Machine Learning with Human Intelligence:

Principled Corner Cutting (PC^2)

Xiao-Li Meng

Harvard University

joint work with Thomas C. M. Lee of University of California at Davis &

Zhan Li of Harvard University

Machine Learning v.s. Statistics

Machine Learning v.s. Statistics

- Shared Grand Task: Separating signal from noise

Machine Learning v.s. Statistics

- Shared Grand Task: Separating signal from noise
- Stereotypical complaint about statisticians:
  Excessive worries over modeling and inferential principles, to a degree of being willing to produce nothing

Machine Learning v.s. Statistics

- Shared Grand Task: Separating signal from noise
- Stereotypical complaint about statisticians:
  Excessive worries over modeling and inferential principles, to a degree of being willing to produce nothing
- Stereotypical complaint about machine learners:
  Strong tendency to let ease of implementation or good performance trump principled justifications, to a point of being willing to deliver anything

Principled Corner Cutting (PC^2)

Principled Corner Cutting (PC^2)

- ***Principle Oriented*** v.s. ***Performance Oriented***

Principled Corner Cutting (PC^2)

- ***Principle Oriented*** v.s. ***Performance Oriented***
- We need BOTH in order to reach a sensible compromise between statistical efficiency and computational efficiency

Principled Corner Cutting (PC^2)

- ***Principle Oriented*** v.s. ***Performance Oriented***
- We need BOTH in order to reach a sensible compromise between statistical efficiency and computational efficiency
- We need to train more ***Principled Corner Cutters*****:** Who can formulate the solution from the soundest principles available but are at ease of cutting corners guided by these principles, to achieve as much statistical efficiency as feasible while maintaining computational efficiency under time and resource constraints.

Let's start with a simple illustration

Let's start with a simple illustration

- Mr. Littlestat was given a **black box** which computes the Least Squares Estimate (LSE) of β for the linear regression

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n, \quad \epsilon_i \; i.i.d. \sim F[0,1].$$

Let’s start with a simple illustration

- Mr. Littlestat was given a **black box** which computes the Least Squares Estimate (LSE) of β for the linear regression

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \; i.i.d. \sim F[0, 1].$$

- And it only works when $n = 2^4 = 16$, outputting

$$\hat{\beta}_{16}(y_1, \ldots, y_{16}) = \frac{\sum_{i=1}^{16} y_i x_i}{\sum_{i=1}^{16} x_i^2}. \qquad (A)$$

Let's start with a simple illustration

- Mr. Littlestat was given a **black box** which computes the Least Squares Estimate (LSE) of β for the linear regression

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \; i.i.d. \sim F[0, 1].$$

- And it only works when $n = 2^4 = 16$, outputting

$$\hat{\beta}_{16}(y_1, \ldots, y_{16}) = \frac{\sum_{i=1}^{16} y_i x_i}{\sum_{i=1}^{16} x_i^2}. \qquad (A)$$

- But Mr. Littlestat only has $n = 13$. Can he still use the same program?

Is it possible?

Is it possible?

- Is it possible to use the **black box** designed for LSE with $n = 16$ to compute the LSE **exactly** with $n = 13$?

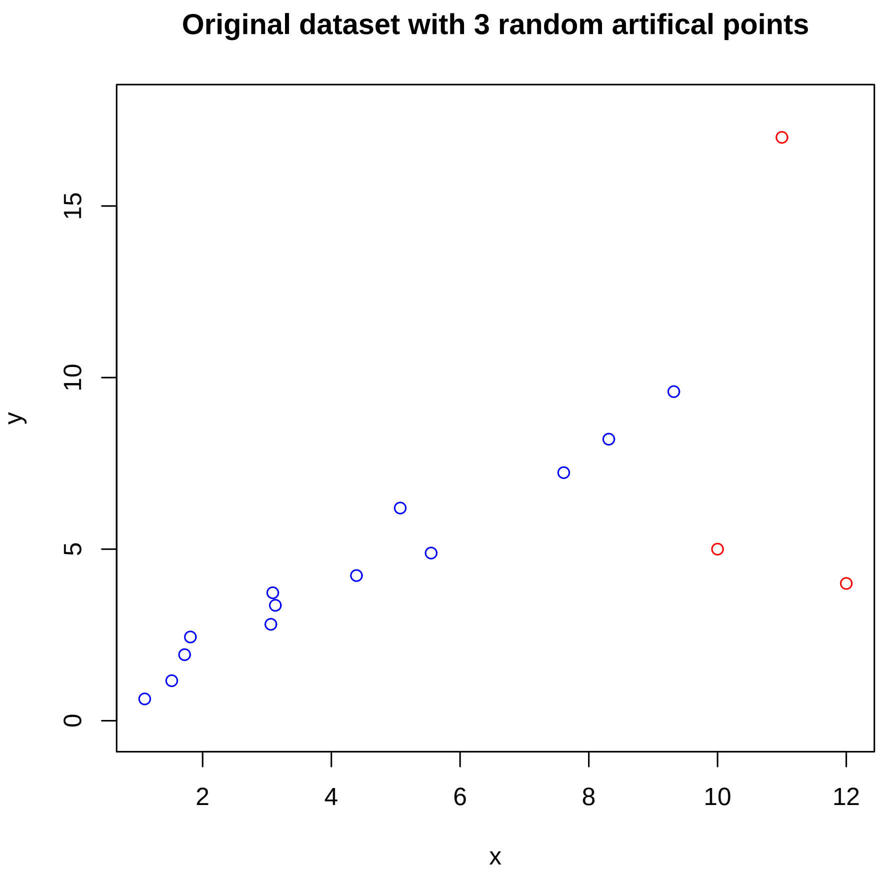
Is it possible?

- Is it possible to use the **black box** designed for LSE with $n = 16$ to compute the LSE **exactly** with $n = 13$?
- The answer has to be **YES** because ...

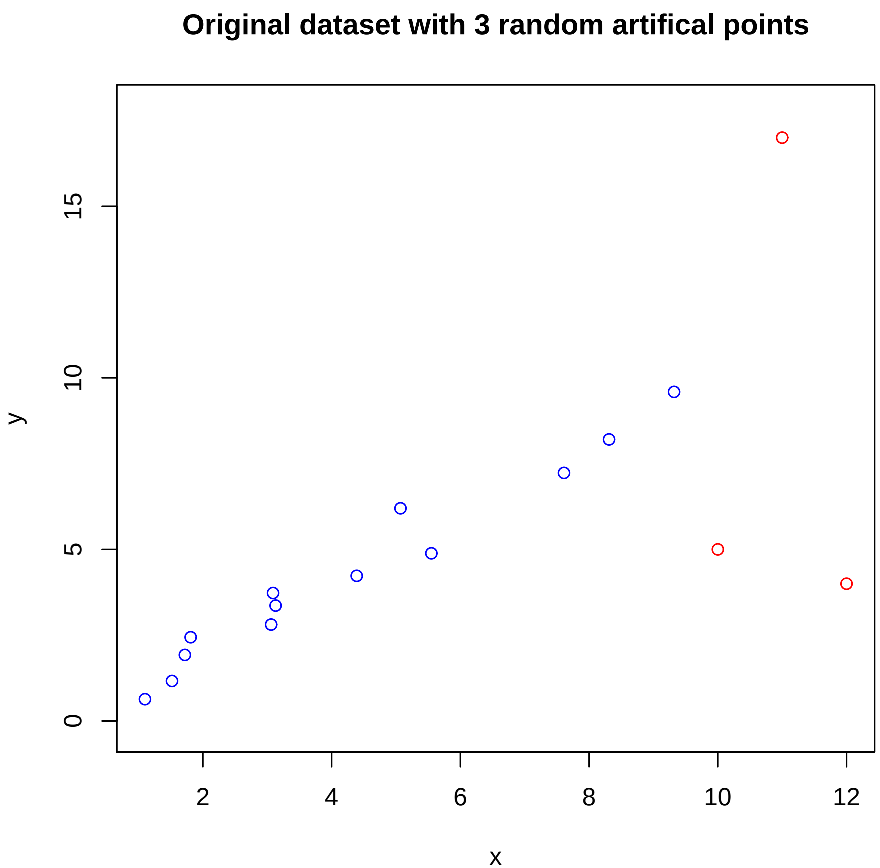
Is it possible?

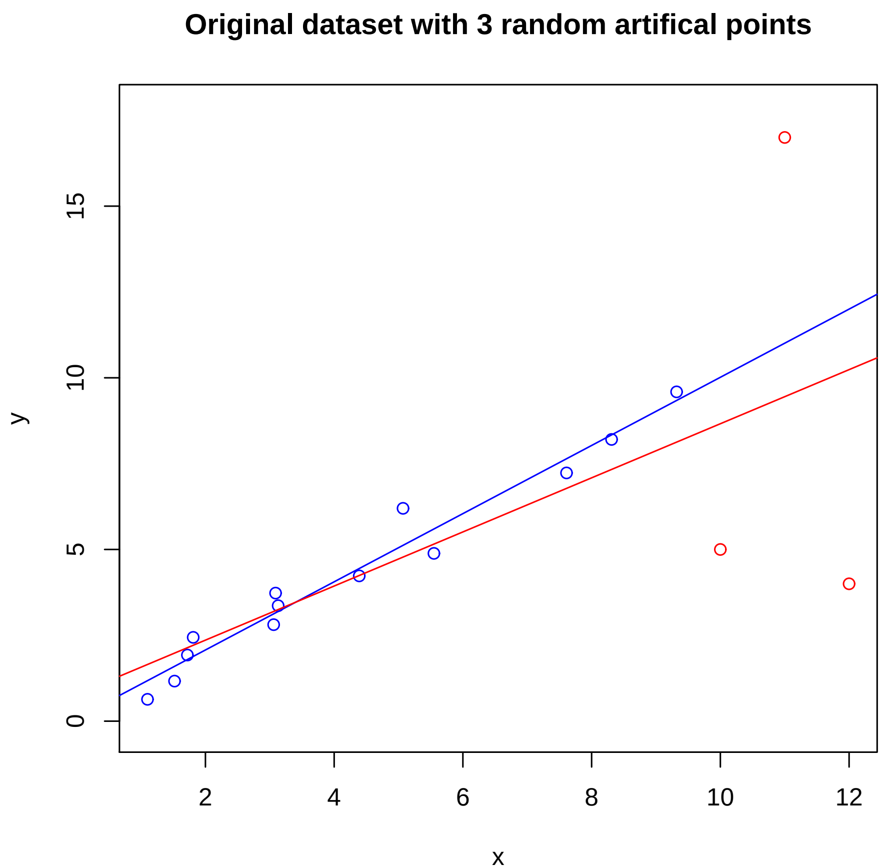
- Is it possible to use the **black box** designed for LSE with $n = 16$ to compute the LSE **exactly** with $n = 13$?
- The answer has to be **YES** because ...
- The ***Principle of Selection Bias***!

