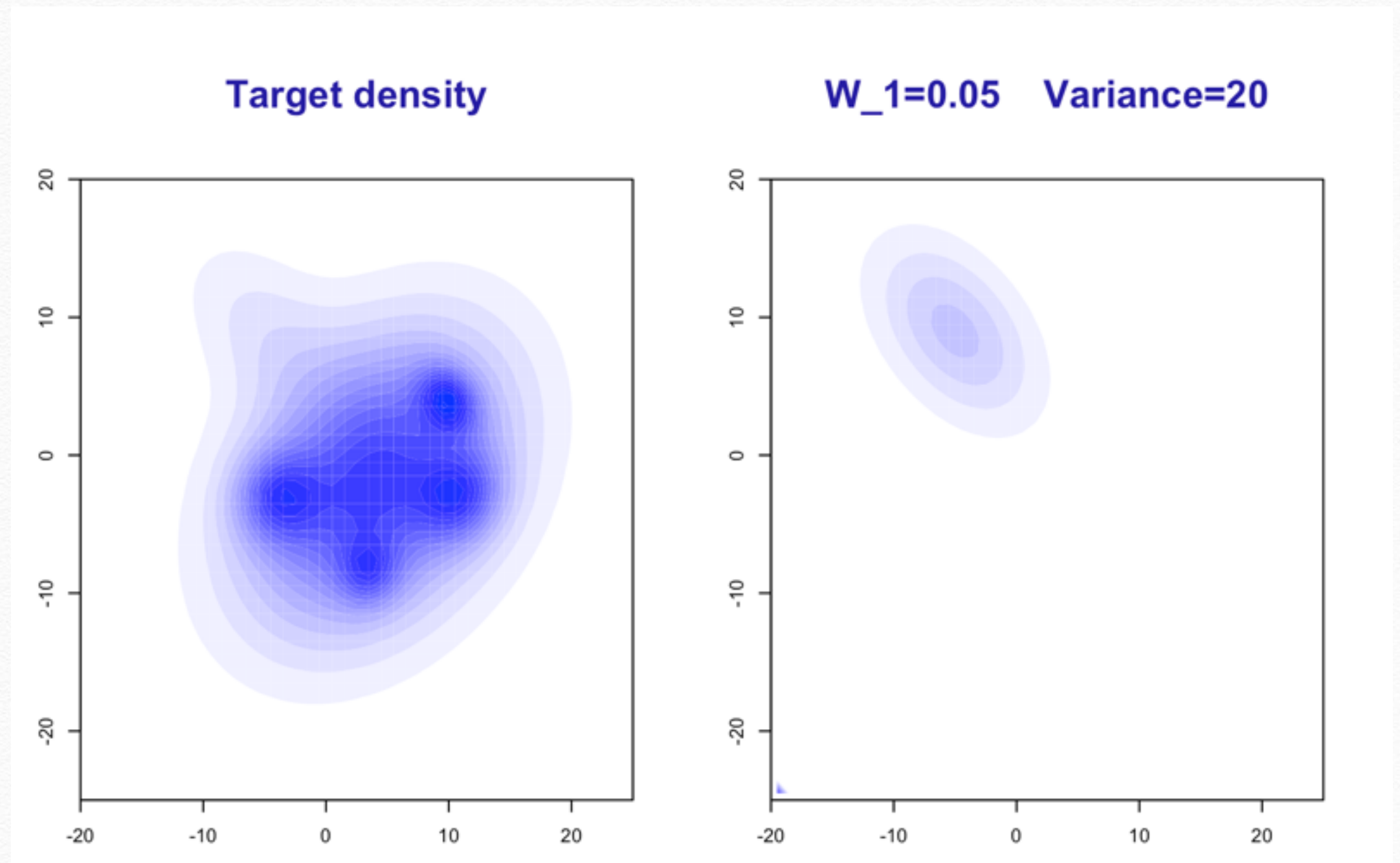
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



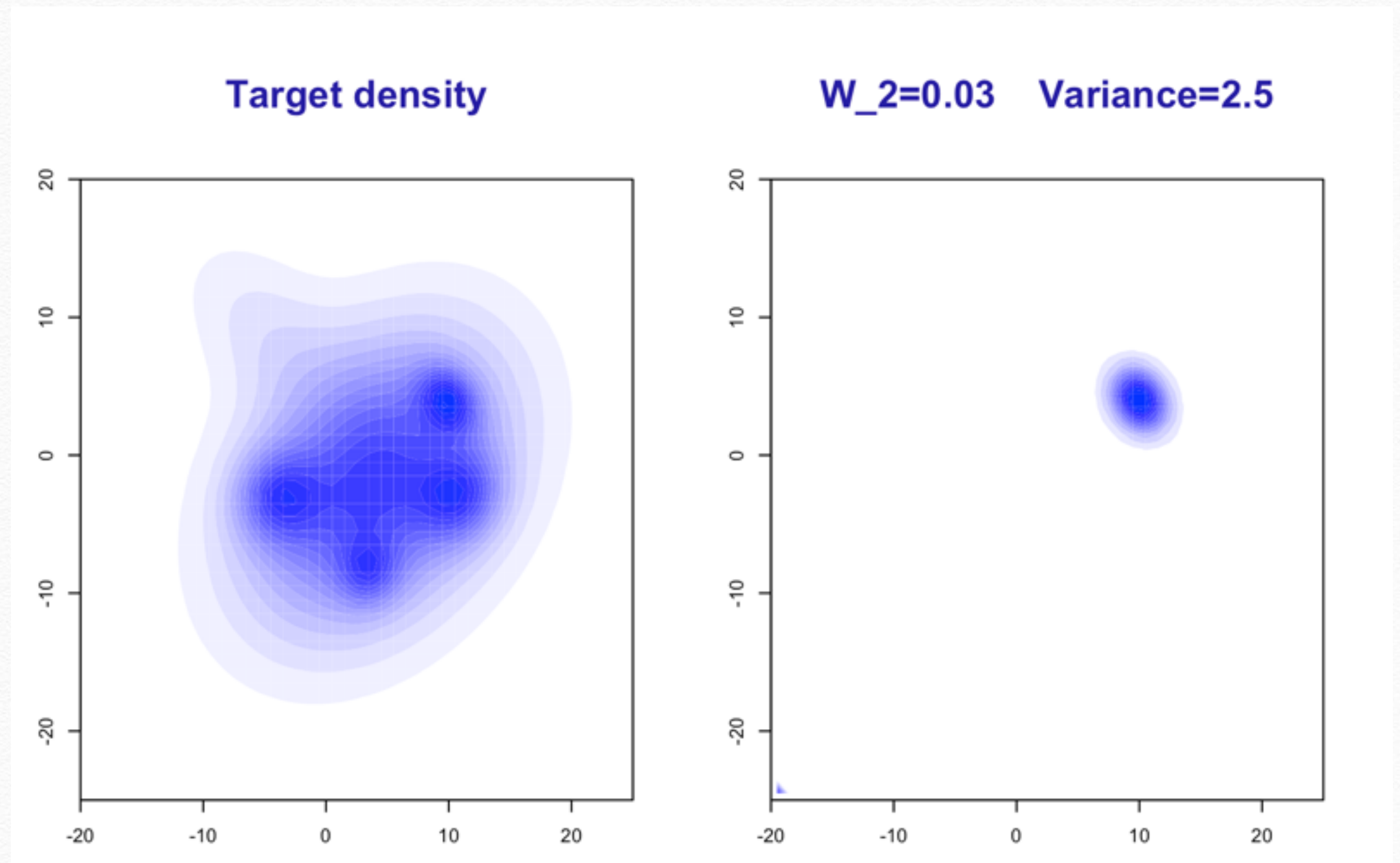
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



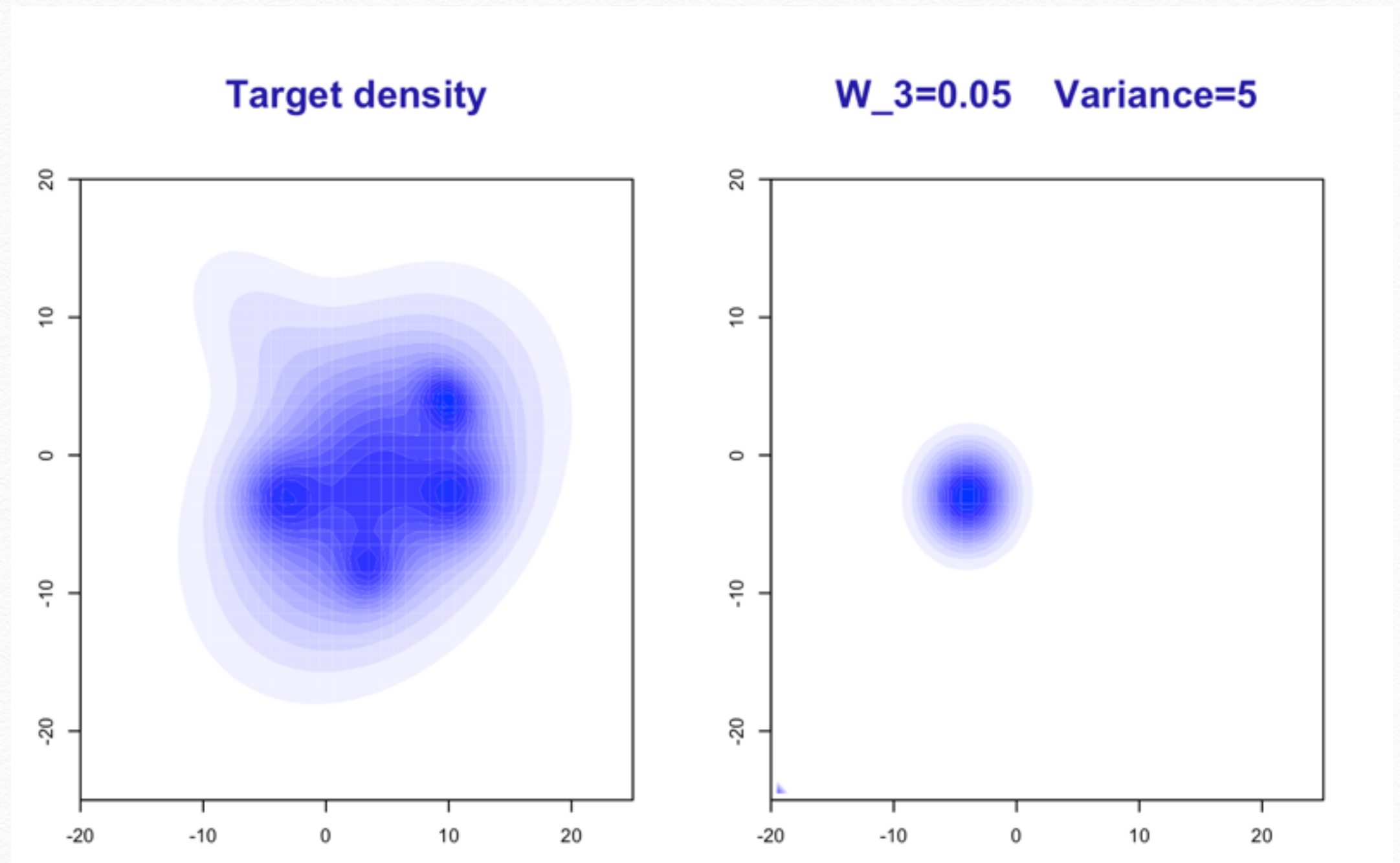
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



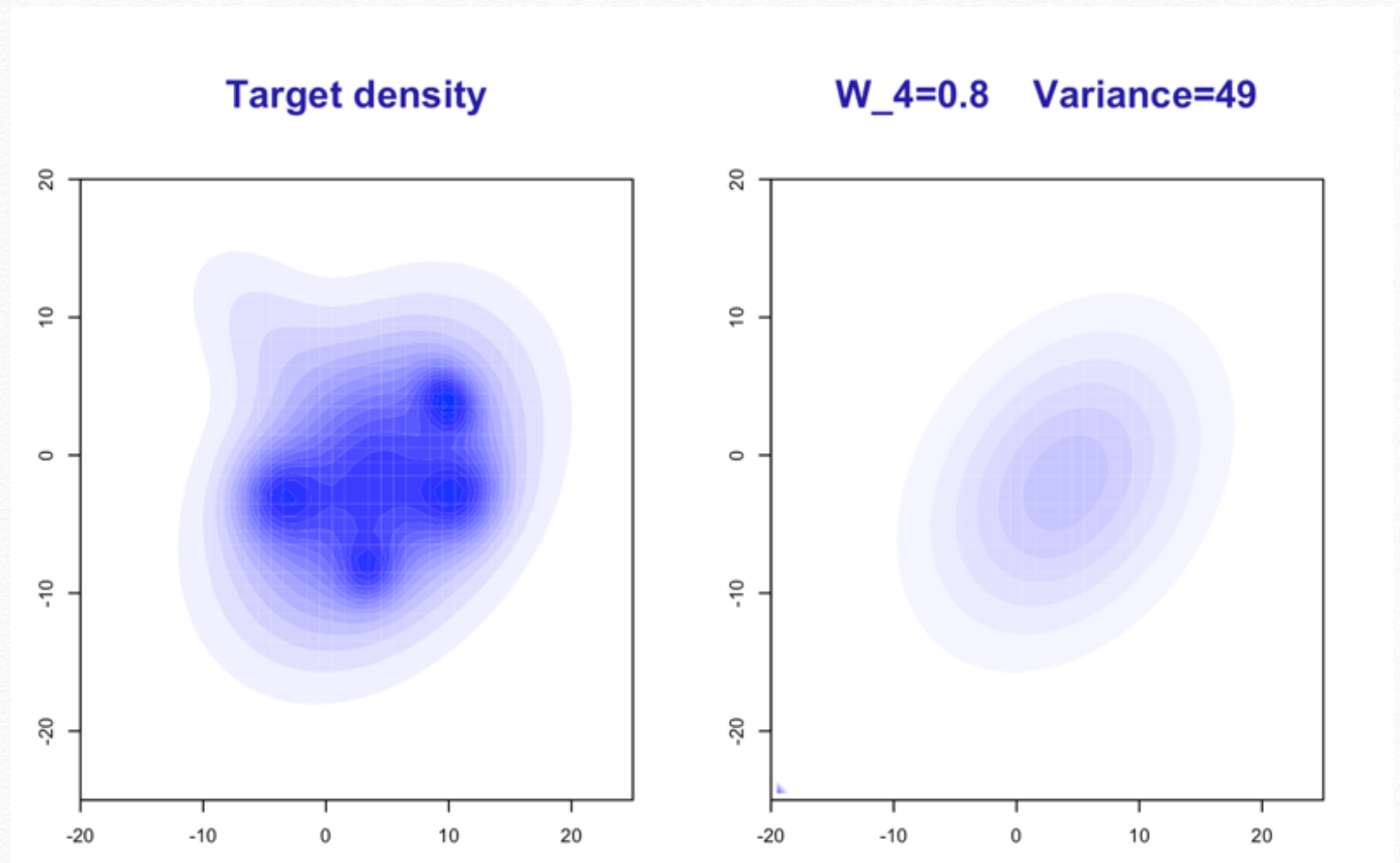
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



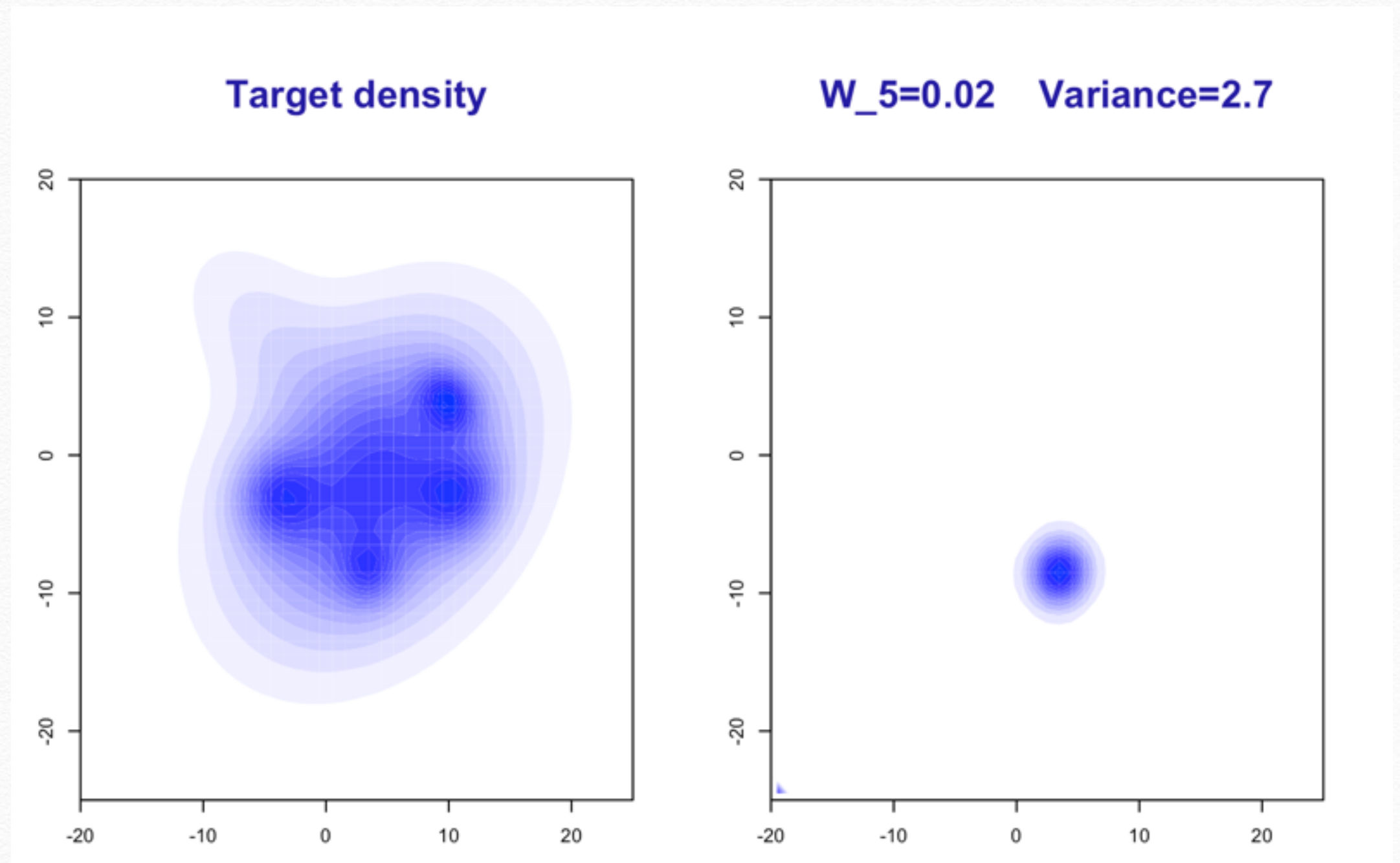
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



