Iowa State University
College of Engineering

General-Purpose Open-Source Program for Ultra Incomplete Data-Oriented Parallel Fractional Hot Deck Imputation (UP-FHDI)

Yicheng Yang, Jae-Kwang Kim, and In Ho Cho

Supported by
NSF Cyberinfrastructure for Sustained Scientific Innovation
Grant #: 1931380

Incomplete Data in Engineering and Science

Incomplete Data in Infrastructure Engineering

- Hybrid Data Set from Bridge and Transportation Sensor Data

Month	Day	Hour	DOW	steelTem	concTemp	airTempA	strainAvg	Total.Traf	Small.Veh	Medium.V	Large.Veh	-350~-345	-345~-340	-340~-335	-335~-330	-330~-325	-325~-320	-320~-315
7	31	10	5	84.08263	85.62893	84.08263	1.81626	611	385	50	33	0	0	0	0	0	0	0
7	31	11	5	85.53422	87.86025	85.53422	2.039314	1295	842	114	64	0	0	0	0	0	0	0
7	31	12	5	87.38675	91.2081	87.38675	2.23224	1311	863	115	58	0	0	0	0	0	0	0
7	31	13	5	88.56247	94.13685	88.56247	4.33039	1286	883	109	56	0	0	0	0	0	0	0
7	31	14	5	89.17667	96.52857	89.17657	9.566355	1530	1063	145	48	0	0	0	0	0	0	0
7	31	15	5	88.9463	97.84863	88.9463	12.80208	1670	1165	130	60	0	0	0	0	0	0	0
7	31	16	5	89.00592	98.78397	89.00592	14.6337	1989	1436	173	51	0	0	0	0	0	0	0
7	31	17	5	88.30502	98.78458	88.30502	17.24283	1784	1230	192	33	0	0	0	0	0	0	0
7	31	18	5	87.61205	99.01495	87.61205	18.2749	1234	819	146	36	0	0	0	0	0	0	0
7	31	19	5	85.4382	97.94607	85.4383	15.883	960	634	119	42	0	0	0	0	0	0	0
7	31	20	5	83.90578	96.76805	83.90578	10.90513	733	470	76	44	0	0	0	0	0	0	0
7	31	21	5	82.69365	94.84543	82.69365	10.36729	621	380	98	43	0	0	0	0	0	0	0
7	31	22	5	80.78168	92.44818	80.78168	9.539218	376	223	69	29	0	0	0	0	0	0	0
7	31	23	5	79.13237	90.35015	79.13237	9.414203	296	184	56	30	0	0	0	0	0	0	0
8	1	0	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	1	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	2	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	3	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	4	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	5	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	6	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	7	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	8	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	9	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	10	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	11	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1	12	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

(raw data from Dr. Phares and Dr. Sharma)

- Shear Wall Structure Database (ACI 445-B; SERIES; BRI Wall Database)

Specimen Name	Vertical Web reinforcement ratio	Horizontal Web Reinforcement Ratio	Steel_Vertical1_fy	Steel_Vertical1_fu	Steel_Vertical1_Spacing	Steel_Vertical1_strain at fu	Steel_Vertical1_Diameter
RW1	3.27E-03	3.27E-03	3.00E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	3.30E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	3.60E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	3.90E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	4.20E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	4.50E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	4.80E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	5.10E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	5.40E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	5.70E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	6.00E+08	4.14E+14	5.08E+01	1.00E-01	9.53E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	1.27E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	1.59E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	1.91E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	2.22E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	2.54E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	2.87E-0
RW1	3.27E-03	3.27E-03	3.95E+08	4.14E+14	5.08E+01	1.00E-01	3.23E-0
TW2	4.45E-03	4.45E-03	4.14E+08	6.41E+08		1.00E-01	9.53E-0
TW2	4.45E-03	4.45E-03	4.14E+08	6.41E+08		1.00E-01	9.53E-0
TW2	4.45E-03	4.45E-03	4.14E+08	6.41E+08		1.00E-01	9.53E-0
TW2	4.45E-03	4.45E-03	4.14E+08	6.41E+08		1.00E-01	9.53E-0
TW2	4.45E-03	4.45E-03	4.14E+08	6.41E+08		1.00E-01	9.53E-0
DP1			6.05E+08				7.00E-0
BW1-O	2.86E-03	4.10E-03	4.14E+08				0.01299988
BW1-O	2.86E-03	4.10E-03	4.14E+08				0.01299988