A Numerical Illustration

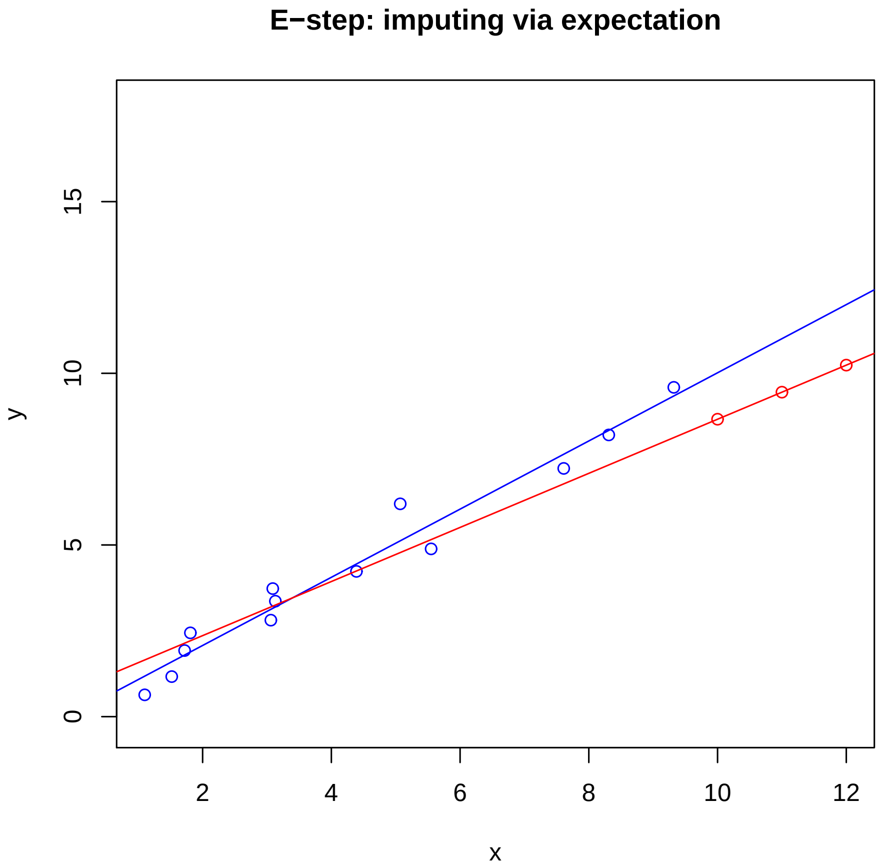


A Numerical Illustration

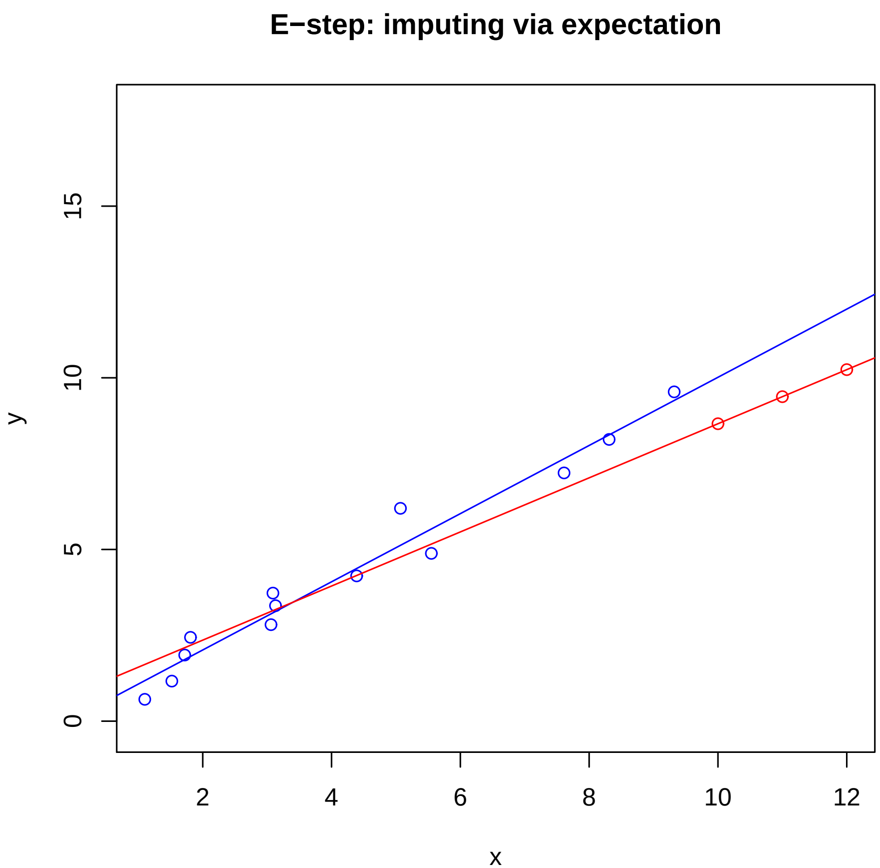




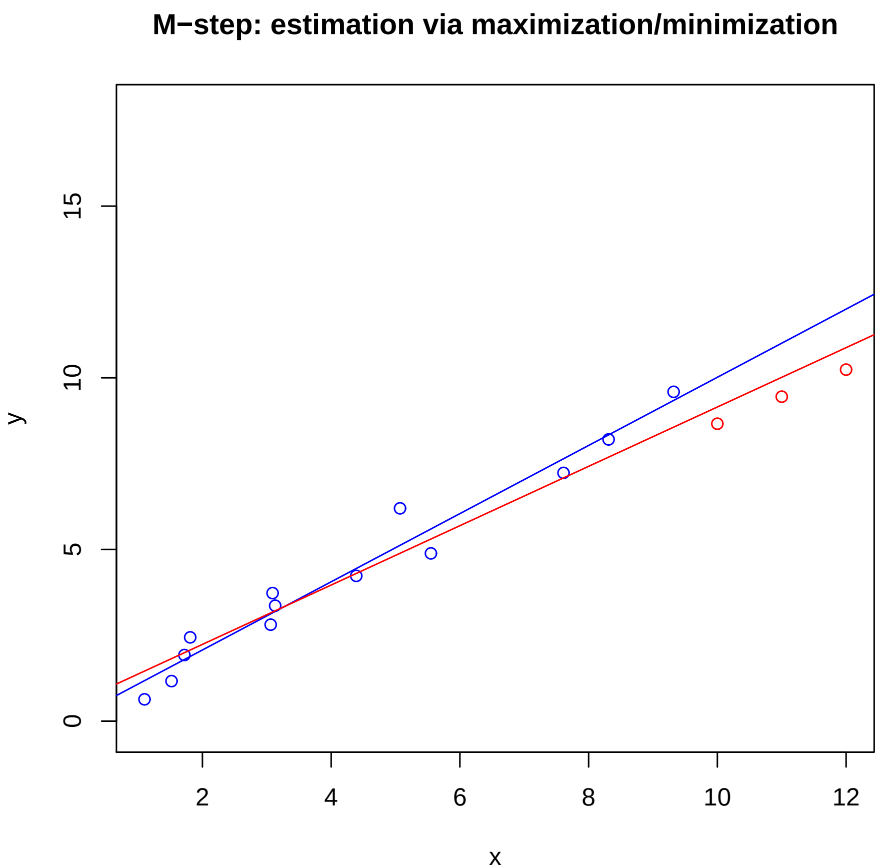
Iteration 1



Iteration 1

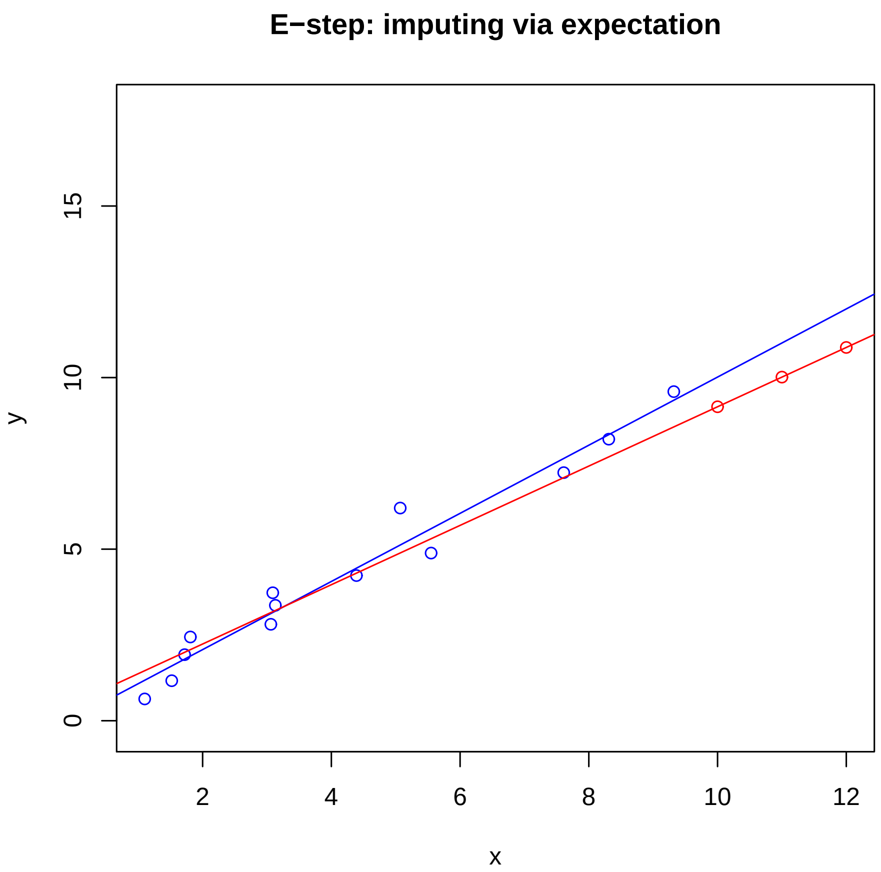


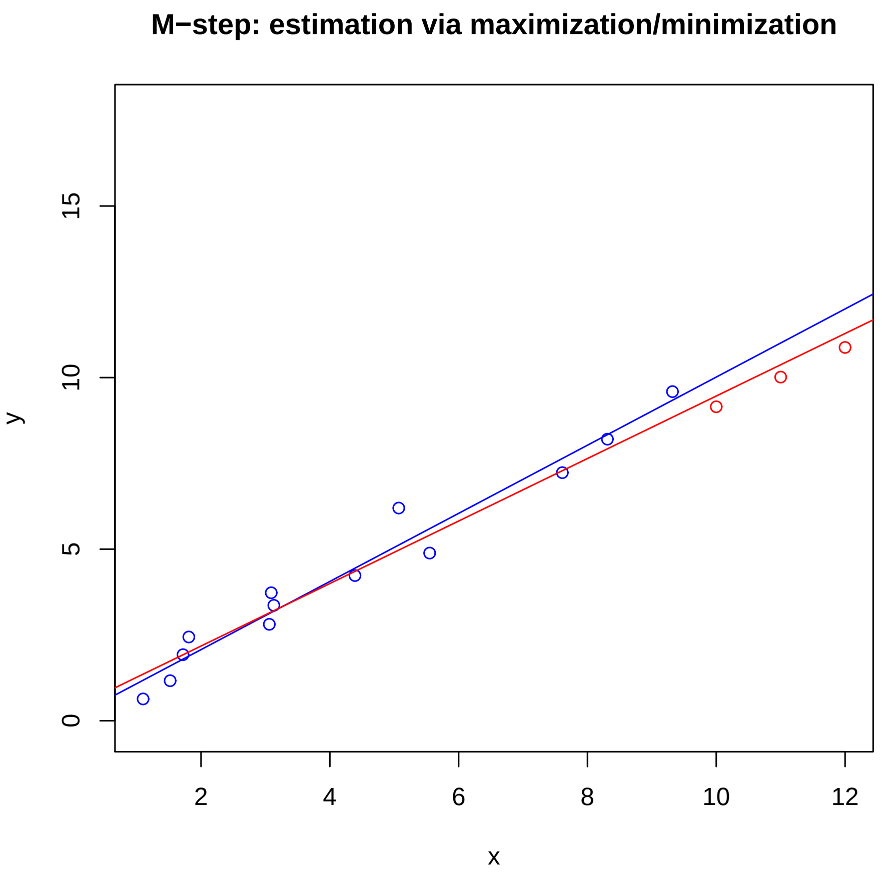
E−step: imputing via expectation
y
x
0
5
10
15
2
4
6
8
10
12