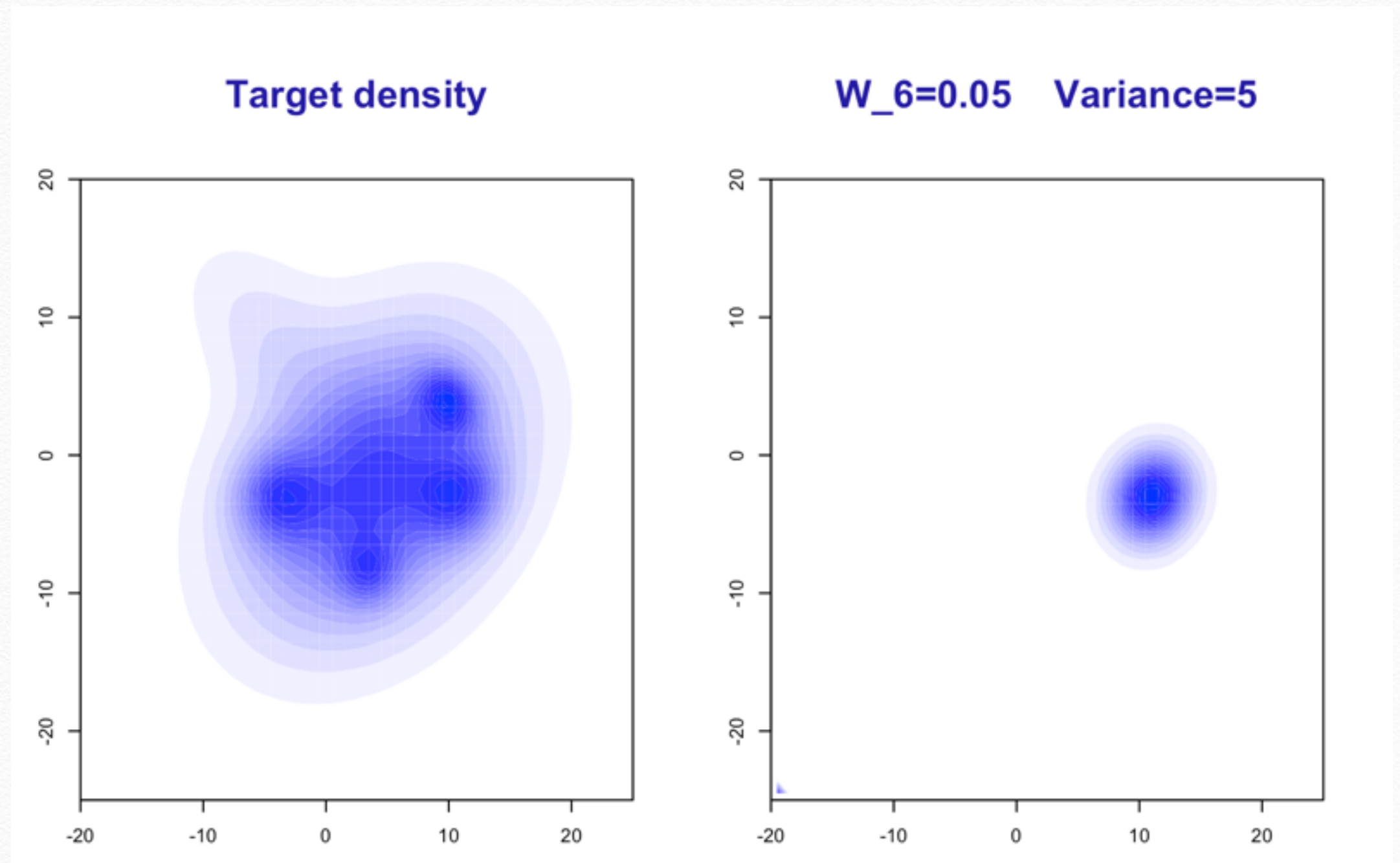
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



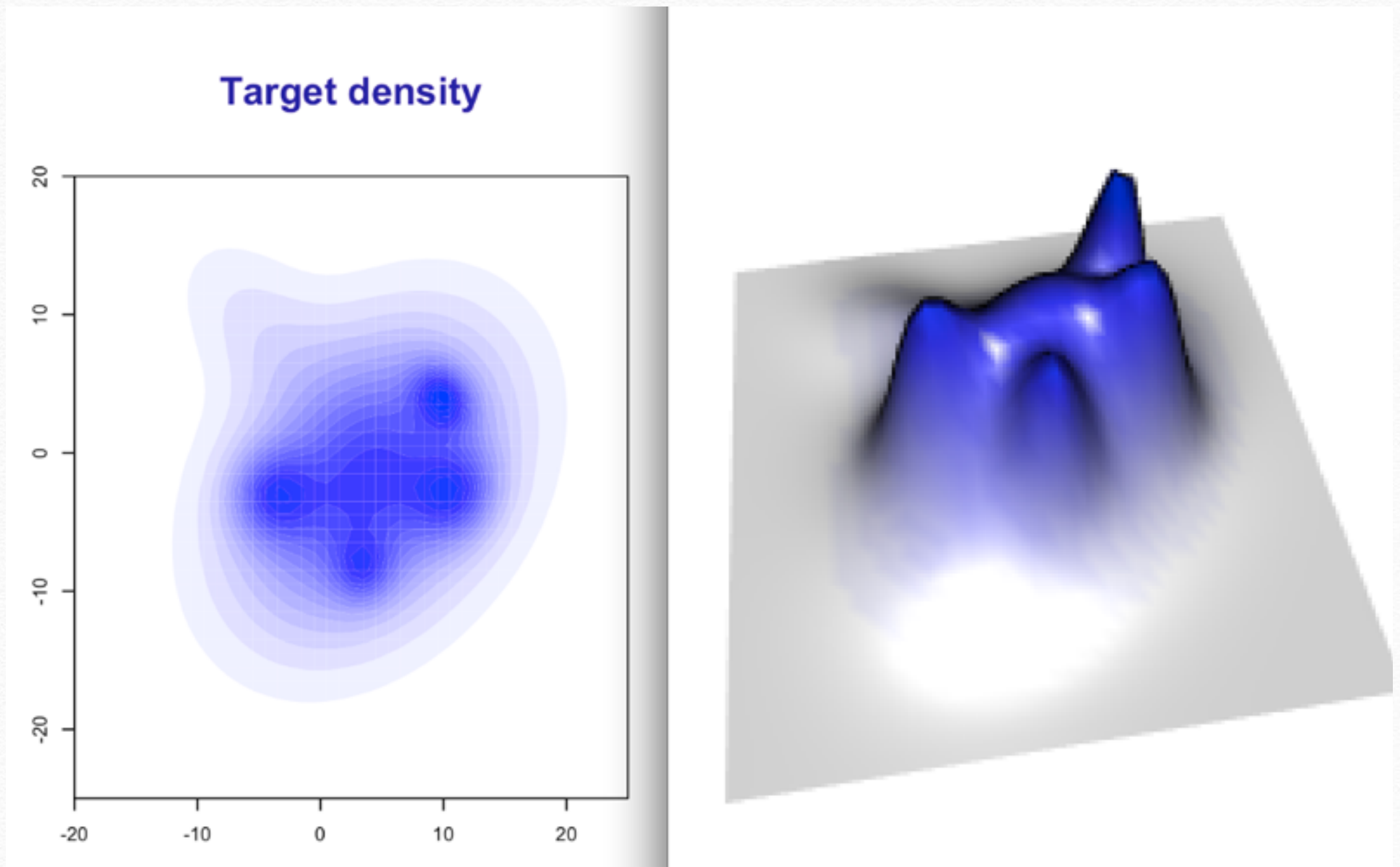
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals



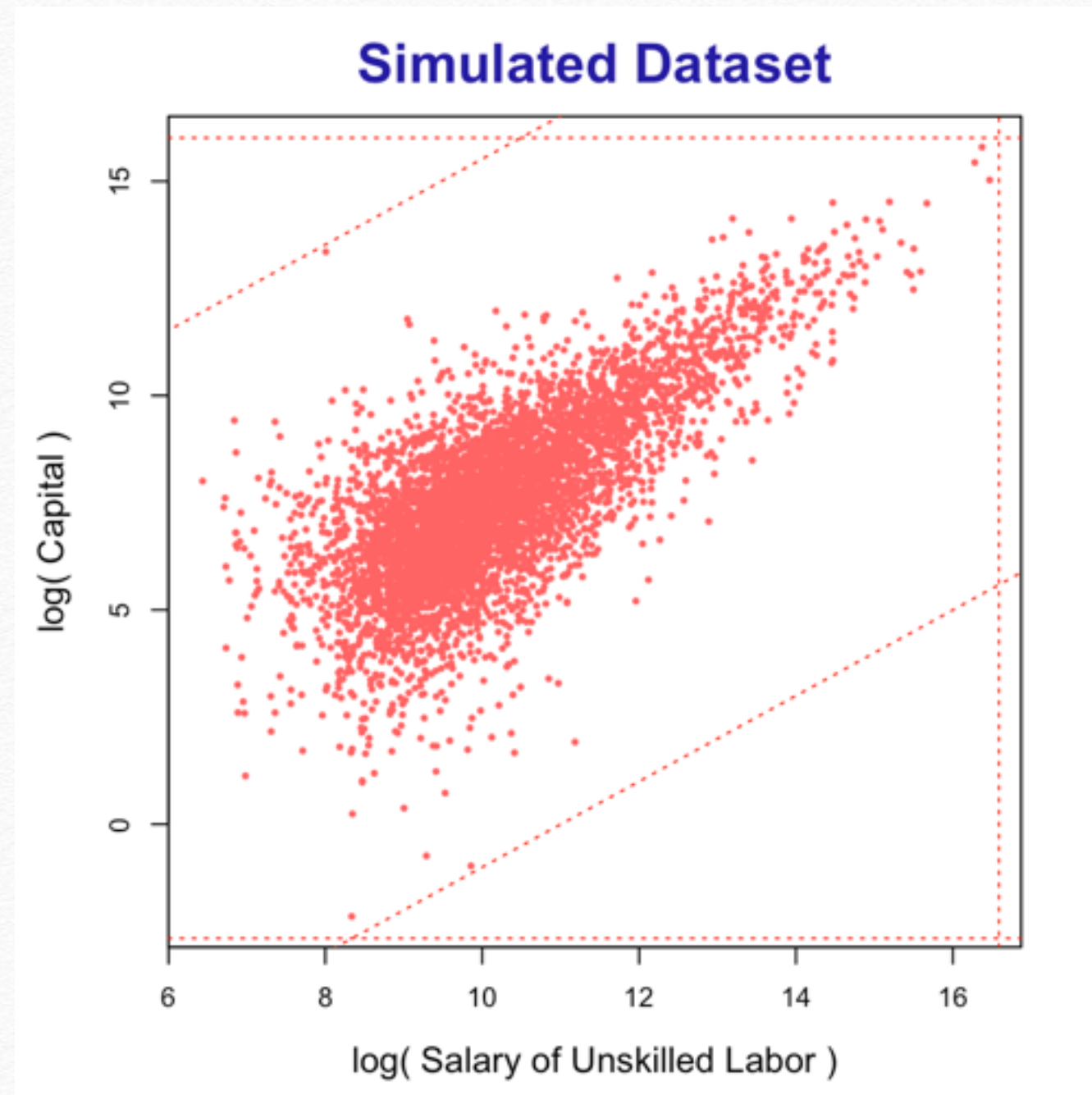
- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Illustration of Mixture Normals

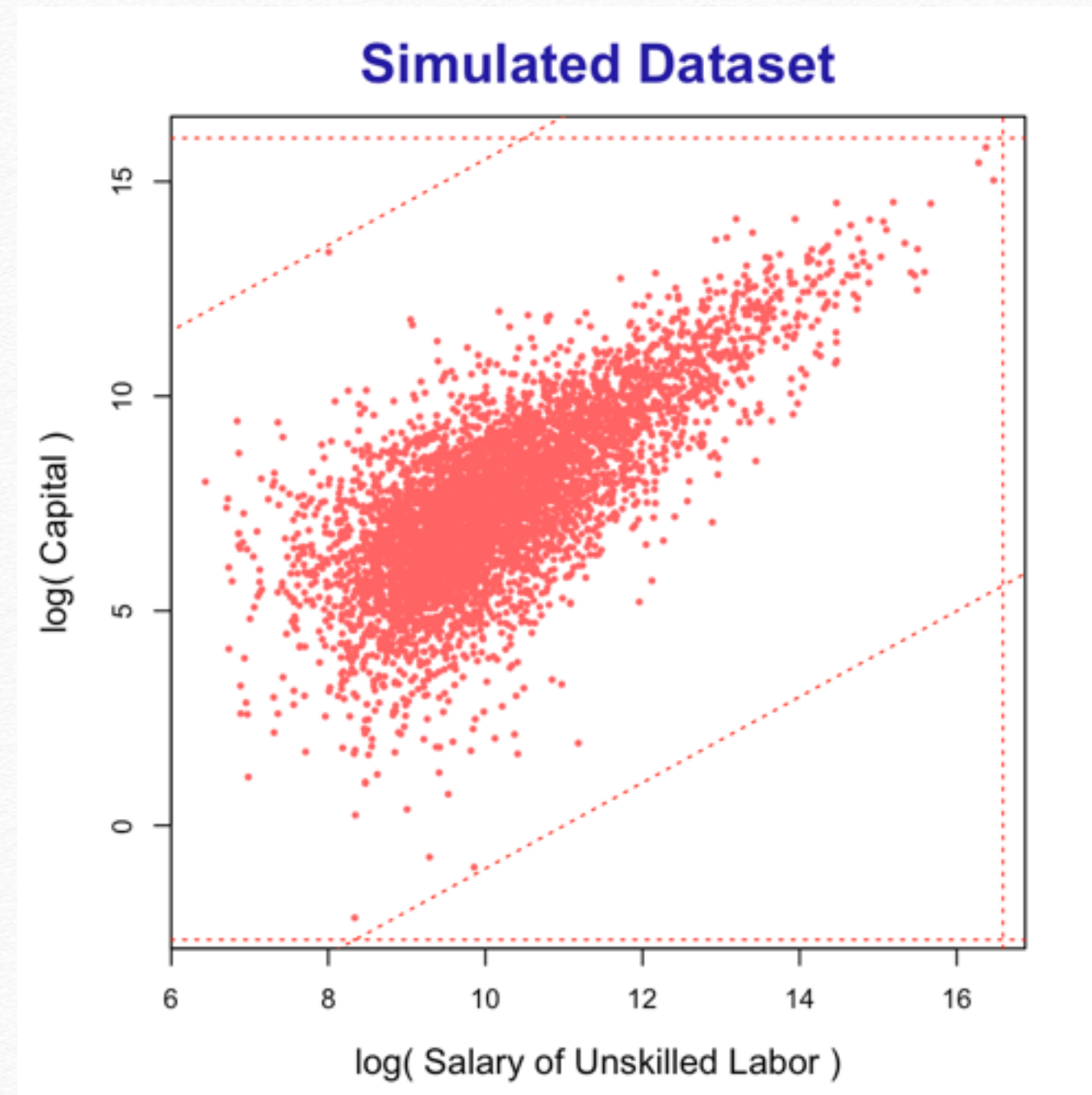


- ❖ Dirichlet process (DP) prior helps the model stochastically decide the number of components and weights