NA

(domain-specific community database)

Naïve Remedy for Imputation

Widely Used Naïve Method in ML Community

- Naive imputation: simply use each variable's **mean** to impute missing values
- **Removal of** the entire unit (instance) which has missing values

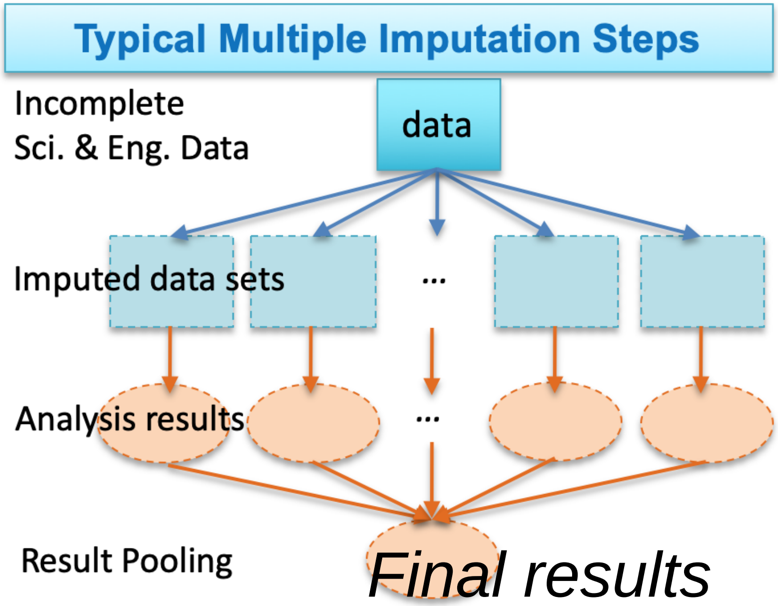
Statistical problems resulting from the naïve remedy

- Loss of substantial information
- May introduce unexpected bias
- May lead to low accuracy in machine learning/statistical predictions
- May mislead incorrect statistical inference

Other Popular Imputation Methods?

Multiple Imputation (MI)

- One of the most popular imputation methods
- Create *M* completed datasets for full imputation uncertainty
- Since Rubin (1976), extensive investigations have been conducted (Rubin 1987, Schafer 1997, Little and Rubin 2002, etc.)



M imputed data sets

Statistical Analysis on each set

Final results

Popular Imputation Methods: MI

Difficulty in General Use of Multiple Imputation

MI requires

- “congeniality” condition (Meng 1994) and
- “self-efficient” estimation (Meng and Romero 2003)

If not, the MI variance estimator may be

- inconsistent (Nielsen 2003; Kim et al. 2006) and
- considerably biased (Beaumont et al. 2011).

Challenges of Existing Imputation Methods for Big Incomplete Data

- They often require statistical and/or distributional assumptions, which are obstacles for general researchers.
- Computational limits of them prevent general applications to large/big incomplete data in broad Eng. or Sci.

Our Choice for Big Data Imputation: Fractional Hot Deck Imputation

Our group developed and shared a public, open-source *R* package "FHDI" (***The R Journal***, 2018 [1])

Strengths of "Hot Deck" Imputation

- Do not require "self-efficient" estimation condition
- Do not create artificial values, instead use the real observations
- Do not need model/distributional assumptions
- Seek to leverage and preserve the joint probability of data available.