M−step: estimation via maximization/minimization
y
x
0
5
10
15
2
4
6
8
10
12

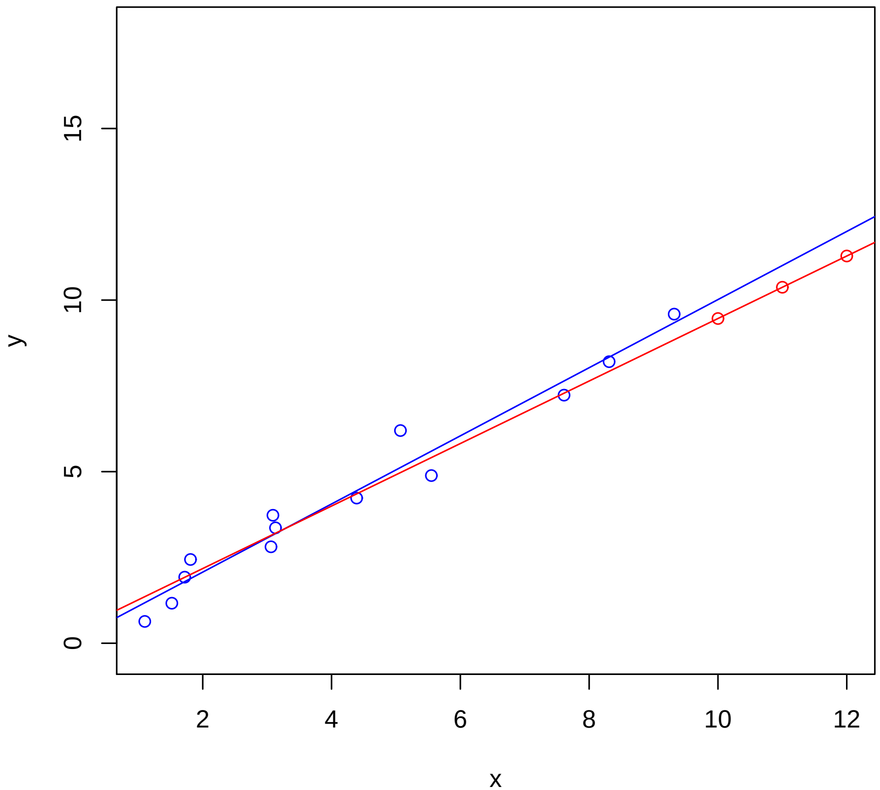
Iteration 2


E−step: imputing via expectation
y
x
0
5
10
15
2
4
6
8
10
12

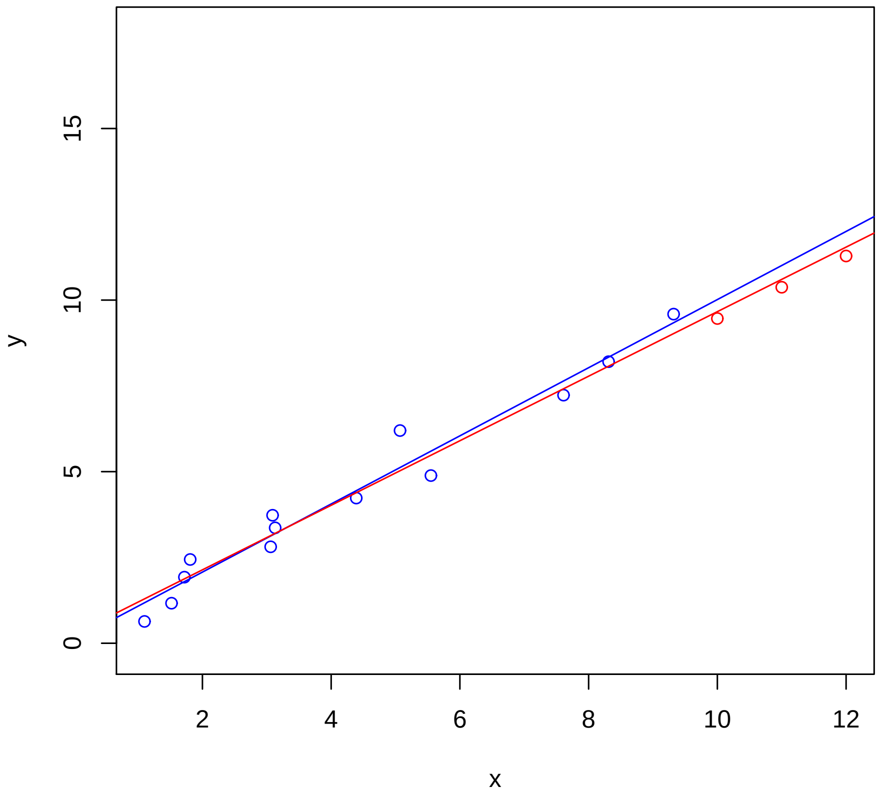

M−step: estimation via maximization/minimization
y
x
0
5
10
15
2
4
6
8
10
12

Iteration 3

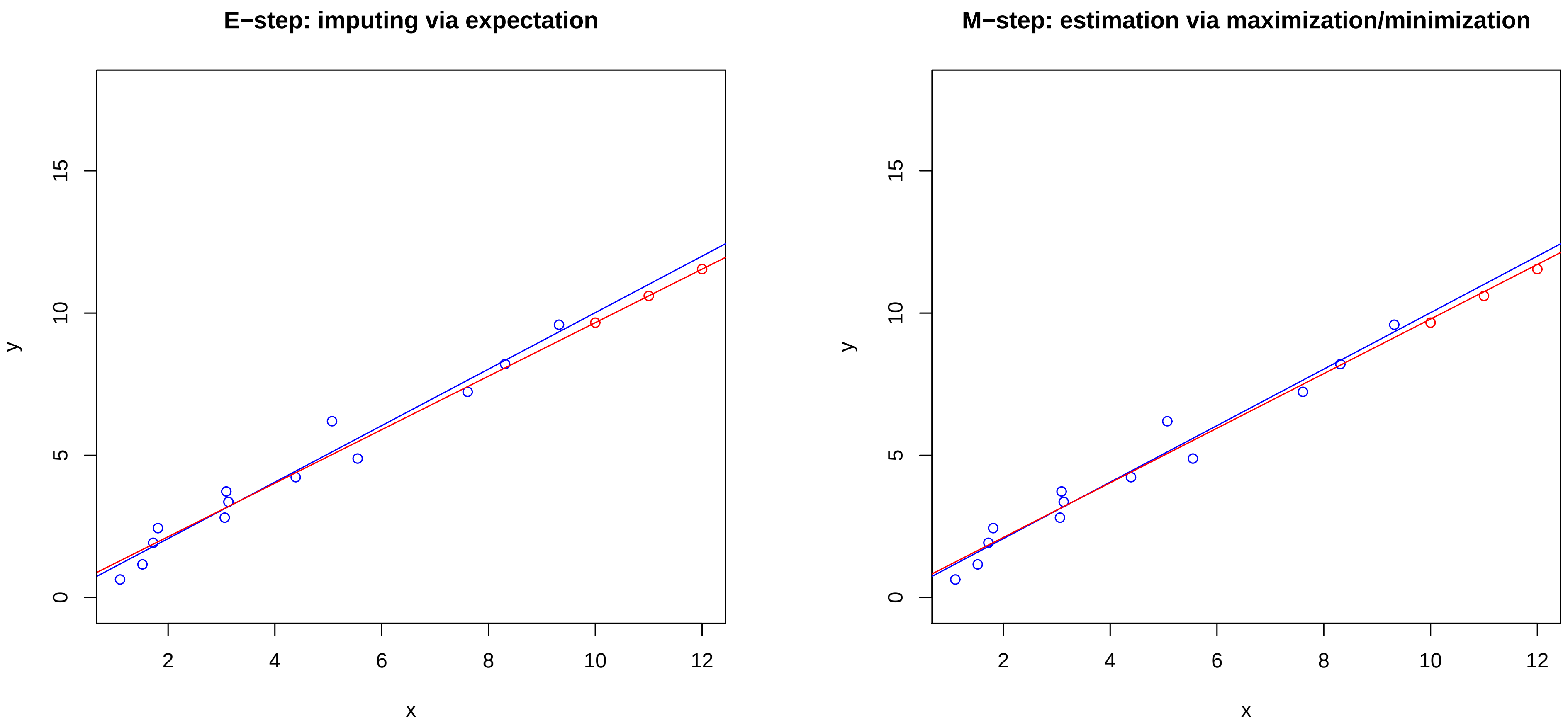
E−step: imputing via expectation



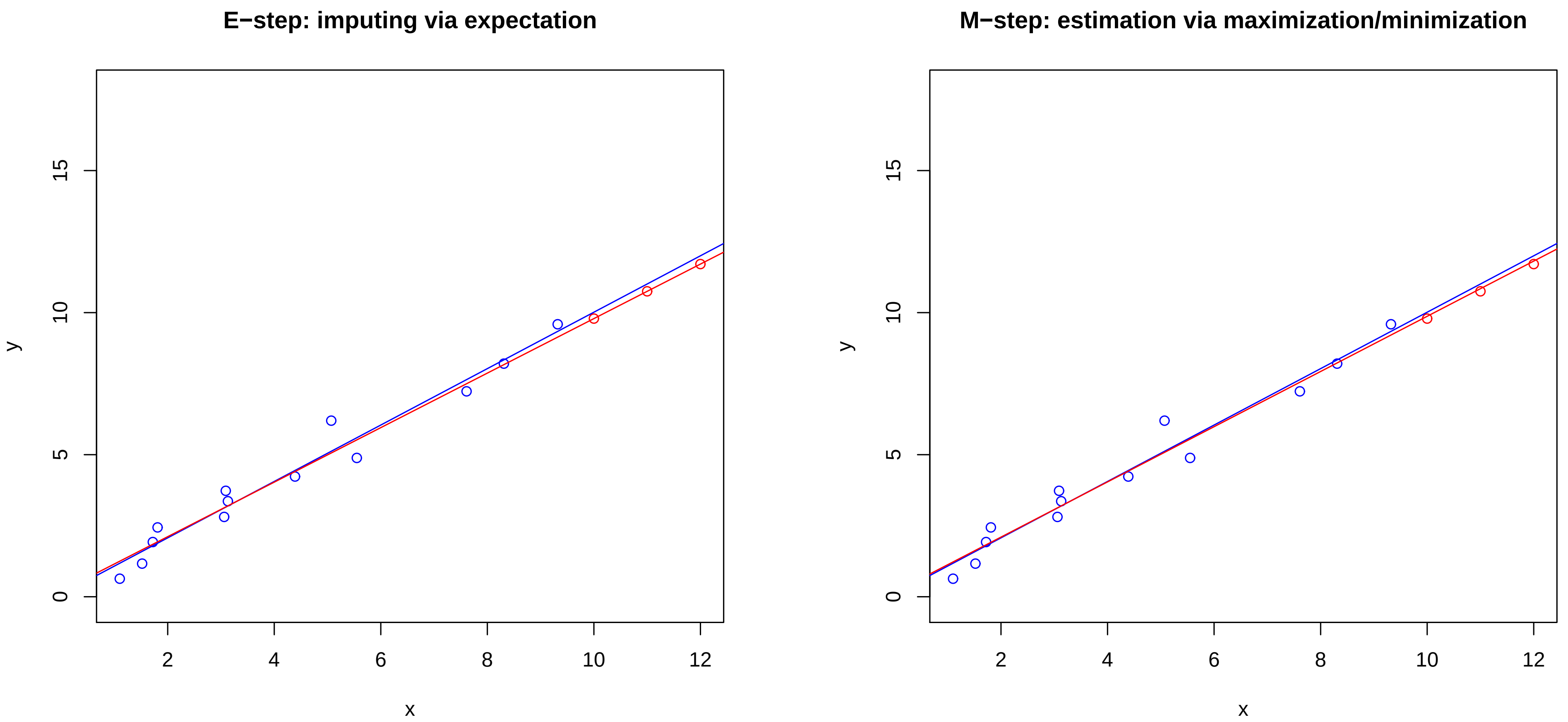
M−step: estimation via maximization/minimization



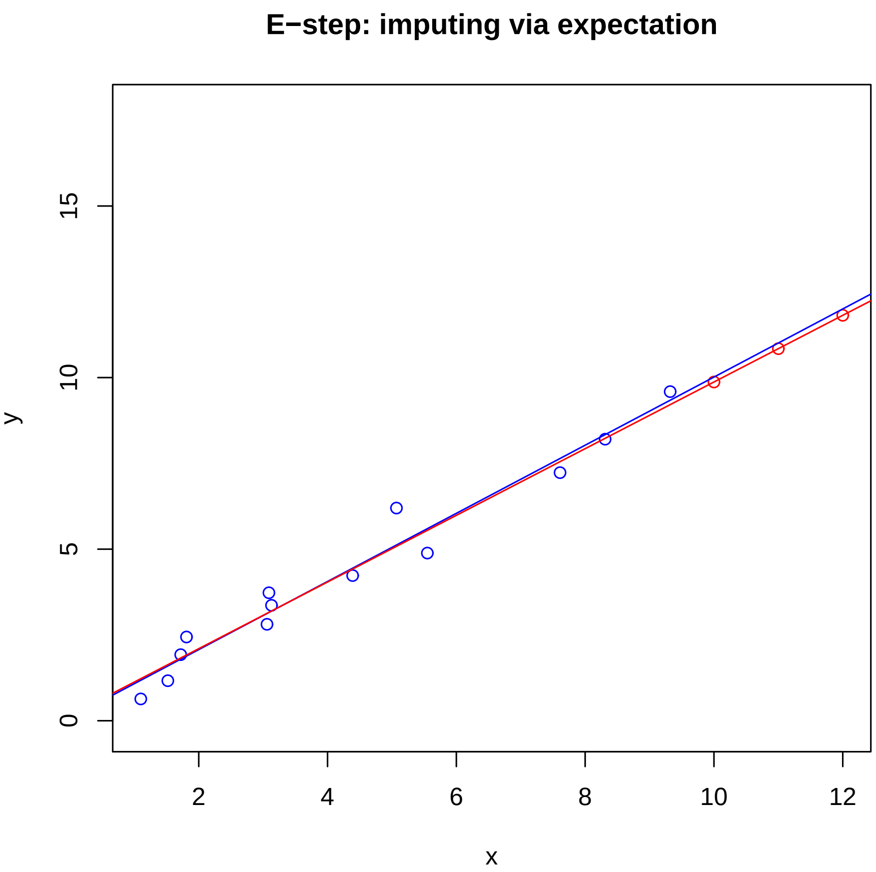
Iteration 4


E−step: imputing via expectation
M−step: estimation via maximization/minimization
y
x
0
5
10
15
2
4
6
8
10
12

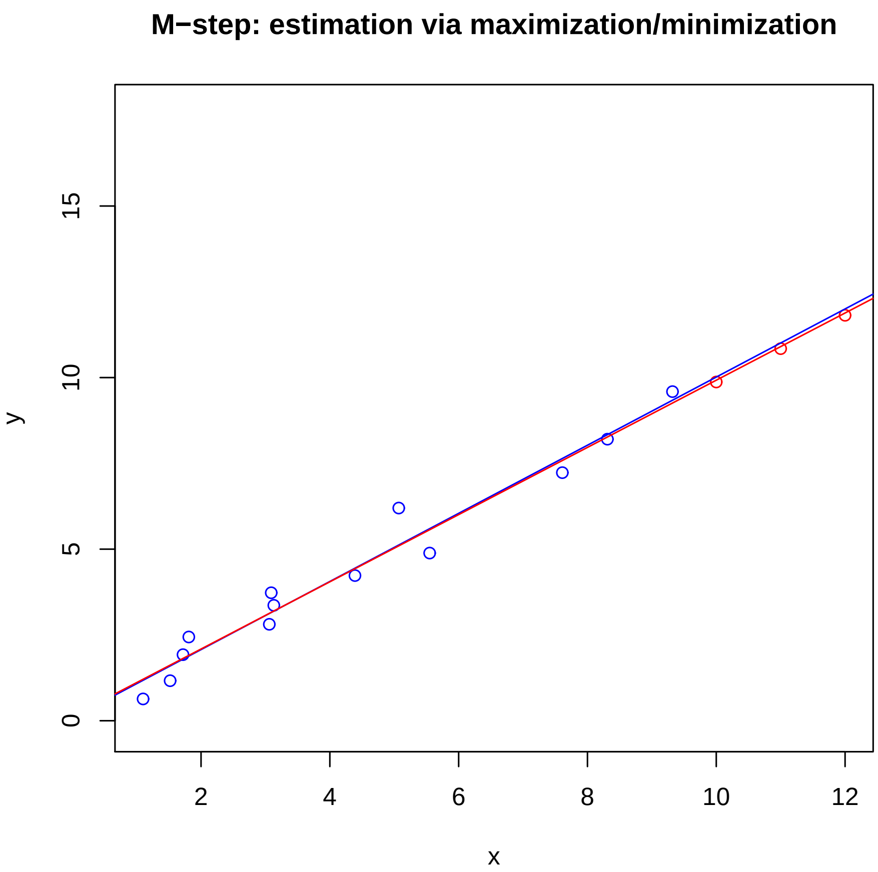
Iteration 5


E−step: imputing via expectation
M−step: estimation via maximization/minimization
y
x
0
5
10
15
2
4
6
8
10
12