Simulation Study using Colombian Manufacturing Survey data

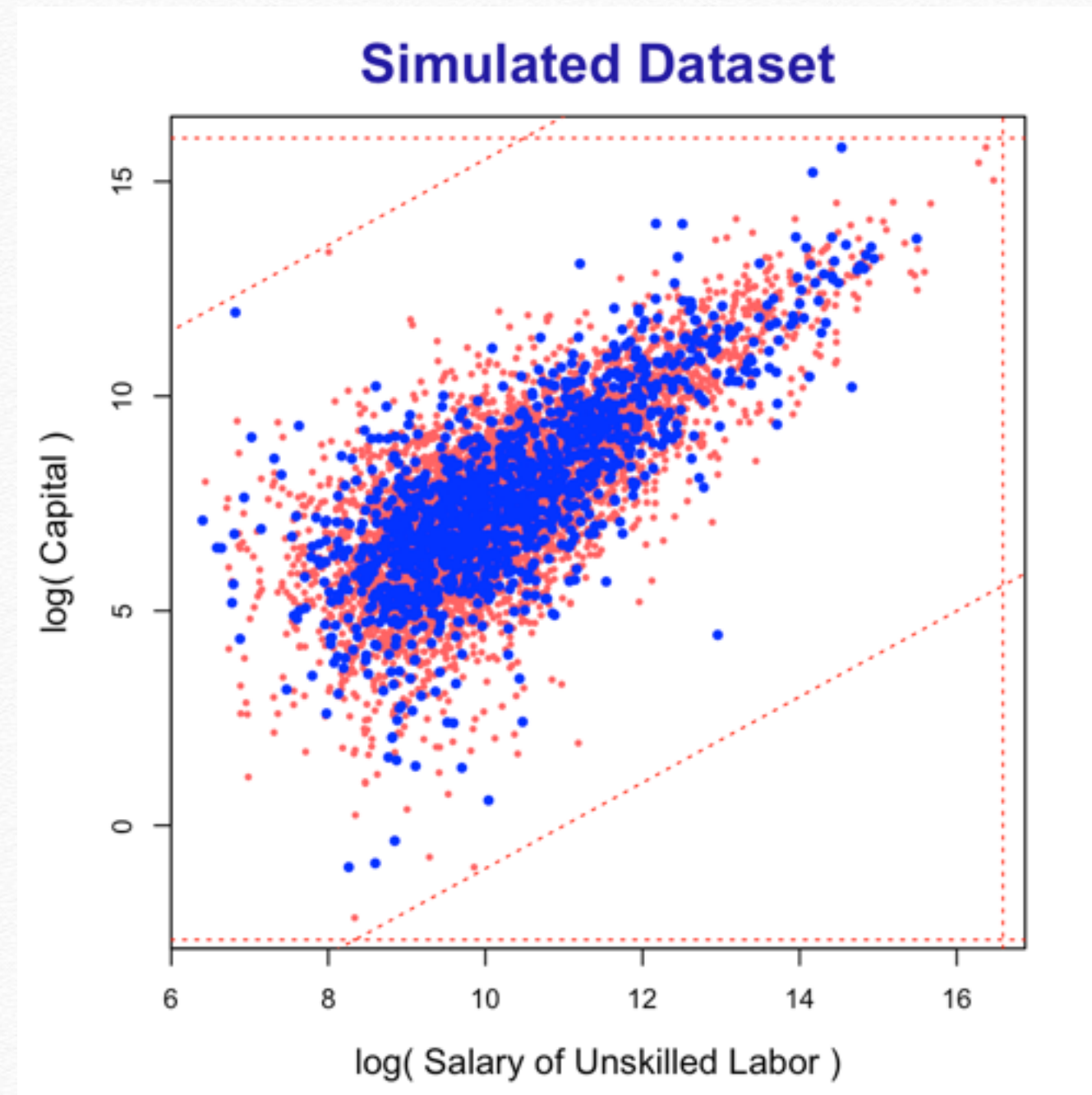


Simulation Study using Colombian Manufacturing Survey data



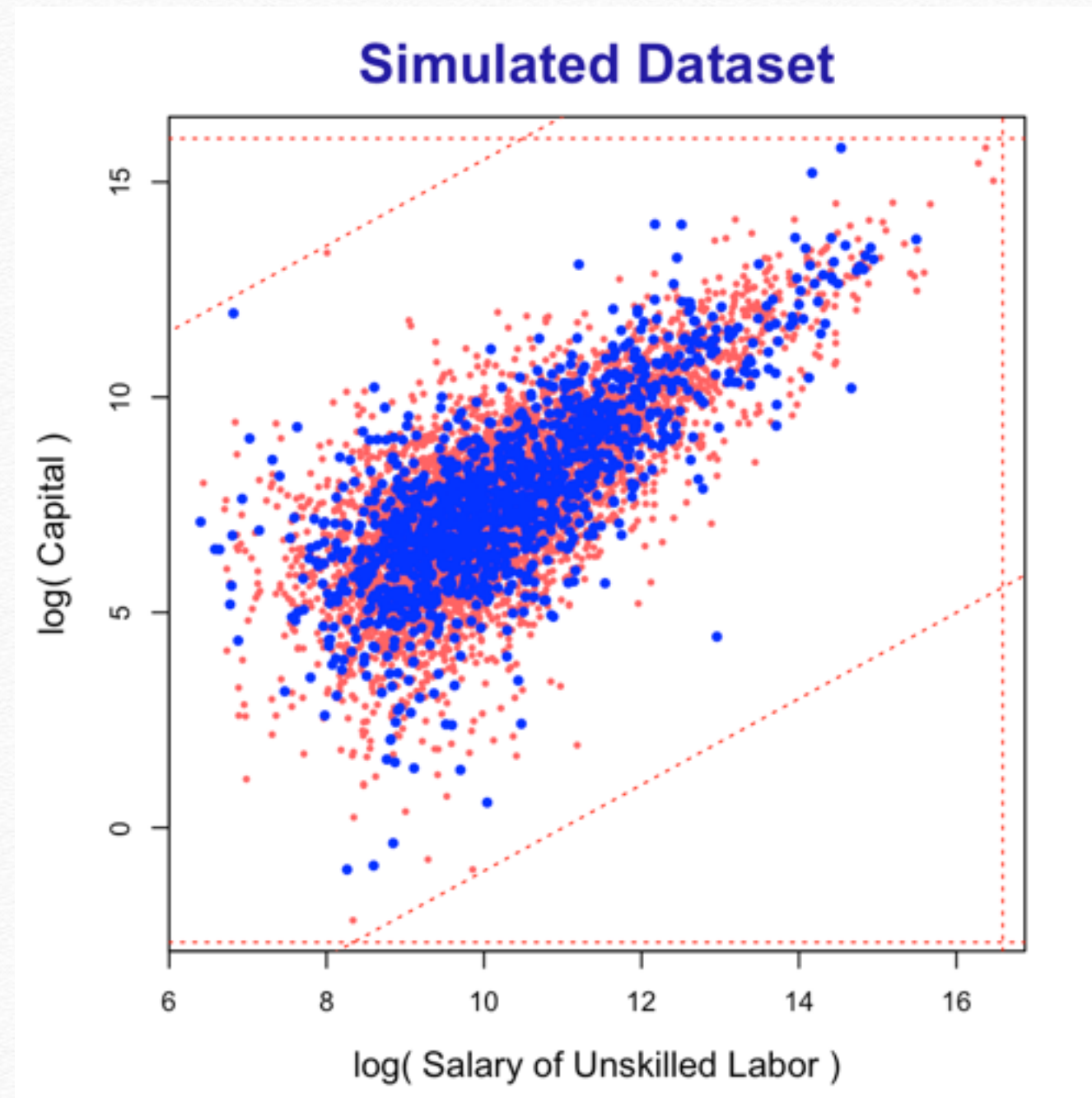
1. Assume data are truly reported values with no missing

Simulation Study using Colombian Manufacturing Survey data



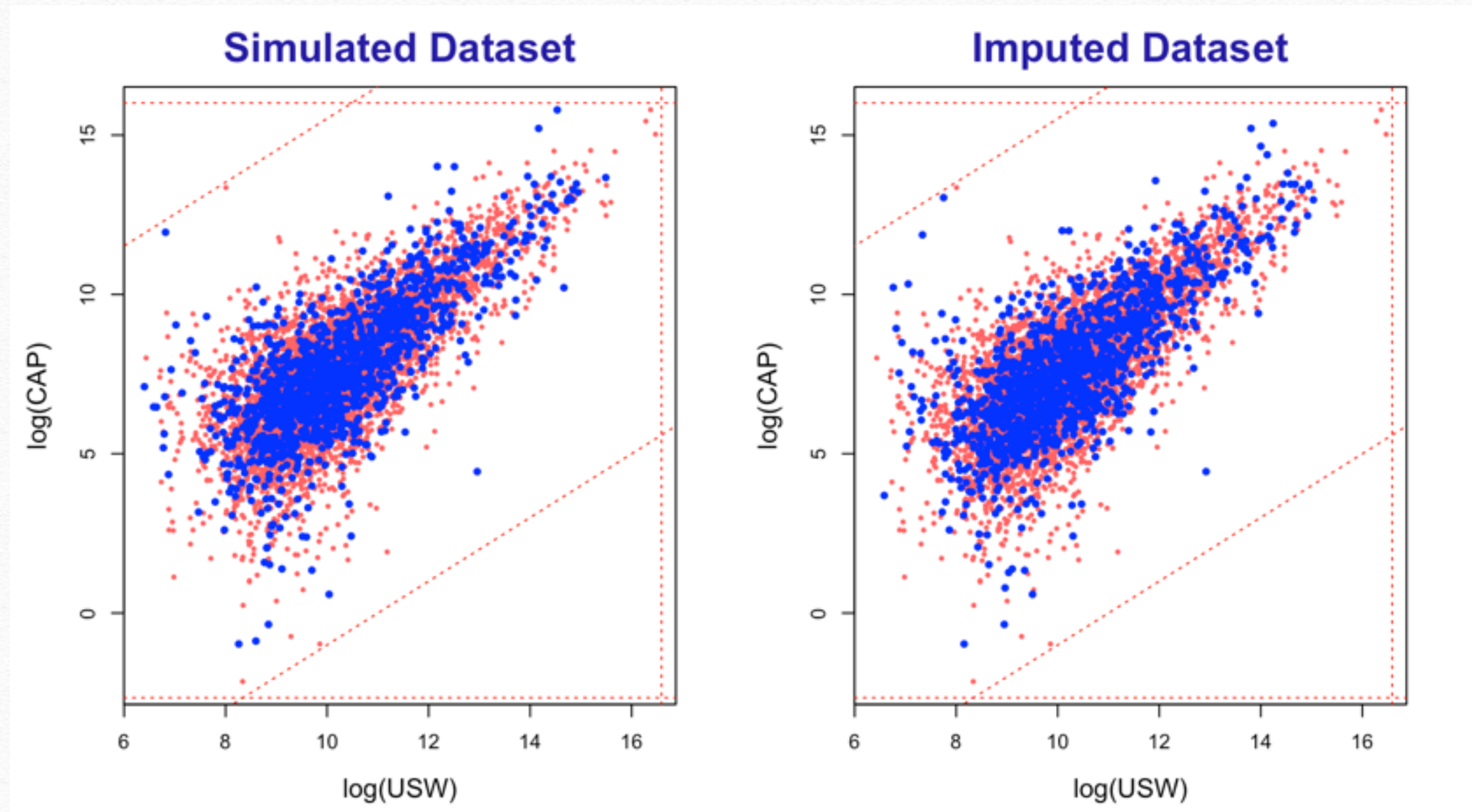
2. Randomly blank some values as simulated nonresponse

Simulation Study using Colombian Manufacturing Survey data



3. Fill in simulated missing values using the suggested method

Simulation Study using Colombian Manufacturing Survey data



- ❖ Pink dots: unchanged values
- ❖ Blue dots: (Left) original values before blanking
(Right) imputed values



Topic II

Simultaneous data editing and imputation

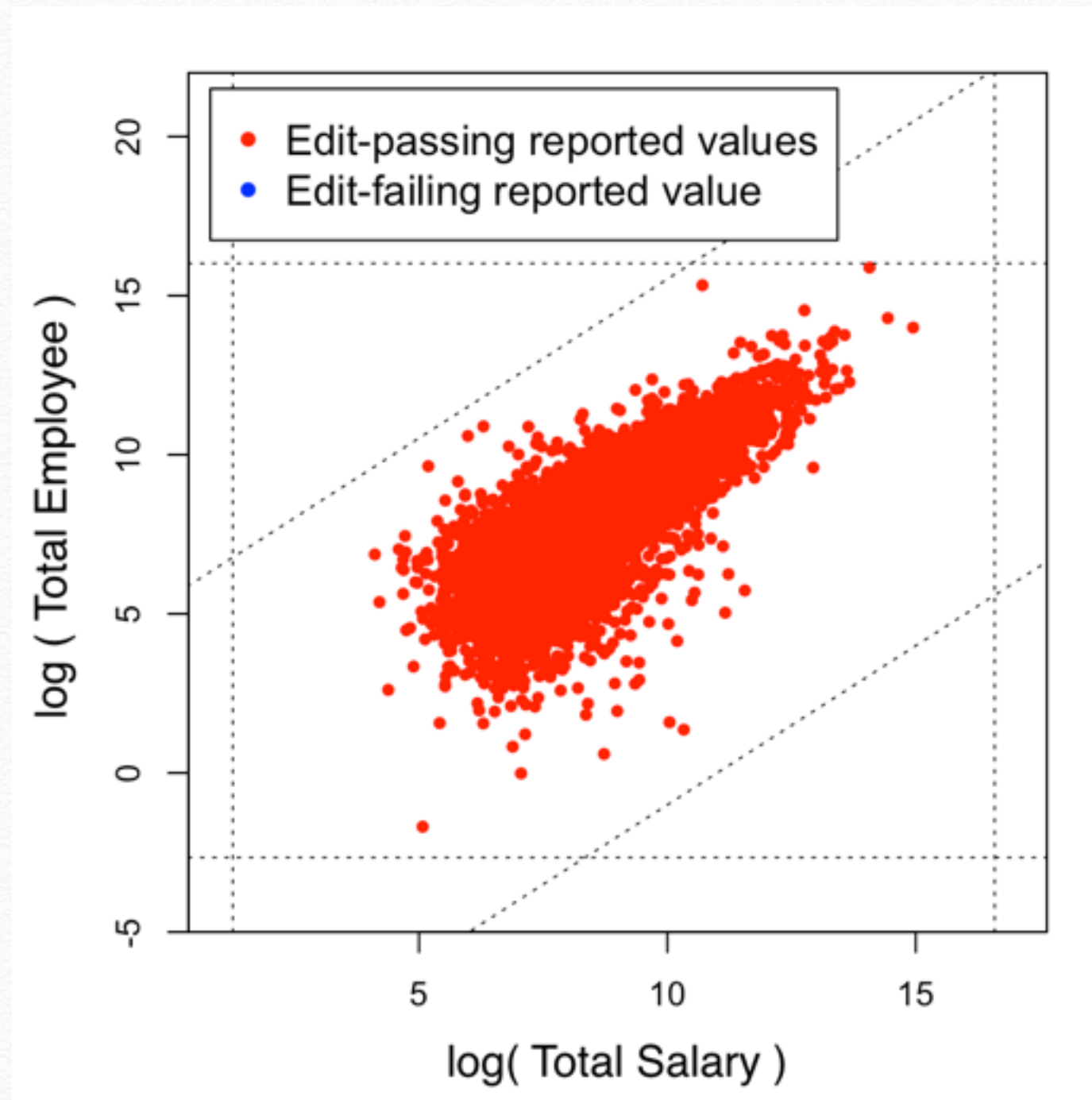
Automatic Data Editing

- Agencies detect and edit unacceptable errors in survey data
 - Manual editing
 - utilizing expert knowledge
 - Automatic editing
 - fast and handling massive datasets
- Automatic editing process
 1. Error localization step
 - Which variable of a record is incorrect?
 2. Imputation step
 - What is a reasonable value to replace the incorrect value?

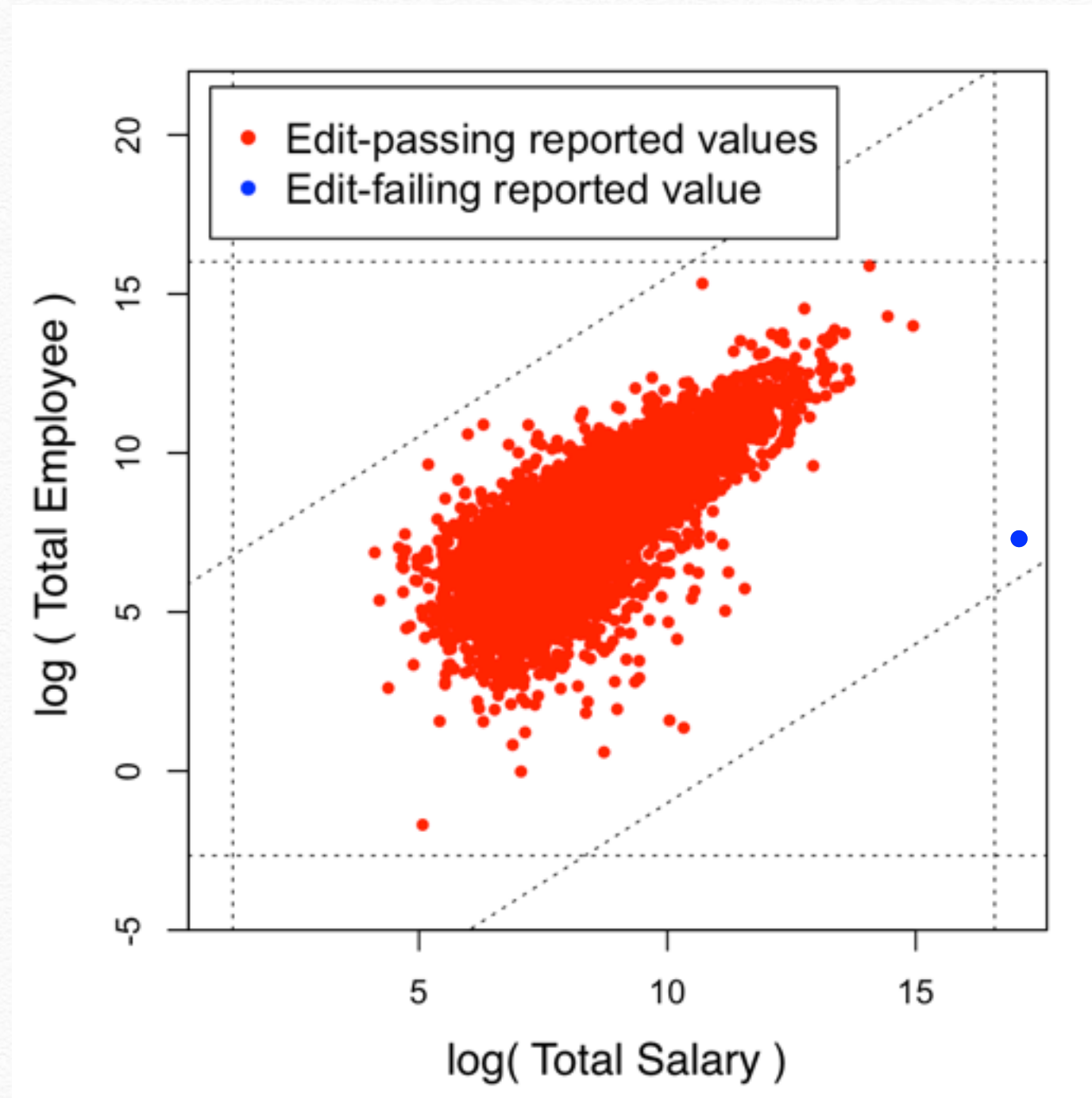
Fellegi-Holt (F-H) Approach

- Since proposed by Fellegi and Holt in 1976, the best-known, most-used guiding principle for automatic data editing
- Mathematical optimization approach
 - Objective function
 - the number of changed variables (to be minimized)
 - Constraints
 - imputed/edited values satisfy edit rules
- ❖ Example
 - If avg. salary $> \$1\text{M}$, need to further review
 - avg. salary = total salary / total employees
 - F-H changes either variable, but not change both variables

