Multivariate Missing Handling for Mixed Data

결측자료분석 Final Project

박수희-이혜원-최홍석

INDEX

1 **Generate Missing**

- Generate 6 types of Missing dataset (R ampute)

2 **Imputation/Inference**

- Expected Maximization
- Data Augmentation
- Chained-Equation Multiple Imputation

3 **Conclusion**

Generate Missing

Data used

wine quality dataset *(from UCI machine learning repository)*

- red wine quality data: 1599 rows, 13 variables
- white wine quality data: 4898 rows, 13 variables

➡ **wine_total dataset** (red wine + white wine) : 6497 rows, 13 variables

Fixed acidity	volatile acidity	Citric.acid	Residual.sugar	chlorides	Free.sulfur.dioxide	Total.sulfulr.dioxide
float	float	float	float	float	integer	integer

density	pH	sulphates	alcohol	quality	kinds
float	float	float	float	categorical	categorical

Generate Missing

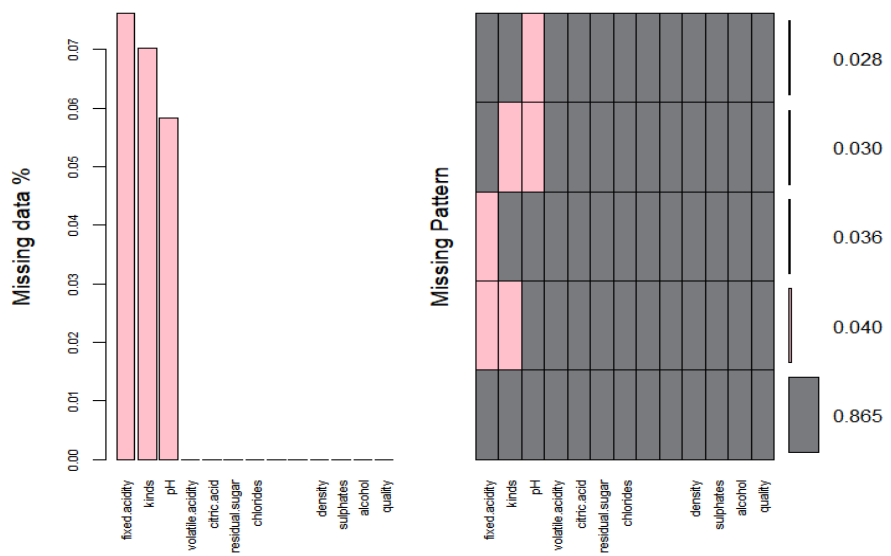
Generate Missing by R function ampute

- Missing variables: **fixed.acidity**(num), **pH**(num), **kinds**(category)
- Missing data mechanism: **Missing at Random(MAR)** ➡ *mech="MAR"*

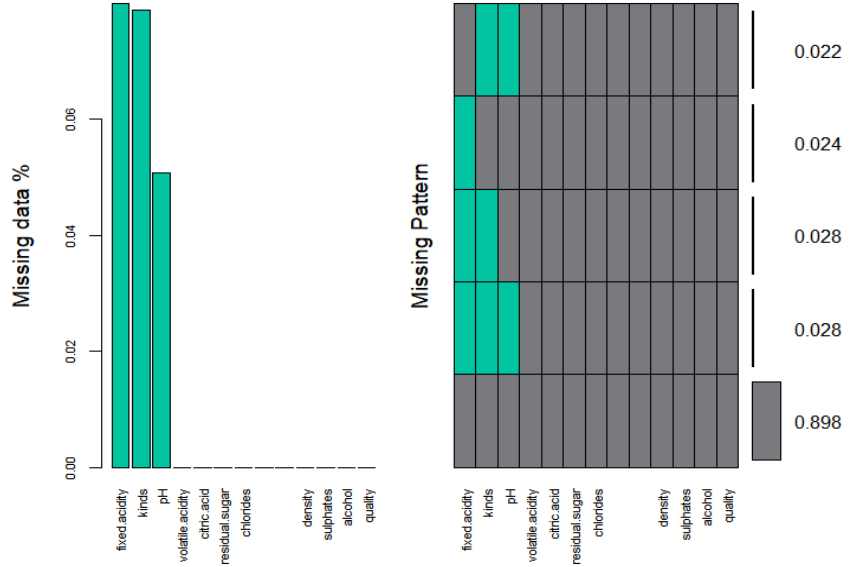
Mice package의 **ampute** 함수를 이용하여 MAR 가정 하에서 missing을 발생시킨다.
이 때 **두 가지의 missing pattern** 을 생성하고, 각 패턴에 대해 **20%, 35%, 50%** 비율로 missing을 만든다.

➡ 총 6가지의 missing data에 대해 imputation 및 평균, 분산에 대한 inference를 진행한다.

Form #1



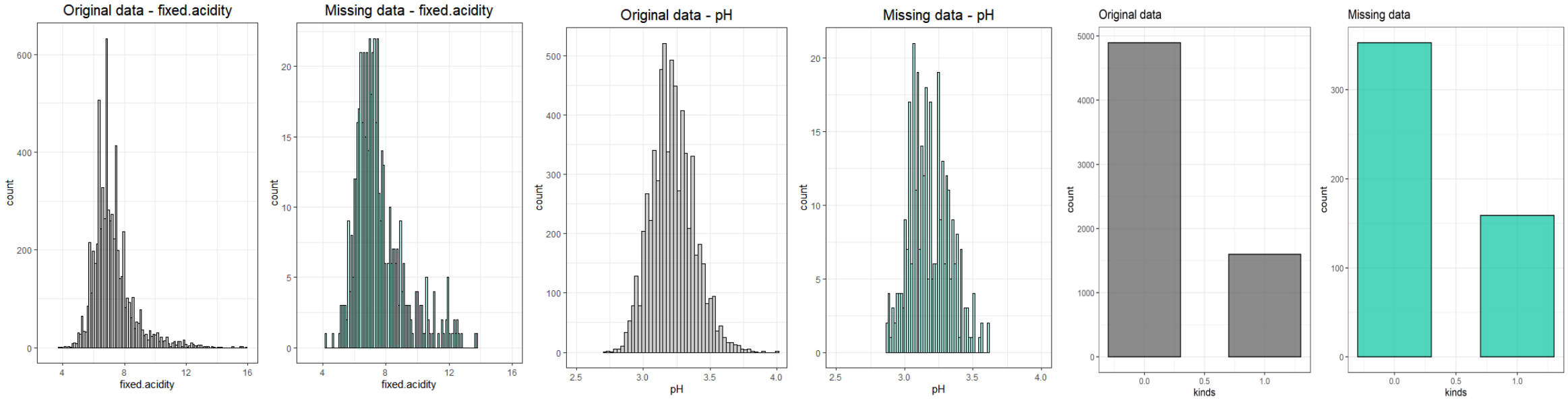
Form #2



Generate Missing

Distributions for Original vs Missing index

각 변수의 결측된 값과 원래 값의 분포를 비교한 결과, 극명한 차이는 아니지만 비교적 상이한 것을 확인할 수 있다.



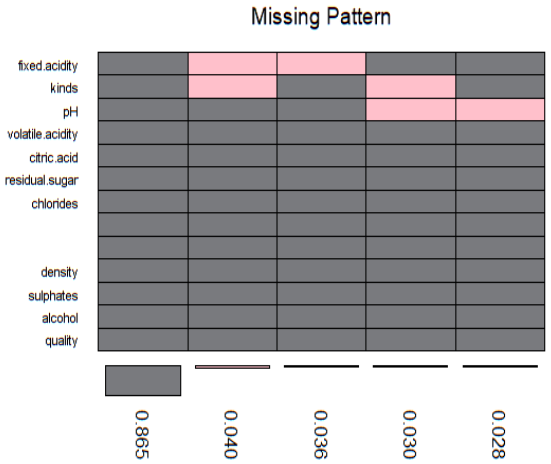
Generate Missing

Missing percentage of each variable for 6 types of data

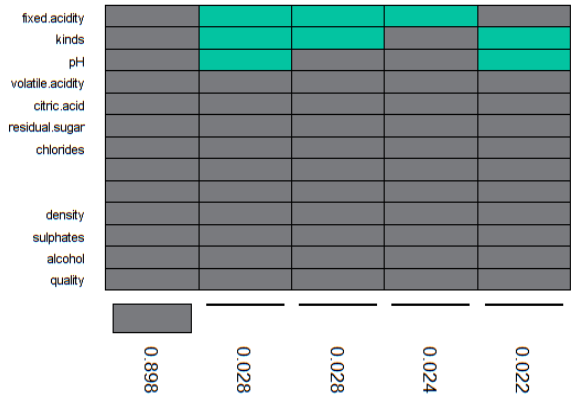
6가지 데이터에서 각 변수의 결측 비율은 다음과 같다.

	fixed.acidity	pH	kinds
Form1-20%	5.83%	7.02%	7.98%
Form1-35%	9.33%	11.24%	13.44%
Form1-50%	20.64%	13.11%	16.62%
Form2-20%	7.98%	5.08%	7.88%
Form2-35%	13.44%	8.77%	13.16%
Form2-50%	18.93%	12.27%	18.72%