Iteration 6

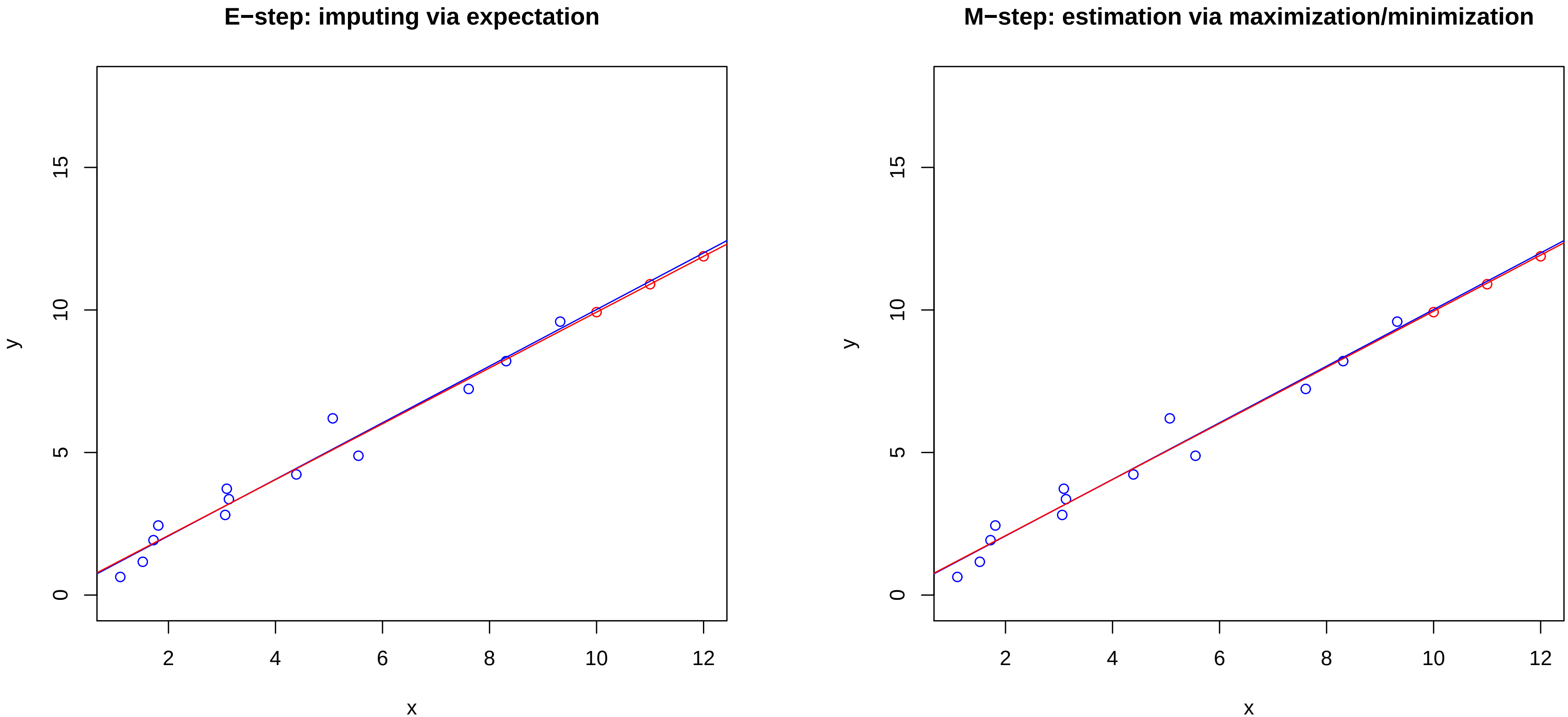


E−step: imputing via expectation
y
0
5
10
15
2
4
6
8
10
12
x



M−step: estimation via maximization/minimization
y
0
5
10
15
2
4
6
8
10
12
x

Iteration 7


E−step: imputing via expectation
M−step: estimation via maximization/minimization
0
5
10
15
y
2
4
6
8
10
12
x

What questions would you ask?

What questions would you ask?

- From all: How general is this method???

What questions would you ask?

- From all: How general is this method???
- From a statistical estimation perspective:
  What's the statistical principle behind it? Is it (asymptotically) efficient in some sense? What assumptions on missing-data mechanism are needed to justify its validity?

What questions would you ask?

- From all: How general is this method???
- From a statistical estimation perspective:
  What's the statistical principle behind it? Is it (asymptotically) efficient in some sense? What assumptions on missing-data mechanism are needed to justify its validity?
- From an algorithmic implementation perspective:
  How many iterations usually does it take? Does the number of iterations depend on where I put the initial points? Does the method scalable to high dimensional data sets? Can it be implemented generically?

The Self-Consistency Principle

The Self-Consistency Principle

- Suppose $\hat{f}_{\mathrm{com}}$ is an estimator for f given complete data $\boldsymbol{y}_{\mathrm{com}}$, but we only observe a subset $\boldsymbol{y}_{\mathrm{obs}}$.

The Self-Consistency Principle

- Suppose $\hat{f}_{\text{com}}$ is an estimator for f given complete data $\boldsymbol{y}_{\text{com}}$, but we only observe a subset $\boldsymbol{y}_{\text{obs}}$.
- Intuitively, the "best" estimate of $\boldsymbol{f}$ given the procedure $\hat{\boldsymbol{f}}_{\text{com}}$ and the imputation model $p(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}}, f)$, $\hat{f}_{\text{obs}}$, should satisfy (exactly or asymptotically)

$$E\left[\hat{\boldsymbol{f}}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}_{\text{obs}}\right] = \hat{\boldsymbol{f}}_{\text{obs}}(\cdot)$$

The Self-Consistency Principle

- Suppose $\hat{f}_{\text{com}}$ is an estimator for f given complete data $\boldsymbol{y}_{\text{com}}$, but we only observe a subset $\boldsymbol{y}_{\text{obs}}$.
- Intuitively, the “best" estimate of $\boldsymbol{f}$ given the procedure $\hat{\boldsymbol{f}}_{\text{com}}$ and the imputation model $p(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}}, f)$, $\hat{f}_{\text{obs}}$, should satisfy (exactly or asymptotically)

$$E\left[\hat{\boldsymbol{f}}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}_{\text{obs}}\right] = \hat{\boldsymbol{f}}_{\text{obs}}(\cdot)$$

- It is a form of Self-Rao-Blackwellization – bring out the best. We will theoretically justify being the “best".

It all started by Efron (1967) ...

It all started by Efron (1967) ...

- For i.i.d. data with independent right censoring, the Kaplan-Meier estimator of CDF F is an NPMLE.

It all started by Efron (1967) ...

- For i.i.d. data with independent right censoring, the Kaplan-Meier estimator of CDF $\boldsymbol{F}$ is an NPMLE.
- Efron (1967) introduced "self-consistency", and shown that the estimator $\hat{\boldsymbol{F}}_{\text{obs}}$ from solving

$$E\left[\hat{\boldsymbol{F}}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}}; \boldsymbol{F} = \hat{\boldsymbol{F}}_{\text{obs}}\right] = \hat{\boldsymbol{F}}_{\text{obs}}(\cdot)$$

is exactly the K-M estimator, where $\hat{\boldsymbol{F}}_{\text{com}}$ is the complete-data empirical CDF.

It all started by Efron (1967) ...

- For i.i.d. data with independent right censoring, the Kaplan-Meier estimator of CDF F is an NPMLE.
- Efron (1967) introduced “self-consistency", and shown that the estimator $\hat{F}_{\rm obs}$ from solving

$$E\left[\hat{\boldsymbol{F}}_{\rm com}(\cdot)\middle|\boldsymbol{y}_{\rm obs}; \boldsymbol{F}=\hat{\boldsymbol{F}}_{\rm obs}\right]=\hat{\boldsymbol{F}}_{\rm obs}(\cdot)$$

  is exactly the K-M estimator, where $\hat{F}_{\rm com}$ is the complete-data empirical CDF.
- Considerable progresses by Turnbull (1974, 1976), Tasi and Crowley (1985), Tasi (1986), Chan and Yang (1987), Ren and Mykland (1996), Van der Laan (1997, 1998, etc. under more general censoring.

Least Squares Estimator is Self-consistent

Least Squares Estimator is Self-consistent

- Self-consistency directs us to seek $\hat{\beta}_{13}$ such that

$$E\left[\hat{\beta}_{16}(y_1,\ldots,y_{16})\middle|y_1,\ldots,y_{13};\beta=\hat{\beta}_{13}\right]=\hat{\beta}_{13}, \qquad (B)$$

Least Squares Estimator is Self-consistent

- Self-consistency directs us to seek $\hat{\beta}_{13}$ such that

$$E\left[\hat{\beta}_{16}(y_1, \ldots, y_{16}) \middle| y_1, \ldots, y_{13}; \beta = \hat{\beta}_{13}\right] = \hat{\beta}_{13}, \qquad (B)$$

- (B) can be solved iteratively without knowing the form of $\hat{\beta}_{16}$. Starting with $\beta_{13}^{(0)}$, at the t^{th} iteration, (1) impute the missing y_i by $y_i^{(t)} = \beta_{13}^{(t)} x_i$ and (2) compute

$$\beta_{13}^{(t+1)} = \hat{\beta}_{16}(y_1, \ldots, y_{13}, y_{14}^{(t)}, y_{15}^{(t)}, y_{16}^{(t)}). \qquad (C)$$

Least Squares Estimator is Self-consistent

- Self-consistency directs us to seek $\hat{\beta}_{13}$ such that

$$E\left[\hat{\beta}_{16}(y_1, \ldots, y_{16}) \middle| y_1, \ldots, y_{13}; \beta = \hat{\beta}_{13}\right] = \hat{\beta}_{13}, \qquad (B)$$

- (B) can be solved iteratively without knowing the form of $\hat{\beta}_{16}$. Starting with $\beta_{13}^{(0)}$, at the t^{th} iteration, (1) impute the missing y_i by $y_i^{(t)} = \beta_{13}^{(t)} x_i$ and (2) compute

$$\beta_{13}^{(t+1)} = \hat{\beta}_{16}(y_1, \ldots, y_{13}, y_{14}^{(t)}, y_{15}^{(t)}, y_{16}^{(t)}). \qquad (C)$$

- The limit of (C), denoted by $\hat{\beta}_{13}$, satisfies

$$\hat{\beta}_{13} = \frac{\sum_{i=1}^{13} y_i x_i + \hat{\beta}_{13} \sum_{i=14}^{16} x_i^2}{\sum_{i=1}^{16} x_i^2} \Longrightarrow \hat{\beta}_{13} = \frac{\sum_{i=1}^{13} y_i x_i}{\sum_{i=1}^{13} x_i^2}$$

... so are (almost) all Parametric MLEs ...

... so are (almost) all Parametric MLEs ...

- log-likelihood $\ell(\theta|\boldsymbol{y}_{\rm com})$; complete-data MLE $\hat{\theta}_{\rm com}$

... so are (almost) all Parametric MLEs ...

- log-likelihood $\ell(\theta|\boldsymbol{y}_{\text{com}})$; complete-data MLE $\hat{\theta}_{\text{com}}$
- score $S(\theta|\boldsymbol{y}_{\text{com}})$ & expected Fisher information $I(\theta)$

$$\hat{\theta}_{\text{com}} - \theta = \frac{S(\theta|\boldsymbol{y}_{\text{com}})}{I(\theta)} + o_p(N^{-1/2}).$$

$$E[\hat{\theta}_{\text{com}}|\boldsymbol{y}_{\text{obs}};\theta] - \theta = \frac{E[S(\theta|\boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}};\theta]}{I(\theta)} + o_p(n^{-1/2})$$

... so are (almost) all Parametric MLEs ...

- log-likelihood $\ell(\theta|\boldsymbol{y}_{\rm com})$; complete-data MLE $\hat{\theta}_{\rm com}$
- score $S(\theta|\boldsymbol{y}_{\rm com})$ & expected Fisher information $I(\theta)$

$$\hat{\theta}_{\rm com} - \theta = \frac{S(\theta|\boldsymbol{y}_{\rm com})}{I(\theta)} + o_p(N^{-1/2}).$$

$$E[\hat{\theta}_{\rm com}|\boldsymbol{y}_{\rm obs};\theta] - \theta = \frac{E[S(\theta|\boldsymbol{y}_{\rm com})|\boldsymbol{y}_{\rm obs};\theta]}{I(\theta)} + o_p(n^{-1/2})$$

- Because of the Fisher's identity

$$E[S(\theta|\boldsymbol{y}_{\rm com})|\boldsymbol{y}_{\rm obs};\theta] = S(\theta|\boldsymbol{y}_{\rm obs})$$

& $S(\hat{\theta}_{\rm obs}|\boldsymbol{y}_{\rm obs}) = 0$, observed-data MLE $\hat{\theta}_{\rm obs}$ must satisfy

$$E[\hat{\theta}_{\rm com}|\boldsymbol{y}_{\rm obs}, \theta = \hat{\theta}_{\rm obs}] = \hat{\theta}_{\rm obs} + o_p(n^{-1/2}).$$

A Multiple Imputation Self-Consistent (MISC) Algorithm

A Multiple Imputation Self-Consistent (MISC) Algorithm

Starting from $\hat{\boldsymbol{f}}^{(0)}$, for $t = 1, \ldots,$ iterating three steps:

A Multiple Imputation Self-Consistent (MISC) Algorithm

Starting from $\hat{\boldsymbol{f}}^{(0)}$, for $t = 1, \ldots,$ iterating three steps:

1. Multiple Imputation: for $\ell = 1, \ldots, m$, draw independently $\boldsymbol{y}_{\text{mis}}^{\ell} \sim P(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)})$

A Multiple Imputation Self-Consistent (MISC) Algorithm

Starting from $\hat{\boldsymbol{f}}^{(0)}$, for $t = 1, \ldots,$ iterating three steps:

1. Multiple Imputation: for $\ell = 1, \ldots, m$, draw independently $\boldsymbol{y}^{\ell}_{\text{mis}} \sim P(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)})$
2. Applying the complete-data procedure to $\boldsymbol{y}^{\ell} = \{\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}^{\ell}_{\text{mis}}\}$ to compute $\hat{\boldsymbol{f}}_{\ell}$, $\ell = 1, \ldots, m$