Edited/Imputed Values Under F-H Approach

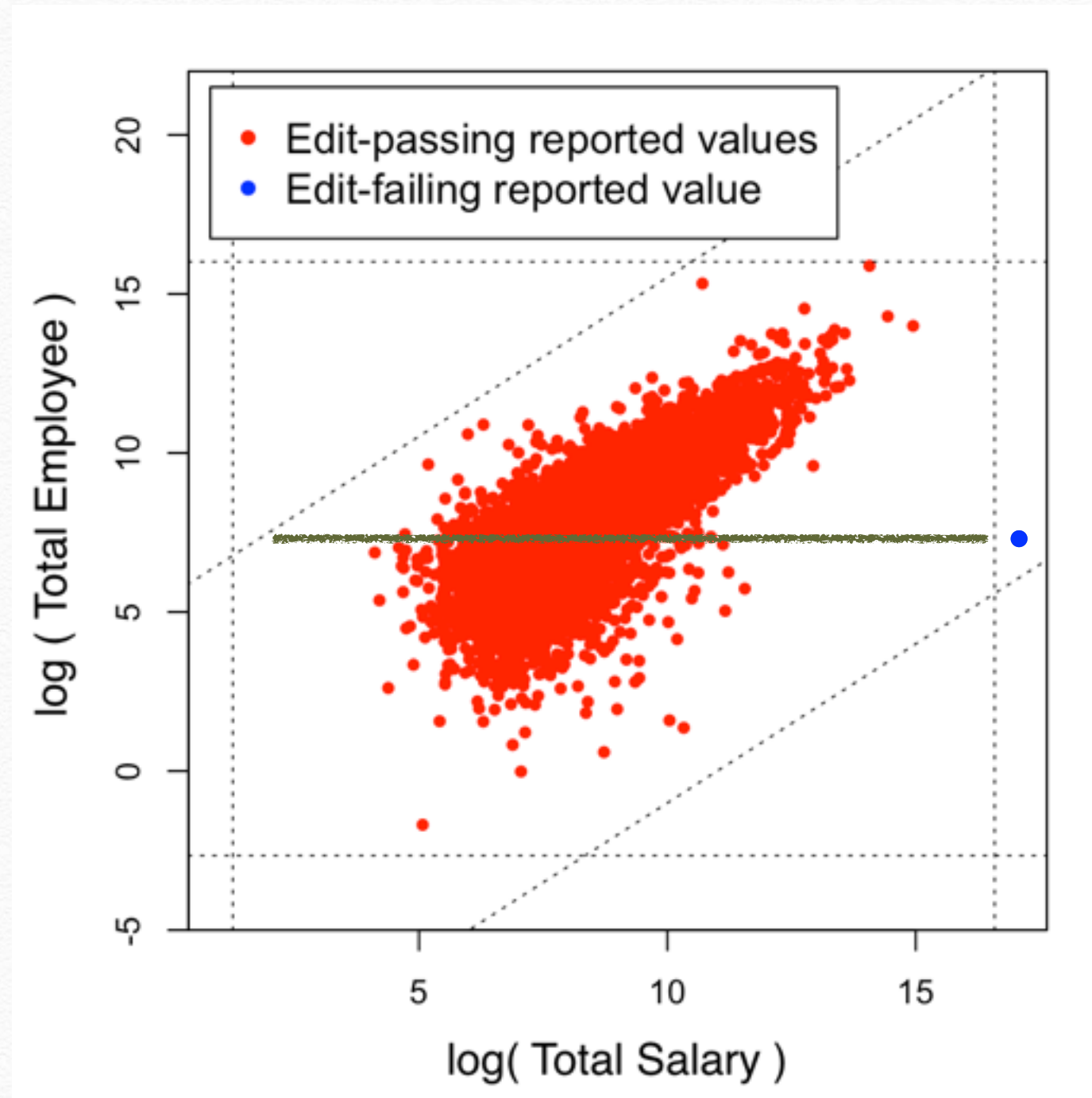


Edited/Imputed Values Under F-H Approach



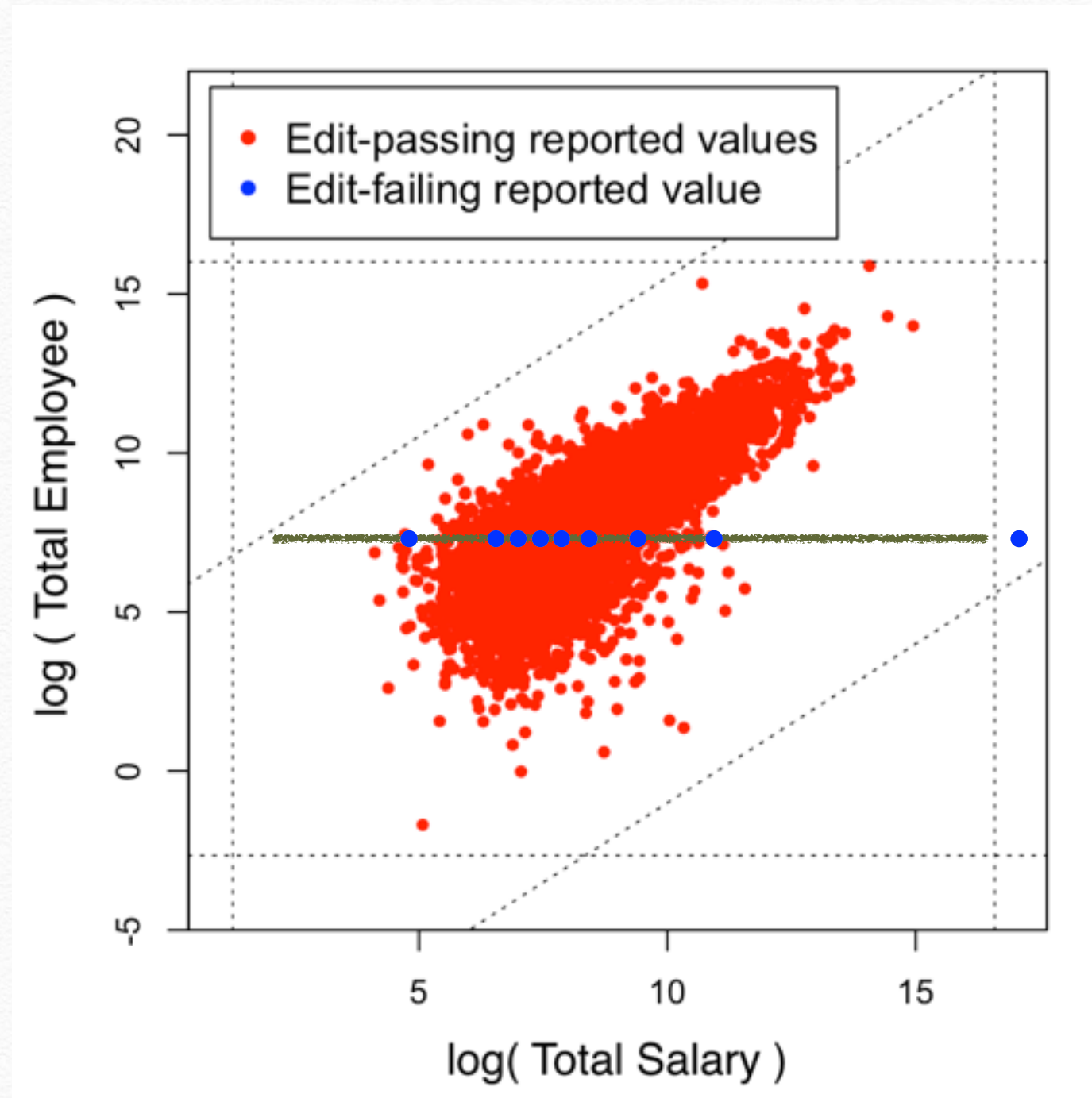
Case 1. assume the observed value failing edit rules

Edited/Imputed Values Under F-H Approach



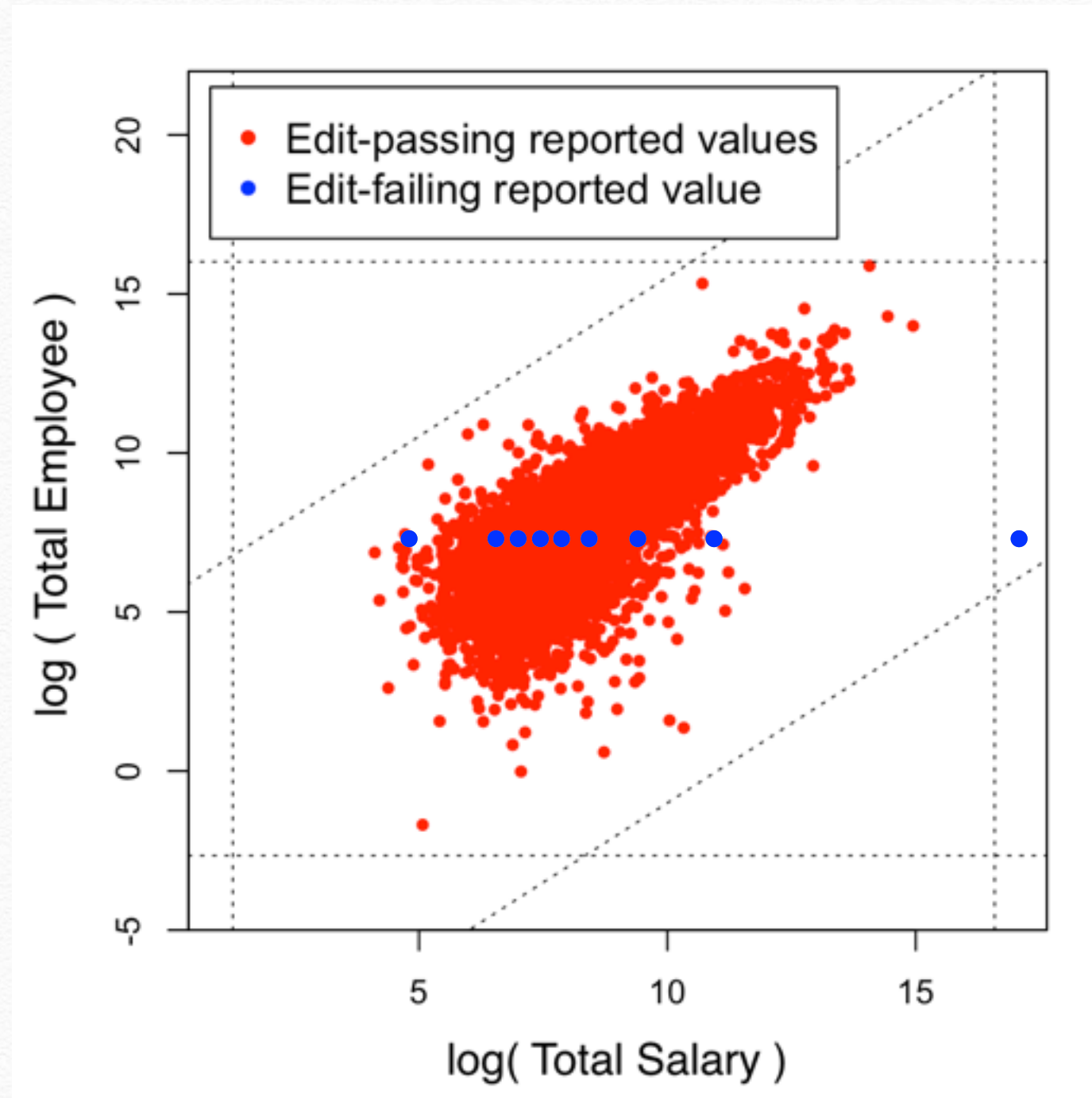
Case 1. no option but changing the value of X_1

Edited/Imputed Values Under F-H Approach



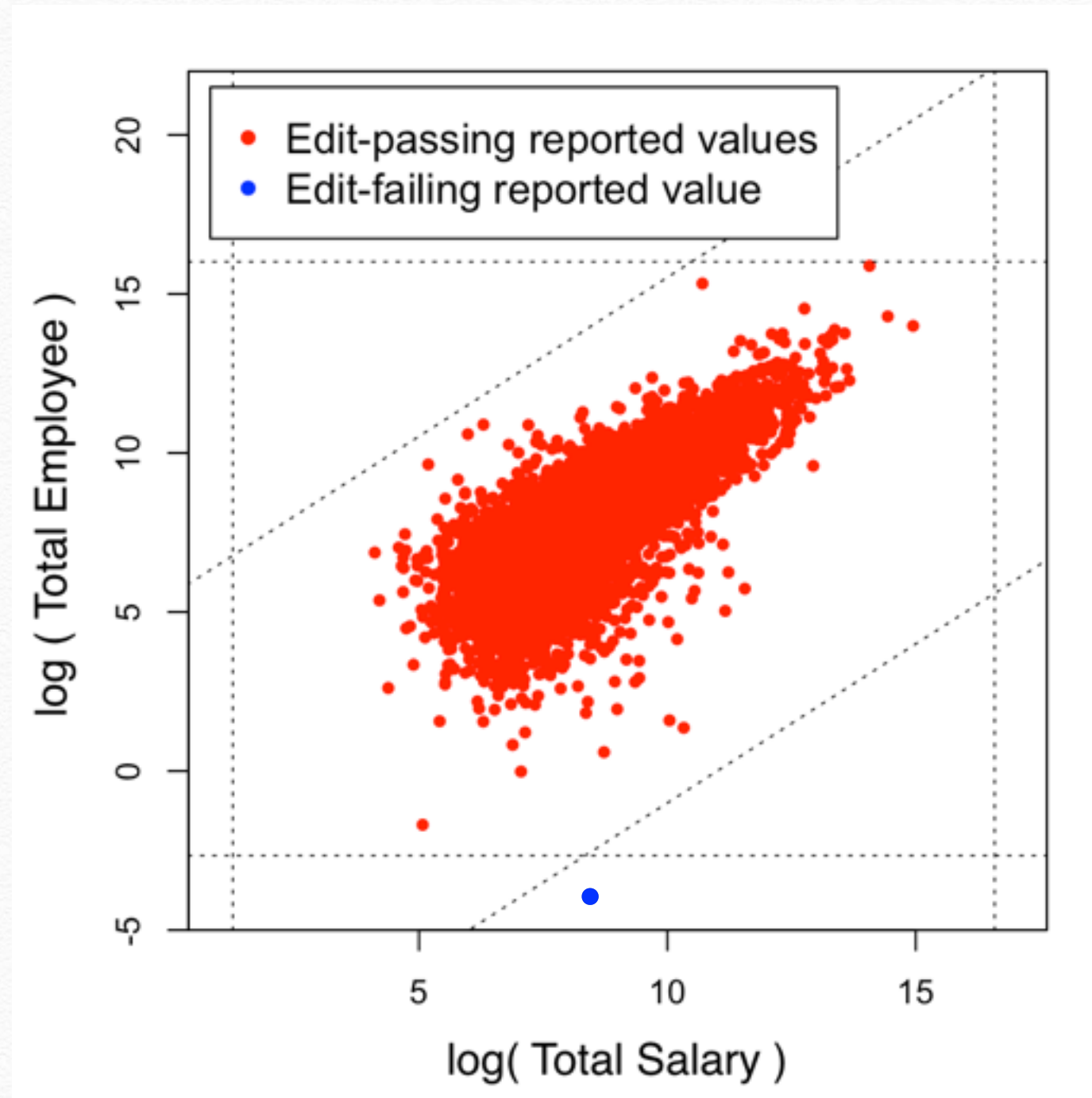
Case 1. can draw imputations from high density region

Edited/Imputed Values Under F-H Approach



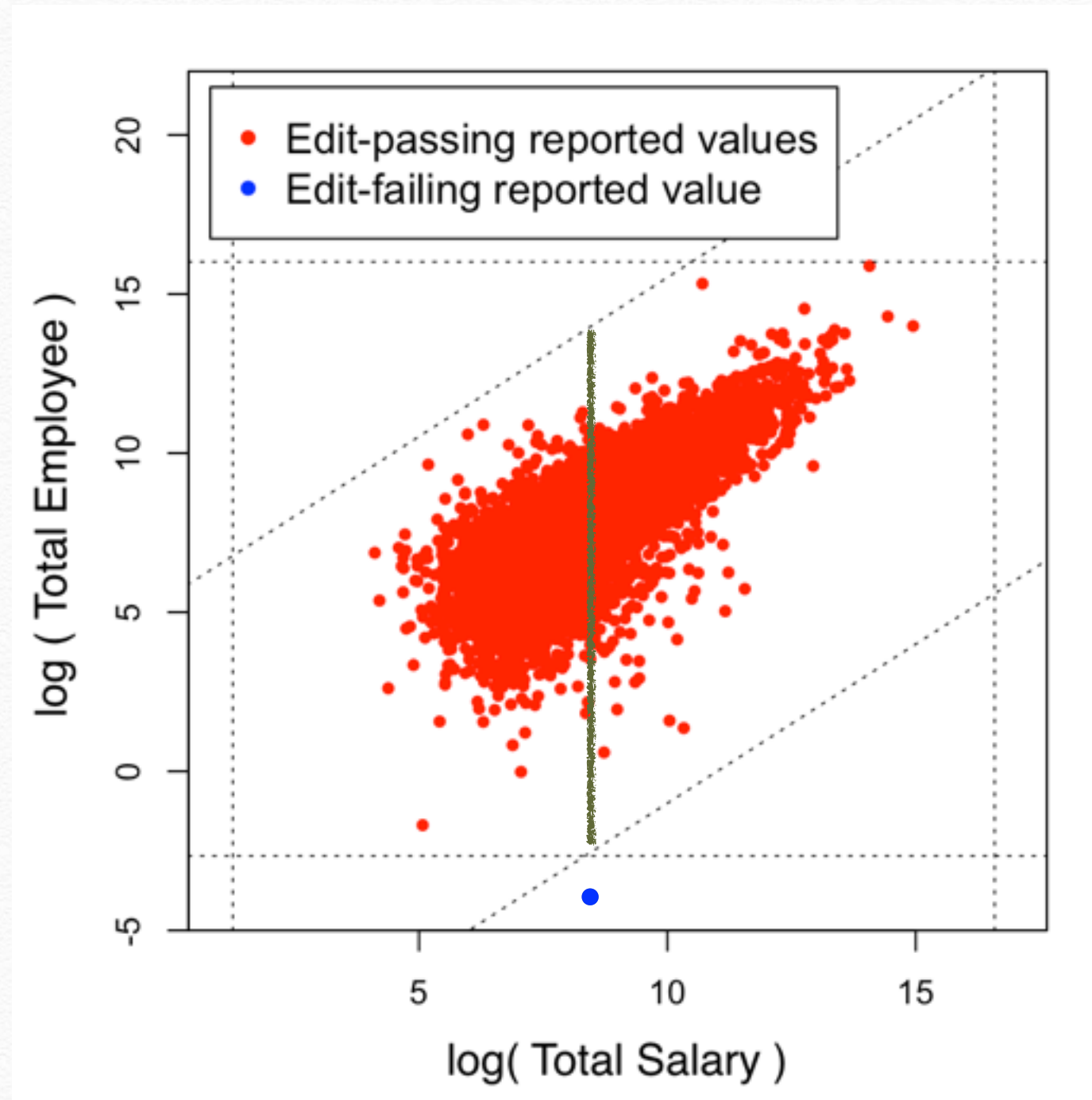
Case 1. can draw imputations from high density region