Form1

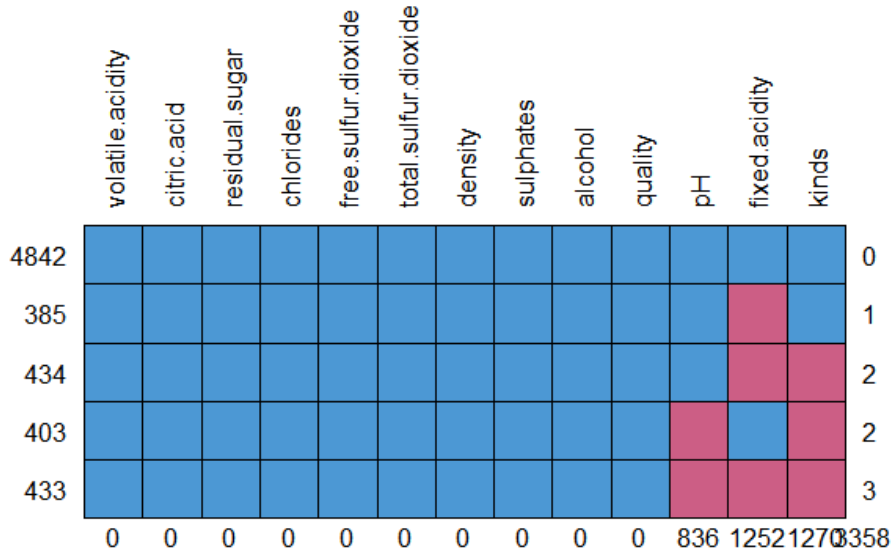


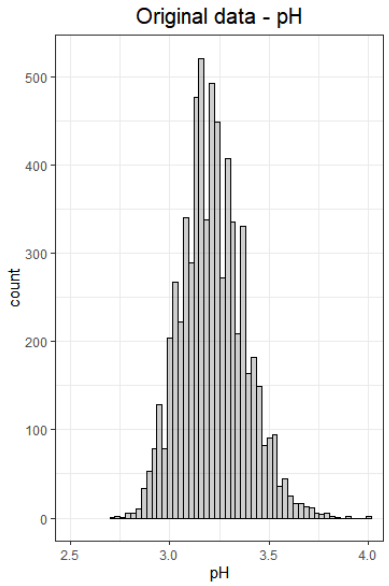
Form2

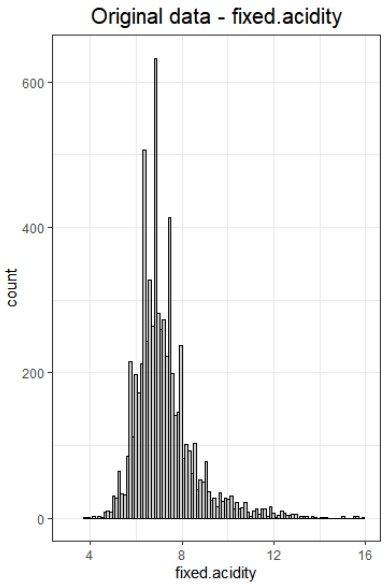


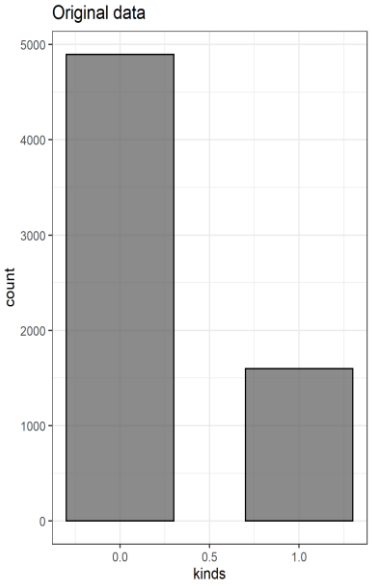
Imputation/Inference

Not Satisfy Multivariate Normal Assumption









1. **EM/DA** under Multivariate Normal assumption (R library: Amelia/Norm)
2. **EM/DA** using General Location Model (R library: mlmi)
3. **Chained-Equation Multiple Imputation**

Imputation/Inference

1

Expected Maximization

- EM under Multivariate Normal Assumption (R Amelia)
- EM using General location model (R mlmi)

2

Data Augmentation

- DA under Multivariate Normal Assumption (R Norm)
- DA using General location model (R mlmi)

3

Chained-Equation MI

- Chained-Equation MI (Mice)

Imputation : (1)EM

Try1) EM under Multivariate Normal Assumption

Library used: Amelia

Missing data에 EM algorithm을 지원하는 패키지

amelia *AMELIA: Multiple Imputation of Incomplete Multivariate Data*

Description

Runs the bootstrap EM algorithm on incomplete data and creates imputed datasets.

Usage

```
amelia(x, ...)
```

Usage

amelia(data, m, noms)

- data: missing(NA) 을 포함하고 있는 데이터 input
- m: EM을 몇 번 시행할 것인지를 설정하는 값. 만약 m=3 이면 EM을 3번 시행
- noms: categorical 변수 입력

"R Library Amelia는 **Multivariate Normal Assumption** 필요"

https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

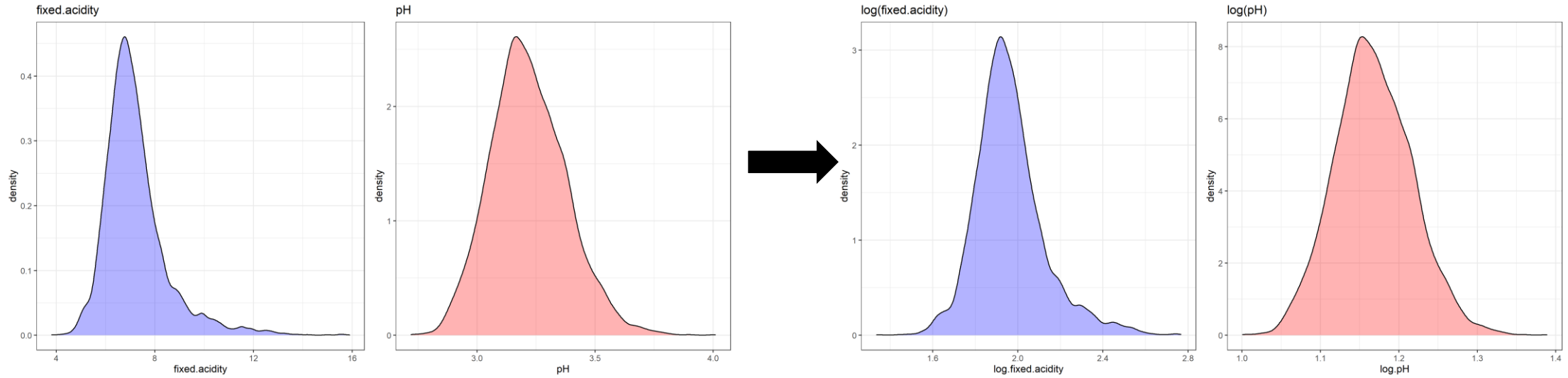
Imputation : (1)EM

Multivariate Normal Assumption for EM/DA

- Schafer (4) suggested that inference from MVNI **may often be reasonable even if multivariate normality does not hold**, and MVNI has been widely applied in contexts where data are clearly not multivariate normal.(1)
- One can often make the normality assumption more tenable by **applying suitable transformations** to one or more of the variables.(2)

Variable Transformation

연속형 변수들은 Log trasformation을 진행하여 skewness를 완화시켰다.



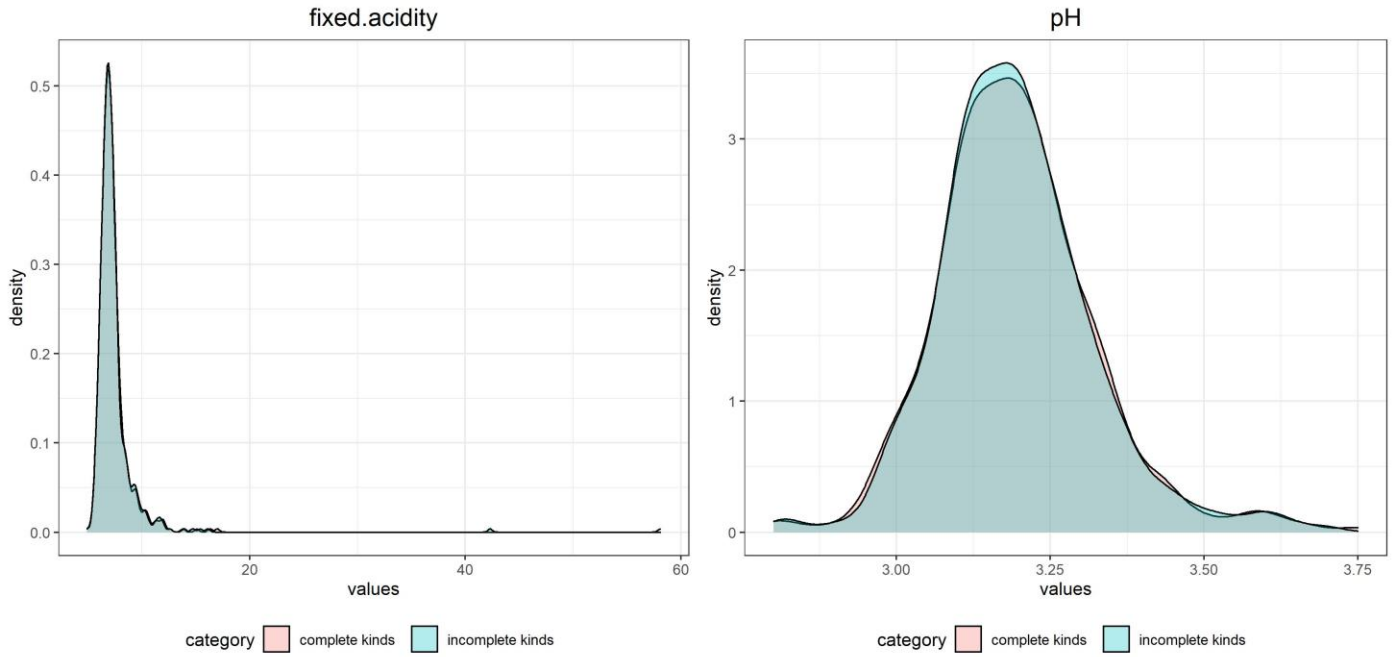
(1)- Katherine J. Lee* and John B. Carlin (2010): Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation
(2) - J.L. Schafer (1997): Analysis of Incomplete Multivariate Data

Imputation : (1)EM

Multivariate Normal Assumption for EM/DA

(2) If some variables are not normal (ex. discrete) but are completely observed, then the multivariate normal model may still be used for inference.