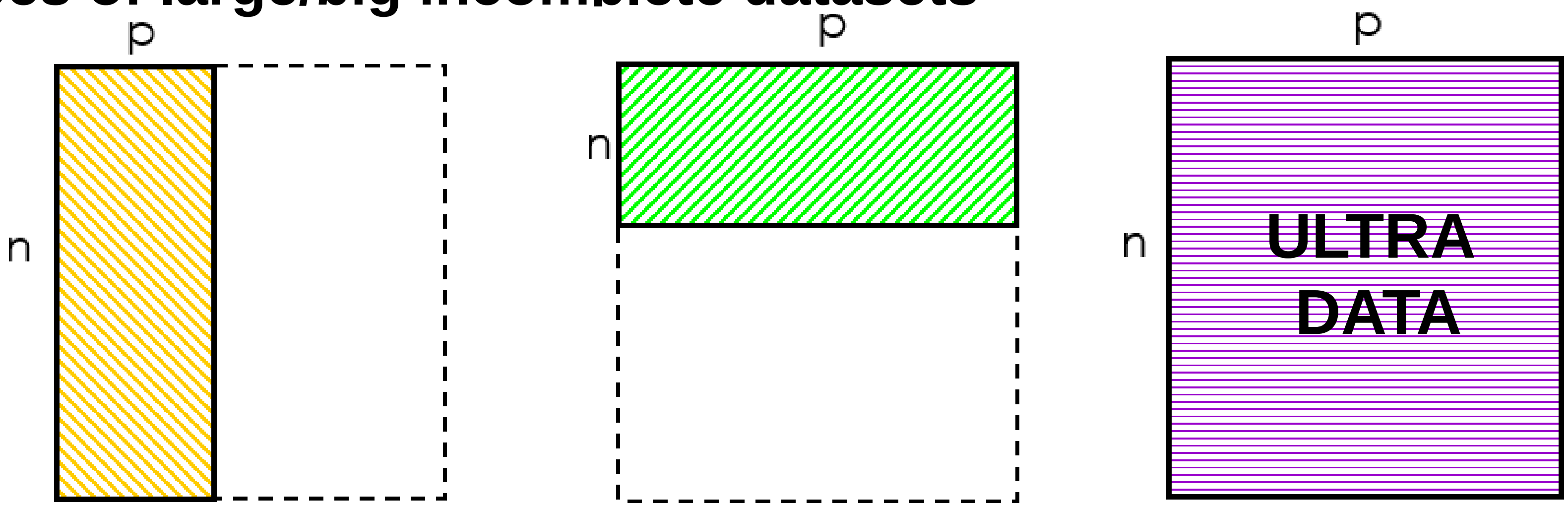
Still, FHDI is not suitable for tackling big incomplete data

Motivations to develop UP-FHDI

- Limitations of the serial version R package FHDI regarding time and memory requirements
- Hard to deal with large/big data with immense volume and/or too many variables
- The positive impact of FHDI on learning and prediction (Cho et al. ***IEEE, TKDE,*** 2019 [2])
- As we enter the era of big data and powerful computing, parallel computing techniques are gradually attempted in imputations.
- Strong need for general-purpose and assumption-free big data (big-n and/or big-p) imputation tools

Parallel Fractional Hot Deck Imputation for Ultra Data

Types of large/big incomplete datasets



big-n data: $n \gg p$
e.g., $n = 1M, p = 4$

big-p data: $n \leq p$
e.g., $n = 1000, p = 10{,}000$

Ultra data: n and p are both large
e.g., $n = 1M, p = 10{,}00$

Tackled by P-FHDI ver. 1.0
(Cho et al. ***IEEE, TKDE,*** 2020 [3])

UP-FHDI
(Cho et al. ***IEEE, TKDE***, 2021 [4]. Under review)

n: number of instances
p: number of variables
η: missing rate

Key Procedures of UP-FHDI

- **Parallel Cell Construction** (denoted as Process 1)

  Categorization of imputation cells

  Donor selection in conjunction with the sure independence screening (SIS) and K-nearest neighbor (KNN) searching
- **Parallel Cell Probability Estimation** (Process 2)