Edited/Imputed Values Under F-H Approach



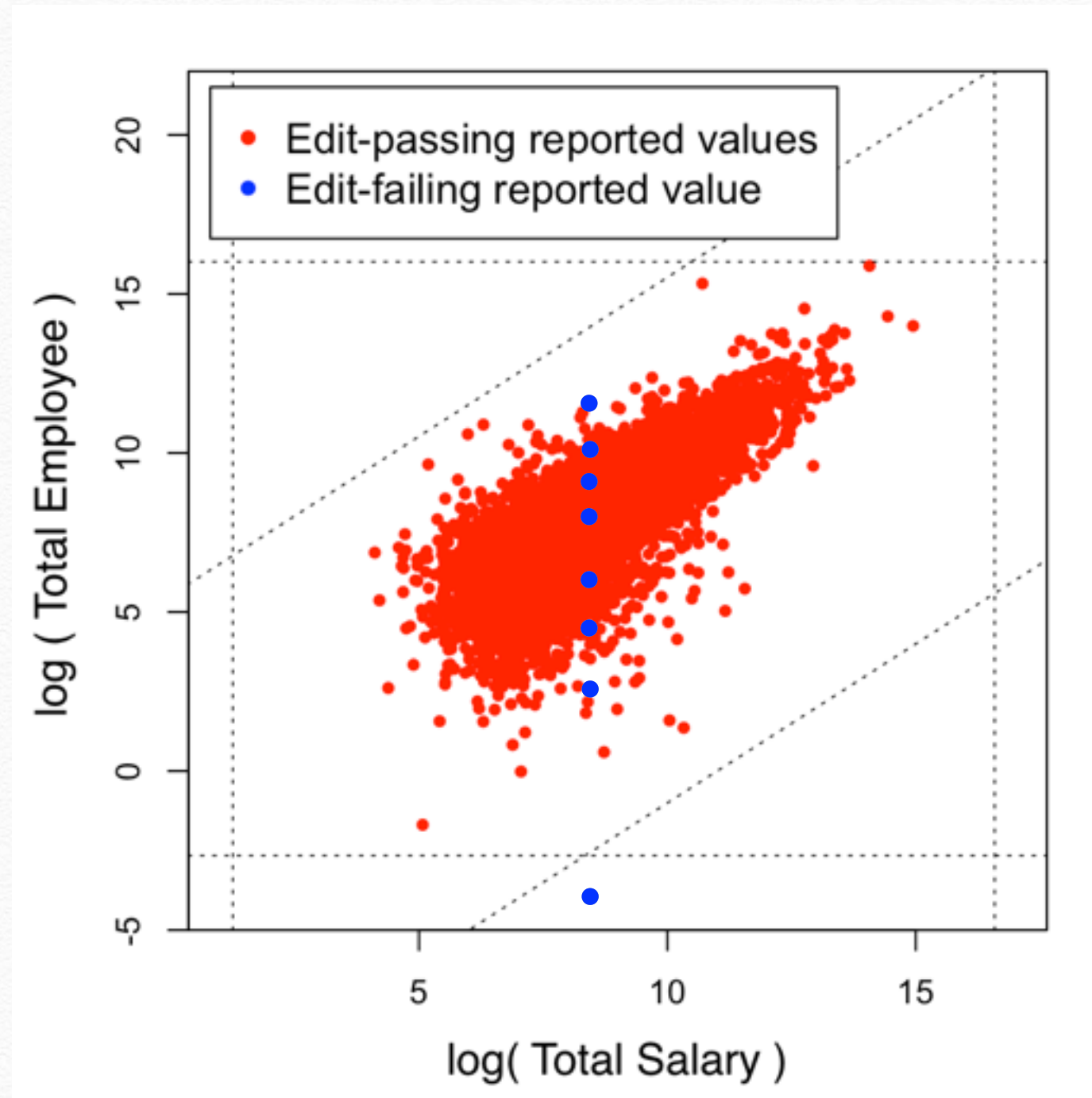
Case 2. no option but changing the value of X_2

Edited/Imputed Values Under F-H Approach



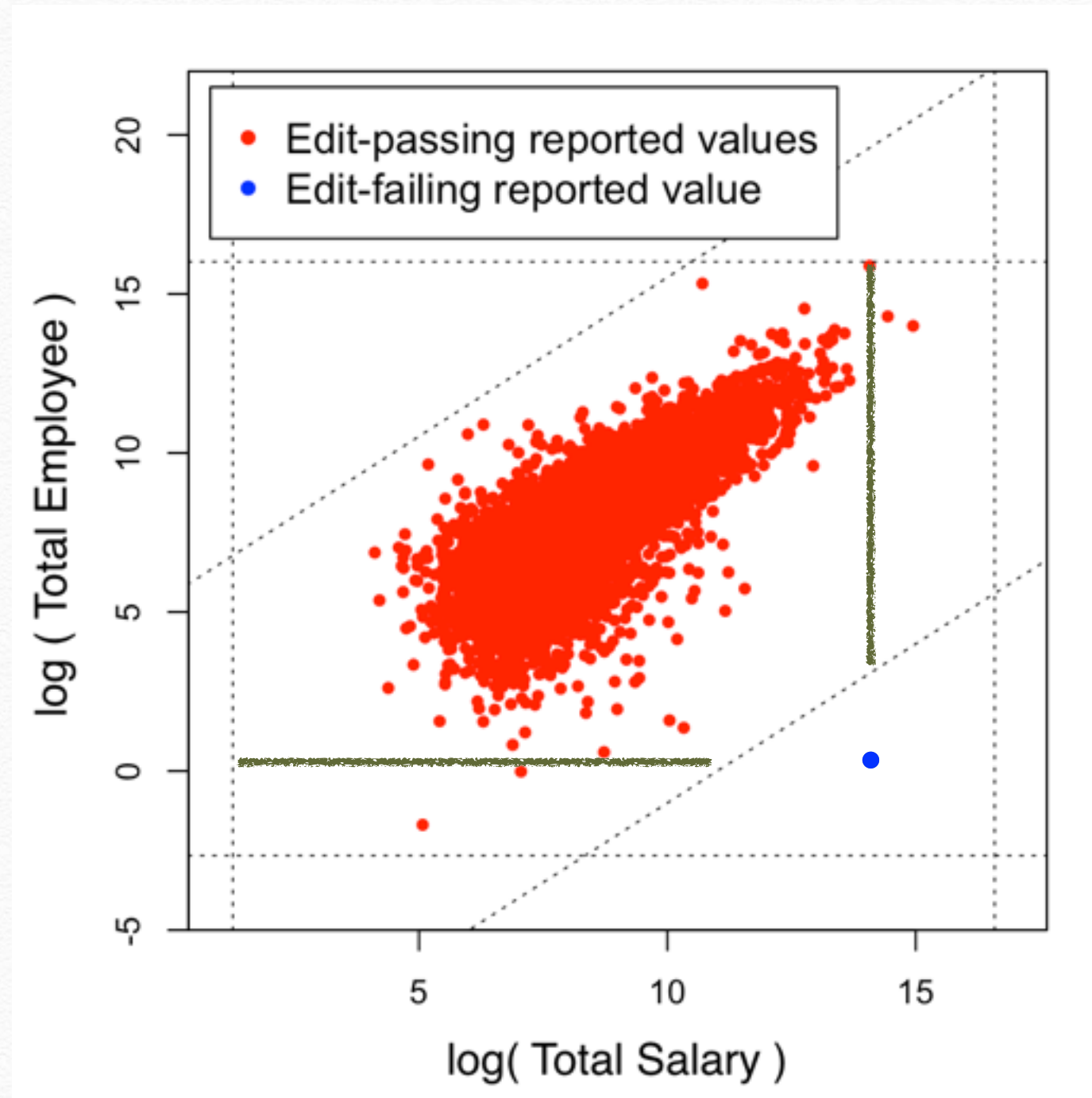
Case 2. no option but changing the value of X_2

Edited/Imputed Values Under F-H Approach



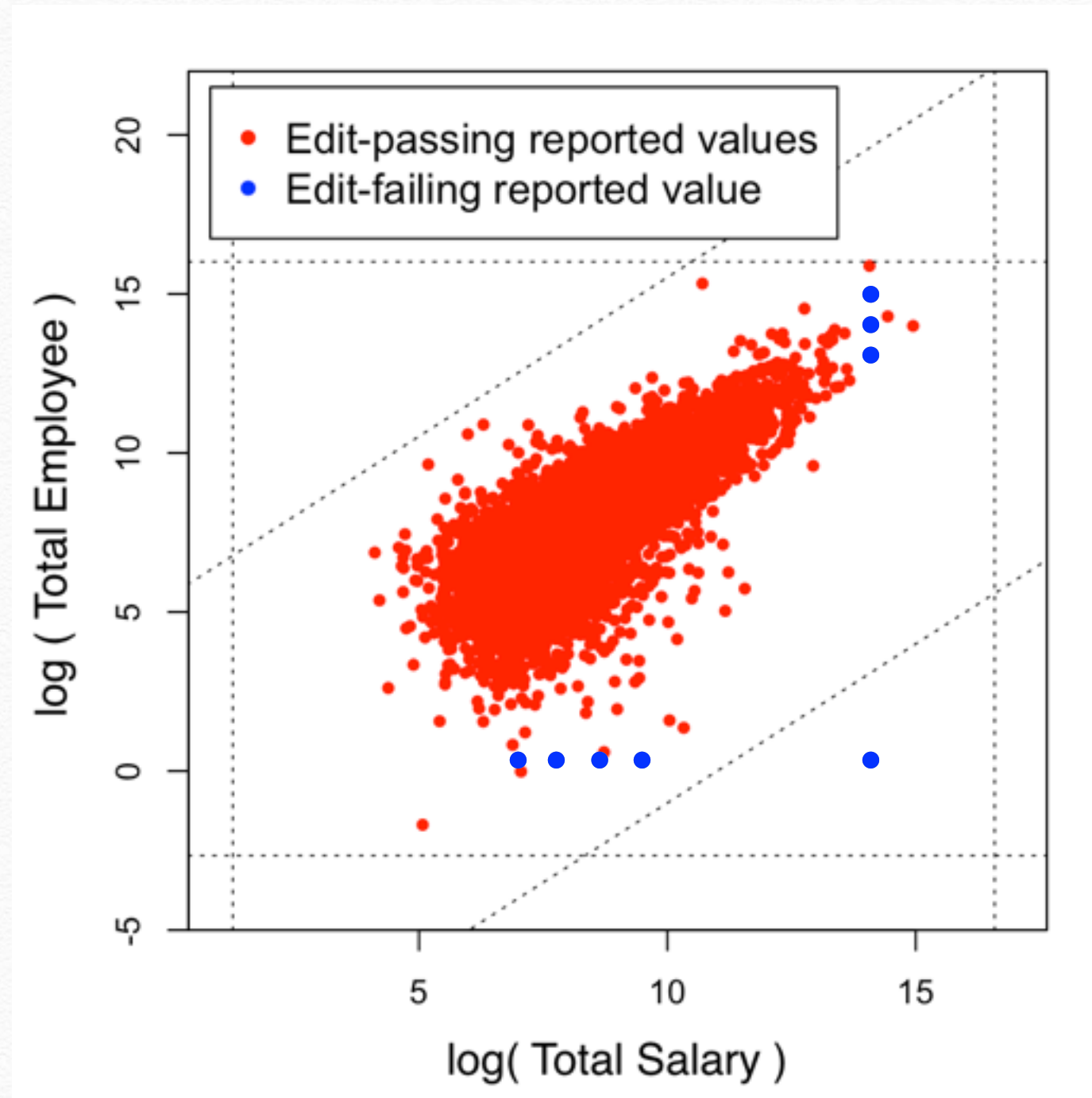
Case 2. can draw imputations from high density region

Edited/Imputed Values Under F-H Approach



Case 3. both options available: changing X_1 or X_2

Edited/Imputed Values Under F-H Approach



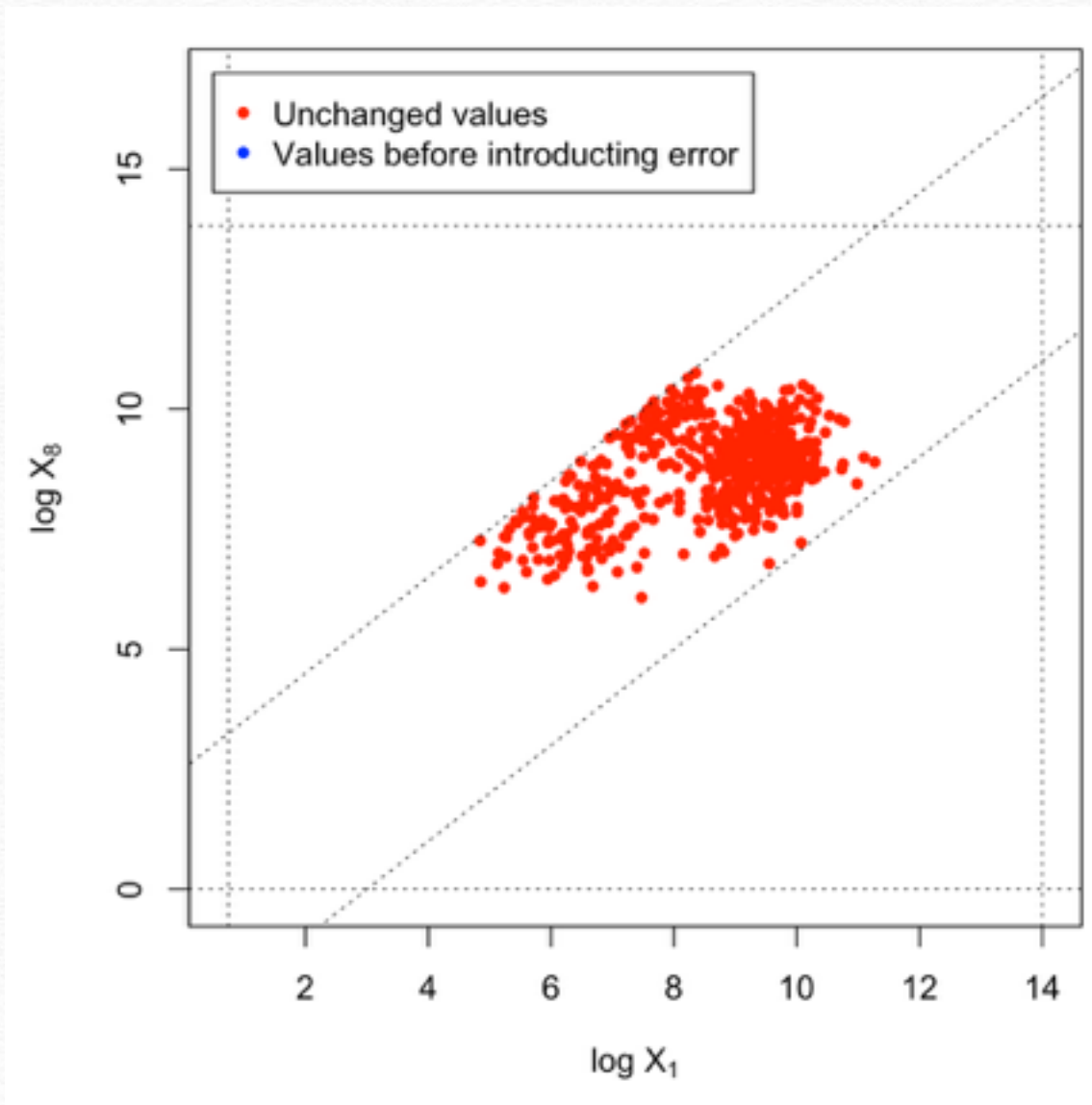
Case 3. draw imputed values from tails of distribution

Bayesian Data Editing (NISS report 189)

- Nonparametric Bayesian Model
 - use Dirichlet process (DP) mixture normals
 - restrict support under constrained regions
 - **balance edits** as well as ratio edits
 - utilize latent indicator to stochastically find the **location of error**
- Multiple Imputation Approach
 - measure uncertainty introduced by imputation process and **data editing process**