A Multiple Imputation Self-Consistent (MISC) Algorithm

Starting from $\hat{\boldsymbol{f}}^{(0)}$, for $t = 1, \ldots,$ iterating three steps:

1. **Multiple Imputation:** for $\ell = 1, \ldots, m$, draw independently $\boldsymbol{y}_{\text{mis}}^{\ell} \sim P(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)})$
2. **Applying the complete-data procedure** to $\boldsymbol{y}^{\ell} = \{\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{\ell}\}$ to compute $\hat{\boldsymbol{f}}_{\ell}$, $\ell = 1, \ldots, m$
3. **Combining Estimates:**

   Under L^2 : $\qquad \hat{\boldsymbol{f}}^{(t)} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\boldsymbol{f}}_{\ell}.$

A Multiple Imputation Self-Consistent (MISC) Algorithm

Starting from $\hat{\boldsymbol{f}}^{(0)}$, for $t = 1, \ldots,$ iterating three steps:

1. **Multiple Imputation:** for $\ell = 1, \ldots, m$, draw independently $\boldsymbol{y}_{\text{mis}}^{\ell} \sim P(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)})$
2. **Applying the complete-data procedure** to $\boldsymbol{y}^{\ell} = \{\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{\ell}\}$ to compute $\hat{\boldsymbol{f}}_{\ell}$, $\ell = 1, \ldots, m$
3. **Combining Estimates:**

   Under L^2 : $\quad \hat{\boldsymbol{f}}^{(t)} = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\boldsymbol{f}}_{\ell}.$

   Under L^1 : $\quad \hat{\boldsymbol{f}}^{(t)} = \text{Median}\{\hat{\boldsymbol{f}}_{\ell},\ \ell = 1, \ldots, m\}$

A Multiple Imputation Self-Consistent (MISC) Algorithm

Starting from $\hat{\boldsymbol{f}}^{(0)}$, for $t = 1, \ldots,$ iterating three steps:

1. **Multiple Imputation:** for $\ell = 1, \ldots, m$, draw independently $\boldsymbol{y}_{\text{mis}}^{\ell} \sim P(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)})$
2. **Applying the complete-data procedure** to $\boldsymbol{y}^{\ell} = \{\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{\ell}\}$ to compute $\hat{\boldsymbol{f}}_{\ell}$, $\ell = 1, \ldots, m$
3. **Combining Estimates:**

   $$\text{Under } L^2: \qquad \hat{\boldsymbol{f}}^{(t)} = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\boldsymbol{f}}_{\ell}.$$

   $$\text{Under } L^1: \qquad \hat{\boldsymbol{f}}^{(t)} = \text{Median}\{\hat{\boldsymbol{f}}_{\ell},\ \ell = 1, \ldots, m\}$$

   (nuisance part of f can be handled differently.)

MISC: No corner cutting, but ...

- Advantages:

MISC: No corner cutting, but ...

- Advantages:
  1. A generic algorithm: can be applied with any complete-data *procedure*;

MISC: No corner cutting, but ...

- Advantages:
  1. A generic algorithm: can be applied with any complete-data *procedure*;
  2. And any error norm: simply modify the combining rule accordingly.

MISC: No corner cutting, but ...

- Advantages:
  1. A generic algorithm: can be applied with any complete-data *procedure*;
  2. And any error norm: simply modify the combining rule accordingly.
  3. Additional programming is often easy.

MISC: No corner cutting, but ...

- Advantages:
  1. A generic algorithm: can be applied with any complete-data *procedure*;
  2. And any error norm: simply modify the combining rule accordingly.
  3. Additional programming is often easy.
  4. Provides a benchmark.

MISC: No corner cutting, but ...

- Advantages:
  1. A generic algorithm: can be applied with any complete-data *procedure*;
  2. And any error norm: simply modify the combining rule accordingly.
  3. Additional programming is often easy.
  4. Provides a benchmark.
- Disadvantage: computationally very expensive, especially when the Monte Carlo size m is large (e.g., $m = 100$).

So What Are The Theoretical Guarantees?

So What Are The Theoretical Guarantees?

- Let $\hat{f}_{\mathrm{com}}$ be an estimator of f based on $\boldsymbol{y}_{\mathrm{com}}$, and the error norm be L^p

$$||\hat{f}_{\mathrm{com}} - f||_p = \left[E \left(\int |\hat{f}_{\mathrm{com}}(t) - f(t)|^p dt \right) \right]^{1/p}$$

So What Are The Theoretical Guarantees?

- Let $\hat{f}_{\rm com}$ be an estimator of f based on $\boldsymbol{y}_{\rm com}$, and the error norm be L^p

$$||\hat{f}_{\rm com} - f||_p = \left[E\left(\int |\hat{f}_{\rm com}(t) - f(t)|^p dt\right)\right]^{1/p}$$

- Let $M(f;\boldsymbol{y}_{\rm com})$ be the projection of $\hat{f}_{\rm com}$ under the *conditionally expected* norm:

$$M(f;\boldsymbol{y}_{\rm obs}) = {\rm argmin}_g E\left[\int |\hat{f}_{\rm com}(t) - g(t)|^p dt \middle| \boldsymbol{y}_{\rm obs}; f\right]$$

E.g., for $p = 2$, $M(f;\boldsymbol{y}_{\rm obs})(t) = E[\hat{f}_{\rm com}(t)|\boldsymbol{y}_{\rm obs}; f]$

So What Are The Theoretical Guarantees?

- Let $\hat{f}_{\rm com}$ be an estimator of f based on $\boldsymbol{y}_{\rm com}$, and the error norm be L^p

$$||\hat{f}_{\rm com} - f||_p = \left[E\left(\int |\hat{f}_{\rm com}(t) - f(t)|^p dt\right)\right]^{1/p}$$

- Let $M(f; \boldsymbol{y}_{\rm com})$ be the projection of $\hat{f}_{\rm com}$ under the *conditionally expected* norm:

$$M(f; \boldsymbol{y}_{\rm obs}) = \mathrm{argmin}_g E\left[\int |\hat{f}_{\rm com}(t) - g(t)|^p dt \middle| \boldsymbol{y}_{\rm obs}; f\right]$$

  E.g., for $p = 2$, $M(f; \boldsymbol{y}_{\rm obs})(t) = E[\hat{f}_{\rm com}(t)|\boldsymbol{y}_{\rm obs}; f]$

- Let $M(\hat{f}) \equiv M(f = \hat{f}; \boldsymbol{y}_{\rm obs})$ be the induced mapping from $\mathcal{F}_{\rm obs}$—a suitably defined sub-space of L^p that includes the true f_0—into itself.

The Power of Contraction Mapping

The Power of Contraction Mapping

- Define $|f|_p = \left[\int |f(t)|^p dt\right]^{1/p}$. Suppose $M(f)$ is (a.s.) a contraction mapping on $\mathcal{F}_{\rm obs}$ with respect to $|f|_p$, then (a.s) there exists a unique solution to $|M(\hat{f}_{\rm obs}) - \hat{f}_{\rm obs}|_p = 0$.

The Power of Contraction Mapping

- Define $|f|_p = \left[\int |f(t)|^p dt\right]^{1/p}$. Suppose $M(f)$ is (a.s.) a contraction mapping on $\mathcal{F}_{\rm obs}$ with respect to $|f|_p$, then (a.s) there exists a unique solution to $|M(\hat{f}_{\rm obs}) - \hat{f}_{\rm obs}|_p = 0$.
- Suppose there exists a $0 < \delta < 1$ such that $\forall \hat{f}_1, \hat{f}_2 \in \mathcal{F}_{\rm obs}$, $||M(\hat{f}_1) - M(\hat{f}_2)||_p \leq \delta||\hat{f}_1 - \hat{f}_2||_p$. Then for any $f \in \mathcal{F}_{\rm obs}$,

$$||\hat{f}_{\rm obs} - f||_p \leq 2\frac{||\hat{f}_{\rm com} - f||_p}{1-\delta}$$

The Power of Contraction Mapping

- Define $|f|_p = \left[\int |f(t)|^p dt\right]^{1/p}$. Suppose $M(f)$ is (a.s.) a contraction mapping on $\mathcal{F}_{\text{obs}}$ with respect to $|f|_p$, then (a.s) there exists a unique solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$.
- Suppose there exists a $0 < \delta < 1$ such that $\forall \hat{f}_1, \hat{f}_2 \in \mathcal{F}_{\text{obs}}$, $||M(\hat{f}_1) - M(\hat{f}_2)||_p \leq \delta||\hat{f}_1 - \hat{f}_2||_p$. Then for any $f \in \mathcal{F}_{\text{obs}}$,

$$||\hat{f}_{\text{obs}} - f||_p \leq 2\frac{||\hat{f}_{\text{com}} - f||_p}{1-\delta}$$

Proof: $||\hat{f}_{\text{obs}} - f||_p \leq ||M(\hat{f}_{\text{obs}}) - M(f)||_p + ||M(f) - f||_p$

The Power of Contraction Mapping

- Define $|f|_p = \left[\int |f(t)|^p dt\right]^{1/p}$. Suppose $M(f)$ is (a.s.) a contraction mapping on $\mathcal{F}_{\text{obs}}$ with respect to $|f|_p$, then (a.s) there exists a unique solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$.
- Suppose there exists a $0 < \delta < 1$ such that $\forall \hat{f}_1, \hat{f}_2 \in \mathcal{F}_{\text{obs}}$, $||M(\hat{f}_1) - M(\hat{f}_2)||_p \le \delta||\hat{f}_1 - \hat{f}_2||_p$. Then for any $f \in \mathcal{F}_{\text{obs}}$,

$$||\hat{f}_{\text{obs}} - f||_p \le 2\frac{||\hat{f}_{\text{com}} - f||_p}{1-\delta}$$

Proof: $||\hat{f}_{\text{obs}} - f||_p \le ||M(\hat{f}_{\text{obs}}) - M(f)||_p + ||M(f) - f||_p$

$$||\hat{f}_{\text{obs}} - f||_p \le \delta||\hat{f}_{\text{obs}} - f||_p + ||M(f) - f||_p$$

The Power of Contraction Mapping

- Define $|f|_p = \left[\int |f(t)|^p dt\right]^{1/p}$. Suppose $M(f)$ is (a.s.) a contraction mapping on $\mathcal{F}_{\text{obs}}$ with respect to $|f|_p$, then (a.s) there exists a unique solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$.
- Suppose there exists a $0 < \delta < 1$ such that $\forall \hat{f}_1, \hat{f}_2 \in \mathcal{F}_{\text{obs}}$, $||M(\hat{f}_1) - M(\hat{f}_2)||_p \leq \delta||\hat{f}_1 - \hat{f}_2||_p$. Then for any $f \in \mathcal{F}_{\text{obs}}$,

$$||\hat{f}_{\text{obs}} - f||_p \leq 2\frac{||\hat{f}_{\text{com}} - f||_p}{1-\delta}$$

Proof: $||\hat{f}_{\text{obs}} - f||_p \leq ||M(\hat{f}_{\text{obs}}) - M(f)||_p + ||M(f) - f||_p$

$$||\hat{f}_{\text{obs}} - f||_p \leq \delta||\hat{f}_{\text{obs}} - f||_p + ||M(f) - f||_p$$

$$||M(f) - f||_p \leq ||M(f) - \hat{f}_{\text{com}}||_p + ||\hat{f}_{\text{com}} - f||_p \leq 2||\hat{f}_{\text{com}} - f||_p$$

Theory without Contraction Mapping

Theory without Contraction Mapping

- If $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous with respect to $|f|_p$, then there exists a solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$ by applying Brouwer's Fixed Point Theorem.

Theory without Contraction Mapping

- If $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous with respect to $|f|_p$, then there exists a solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$ by applying Brouwer's Fixed Point Theorem.
- Suppose

(1) $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous w.r.t $||\cdot||_p$;

(2) $\hat{\psi}_n(f) = f - M(f)$ uniformly converges on $\mathcal{F}_{\text{obs}}$ to some $\psi(f)$ with respect to $||\cdot||_p$ as the sample size $n \to \infty$;

(3) and the true f_0 is the only solution to $\psi(f) = 0$.

Theory without Contraction Mapping

- If $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous with respect to $|f|_p$, then there exists a solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$ by applying Brouwer's Fixed Point Theorem.
- Suppose

  (1) $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous w.r.t $||\cdot||_p$;

  (2) $\hat{\psi}_n(f) = f - M(f)$ uniformly converges on $\mathcal{F}_{\text{obs}}$ to some $\psi(f)$ with respect to $||\cdot||_p$ as the sample size $n \to \infty$;

  (3) and the true f_0 is the only solution to $\psi(f) = 0$.

  Then any solution of $M(f) = f$ converges to the true f_0 w.r.t $||\cdot||_p$, as $n \to \infty$.

Theory without Contraction Mapping

- If $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous with respect to $|f|_p$, then there exists a solution to $|M(\hat{f}_{\text{obs}}) - \hat{f}_{\text{obs}}|_p = 0$ by applying Brouwer's Fixed Point Theorem.
- Suppose

  (1) $\mathcal{F}_{\text{obs}}$ is compact and $M(f)$ is continuous w.r.t $||\cdot||_p$;

  (2) $\hat{\psi}_n(f) = f - M(f)$ uniformly converges on $\mathcal{F}_{\text{obs}}$ to some $\psi(f)$ with respect to $||\cdot||_p$ as the sample size $n \to \infty$;

  (3) and the true f_0 is the only solution to $\psi(f) = 0$.

  Then any solution of $M(f) = f$ converges to the true f_0 w.r.t $||\cdot||_p$, as $n \to \infty$.
- Many generalizations/refinements are possible ...