Discrete variable의 fully observed 여부에 따른 연속형 변수 Imputation 후 분포

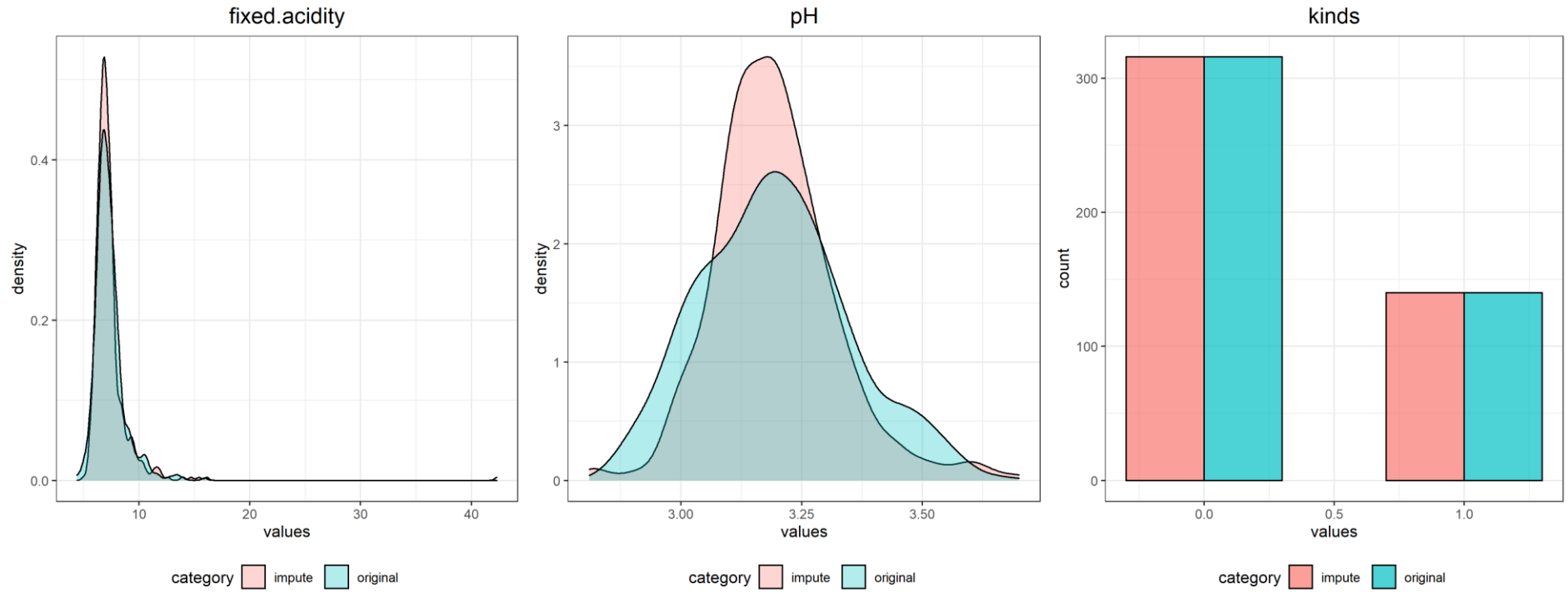


➡ 해당 데이터에서는, Discrete variable이 완전히 관측된 경우와 결측이 존재하는 경우 각각에서,
Multivariate Normal Assumption을 가정하고 연속형 변수들을 대체한 결과의 차이가 크지 않았다.
이는 kinds를 제외하고도 관련성이 높은 변수들이 많았던 것으로 판단되었다.

Imputation : (1)EM

Distributions for Original vs Inputed index

For data Form1-20%



Impute 값들이 중앙에 밀집해 있다.

Imputation : (1)EM

Estimated Mean/Variance

<fixed.acidity>

	Mean	Variance
Form1-20%	7.32202	1.94909
Form1-35%	7.28918	1.51745
Form1-50%	7.25377	1.79098
Form2-20%	7.45453	2.07615
Form2-35%	7.46976	1.78551
Form2-50%	7.39260	1.88952
Original	7.21531	1.68074

<pH>

	Mean	Variance
Form1-20%	3.19116	0.01729
Form1-35%	3.20119	0.02621
Form1-50%	3.19570	0.02327
Form2-20%	3.19470	0.01289
Form2-35%	3.21065	0.02111
Form2-50%	3.20418	0.01368
Original	3.21850	0.02585

<kinds>

	Accuracy
Form1-20%	0.9934
Form1-35%	0.99178
Form1-50%	0.99167
Form2-20%	0.99414
Form2-35%	0.99298
Form2-50%	0.99256

Imputation : (1)EM

Sample Data

	original.fixed.acidity	impute.fixed.acidity	original.pH	impute.pH	original.kinds	original.kinds.1
1	8.9	8.1	3.11	2.98	1	1
2	8.9	8.1	3.45	3.33	1	1
3	8.5	8.8	3.16	3.05	1	1
4	7.9	7.8	3	2.98	1	1
5	7.5	7.5	3.26	3.18	1	1
6	7.8	7.3	3.28	3.34	1	1
7	8.6	8.5	3.57	3.65	1	1
8	7.8	7.5	3.35	3.29	1	1
9	7.3	6.5	3.31	3.35	1	1
10	8.2	8.1	3.07	2.97	1	1

Imputation : (1)EM

Try2) EM using General Location Model

Library used: mlmi

연속형과 범주형이 섞인 데이터에서 GLOM을 이용하여
EM, DA 등의 Maximum likelihood Multiple Imputation을
지원하는 패키지

mixImp — *Imputation for a mixture of continuous and categorical variables using the general location model.*

Description

This function performs multiple imputation under a general location model as described by Schafer (1997), using the `mix` package. Imputation can either be performed using posterior draws (`pd=TRUE`) or conditonal on the maximum likelihood estimate of the model parameters (`pd=FALSE`), referred to as maximum likelihood multiple imputation by von Hippel (2018).

Usage

```
mixImp(obsData, nCat, M = 10, pd = FALSE, marginsType = 1,
  margins = NULL, designType = 1, design = NULL, steps = 100,
  rseed)
```

Usage

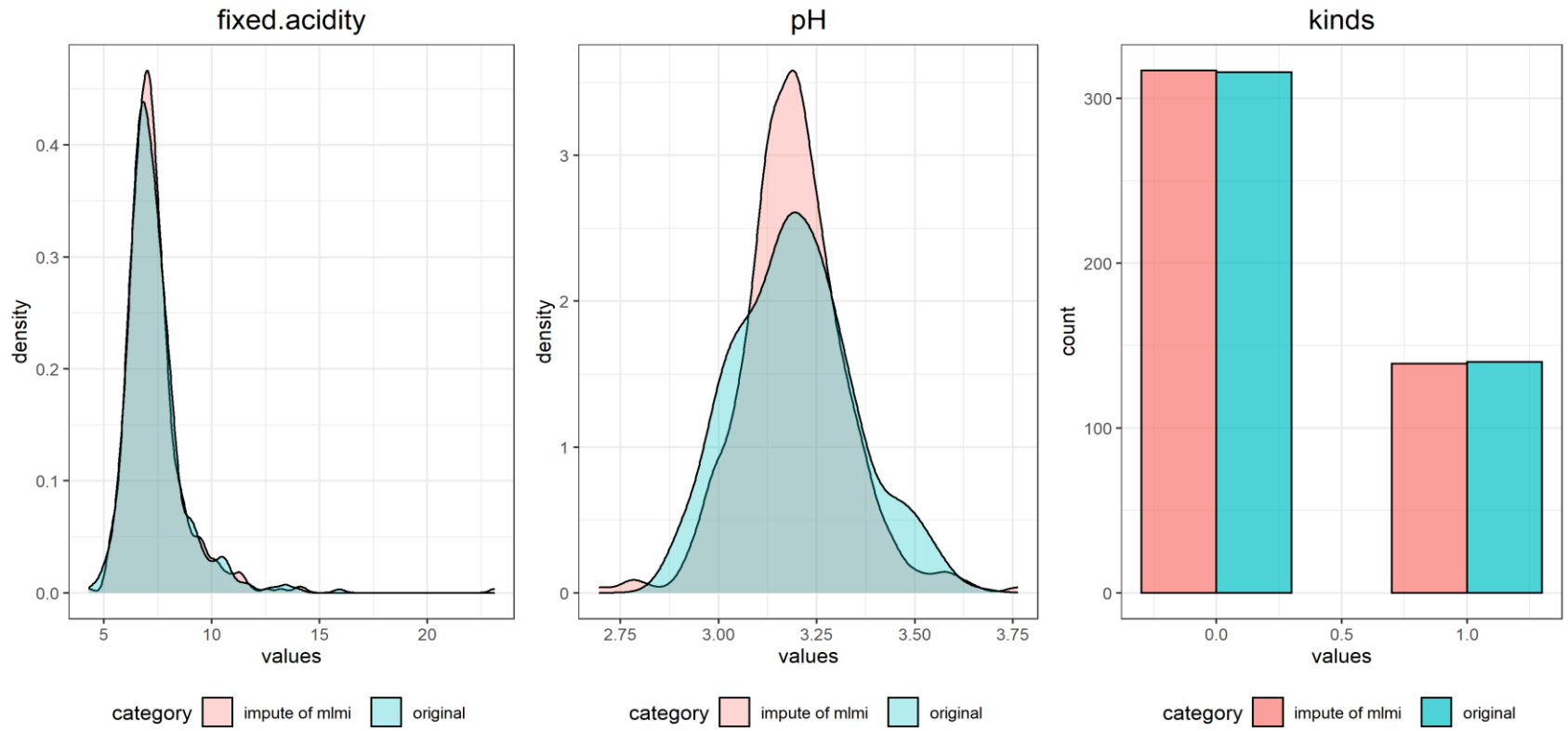
mixImp(data, M, steps, nCat, pd=FALSE)

- data: missing(NA) 을 포함하고 있는 데이터 input
- M: EM을 몇 번 시행할 것인지를 설정하는 값. 만약 m=3 이면 EM을 3번 시행
- Steps: E와 M사이를 converge하기 위해 몇 번 반복할 것인지 설정하는 값
- pd: FALSE=EM algorithm / TRUE=DA algorithm
- nCat: category 개수 (category 변수들은 데이터에서 앞 열들에 위치)