Simulation Study

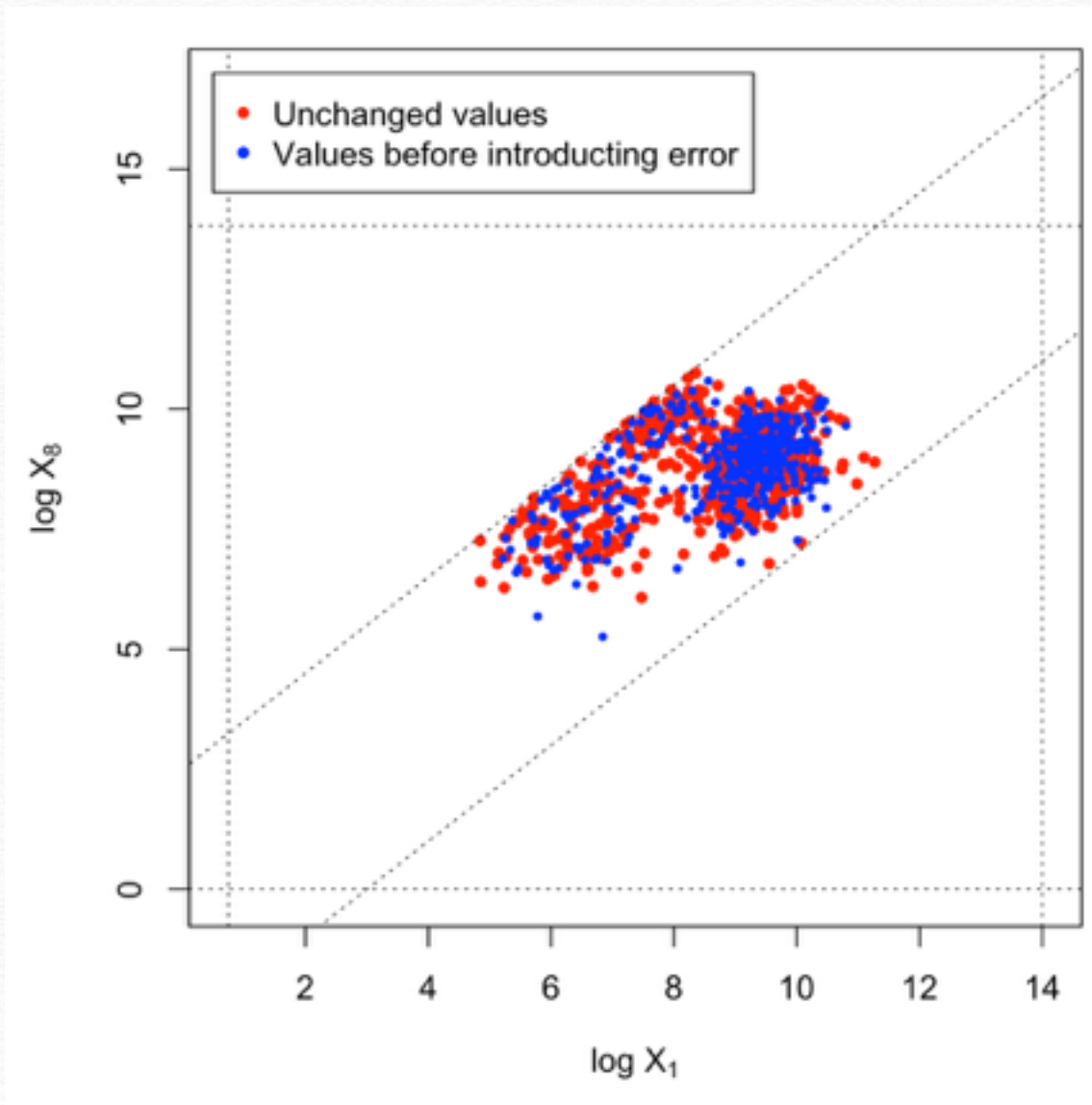
True simulated values $\rightarrow$ Reported values with errors



- ❖ Generate simulated reported values with introduced errors

Simulation Study

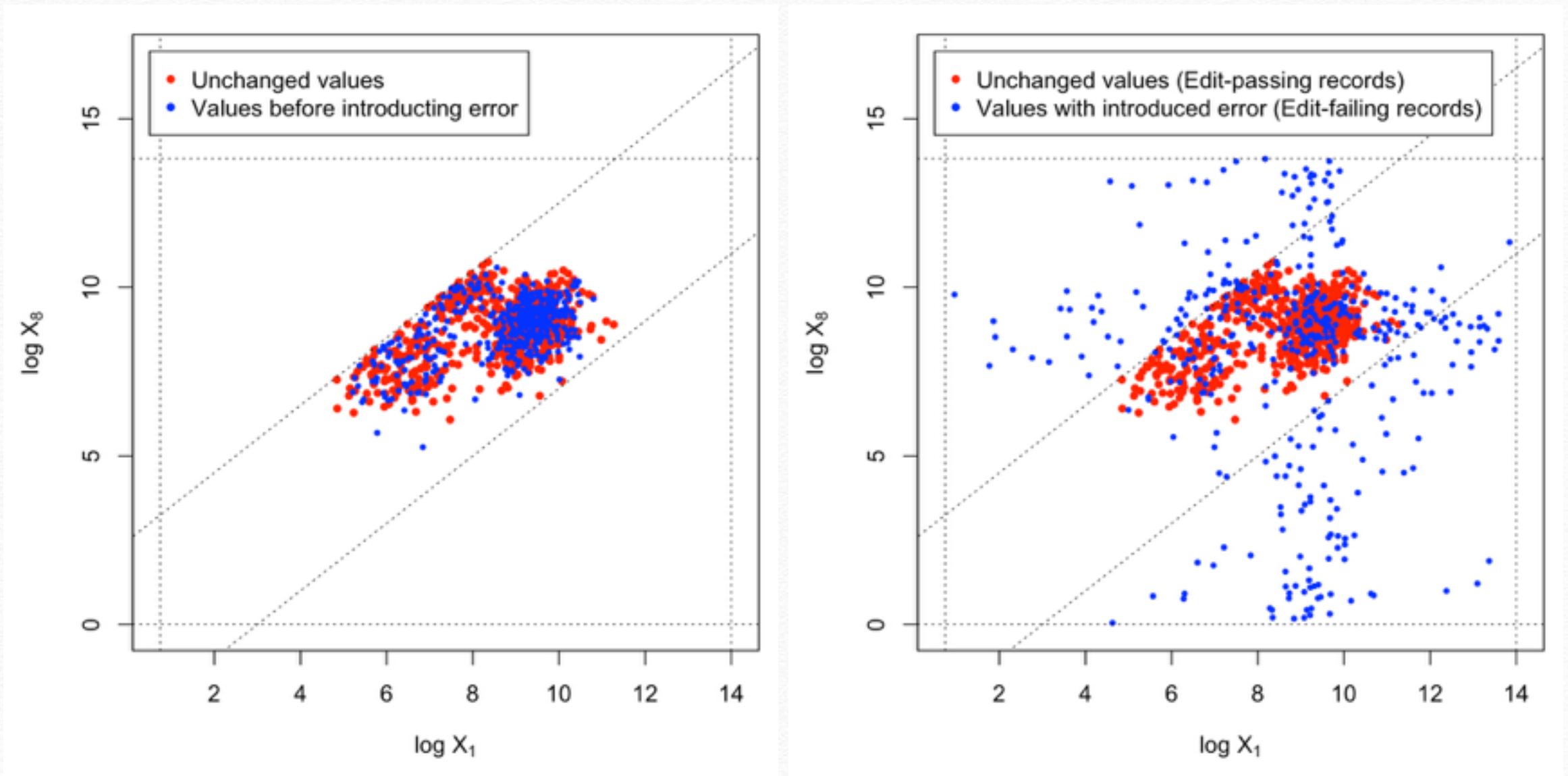
True simulated values $\rightarrow$ Reported values with errors



- ❖ Generate simulated reported values with introduced errors

Simulation Study

True simulated values $\rightarrow$ Reported values with errors



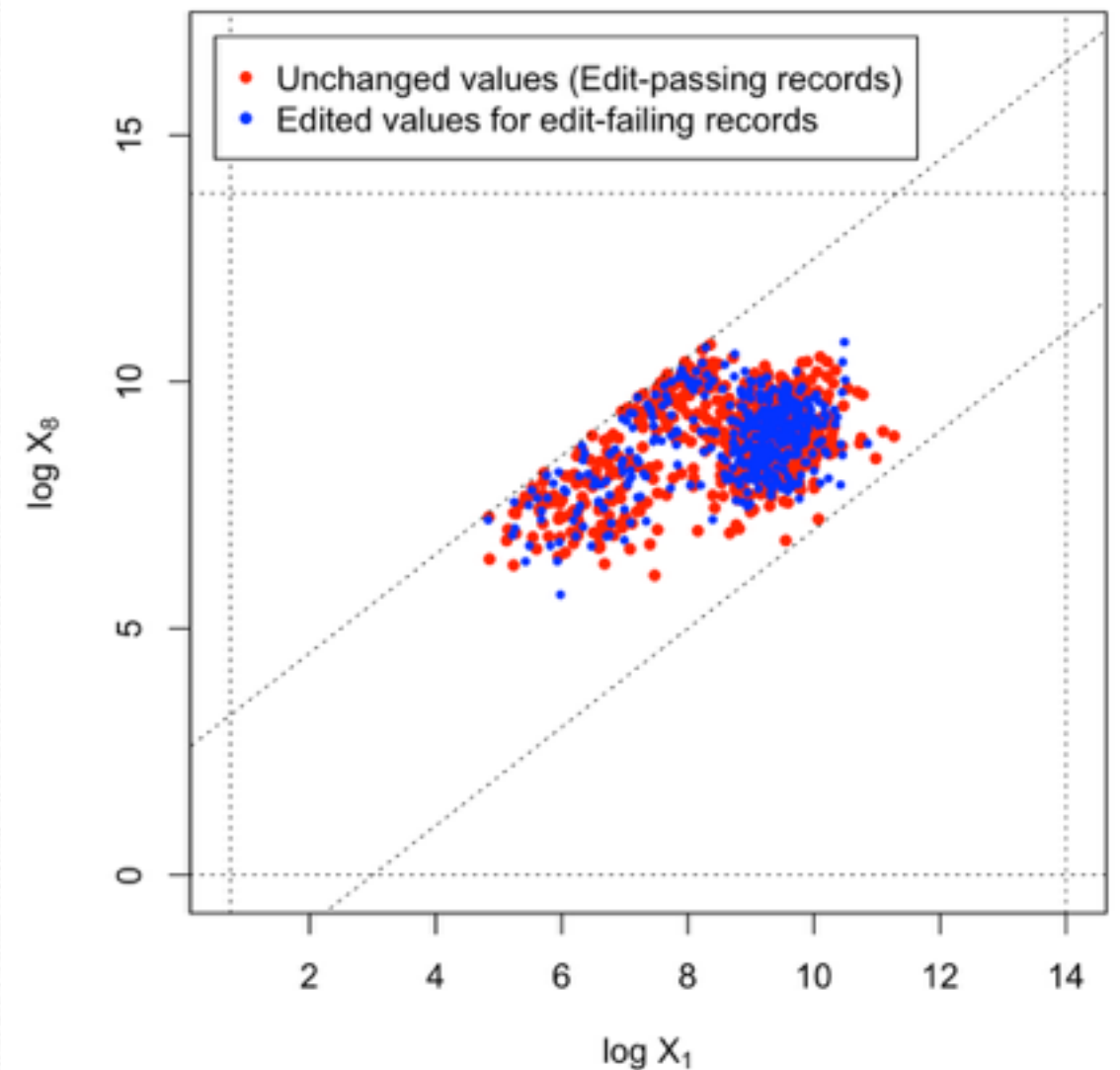
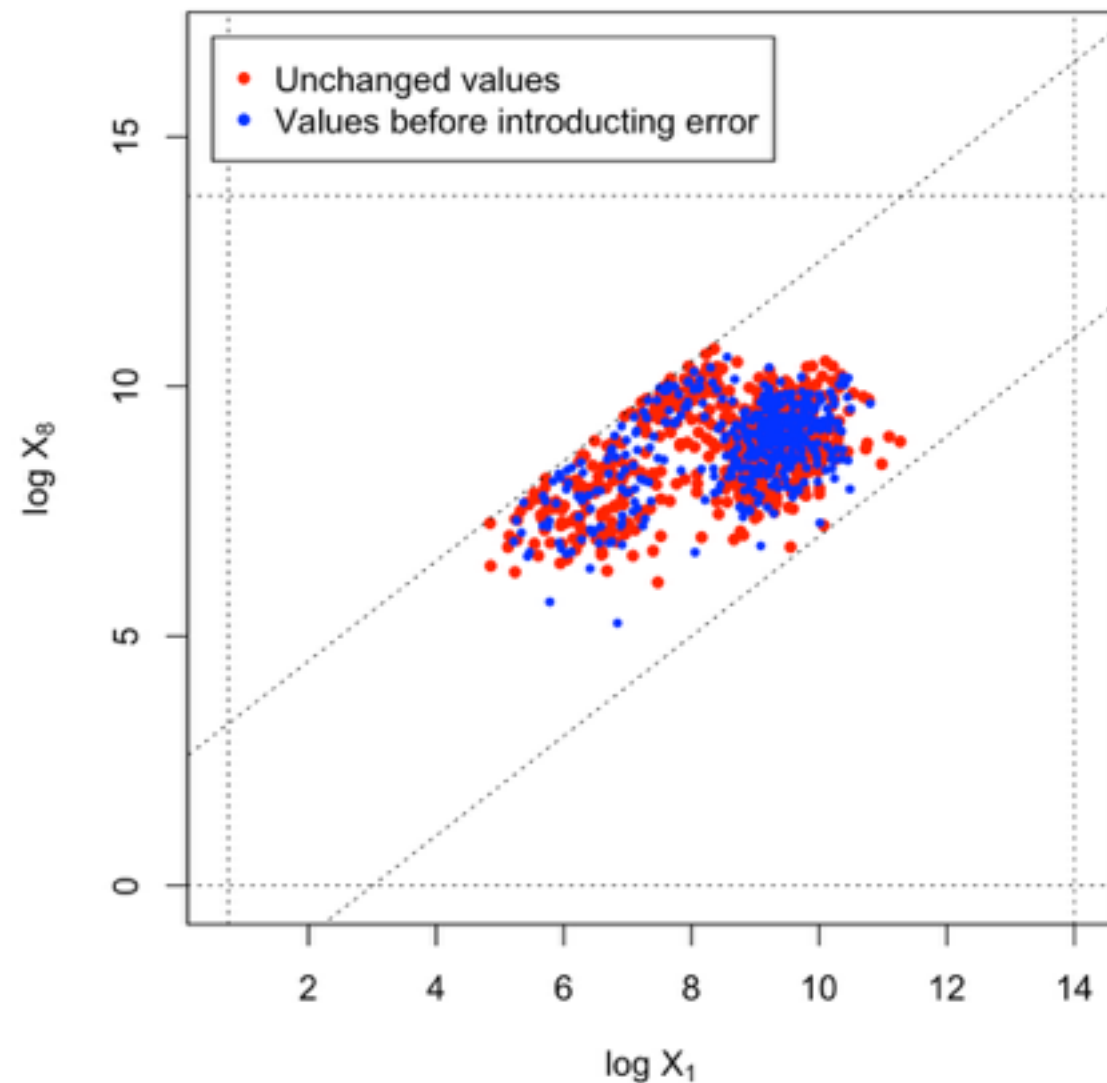
- ❖ Generate simulated reported values with introduced errors

Result with Bayesian Editing

True simulated values



Edited values by BE



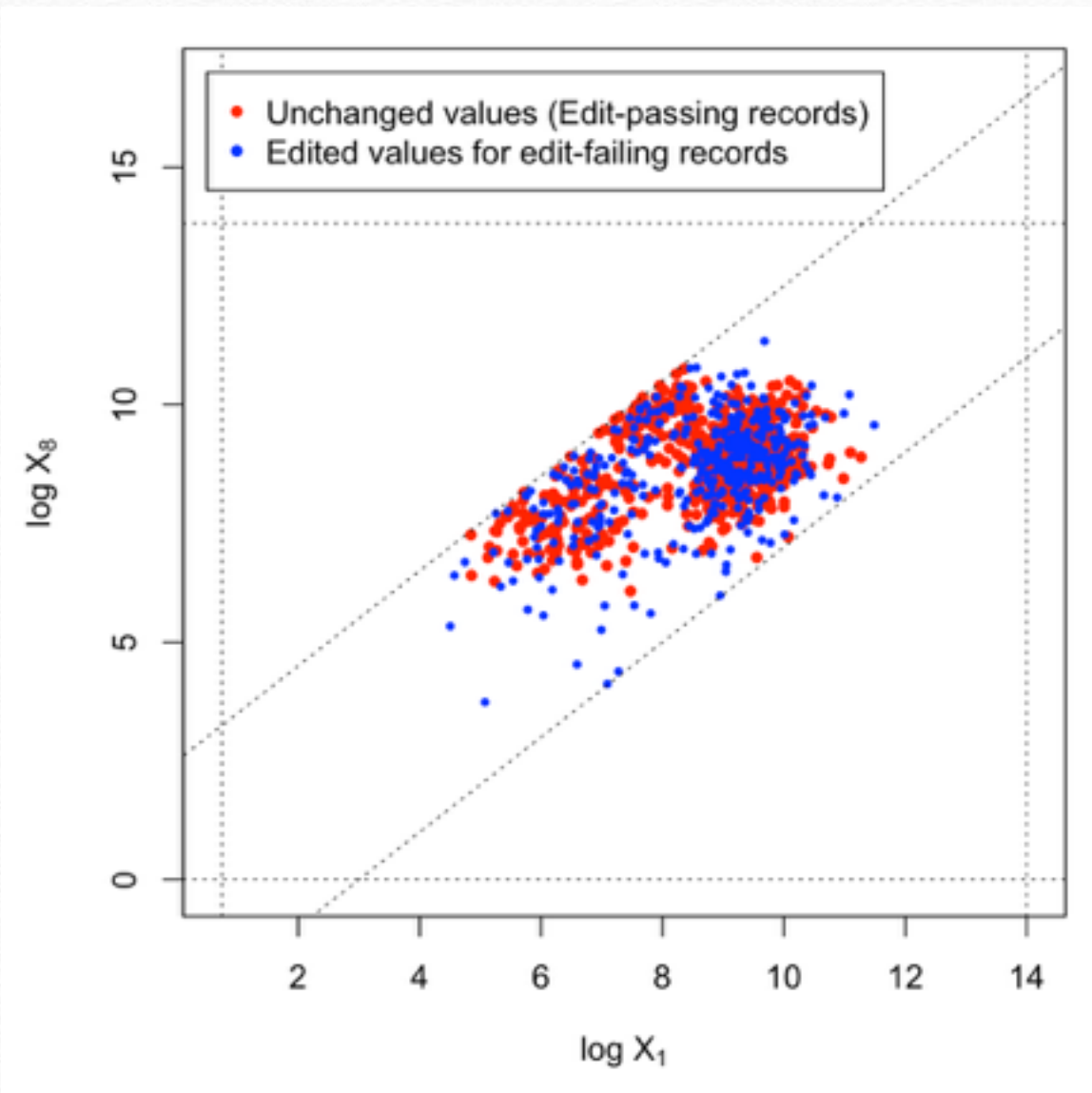
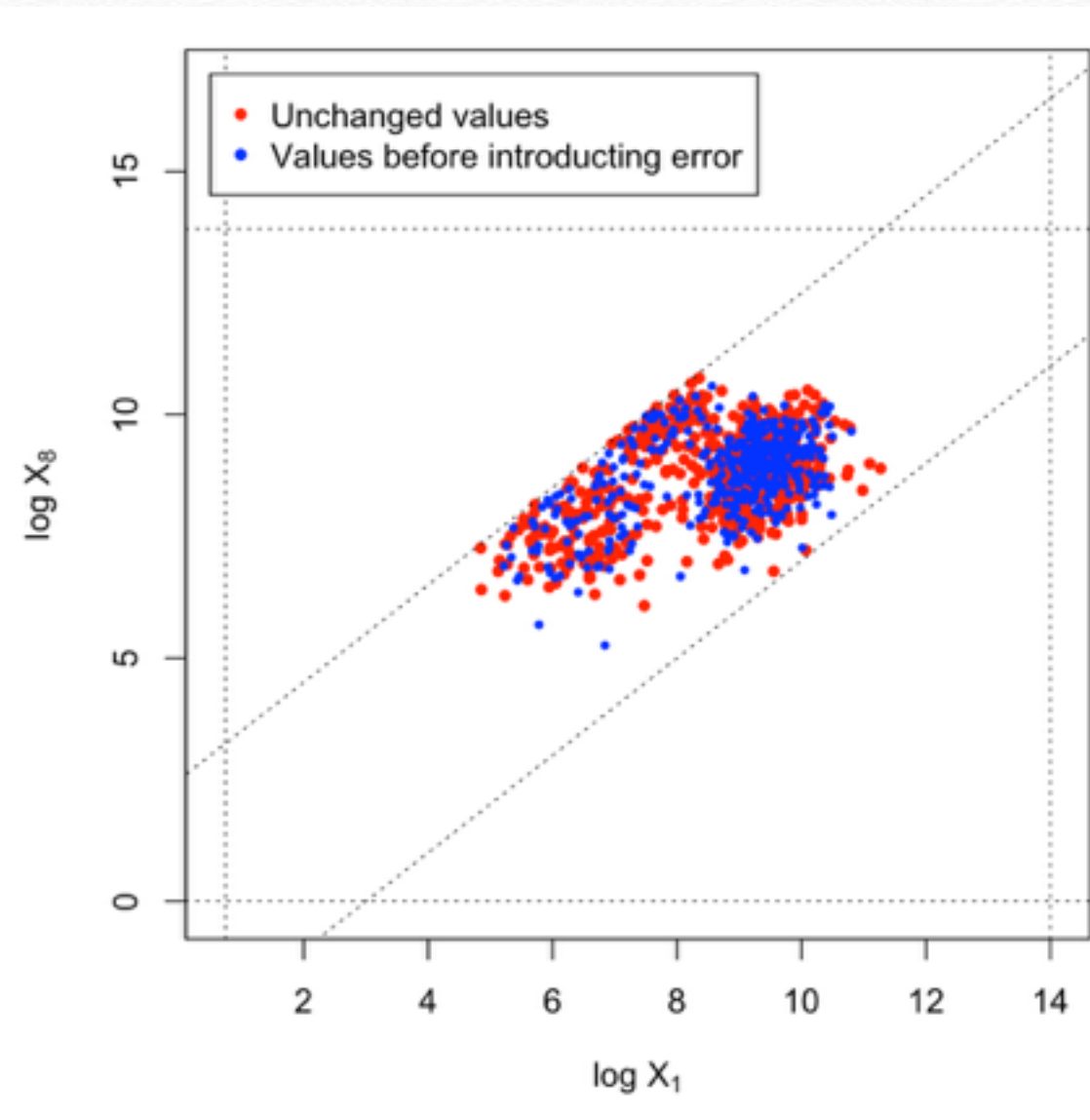
- ❖ Bayes. Editing successfully estimates the distribution of simulated true values