Generality and Implications

Generality and Implications

- The result holds for any $p \geq 1$. Important for LASSO, L^1 regressions, etc.

Generality and Implications

- The result holds for any $p \geq 1$. Important for LASSO, L^1 regressions, etc.
- Potentially a useful theoretical tool, ensuring $\hat{f}_{\text{obs}}$ and $\hat{f}_{\text{com}}$ have the same order of *rate of convergence*, as long as we can show $M(f)$ is a contraction mapping.

Generality and Implications

- The result holds for any $p \geq 1$. Important for LASSO, L^1 regressions, etc.
- Potentially a useful theoretical tool, ensuring $\hat{f}_{\text{obs}}$ and $\hat{f}_{\text{com}}$ have the same order of *rate of convergence*, as long as we can show $M(f)$ is a contraction mapping.
- For wavelets soft thresholding and with $p = 2$, under normality and random missingness,

$$\delta = \sqrt{\text{\% of missing data}}$$

Generality and Implications

- The result holds for any $p \geq 1$. Important for LASSO, L^1 regressions, etc.
- Potentially a useful theoretical tool, ensuring $\hat{f}_{\text{obs}}$ and $\hat{f}_{\text{com}}$ have the same order of *rate of convergence*, as long as we can show $M(f)$ is a contraction mapping.
- For wavelets soft thresholding and with $p = 2$, under normality and random missingness,

$$\delta = \sqrt{\text{\% of missing data}}$$

- $M(f)$ is *not* a contraction map for hard thresholding.

What is the connection with the EM algorithm?

What is the connection with the EM algorithm?

- EM builds on Fisher's identity (Efron, 1977)

$$E[S(\theta; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}; \theta] = S(\theta; \boldsymbol{y}_{\text{obs}}),$$

  and solves

$$E[S(\theta^{(t+1)}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}; \theta^{(t)}] = 0.$$

What is the connection with the EM algorithm?

- EM builds on Fisher's identity (Efron, 1977)

$$E[S(\theta; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}; \theta] = S(\theta; \boldsymbol{y}_{\text{obs}}),$$

  and solves

$$E[S(\theta^{(t+1)}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}; \theta^{(t)}] = 0.$$

- Elashoff and Ryan's (2004) ES (Expectation-Solve) replaces $S(\theta|\boldsymbol{y}_{\text{com}})$ with a complete-data Estimating Equation $U_{\text{com}}(\theta; \boldsymbol{y}_{\text{com}})$:

$$E[U_{\text{com}}(\theta^{(t+1)}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}; \theta^{(t)}] = 0.$$

Moving from Algorithmic Principle to Estimation Principle

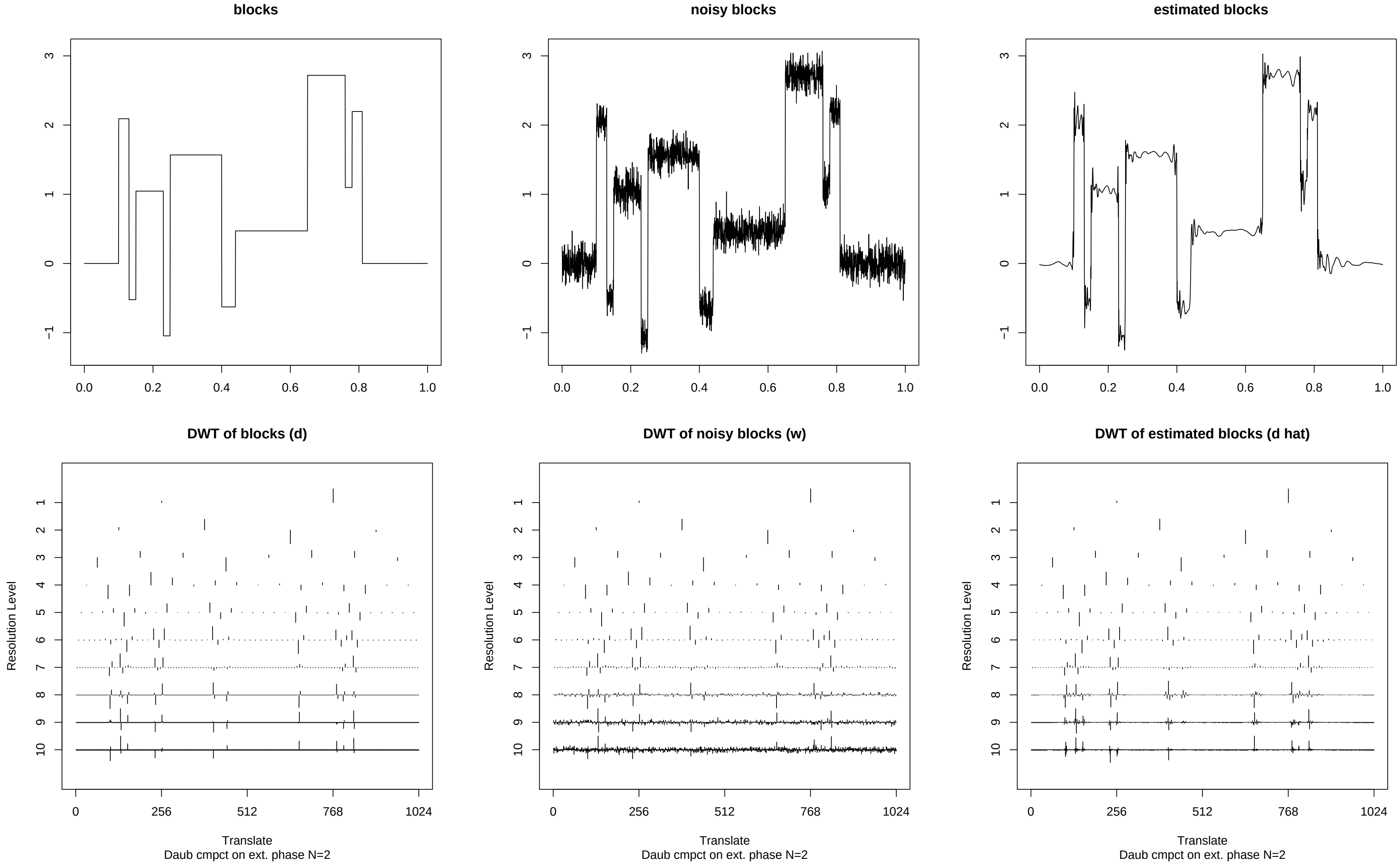
Moving from Algorithmic Principle to Estimation Principle

- For quasi-likelihood, Heyde and Morton (1996) emphasized viewing the E-step as a *projection*, providing an *estimation* principle.

Moving from Algorithmic Principle to Estimation Principle

- For quasi-likelihood, Heyde and Morton (1996) emphasized viewing the E-step as a *projection*, providing an *estimation* principle.
- Self-consistency offers a general principle for *defining* an incomplete-data estimator for f when given
  - an arbitrary *complete-data procedure*;
  - a missing-data mechanism $P(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}}; f)$;
  - an error norm.

Wavelet Denoising (Donoho and Johnstone, 1994)



Incomplete Designs

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^n$ that satisfy

$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\,\mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$ that satisfy

$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\, \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^{n}$ is a subset of $\boldsymbol{X}_N = \{\frac{i}{N}\}_{i=0}^{N-1}$.

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$ that satisfy

$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\, \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^{n}$ is a subset of $\boldsymbol{X}_N = \{\frac{i}{N}\}_{i=0}^{N-1}$.
- Aim: estimate f via wavelet regression given $\boldsymbol{y}_{\text{obs}}$.

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$ that satisfy

$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\, \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^{n}$ is a subset of $\boldsymbol{X}_N = \{\frac{i}{N}\}_{i=0}^{N-1}$.
- Aim: estimate f via wavelet regression given $\boldsymbol{y}_{\text{obs}}$.
- **Key idea:** View $\boldsymbol{y}_{\text{obs}}$ as incomplete data from $\boldsymbol{y}_{\text{com}} = \{x_i = \frac{i}{N}, y_i\}_{i=0}^{N-1}$ with y_i missing when $x_i \notin \boldsymbol{X}_{\text{obs}}$.

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$ that satisfy

$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\, \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^{n}$ is a subset of $\boldsymbol{X}_N = \{\frac{i}{N}\}_{i=0}^{N-1}$.
- Aim: estimate f via wavelet regression given $\boldsymbol{y}_{\text{obs}}$.
- **Key idea:** View $\boldsymbol{y}_{\text{obs}}$ as incomplete data from $\boldsymbol{y}_{\text{com}} = \{x_i = \frac{i}{N}, y_i\}_{i=0}^{N-1}$ with y_i missing when $x_i \notin \boldsymbol{X}_{\text{obs}}$.
- Applications:
  1. Actual missing y's with a regular design.

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^n$ that satisfy

$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\, \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^n$ is a subset of $\boldsymbol{X}_N = \{\frac{i}{N}\}_{i=0}^{N-1}$.
- Aim: estimate f via wavelet regression given $\boldsymbol{y}_{\text{obs}}$.
- **Key idea:** View $\boldsymbol{y}_{\text{obs}}$ as incomplete data from $\boldsymbol{y}_{\text{com}} = \{x_i = \frac{i}{N}, y_i\}_{i=0}^{N-1}$ with y_i missing when $x_i \notin \boldsymbol{X}_{\text{obs}}$.
- Applications:
  1. Actual missing y's with a regular design.
  2. Deleting outliers from a regular design data set.

Incomplete Designs

- Suppose we observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^n$ that satisfy

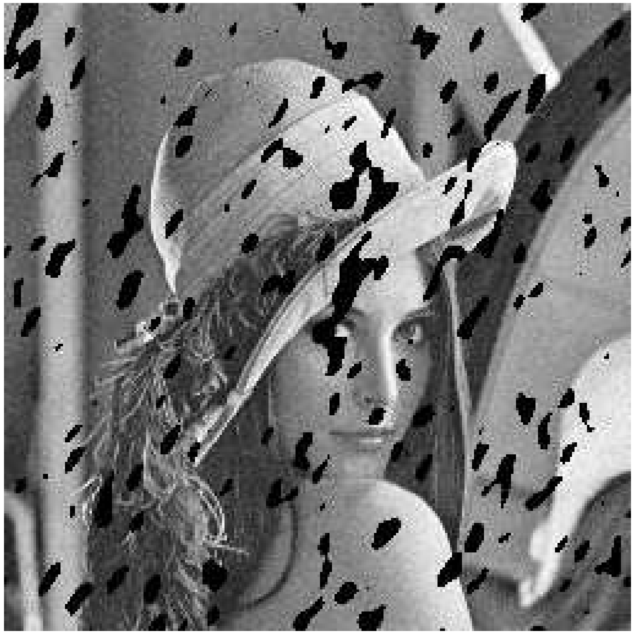
$$y_i = f(x_i) + e_i, \quad e_i \sim i.i.d\, \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^n$ is a subset of $\boldsymbol{X}_N = \{\frac{i}{N}\}_{i=0}^{N-1}$.
- Aim: estimate f via wavelet regression given $\boldsymbol{y}_{\text{obs}}$.
- **Key idea:** View $\boldsymbol{y}_{\text{obs}}$ as incomplete data from $\boldsymbol{y}_{\text{com}} = \{x_i = \frac{i}{N}, y_i\}_{i=0}^{N-1}$ with y_i missing when $x_i \notin \boldsymbol{X}_{\text{obs}}$.
- Applications:
  1. Actual missing y's with a regular design.
  2. Deleting outliers from a regular design data set.
  3. Cross-validation for a regular design problem.

Incomplete/Missing Data in 2D

- instrument malfunction, damaged photos, etc.

missing at random



clustering

A Simple (SIM) Approximated Algorithm

A Simple (SIM) Approximated Algorithm

- Starting with $\hat{\boldsymbol{f}}^{(0)}$ and $\hat{\sigma}^{(0)}$, for $t = 1, \ldots$, iterating:

A Simple (SIM) Approximated Algorithm

- Starting with $\hat{\boldsymbol{f}}^{(0)}$ and $\hat{\sigma}^{(0)}$, for $t = 1, \ldots$, iterating:
  1. Impute the missing y_i by $y_i^{(t)} = \hat{f}_i^{(t-1)}$ and create $\boldsymbol{y}^{(t)} = \{y_i : y_i \text{ is observed}\} \cup \{y_i^{(t)} : y_i \text{ is missing}\}$

A Simple (SIM) Approximated Algorithm

- Starting with $\hat{\boldsymbol{f}}^{(0)}$ and $\hat{\sigma}^{(0)}$, for $t = 1, \ldots$, iterating:
  1. Impute the missing y_i by $y_i^{(t)} = \hat{f}_i^{(t-1)}$ and create $\boldsymbol{y}^{(t)} = \{y_i : y_i \text{ is observed}\} \cup \{y_i^{(t)} : y_i \text{ is missing}\}$
  2. Obtain $\boldsymbol{w}^{(t)} = \boldsymbol{W}\boldsymbol{y}^{(t)}$ & “finest scale" estimate $\tilde{\sigma}^{(t)}$

A Simple (SIM) Approximated Algorithm

- Starting with $\hat{\boldsymbol{f}}^{(0)}$ and $\hat{\sigma}^{(0)}$, for $t = 1, \ldots$, iterating:
  1. Impute the missing y_i by $y_i^{(t)} = \hat{f}_i^{(t-1)}$ and create $\boldsymbol{y}^{(t)} = \{y_i : y_i \text{ is observed}\} \cup \{y_i^{(t)} : y_i \text{ is missing}\}$
  2. Obtain $\boldsymbol{w}^{(t)} = \boldsymbol{W}\boldsymbol{y}^{(t)}$ & “finest scale" estimate $\tilde{\sigma}^{(t)}$
  3. Use the variance inflation formula to compute

$$\hat{\sigma}^{(t)} = \sqrt{[\tilde{\sigma}^{(t)}]^2 + C_m[\hat{\sigma}^{(t-1)}]^2},$$

where $C_m = 1 - \frac{n}{N}$ is fraction of missing data

A Simple (SIM) Approximated Algorithm

- Starting with $\hat{\boldsymbol{f}}^{(0)}$ and $\hat{\sigma}^{(0)}$, for $t = 1, \ldots,$ iterating:
  1. Impute the missing y_i by $y_i^{(t)} = \hat{f}_i^{(t-1)}$ and create $\boldsymbol{y}^{(t)} = \{y_i : y_i$ is observed$\} \cup \{y_i^{(t)} : y_i$ is missing$\}$
  2. Obtain $\boldsymbol{w}^{(t)} = \boldsymbol{W}\boldsymbol{y}^{(t)}$ & “finest scale" estimate $\tilde{\sigma}^{(t)}$
  3. Use the variance inflation formula to compute

$$\hat{\sigma}^{(t)} = \sqrt{[\tilde{\sigma}^{(t)}]^2 + C_m[\hat{\sigma}^{(t-1)}]^2},$$

     where $C_m = 1 - \frac{n}{N}$ is fraction of missing data
  4. Threshold $\boldsymbol{w}^{(t)}$ with $g(\hat{\sigma}^{(t)})$ (e.g. $g(\sigma) = \sigma\sqrt{2\log N}$) to obtain $\hat{\boldsymbol{w}}^{(t)}$, and then $\hat{\boldsymbol{f}}^{(t)} = \boldsymbol{W}^T\hat{\boldsymbol{w}}^{(t)}$

SIM: Extreme Corner Cutting

SIM: Extreme Corner Cutting

- It is fast, and it works very well when $C_m << 1$ — Quick and Dirty, but it can be filthy!