Imputation : (1)EM

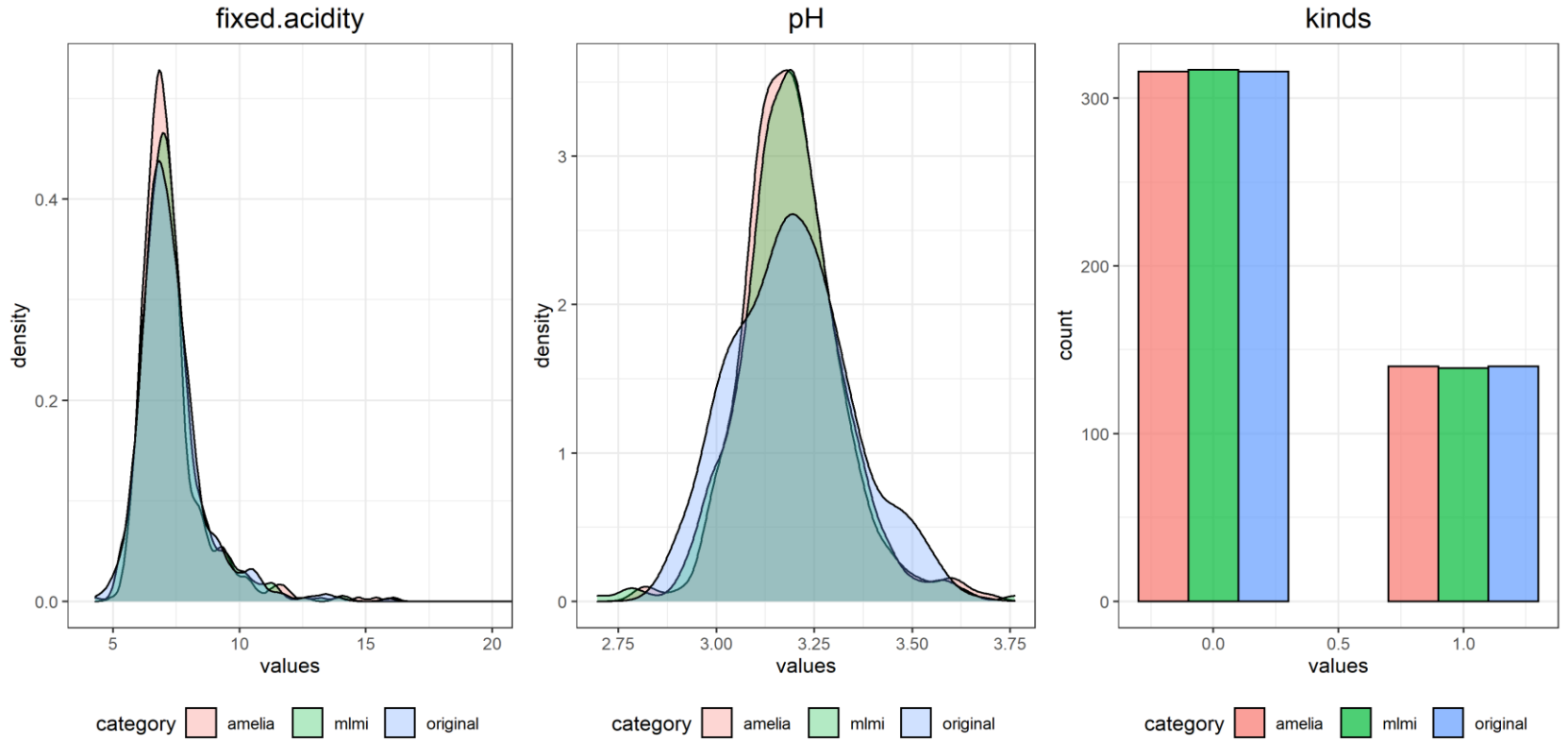
Distributions for Original vs Inputed index

For data Form1-20%



Imputation : (1)EM

Distributions for Original vs Amelia Inputed vs mlmi Imputed index



➡ 전반적으로는 Mlmi와 amelia가 유사한 분포를 보이고 있는데,
Fixed.acidity 경우에서 amelia보다 mlmi가 더 실제 분포에 가까운 모습을 보인다.

Imputation : (2)DA

Try1) DA under Multivariate Normal Assumption

Library used: NORM

Incomplete Multivariate Normal data에서
DA algorithm을 지원하는 패키지

Usage

da.norm(s, start, prior, steps=1, …)

- s: plelim.norm을 통해 얻은 incomplete data에 대한 정보
- start: 초기 파라미터 값
- steps: I와 P사이를 converge하기 위해 몇 번 반복할 것인지 설정하는 값

  ⟶ **steps=3, 100, 1000** 의 3가지 경우 진행

da.norm — *Data augmentation for incomplete multivariate normal data*

Description

Data augmentation under a normal-inverted Wishart prior. If no prior is specified by the user, the usual "noninformative" prior for the multivariate normal distribution is used. This function simulates one or more iterations of a single Markov chain. Each iteration consists of a random imputation of the missing data given the observed data and the current parameter value (I-step), followed by a draw from the posterior distribution of the parameter given the observed data and the imputed data (P-step).

Usage

```
da.norm(s, start, prior, steps=1, showits=FALSE, return.ymis=FALSE)
```

$For\ \ t\ \ in\ \ 1{:}\,steps,$

$$I-step{:}\ \ Y_{(0)}^{(t+1)} \sim P\left(Y_{(0)} \middle| Y_{(1)}, \theta^{(t)}\right)$$

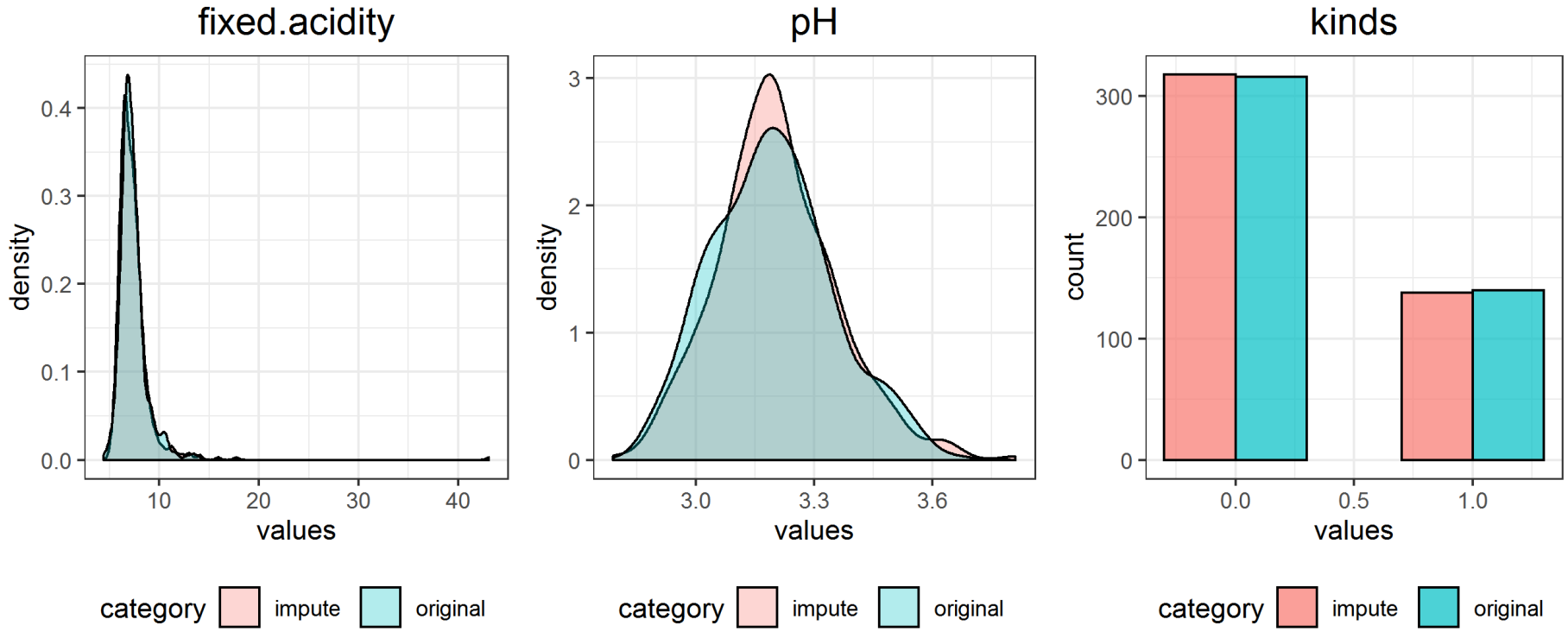
$$P-step{:}\ \ \theta^{(t+1)} \sim P\left(\theta \middle| Y_{(1)}, Y_{(0)}^{(t+1)}\right)$$

https://cran.r-project.org/web/packages/norm/norm.pdf
https://datashoptalk.com/missing-value-imputation-with-data-augmentation-in-r/

Imputation : (2)DA

Distributions for Original vs Inputed index

After 1000 iterative sequences for data Form1-20%



Imputation : (2)DA

Estimated Mean/Variance

After 1000 iterative sequences,

<PH>

	Mean	Variance
Form1-20%	3.2187	0.02611
Form1-35%	3.2194	0.02733
Form1-50%	3.2181	0.02793
Form2-20%	3.2185	0.02599
Form2-35%	3.2179	0.02700
Form2-50%	3.2194	0.02660
Original	3.21850	0.02585

<fixed.acidity>

	Mean	Variance
Form1-20%	7.2202	1.8758
Form1-35%	7.2151	1.6993
Form1-50%	7.2096	1.7288
Form2-20%	7.2117	1.7902
Form2-35%	7.2152	1.6886
Form2-50%	7.2187	2.1957
Original	7.2153	1.6807

<kinds>

	Accuracy
Form1-20%	0.9972
Form1-35%	0.9974
Form1-50%	0.9974
Form2-20%	0.9954
Form2-35%	0.9962
Form2-50%	0.9958