Result with Fellegi-Holt

True simulated values



Edited values by FH



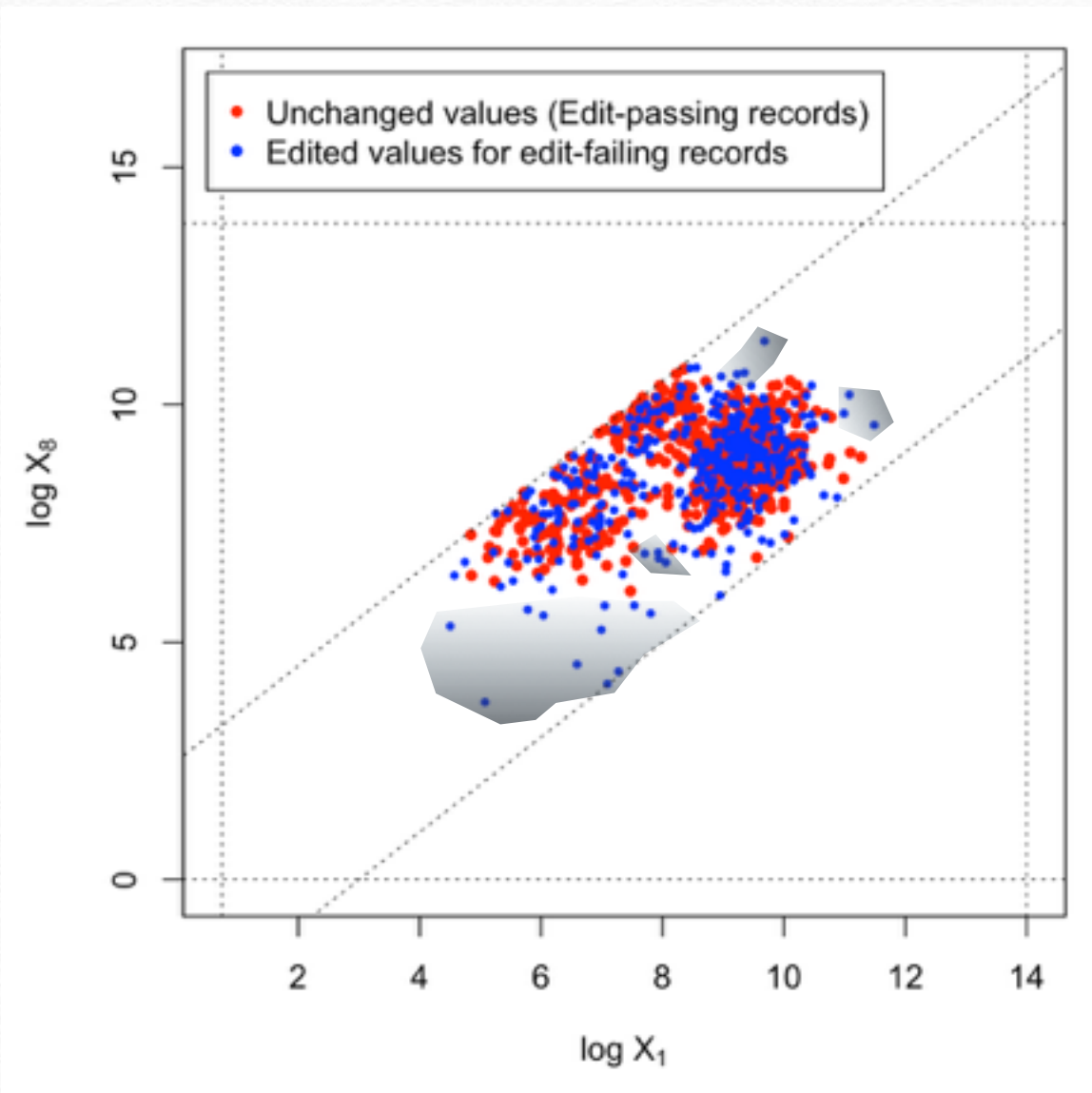
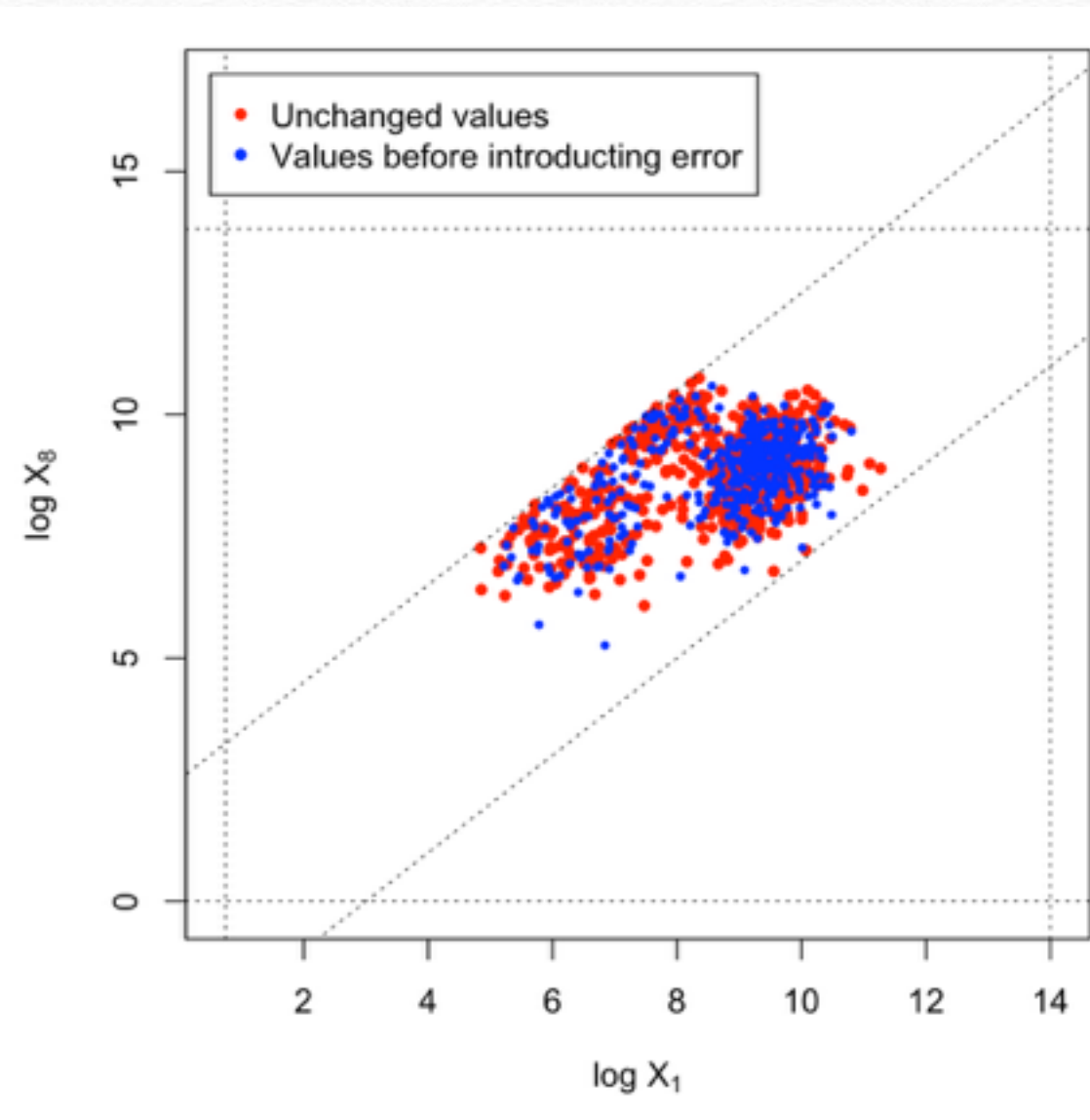
- ❖ F-H approach results in some edited values at tails of the distribution of true values

Result with Fellegi-Holt

True simulated values



Edited values by FH



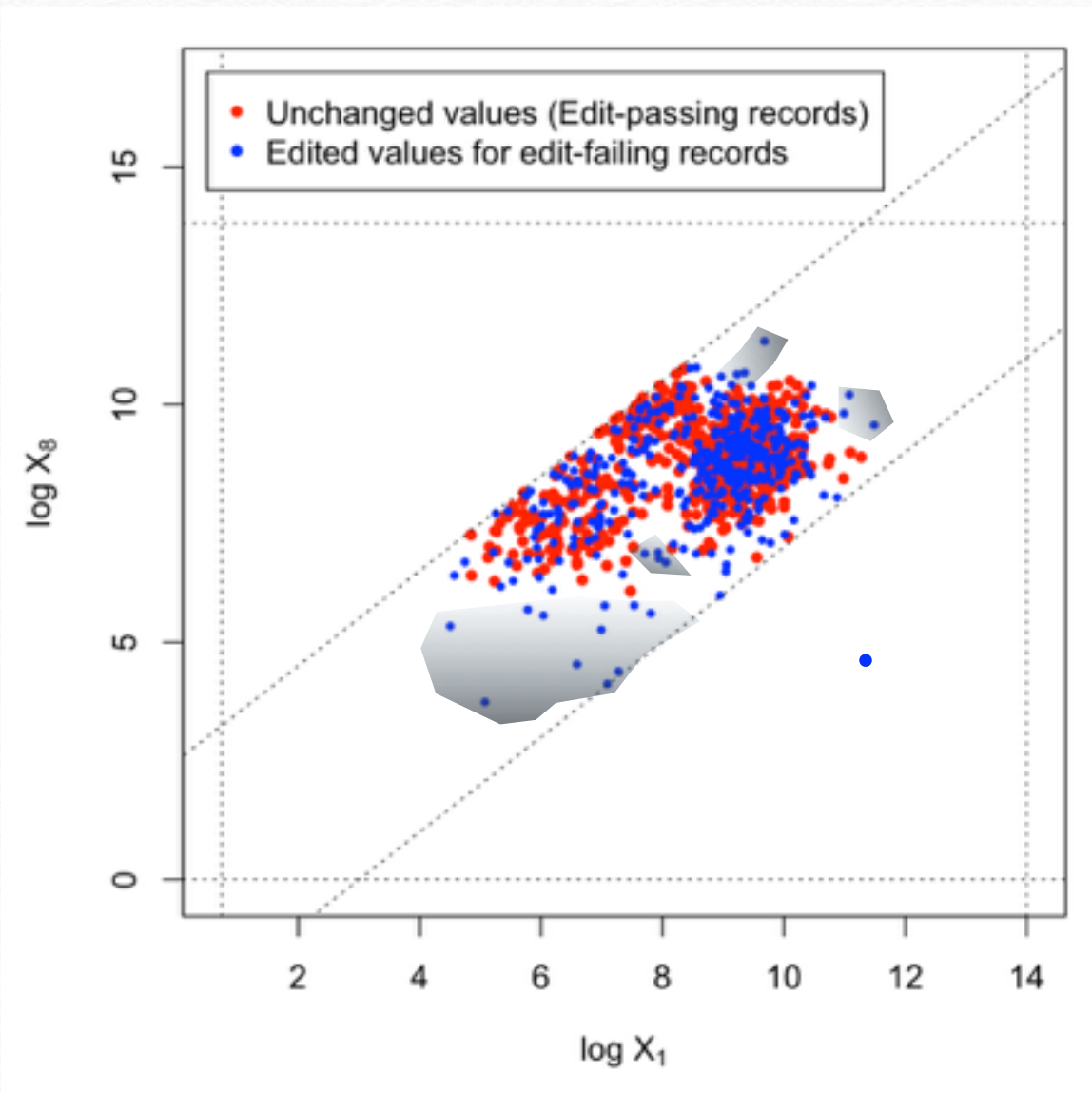
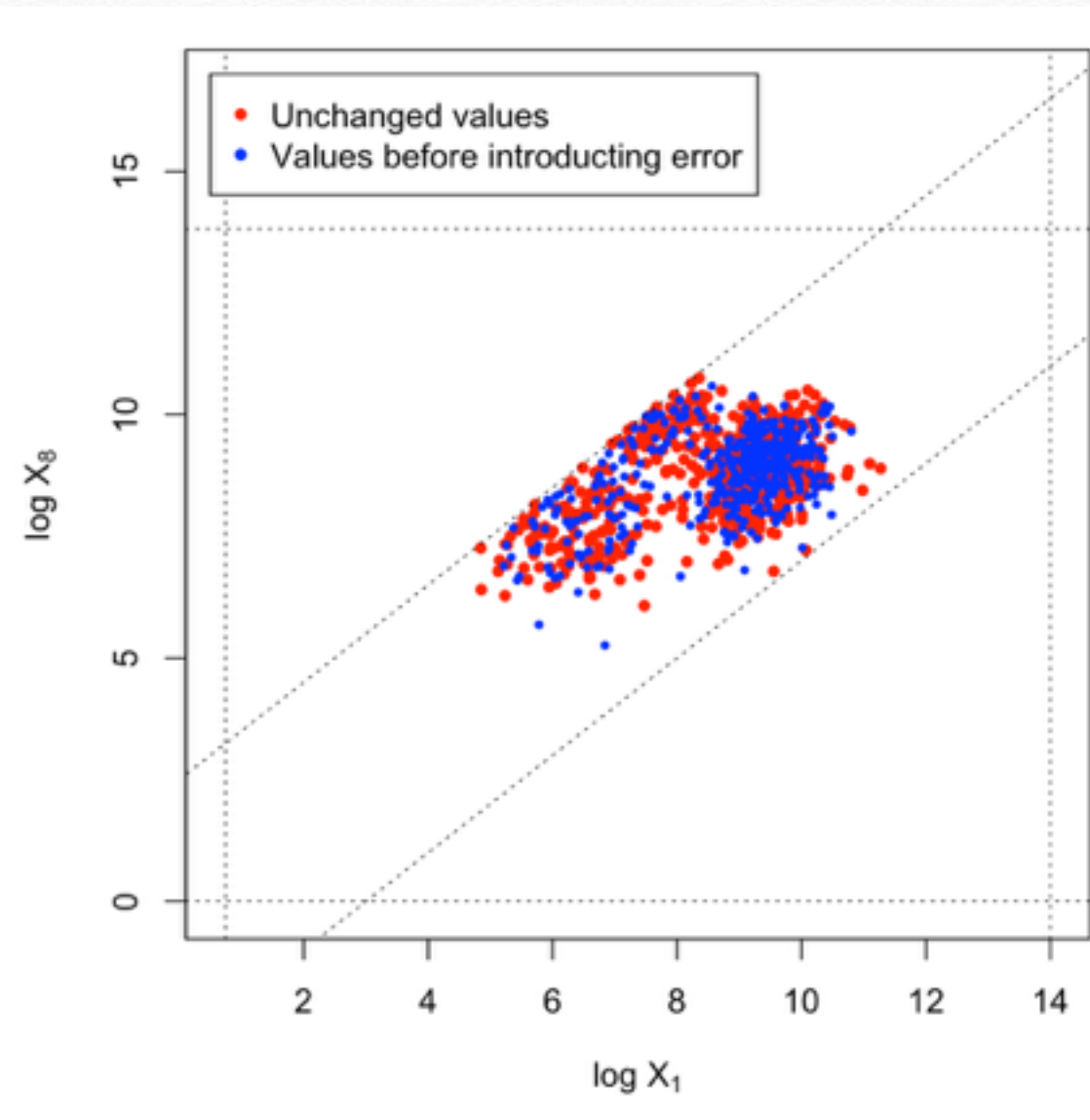
- ❖ F-H approach results in some edited values at tails of the distribution of true values