SIM: Extreme Corner Cutting

- It is fast, and it works very well when $C_m << 1$ — Quick and Dirty, but it can be filthy!
- Key component: variance inflation formula, which accounts for the effect of those imputed $y_i^{(t)}$'s on the estimation of σ^2.

SIM: Extreme Corner Cutting

- It is fast, and it works very well when $C_m << 1$ — Quick and Dirty, but it can be filthy!
- Key component: variance inflation formula, which accounts for the effect of those imputed $y_i^{(t)}$'s on the estimation of σ^2.
- Derived by assuming the conditional expectation

$$E\left[1_{|w_l|\geq g(\tilde{\sigma})} w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right] \approx 1_{\left|E\left[w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f}=\hat{\boldsymbol{f}}^{(t-1)}\right]\right| \geq g(\hat{\sigma})}$$

SIM: Extreme Corner Cutting

- It is fast, and it works very well when $C_m << 1$ — Quick and Dirty, but it can be filthy!
- Key component: variance inflation formula, which accounts for the effect of those imputed $y_i^{(t)}$'s on the estimation of σ^2.
- Derived by assuming the conditional expectation

$$E\left[1_{|w_l|\geq g(\tilde{\sigma})} w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right] \approx 1_{\left|E\left[w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f}=\hat{\boldsymbol{f}}^{(t-1)}\right]\right| \geq g(\hat{\sigma})}$$

- Extreme corner cutting, but we understand when it can help and when it will do great harm.

A Refined (REF) Algorithm: Much Better Corner Cutting

A Refined (REF) Algorithm: Much Better Corner Cutting

- Similar to SIM, but much better approximation to the E-step $\hat{w}_l^{(t)} \equiv E\left[1_{|w_l|\geq g(\tilde{\sigma})} w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right]$ pretending $c = g(\tilde{\sigma})$ is fixed. Under normality, $\hat{w}_l^{(t)}$ is expressible via normal pdf ϕ and CDF Φ:

A Refined (REF) Algorithm: Much Better Corner Cutting

- Similar to SIM, but much better approximation to the E-step $\hat{w}_l^{(t)} \equiv E\left[1_{|w_l| \geq g(\tilde{\sigma})} w_l | \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right]$ pretending $c = g(\tilde{\sigma})$ is fixed. Under normality, $\hat{w}_l^{(t)}$ is expressible via normal pdf ϕ and CDF Φ:

$$
\begin{aligned}
\hat{w}_l^{(t)} &= \alpha(w_l^{(t)}, \eta_l) + \beta(w_l^{(t)}, \eta_l) \times w_l^{(t)} \\
\text{with } \alpha(w, \eta) &= \eta\sigma \left[\phi\left(\frac{c+w}{\eta\sigma}\right) - \phi\left(\frac{c-w}{\eta\sigma}\right)\right], \\
\beta(w, \eta) &= 2 - \Phi\left(\frac{c+w}{\eta\sigma}\right) - \Phi\left(\frac{c-w}{\eta\sigma}\right)
\end{aligned}
$$

A Refined (REF) Algorithm: Much Better Corner Cutting

- Similar to SIM, but much better approximation to the E-step $\hat{w}_l^{(t)} \equiv E\left[1_{|w_l|\geq g(\tilde{\sigma})} w_l | \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right]$ pretending $c = g(\tilde{\sigma})$ is fixed. Under normality, $\hat{w}_l^{(t)}$ is expressible via normal pdf ϕ and CDF Φ:

$$\begin{aligned}
\hat{w}_l^{(t)} &= \alpha(w_l^{(t)}, \eta_l) + \beta(w_l^{(t)}, \eta_l) \times w_l^{(t)} \\
\text{with } \alpha(w, \eta) &= \eta\sigma \left[\phi\left(\frac{c+w}{\eta\sigma}\right) - \phi\left(\frac{c-w}{\eta\sigma}\right)\right], \\
\beta(w, \eta) &= 2 - \Phi\left(\frac{c+w}{\eta\sigma}\right) - \Phi\left(\frac{c-w}{\eta\sigma}\right)
\end{aligned}$$

- η_l can be approximated by $C_m = 1 - \frac{n}{N}$, and $c = g(\hat{\sigma}^{(t)})$

A Refined (REF) Algorithm: Much Better Corner Cutting

- Similar to SIM, but much better approximation to the E-step $\hat{w}_l^{(t)} \equiv E\left[1_{|w_l| \geq g(\tilde{\sigma})} w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right]$ pretending $c = g(\tilde{\sigma})$ is fixed. Under normality, $\hat{w}_l^{(t)}$ is expressible via normal pdf ϕ and CDF Φ:

$$\begin{aligned} \hat{w}_l^{(t)} &= \alpha(w_l^{(t)}, \eta_l) + \beta(w_l^{(t)}, \eta_l) \times w_l^{(t)} \\ \text{with } \alpha(w, \eta) &= \eta\sigma \left[\phi\left(\frac{c+w}{\eta\sigma}\right) - \phi\left(\frac{c-w}{\eta\sigma}\right)\right], \\ \beta(w, \eta) &= 2 - \Phi\left(\frac{c+w}{\eta\sigma}\right) - \Phi\left(\frac{c-w}{\eta\sigma}\right) \end{aligned}$$

- η_l can be approximated by $C_m = 1 - \frac{n}{N}$, and $c = g(\hat{\sigma}^{(t)})$
- A form of “soft thresholding": $\beta(w, \eta) \in (0, 1)$.

Hard Thresholding Perhaps Should be Avoided ...

Hard Thresholding Perhaps Should be Avoided ...

- For soft-thresholding, $1_{(|w_l|\geq c)}\text{sign}(w_l)\{|w_l|-c\}$:

$$\hat{w}_{l,soft}^{(t)} = \hat{w}_{l,hard}^{(t)} + c\left[\Phi\left(\frac{c-w_l^{(t)}}{\eta_l\sigma}\right) - \Phi\left(\frac{c+w_l^{(t)}}{\eta_l\sigma}\right)\right]$$

Hard Thresholding Perhaps Should be Avoided ...

- For soft-thresholding, $1_{(|w_l| \geq c)}$sign$(w_l)\{|w_l| - c\}$:

$$\hat{w}_{l,soft}^{(t)} = \hat{w}_{l,hard}^{(t)} + c\left[\Phi\left(\frac{c - w_l^{(t)}}{\eta_l \sigma}\right) - \Phi\left(\frac{c + w_l^{(t)}}{\eta_l \sigma}\right)\right]$$

- This blue term ensures the contraction property of the self-consistency map, $M(f)$, because for

$$\mu(w) = \alpha(w,\eta) + w\beta(w,\eta) + c\left[\Phi\left(\frac{c-w}{\eta\sigma}\right) - \Phi\left(\frac{c+w}{\eta\sigma}\right)\right],$$

$$\frac{d\mu(w)}{dw} = \beta(w,\eta) \in (0,1).$$

Hard Thresholding Perhaps Should be Avoided ...

- For soft-thresholding, $1_{(|w_l|\geq c)}\text{sign}(w_l)\{|w_l| - c\}$:

$$\hat{w}^{(t)}_{l,soft} = \hat{w}^{(t)}_{l,hard} + c\left[\Phi\left(\frac{c - w^{(t)}_l}{\eta_l\sigma}\right) - \Phi\left(\frac{c + w^{(t)}_l}{\eta_l\sigma}\right)\right]$$

- This blue term ensures the contraction property of the self-consistency map, $M(f)$, because for

$$\mu(w) = \alpha(w,\eta) + w\beta(w,\eta) + c\left[\Phi\left(\frac{c - w}{\eta\sigma}\right) - \Phi\left(\frac{c + w}{\eta\sigma}\right)\right],$$

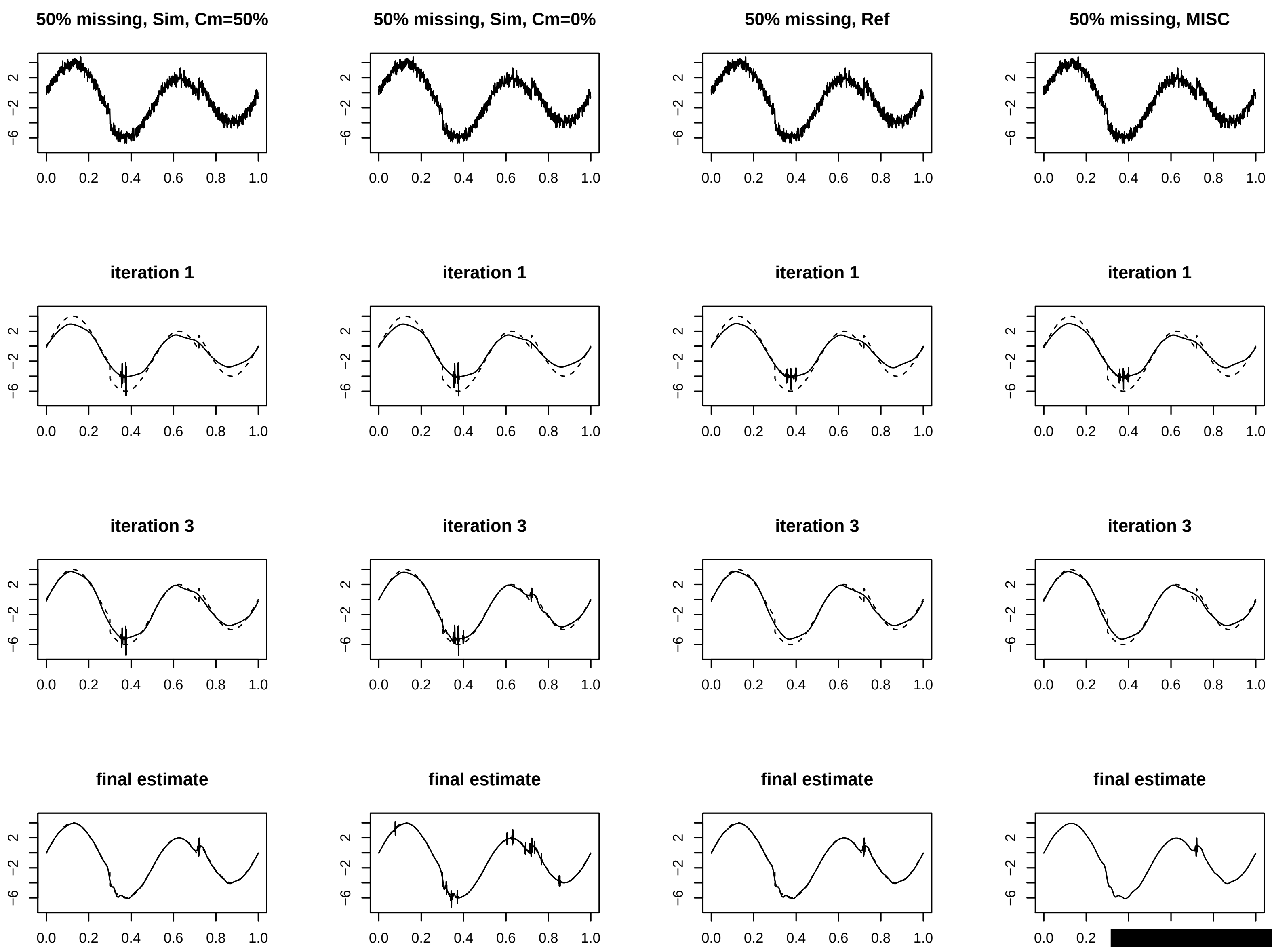
$$\frac{d\mu(w)}{dw} = \beta(w,\eta) \in (0,1).$$

- Not true without the blue term.