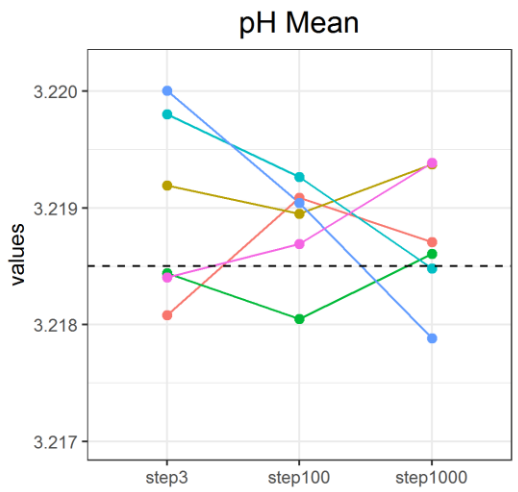
Imputation : (2)DA

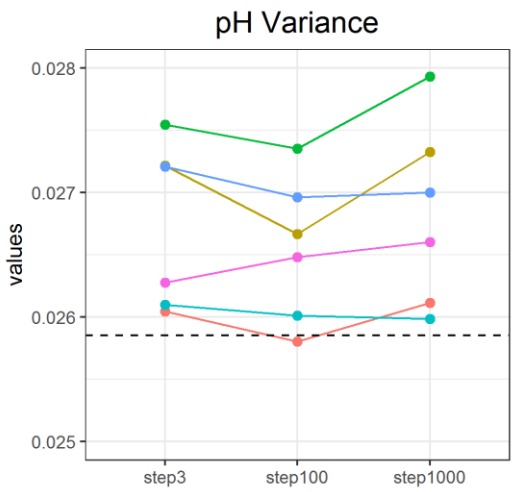
Sample Data

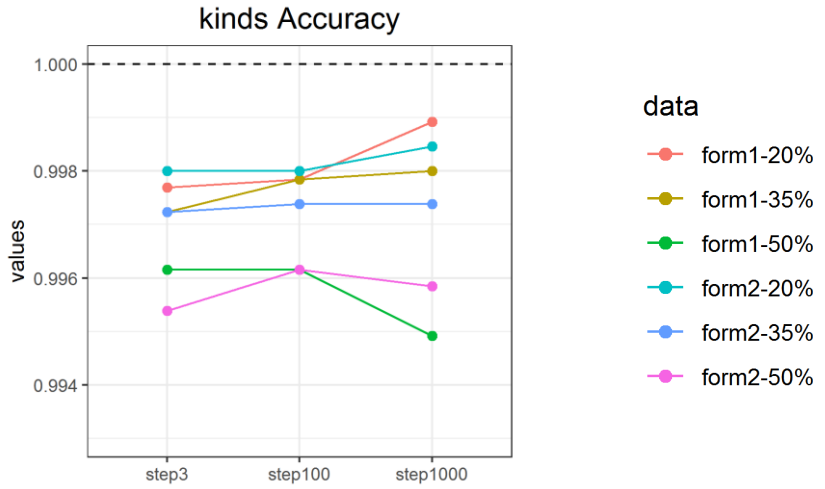
	original.fixed.acidity	impute.fixed.acidity	original.ph	impute.ph	original.kinds	impute.kinds
1	7.5	8.6	3.35	3.35	1	1
2	7.7	8.8	3.25	3.00	1	1
3	8.6	8.3	2.93	3.18	1	1
4	7.8	7.4	3.19	3.24	1	1
5	7.0	7.7	3.34	3.22	1	1
6	7.4	7.4	3.15	3.10	1	1
7	8.9	6.5	3.04	3.10	1	1
8	8.7	7.8	3.34	3.31	1	1
9	10.3	10.9	3.12	3.14	1	1
10	10.3	8.9	3.12	3.44	1	1

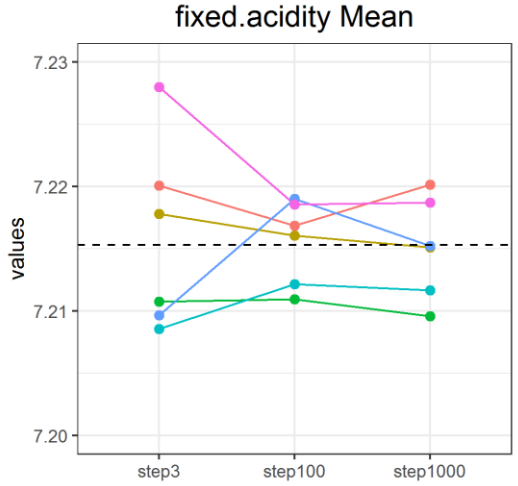
Imputation : (2)DA

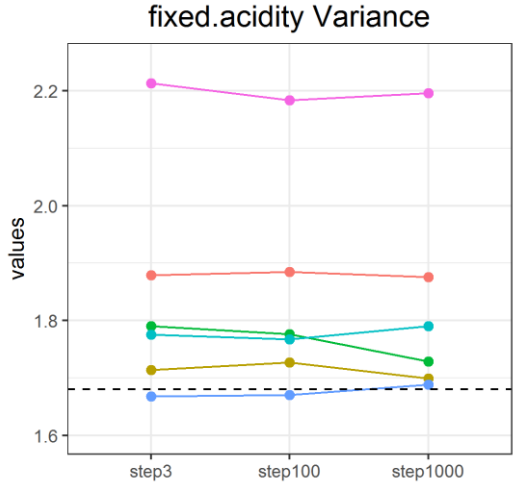
Compare number of sequences: 3, 100, 1000











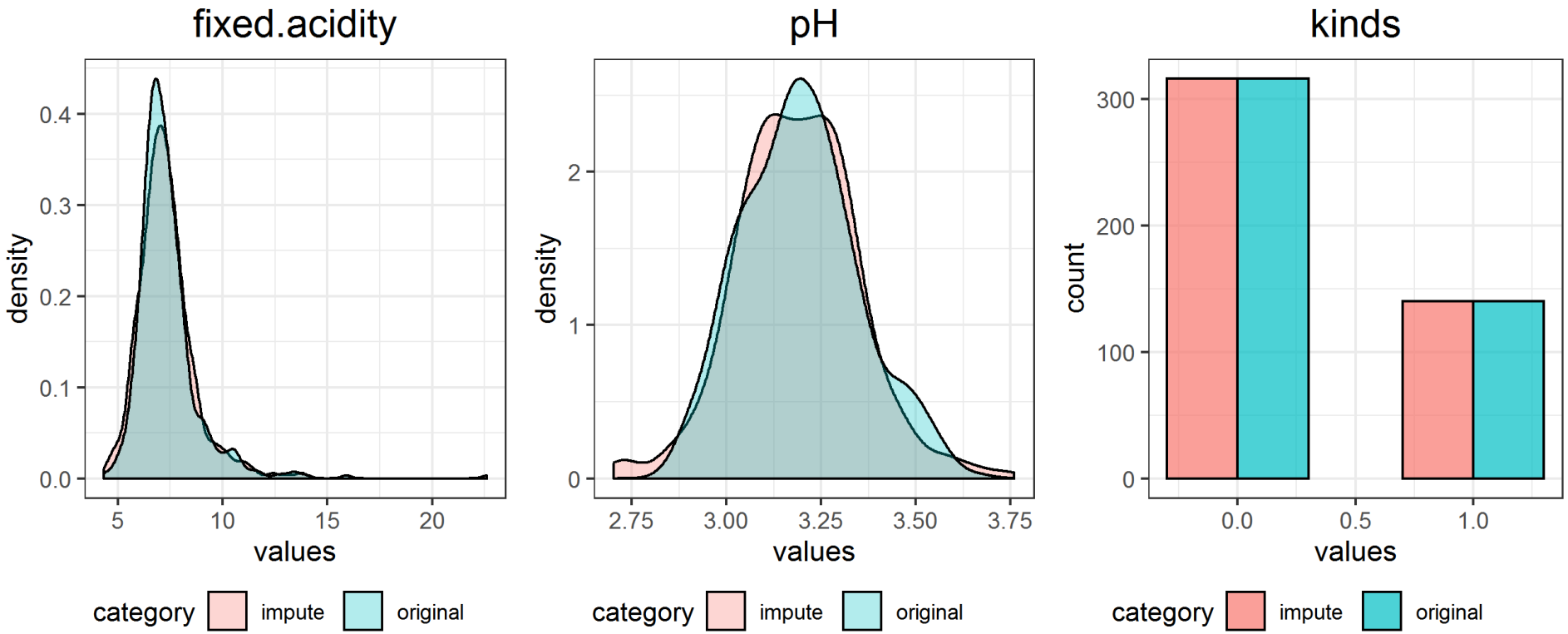
각 데이터에서 sequence에 따라 추정값의 유의한 차이를 보인다고 볼 수 없었다.

한편, 평균의 경우 각 데이터에 따라 유의한 차이를 보이지 않았지만, 분산의 경우에는 결측을 50% 발생시킨 두 데이터의 분산이 높게 추정된 편이며, 범주형 변수의 정확도는 비교적 낮은 편이다.