Result with Fellegi-Holt

True simulated values



Edited values by FH



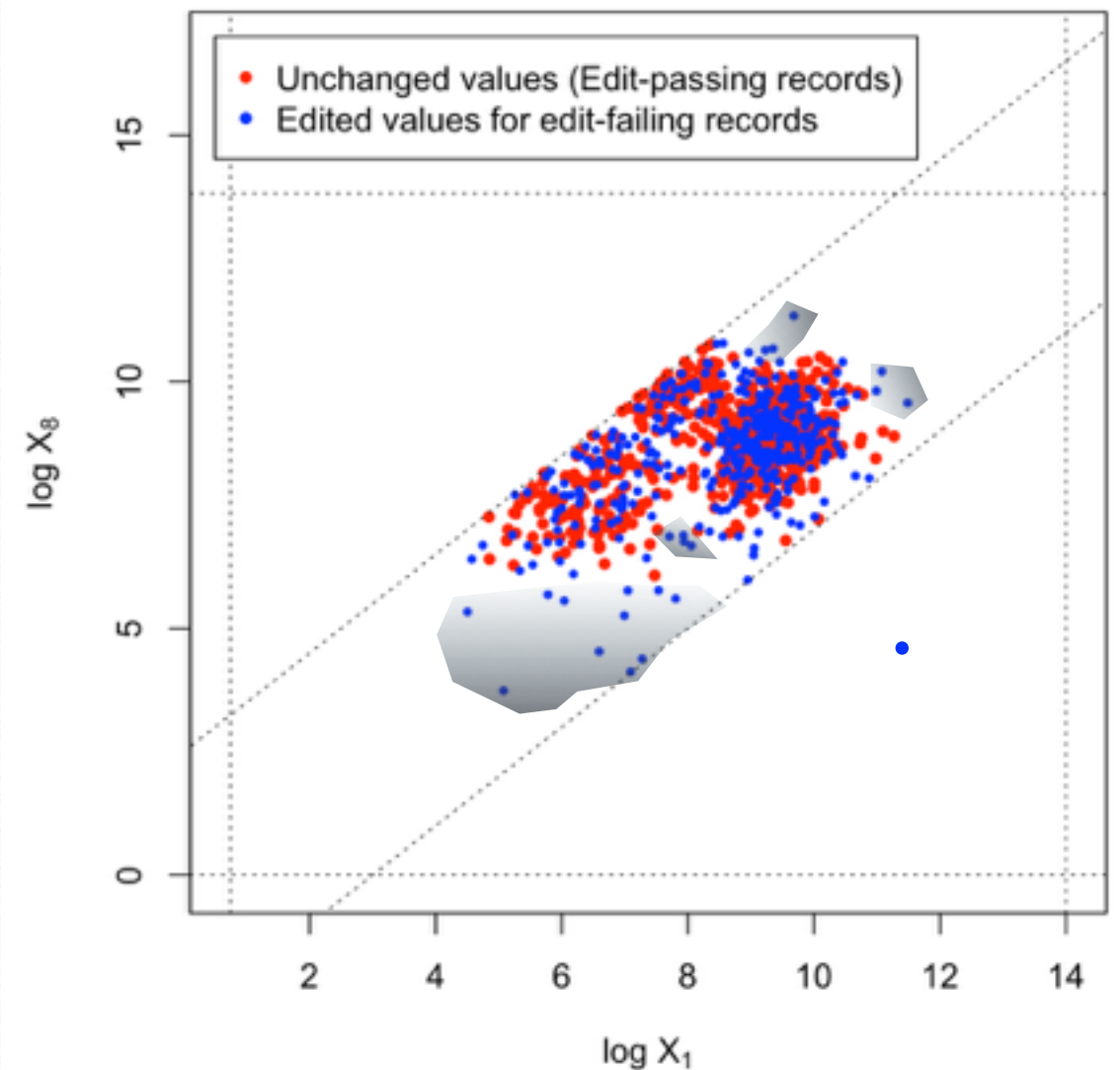
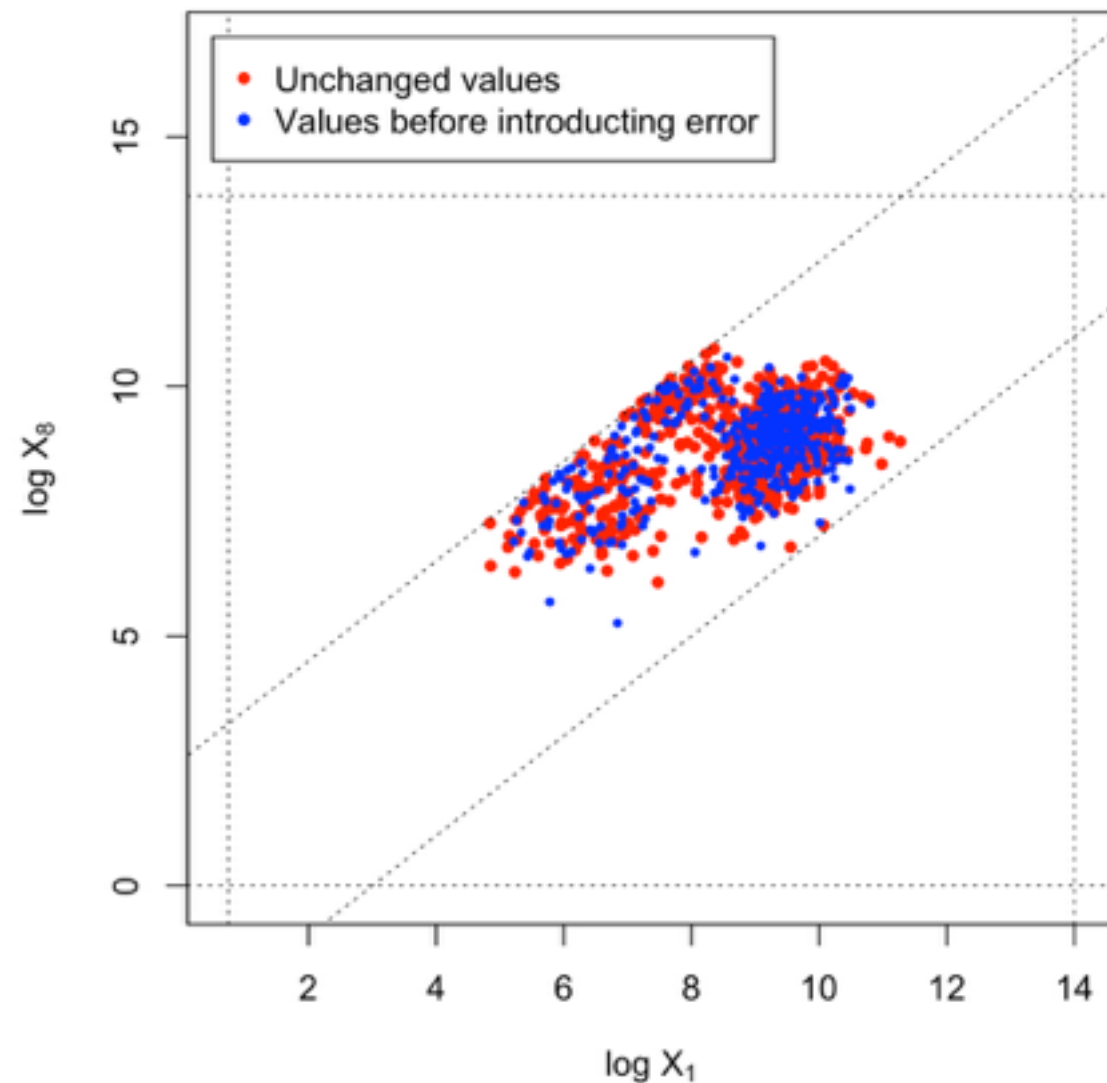
- ❖ F-H approach results in some edited values at tails of the distribution of true values

Result with Fellegi-Holt

True simulated values

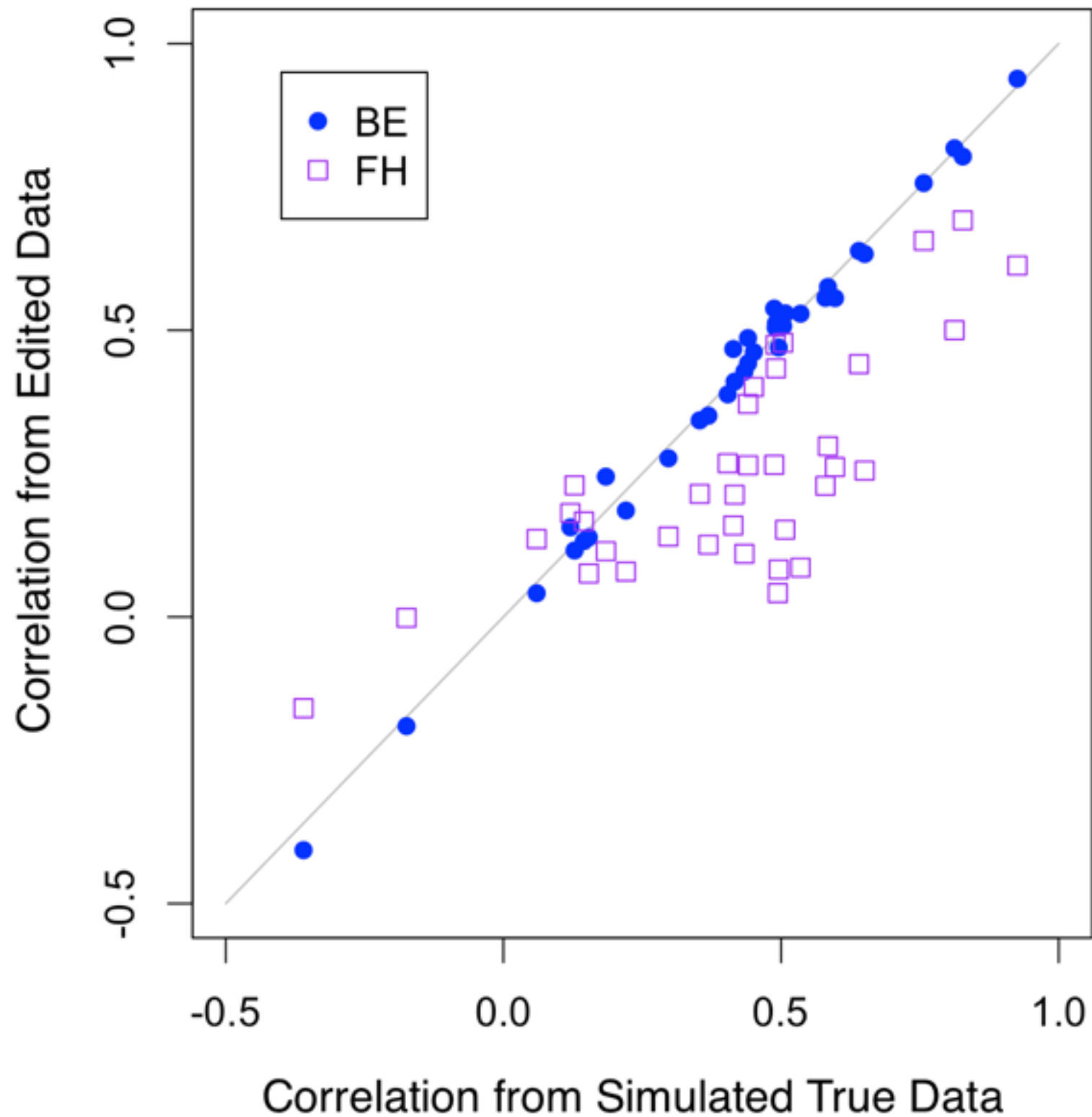


Edited values by FH



- ❖ F-H approach results in some edited values at tails of the distribution of true values

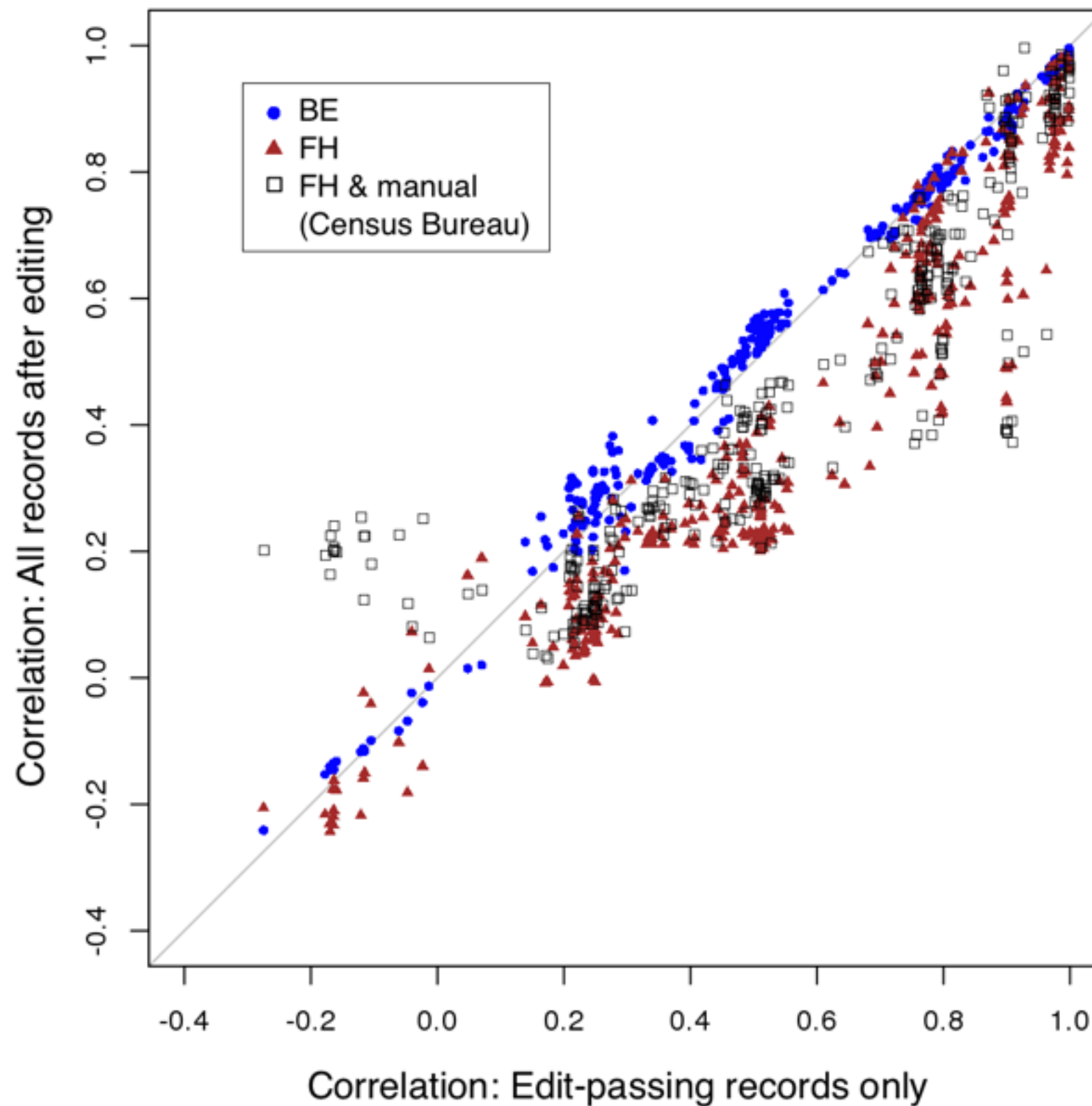
Simulation Study: Comparison of Pairwise Correlations from Edited Data



Application: U.S. Census of Manufactures

- Economic census for manufacturing industries, conducted by the U.S. Census Bureau every five years
 - variables: cost of materials, total emp., total value of shipments
 - widely used by researchers, e.g., interested in plant-level productivity
- **Current editing practice**
 - F-H based automatic editing system
 - using ratio edits and balance edits
 - additional (separated) manual editing processes
- We **compare three editing approaches** with pairwise correlation
 - BE: Bayesian Editing
 - FH: Fellegi-Holt based editing
 - FH & manual: Final edited data produced by the Census Bureau

2007 Census of Manufactures: Comparison of Pairwise Correlations from Edited Data



Topic III

Synthetic microdata for
the U.S. Census of
Manufactures



Integration of Imputation, Editing and Synthesizer (in progress)

- Two-stage Multiple Imputation Approach
 1. impute/edit survey data X given a size measure z
 - resulting in m copies of edited/imputed datasets, $X^{(1)}, \dots, X^{(m)}$
 2. generate synthetic data given $X^{(l)}$ and z
 - resulting in r synthetic file $X_1^{(l)}, \dots, X_r^{(l)}$ for $l=1, \dots, m$
- Inferences
 - based on mr complete-data analyses and combining rules
- Compared to current practices (with separate steps)
 - correctly estimate variance of final synthetic data
 - all benefits enjoyed by Bayesian editing/imputation



Concluding Remarks

- We proposed a Bayesian framework to integrate the currently-separated processes: imputation, data editing, and disclosure limitation
- A future research topic is simultaneous data editing/imputation methods for mixed type data, such as American Community Survey
- R package for Bayesian editing/imputation of continuous variables will be published on CRAN soon
- Technical reports are available at NISS websites (<http://www.niss.org/publications/technical-reports>)

“Thank you”

– Hang J. Kim (hangkim@niss.org)