  Estimate probability for each unique observed cell pattern using EM algorithm
- **Parallel Imputation** (Process 3)

  Missing values are imputed by donors
- **Parallel Variance estimation** (Process 4)
  - Jackknife method for moderately large data
  - Linearized variance estimation for ultra data

Parallel Computing Techniques

Library: MPI (Message Passing Interface)

- Suitable for distributed memory
- Specifies names, calling sequences, and results of functions/subroutines needed to communicate via message passing
- Language bindings for C/C++ and Fortran

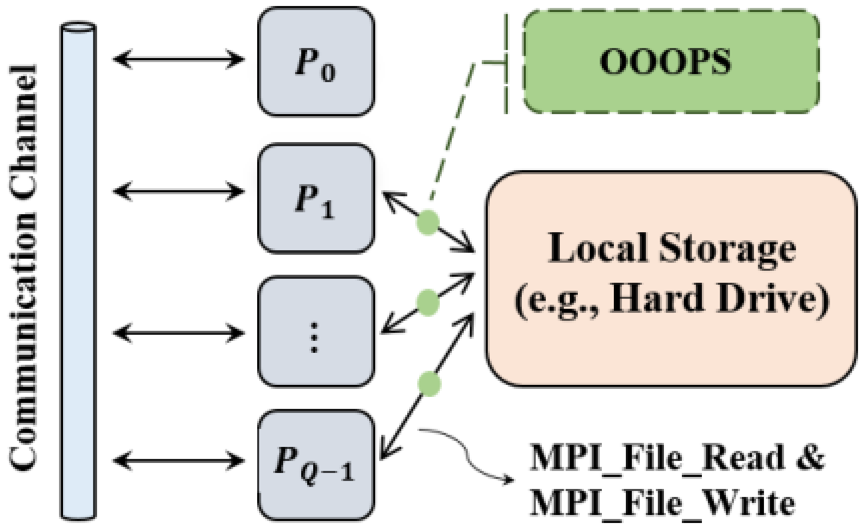
High-performance computing (HPC) facilities:

- Condo2017 [5]: 158 servers. Each server has two 8-core Intel Haswell processors, 128 GB of memory and 2.5 TB local storage
- TACC Stampede2 [6]: 4704 servers. Each server has 192 GB of memory and no quota for local storage

How UP-FHDI Process Big Data?

Parallel file system on Q processors indexed by $0, \cdots, Q-1$.

Intensive IO may harm global distributed system of HPC



Adapted from Cho et al. ***IEEE, TKDE***, 2021 [4]. Under review

- Processers can communicate via communication channel
- Store input data and temporary data in local storage provided by HPC facilities
- Slave processors only fetch required data to their memory
- Optimal Overload IO Protection System (OOOPS) adjusts intensive IO workload

Available Example Datasets for UP-FHDI

- Following example big-n or big-p datasets are available at IEEE DataPort [7]

Dataset	Variable type	Dimension
Synthetic data 1	Continuous	$\mathbf{U}(1000, 4, 0.25)$
Synthetic data 2	Continuous	$\mathbf{U}(10^6, 4, 0.25)$
Air Quality	Hybrid	$\mathbf{U}(41757, 4, 0.1)$
Nursery	Categorical	$\mathbf{U}(12960, 5, 0.3)$
Synthetic data 3	Continuous	$\mathbf{U}(15000, 12, 0.15)$
Synthetic data 4	Continuous	$\mathbf{U}(15000, 16, 0.15)$
Synthetic data 5	Continuous	$\mathbf{U}(15000, 100, 0.15)$
Synthetic data 6	Continuous	$\mathbf{U}(1000, 100, 0.3)$
Synthetic data 7	Continuous	$\mathbf{U}(1000, 1000, 0.3)$
Synthetic data 8	Continuous	$\mathbf{U}(1000, 10000, 0.3)$
Appliance Energy	Continuous	$\mathbf{U}(19735, 26, 0.15)$

Adapted from Cho et al. ***IEEE, TKDE***, 2020 [3].