Imputation : (2)DA

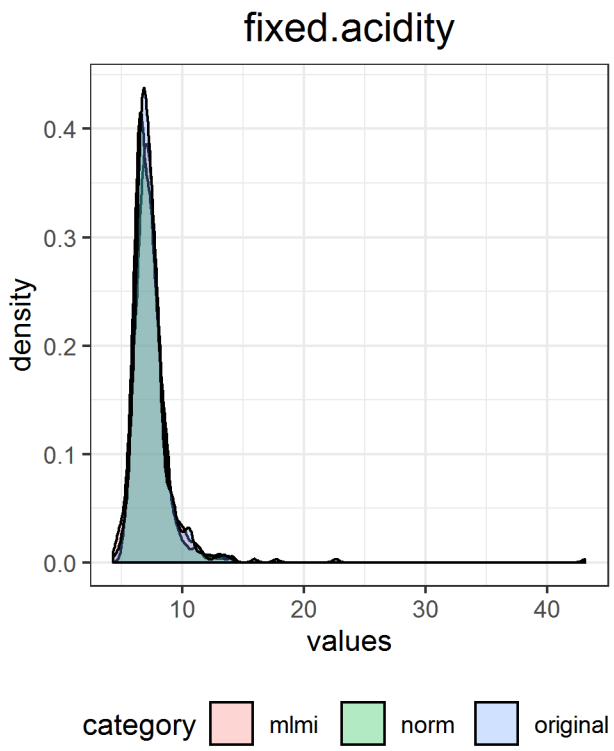
Try2) DA using General Location Model

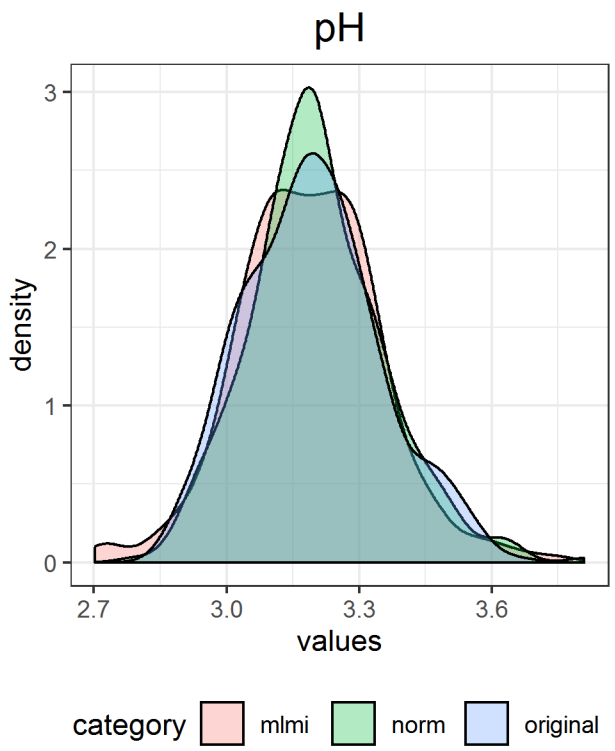
Library used: mlmi

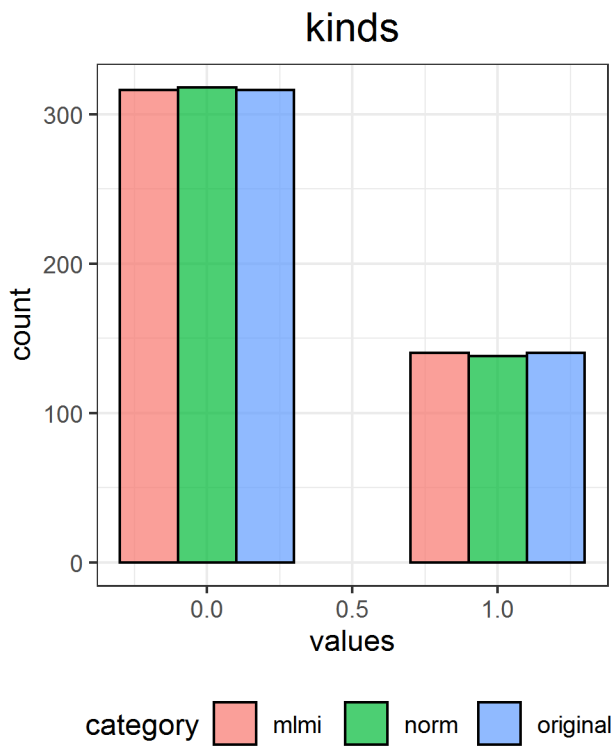


Imputation : (2)DA

Distributions for Original vs Amelia Inputed vs mlmi Imputed index

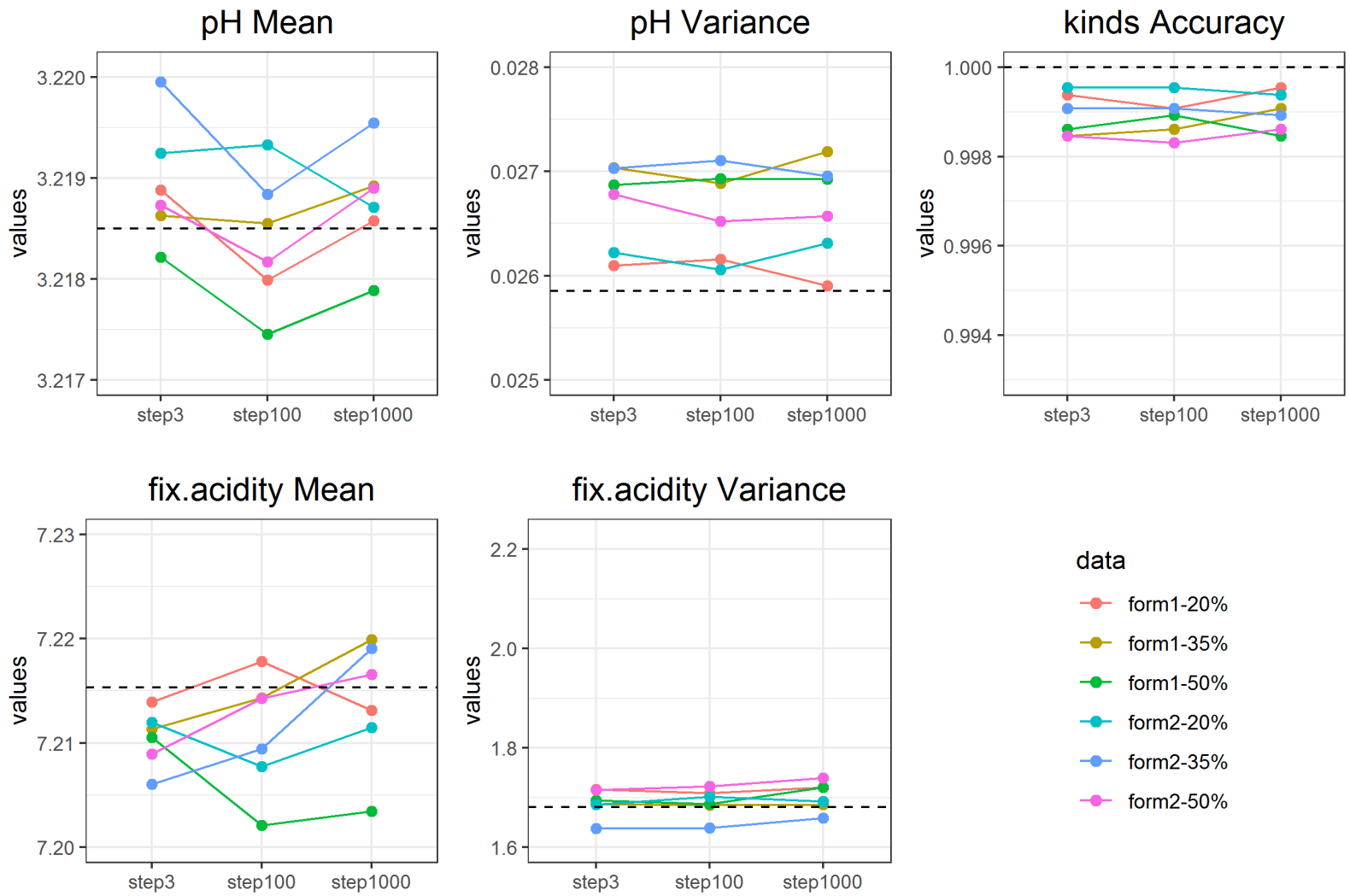






Imputation : (2)DA

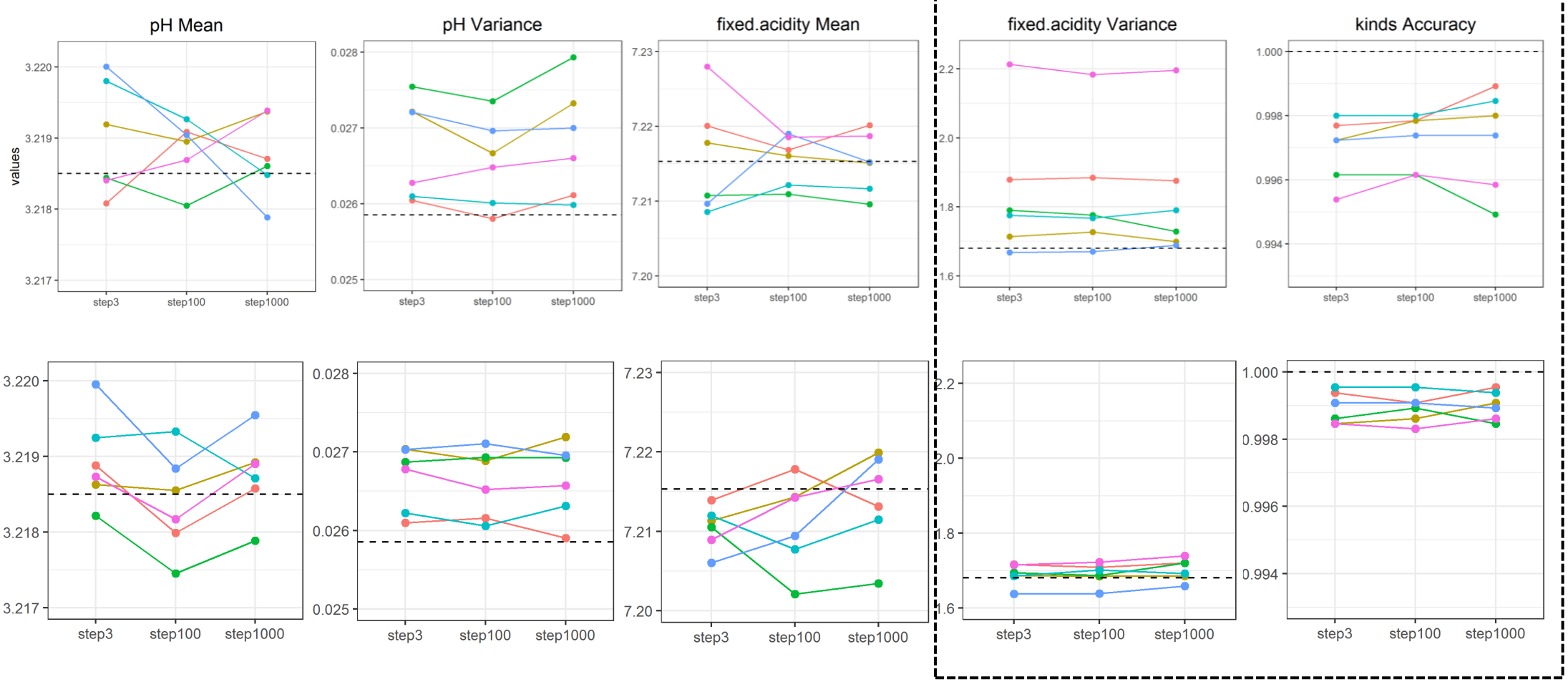
Compare number of sequences: 3, 100, 1000



Imputation : (2)DA

Compare NORM and mlmi for the number of sequences,

library **NORM**



library **mlmi**

➡ 연속형 변수들의 평균은 두 경우의 차이가 거의 없지만, 분산과 범주형 변수의 정확도에서 차이를 보이고 있다.
특히, fixed.acidity의 분산 추정이 모든 데이터에서 안정화되었고, kinds의 정확도는 증가했다.

Imputation : (3)Chain-Equation MI

Library used: MICE

Missing data에 대한 multiple imputation을 지원하는 패키지

Usage

mice(data, m, method, seed)

- data: missing(NA) 을 포함하고 있는 데이터 input
- m: MI를 몇 번 시행할 것인지를 설정하는 값. 만약 m=3 이면 MI를 3번 시행
- method: input data의 각 변수에 어떤 imputation 방법을 적용할 것인지를 설정. 만약 imputation이 필요 없는 변수라면 “” 으로 설정한다.