Note that $\mathbf{U}(n, p, \eta)$ represents incomplete data with n rows and p columns with η missing rate

- Please refer to (Cho et al. ***IEEE, TKDE,*** 2020 [3]) for more details
- Source codes of parallel FHDI are available and executable on local HPC or NSF Cloud Computing (e.g. NSF XSEDE).

Available Example Datasets for UP-FHDI

- Following real-world ULTRA datasets are available at IEEE Dataport [7]

Dataset name	# Instances	# Variables	Category	Source
Swarm	24016	2400	Biology	UCI
CT	53500	380	Medicine	UCI
P53	31159	5408	Genetics	UCI
Radar	325834	175	Agriculture	UCI
Travel	23772	50	Transportation	IEEE DataPort
Bridge	492641	31	Civil	Dr. Cho
Earthquake	901512	15	Civil	USGS

- Please refer to (Cho et al. ***IEEE, TKDE,*** 2021 [4]) for more details
- Source codes of UP-FHDI are available and executable on local HPC or NSF Cloud Computing (e.g. NSF XSEDE).

Strength of UP-FHDI

- The UP-FHDI inherits all strengths of the general-purpose, assumption-free FHDI
- UP-FHDI can cure incomplete synthetic data with one million instances and $10,000$ variables of 80 GB (30% missing rate) in 35 hours with 240 processors
- UP-FHDI positively improves the subsequent machine learning
- The UP-FHDI is now publicly available
- Researchers in broad engineering and science can cure general, large/big data sets with ease

Reference

[1] J. Im, I. Cho, and J. K. Kim, "An R package for fractional hot deck imputation," *The R Journal*, vol. 10, no. 1, pp. 140–154, 2018.

[2] I. Song, **Y. Yang**, J. Im, T. Tong, C. Halil, and I. Cho. "Impacts of fractional hot deck imputation on learning and prediction of engineering data," *IEEE Transactions on Knowledge and Data Engineering* (in-press). [10.1109/TKDE.2019.2922638], 2019.

[3] **Y. Yang**, J. K. Kim, and I. Cho. "Parallel fractional hot deck imputation and variance estimation for big incomplete data curing," *IEEE Transactions on Knowledge and Data Engineering,* 2020 (in-press).

[4] **Y. Yang**, J. K. Kim, and I. Cho. "Ultra data-oriented parallel fractional hot-deck imputation with efficient linearized variance estimation," *IEEE Transactions on Knowledge and Data Engineering,* 2021 (under review).

[5] Condo, "Condo2017: Iowa state university high-performance computing cluster system," 2017. [Online]. Available: https://www.hpc.iastate.edu/guides/condo-2017

[6] TACC, "Texas advanced computing center (tacc) at the university of texas at austin," 2017. [Online]. Available: http://www.tacc.utexas.edu

[7] **Y. Yang**, J. K. Kim, and I. Cho. "Incomplete big datasets for ultra data-oriented parallel fractional hot-deck imputation," *IEEE DataPort*, 2021.

Thank you!

For programs, data sets, and discussion
feel free to contact
icho@iastate.edu

Acknowledgement: Generous support from
NSF CSSI, OAC-1931380

Supplementary Materials

Generate High-Dimensional Synthetic Data

Let $i = 0$ and repeat the following by setting $i = i + 4$ until we obtain p variables:

$$Y_i = \begin{cases} 1 + e_i, & \text{if } i = 0 \;||\; i\%8 = 0 \\ Y_{i-1} + e_i & \text{if } i\%8 \neq 0 \end{cases}$$

$$Y_{i+1} = Y_i + 2 + \rho \times e_i + \sqrt{1-\rho^2}e_{i+1}$$

$$Y_{i+2} = Y_{i+1} + e_{i+2}$$

$$Y_{i+3} = -1 + Y_i + 0.25Y_{i+1} + e_{i+3}$$

where $\rho = 0.5$ and e_i, e_{i+1}, e_{i+3} are randomly generated by normal distribution. And e_{i+2} is generated by gamma distribution.