Mice steps

Step 1 : Simple imputation을 이용해 모든 missing value 값을 임의의 값으로 대체한다.

Step 2 : Missing이 발생한 변수 중 첫번째 변수의 대체된 임의의 값을 다시 missing으로 만든다.

Step 3 : 첫번째 변수를 제외한 나머지 변수들을 이용해서 method에 지정된 imputation 방법을 적용해 첫번째 변수의 missing 값을 impute 한다.
→ Step 3가 끝나면 첫번째 변수에는 missing 없음!

Step 2와 Step 3를 나머지 결측이 존재하는 변수들에 대해서 모두 적용한다.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

Imputation : (3)Chain-Equation MI

How to apply MICE

1. MICE Parameters

(1) Input data: form1(20%,35%,50%), form2(20%,35%,50%) → 6 dataset

(2) Imputation method:

fixed.acidity (연속형)	pH (연속형)	Kinds (범주형)
Stochastic regression	Stochastic regression	Logistic regression

(3) MI 횟수: m=5

2. Imputation

(1) Mice package 에 있는 complete 함수를 이용해 imputed set 확보 후 연속형 변수는 평균값, 범주형 변수는 최빈값으로 대체한다.

(2) Missing data를 모두 imputation 한 후, 기존의 데이터 형식과 동일하게 맞춘다.

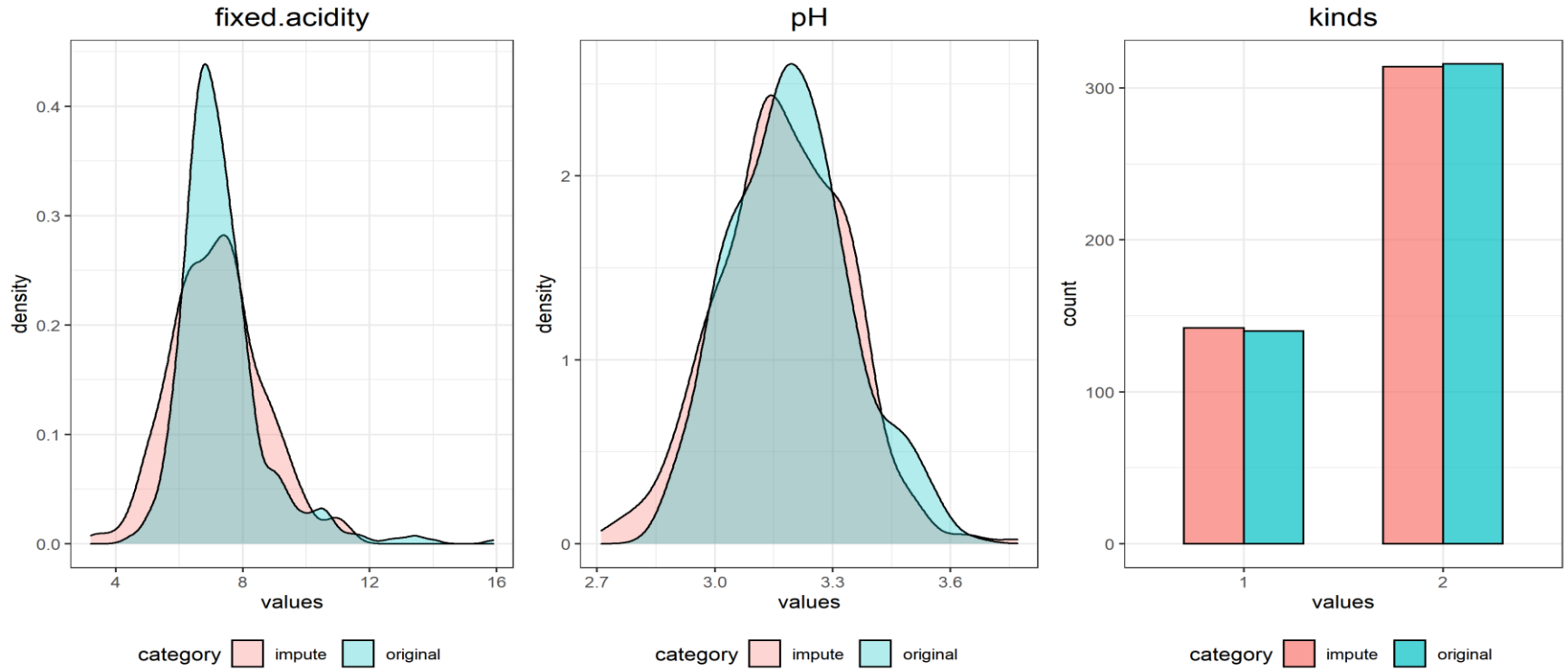
Sample data

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	fixed.acidity	volatile.acid	citric.acid	residual.sug	chlorides	free.sulfur.d	total.sulfur.c	density	pH	sulphates	alcohol	quality	kinds
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	1
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5	1
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	1
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	1
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	1
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	1
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5	1

Imputation : (3)Chain-Equation MI

Distributions for Original vs Inputed index

For data Form1-20%



Imputation : (3)Chain-Equation MI

Estimated Mean/Variance

<fixed.acidity>

	Mean	Variance
Form1-20%	7.20608	1.69526
Form1-35%	7.21165	1.70052
Form1-50%	7.19691	1.63235
Form2-20%	7.21673	1.64431
Form2-35%	7.20965	1.60196
Form2-50%	7.21141	1.65721
Original	7.21531	1.68074

<pH>

	Mean	Variance
Form1-20%	3.21727	0.02609
Form1-35%	3.21823	0.02599
Form1-50%	3.21804	0.02601
Form2-20%	3.21933	0.02620
Form2-35%	3.21724	0.02597
Form2-50%	3.21818	0.02642
Original	3.21850	0.02585

<kinds>

	Accuracy
Form1-20%	0.998
Form1-35%	0.997
Form1-50%	0.994
Form2-20%	0.997
Form2-35%	0.995
Form2-50%	0.994

Imputation : (3)Chain-Equation MI

Analysis of multiply imputed dataset

For data Form1-20%

Estimated value	Fixed.acidity	pH
Mean	7.2105	3.2175
Within imputation variance(wd)	1.6872	0.0260
Between imputation variance(bd)	0.0000	0.0000
Total variance(td)	1.6873	0.0260

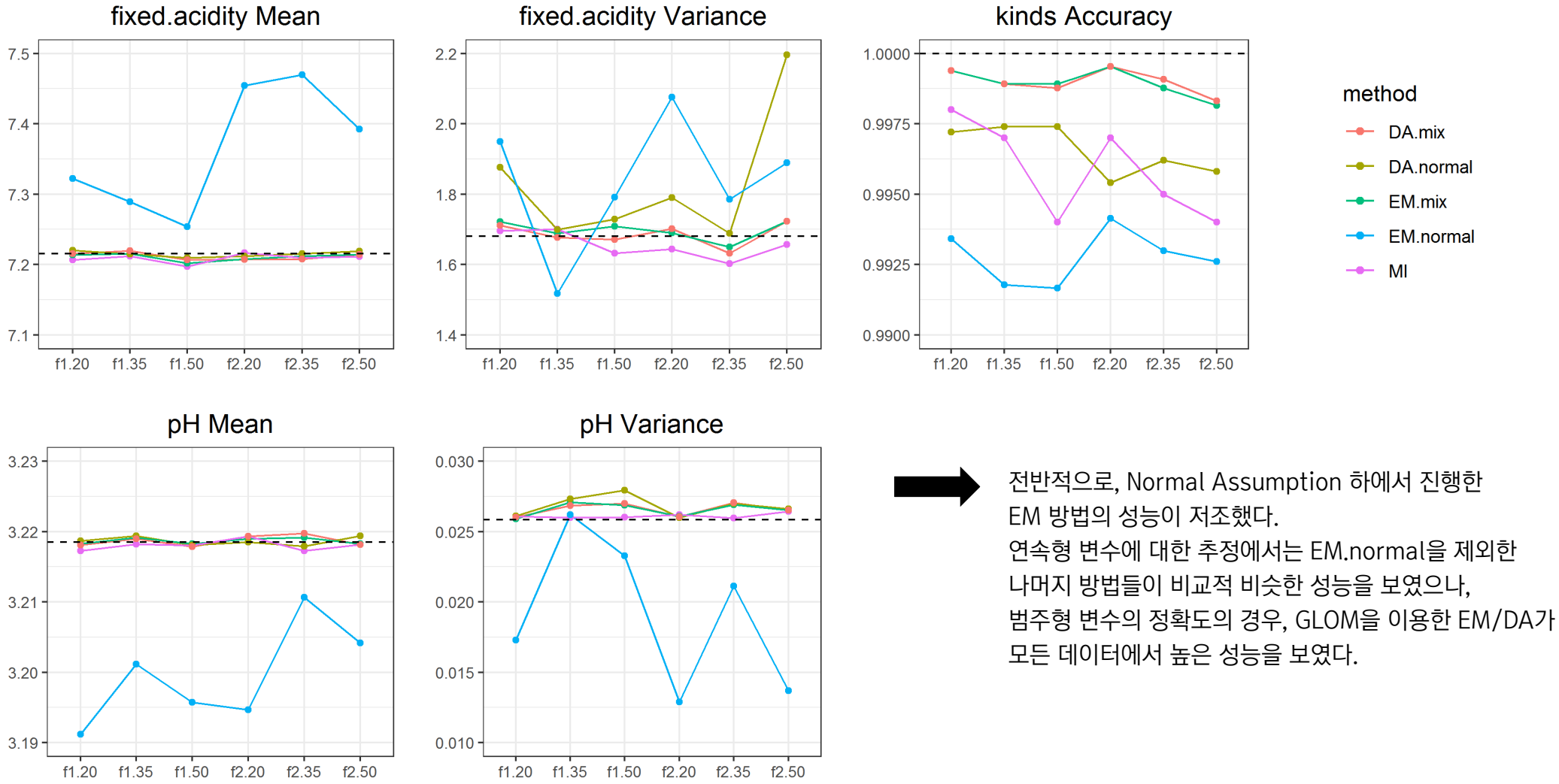
Between Variance가 0에 수렴

→ 해당 데이터에서는 Imputed estimate 간 Variability가 크지 않았다.

→ 이에 대한 이유로는, 결측 변수들을 제외한 나머지 변수들의 설명력이 높았으며,
연속형 값들의 단위도 비교적 낮아(일의 자리 수 이하), 계산된 Variance의 값이 크지 않았던 것으로 판단된다.
또 한, 거듭된 시퀀스로 인해 추정치들 간 Variability가 감소한 것이라고 짐작할 수 있었다.

Conclusion

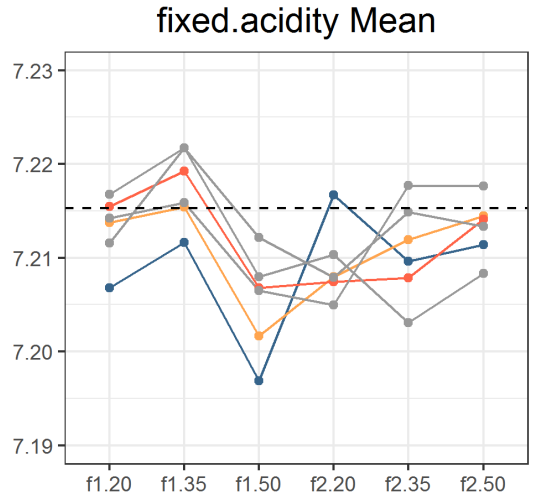
Compare all the methods

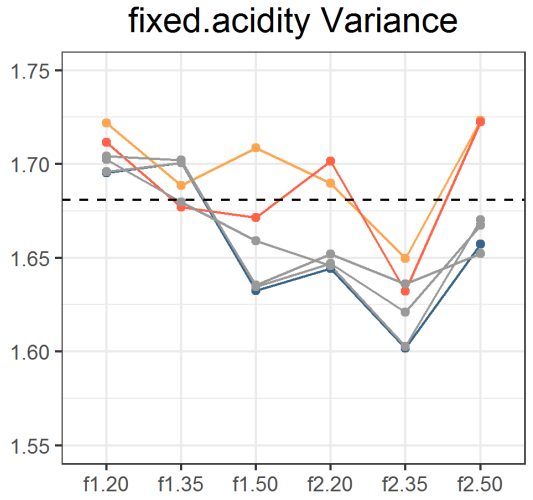


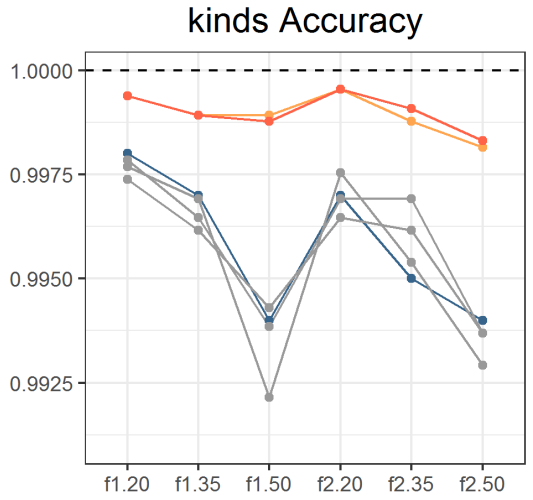
전반적으로, Normal Assumption 하에서 진행한
EM 방법의 성능이 저조했다.
연속형 변수에 대한 추정에서는 EM.normal을 제외한
나머지 방법들이 비교적 비슷한 성능을 보였으나,
범주형 변수의 정확도의 경우, GLOM을 이용한 EM/DA가
모든 데이터에서 높은 성능을 보였다.

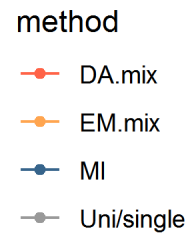
Conclusion

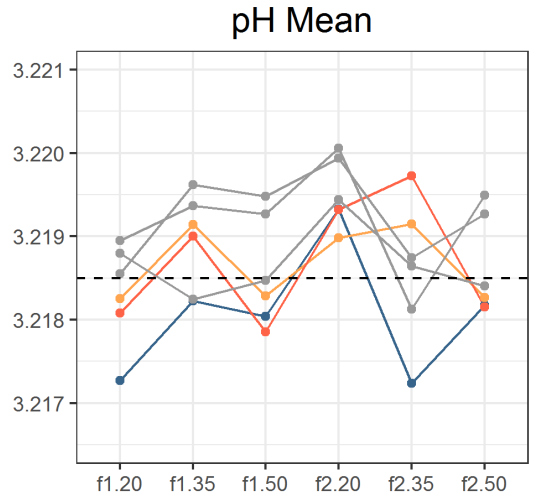
Compare all the methods: Control group – Single/Univariate Stochastic Regression

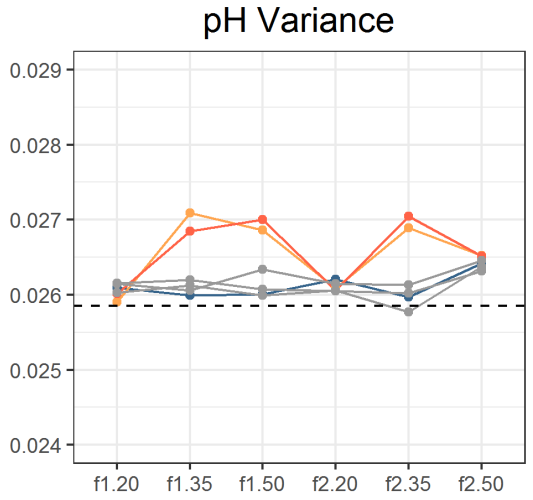












GLOM을 이용한 EM, DA 방법과 MI 방법, 그리고
3번의 Single Stochastic Regression의 결과이다.
4가지 방법 모두 평균, 분산, 정확도의 높은 성능을 보이며,
특히 GLOM을 이용한 DA/EM이 범주형 변수에서
더 높은 정확도를 보인다.

Conclusion

- 해당 데이터에서 결측이 발생한 변수들에는 연속형, 범주형이 섞여 있으며, Normal assumption을 만족하지 않았다.
  이 때, mlmi 패키지의 mixImpute을 이용한 방법이 본 프로젝트에서 전반적으로 우수한 성능을 보였으며,
  특히 범주형 변수의 정확도가 모든 6개의 데이터셋에서 가장 높았다.

- 하지만, 사용했던 모든 방법이 성능이 대체로 좋았기 때문에, 방법들 간 극명한 차이가 존재한다고 보기는 어려웠다.
  결측비율과 패턴에 따라 6가지의 다양한 결측 데이터셋을 생성했음에도 불구하고, 모든 데이터에서 대체로 성능이 높았다.
  이 다음에는 Outlier가 존재하거나 더욱 복잡한 데이터에 대해서도 Evaluation을 해보고 싶다는 생각을 하게 되었다.

- 그럼에도, 연속형, 범주형 변수 모두에 결측이 발생한 상황에서, Multivariate Handling을 할 수 있는 방법들을 충분히 고민하고,
  이러한 방법들이 6가지 종류의 데이터셋에 대해서 일관적으로 우수한 성능을 보였다는 점에서 프로젝트의 의의를 찾을 수 있었다.

감사합니다

Appendix

Single imputation results summary

<fixed.acidity>

	Mean	Variance
Form1-20%	7.214001	1.691999
Form1-35%	7.213864	1.701972
Form1-50%	7.213506	1.647201
Form2-20%	7.207874	1.666728
Form2-35%	7.213618	1.608692
Form2-50%	7.214793	1.68218

<pH>

	Mean	Variance
Form1-20%	3.2179	0.02631709
Form1-35%	3.218774	0.02615364
Form1-50%	3.219278	0.02608328
Form2-20%	3.219659	0.02605762
Form2-35%	3.219287	0.02601558
Form2-50%	3.218478	0.02636923

<kinds>

	Accuracy
Form1-20%	0.998
Form1-35%	0.997
Form1-50%	0.995
Form2-20%	0.997
Form2-35%	0.996
Form2-50%	0.993

Appendix

Amelia Steps

Step 1

명목형 변수: p개의 변수 가정 → p-1 개의 dummy variable 형성 → p-1 개의 값을 continuous값으로 impute
→ p-1 개의 확률로 변환 (실질적으로 p개의 확률) → 확률에 맞춘 draw 로 변수 impute

연속형 변수: 정규분포 형태가 될 수 있게끔 변환 (option 으로 설정)

Step 2

총 m번의 EM 알고리즘 시행.

Step 3

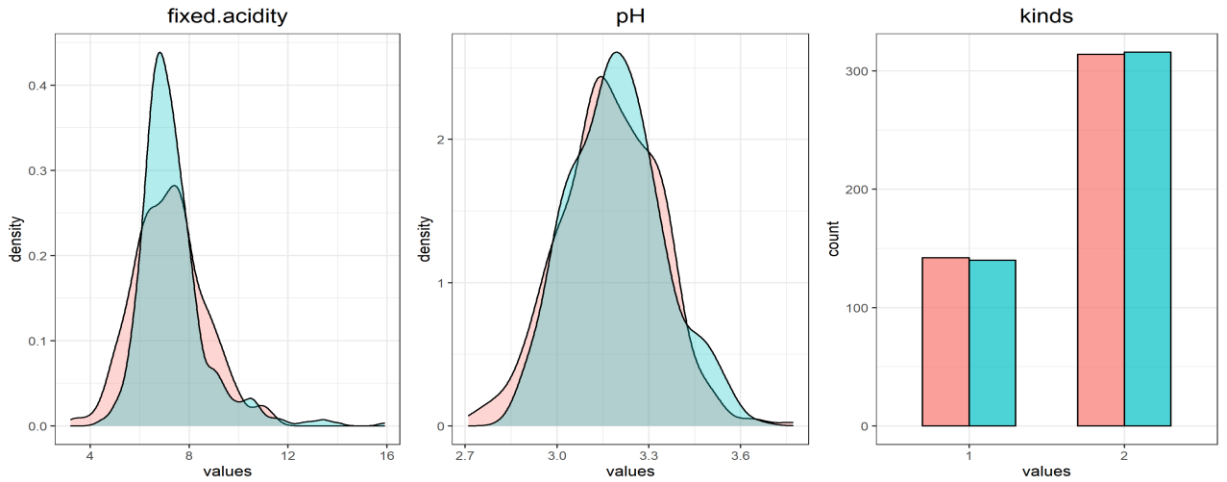
m번 시행의 평균값을 각 missing data에 impute

https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

Appendix

Multiple vs Single Imputation

Multiple Imputation (m=5)



Single Imputation (m=1)