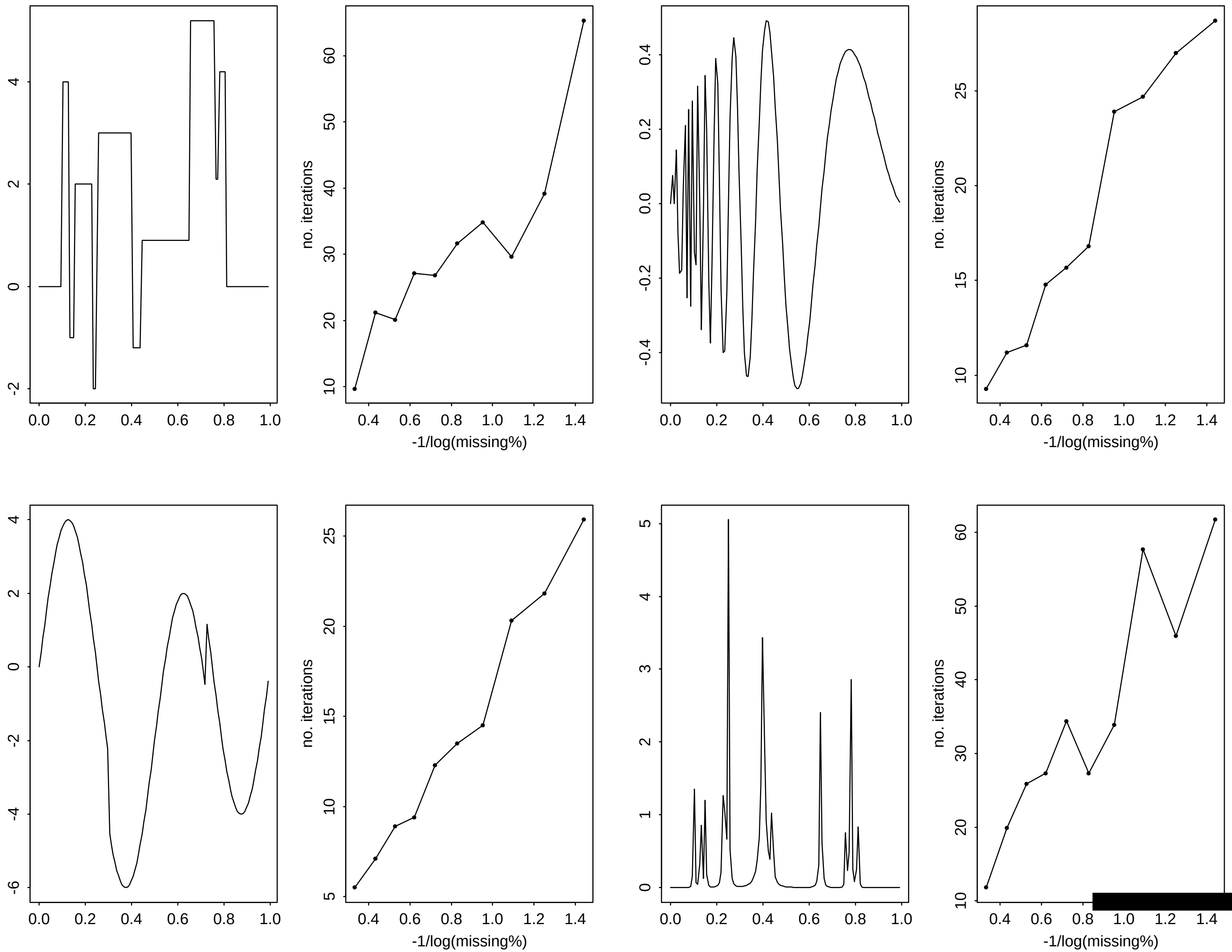
Visual Inspection: Simulation Configurations

- Using four test functions of Donoho & Johnstone (1994).
- Hard universal thresholding: $|w_{jk}| \geq \hat{\sigma}\sqrt{2\log N}$.
- Mother wavelet: D5; primary resolution = 3.
- Signal-to-noise ratio: snr$= \|f\|/\sigma = 7$.
- Complete data size N=2048.
- Random deletion percentage: 10%, 30%, 50%.
- Initial values: $\hat{f}^{(0)} = Lowess$; $\hat{\sigma}^{(0)}$: from residuals.
- Stopping criterion: $|\hat{\sigma}^{(t+1)} - \hat{\sigma}^{(t)}|/\hat{\sigma}^{(t)} < 0.0001$.

SIM ($C_m = \rho = .5$), SIM ($\rho = .5, C_m = 0$), REF ($\rho = .5$), MISC ($\rho = .5$)



Number of Iterations $\propto [-\log(\rho)]^{-1}$



L^1 Generalization: Applications to Variable Selection

L^1 Generalization: Applications to Variable Selection

- Many variable selection methods (e.g., LASSO) emphasize estimates being exactly zero (e.g., $\hat{\beta}_1 = 0$).

L^1 Generalization: Applications to Variable Selection

- Many variable selection methods (e.g., LASSO) emphasize estimates being exactly zero (e.g., $\hat{\beta}_1 = 0$).
- L^2 combining rule, i.e., averaging does not preserve this property.

L^1 Generalization: Applications to Variable Selection

- Many variable selection methods (e.g., LASSO) emphasize estimates being exactly zero (e.g., $\hat{\beta}_1 = 0$).
- L^2 combining rule, i.e., averaging does not preserve this property.
- But L^1 combining rule does. It works like a "voting method": if more than 50% of $\{\hat{\beta}_{1,\ell}, \ell = 1, \ldots, m\}$ are zero, then the next iterate $\hat{\beta}_1^{(t+1)} = 0$.

L^1 Generalization: Applications to Variable Selection

- Many variable selection methods (e.g., LASSO) emphasize estimates being exactly zero (e.g., $\hat{\beta}_1 = 0$).
- L^2 combining rule, i.e., averaging does not preserve this property.
- But L^1 combining rule does. It works like a “voting method”: if more than 50% of $\{\hat{\beta}_{1,\ell}, \ell = 1, \ldots, m\}$ are zero, then the next iterate $\hat{\beta}_1^{(t+1)} = 0$.
- We illustrate this with adaptive LASSO (the same can be applied to other methods such as SCAD).

Variable Selection in a Linear Model

Variable Selection in a Linear Model

- univariate response: y_i

Variable Selection in a Linear Model

- univariate response: y_i
- p-variate explanatory variable: $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$

Variable Selection in a Linear Model

- univariate response: y_i
- p-variate explanatory variable: $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$
- model: $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad e_i \sim i.i.d.\, \mathcal{N}(0, \sigma^2)$

Variable Selection in a Linear Model

- univariate response: y_i
- p-variate explanatory variable: $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$
- model: $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad e_i \sim i.i.d.\ \mathcal{N}(0, \sigma^2)$
- Model parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

Variable Selection in a Linear Model

- univariate response: y_i
- p-variate explanatory variable: $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$
- model: $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad e_i \sim i.i.d.\ \mathcal{N}(0, \sigma^2)$
- Model parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.
- Aim: identify and estimate those non-zero β_j's when some of the entries in $\{x_{i1}, \ldots, x_{ip}, y_i\}_{i=1}^{n}$ are missing.

L^1 Example: Adaptive LASSO (Tibshirani, 1996; Zou, 2006)

L^1 Example: Adaptive LASSO (Tibshirani, 1996; Zou, 2006)

- When there is no missing data:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j |\beta_j| \right\}$$

L^1 Example: Adaptive LASSO (Tibshirani, 1996; Zou, 2006)

- When there is no missing data:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j |\beta_j| \right\}$$

- λ: tuning parameter, selected via BIC.

L^1 Example: Adaptive LASSO (Tibshirani, 1996; Zou, 2006)

- When there is no missing data:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j |\beta_j| \right\}$$

- λ: tuning parameter, selected via BIC.
- α_j: pre-chosen fixed weights; we use $\alpha_j = 1/\hat{\beta}_j^{\text{ols}}$

L^1 Example: Adaptive LASSO (Tibshirani, 1996; Zou, 2006)

- When there is no missing data:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j |\beta_j| \right\}$$

- λ: tuning parameter, selected via BIC.
- α_j: pre-chosen fixed weights; we use $\alpha_j = 1/\hat{\beta}_j^{\text{ols}}$
- We need a model to impute the missing x_{ij}'s given all observed data (both x's and y); we used Joe Shafer's imputation software based on multivariate normal.

L^1 Example: Adaptive LASSO (Tibshirani, 1996; Zou, 2006)

- When there is no missing data:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j |\beta_j| \right\}$$

- λ: tuning parameter, selected via BIC.
- α_j: pre-chosen fixed weights; we use $\alpha_j = 1/\hat{\beta}_j^{\text{ols}}$
- We need a model to impute the missing x_{ij}'s given all observed data (both x's and y); we used Joe Shafer's imputation software based on multivariate normal.
- We applied LASSO to each imputed data set, and then used the L^1 and L^2 combining rules.

Numerical Experiment

Numerical Experiment

- True model: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, with $\sigma = 3$.

Numerical Experiment

- True model: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, with $\sigma = 3$.
- x_{ij} and x_{ik} normal with correlation $0.5^{|j-k|}$.

Numerical Experiment

- True model: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, with $\sigma = 3$.
- x_{ij} and x_{ik} normal with correlation $0.5^{|j-k|}$.
- Sample sizes: $n =$20 and 60.

Numerical Experiment

- True model: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, with $\sigma = 3$.
- x_{ij} and x_{ik} normal with correlation $0.5^{|j-k|}$.
- Sample sizes: $n =$20 and 60.
- Random deletion missing percentages: 10% and 30%.

Numerical Experiment

- True model: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, with $\sigma = 3$.
- x_{ij} and x_{ik} normal with correlation $0.5^{|j-k|}$.
- Sample sizes: $n =$20 and 60.
- Random deletion missing percentages: 10% and 30%.
- 500 replicates, and each uses $m = 100$ imputations.

Numerical Experiment

- True model: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, with $\sigma = 3$.
- x_{ij} and x_{ik} normal with correlation $0.5^{|j-k|}$.
- Sample sizes: $n =$20 and 60.
- Random deletion missing percentages: 10% and 30%.
- 500 replicates, and each uses $m = 100$ imputations.
- For comparisons, we include the complete-data results, and the results from stacking all m imputed data sets to form a size mN data set, but using effective sample size (ESS) for BIC.

Simulation Results with $n = 20$

algorithm	missing	P_{C}	P_{S}	MSER
Median-Combining	10%	9.8	16	1.26
Mean-Combining		0.2	75.6	1.38
Stacking with ESS		6.6	10.6	0.923
Median-Combining	30%	0.6	0.6	3.32
Mean-Combining		0	99.6	3.08
Stacking with ESS		0.6	0.6	0.662
complete data		16.6	39.6	1.0

P_{C} is % the correct model was recovered, P_{S} is % the selected model was a superset of the true, and MSER is the MSE ratio relative to the complete data procedure.

Simulation Results with $n = 60$

algorithm	missing	P_{C}	P_{S}	MSER
Median-Combining	10%	54.6	72.4	1.06
Mean-Combining		5	95.4	1.11
Stacking with ESS		53	73.2	0.833
Median-Combining	30%	17.2	19	2.51
Mean-Combining		0	99.4	2.31
Stacking with ESS		19.4	21.4	0.382
complete data		57.2	88.4	1.0

P_{C} is % the correct model was recovered, P_{S} is % the selected model was a superset of the true, and MSER is the MSE ratio relative to the complete data procedure.

Summary of Key Contributions

Summary of Key Contributions

- Formulated the Self-consistency Principle for any complete-data procedure.

Summary of Key Contributions

- Formulated the Self-consistency Principle for any complete-data procedure.
- Generalized self-consistency methods beyond L^2 norm, especially the median combining rule for multiple imputation inference with discrete parameters.

Summary of Key Contributions

- Formulated the Self-consistency Principle for any complete-data procedure.
- Generalized self-consistency methods beyond L^2 norm, especially the median combining rule for multiple imputation inference with discrete parameters.
- Provided an initial unified theory via contraction mapping and fixed-point theorems.

Summary of Key Contributions

- Formulated the Self-consistency Principle for any complete-data procedure.
- Generalized self-consistency methods beyond L^2 norm, especially the median combining rule for multiple imputation inference with discrete parameters.
- Provided an initial unified theory via contraction mapping and fixed-point theorems.
- Obtained Refined Algorithm for good compromise between statistical and computational efficiency for wavelet applications.

Summary of Key Contributions

- Formulated the Self-consistency Principle for any complete-data procedure.
- Generalized self-consistency methods beyond L^2 norm, especially the median combining rule for multiple imputation inference with discrete parameters.
- Provided an initial unified theory via contraction mapping and fixed-point theorems.
- Obtained Refined Algorithm for good compromise between statistical and computational efficiency for wavelet applications.
- BUT, there are a lot more to be done ...

If you still want more ...

If you still want more ...

- Lee, Thomas C. M. and Meng, Xiao-Li (2005), "A Self-Consistent Wavelet Method for Denoising Images with Missing Pixels", *Proceedings of the 30th IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing Vol II*, 41-44.

If you still want more ...

- Lee, Thomas C. M. and Meng, Xiao-Li (2005), "A Self-Consistent Wavelet Method for Denoising Images with Missing Pixels", *Proceedings of the 30th IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing Vol II*, 41-44.
- Meng, Xiao-Li (2007) "A Helicopter View of The Self-Consistency Framework for Wavelets and Other Signal Extraction Methods In the Presence of Missing and Irregularly Spaced Data", *Wavelets XII, Proceedings of SPIE Vol. 6701 (Bellingham, WA, 2007).*

If you still want more ...

- Lee, Thomas C. M. and Meng, Xiao-Li (2005), "A Self-Consistent Wavelet Method for Denoising Images with Missing Pixels", *Proceedings of the 30th IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing Vol II*, 41-44.
- Meng, Xiao-Li (2007) "A Helicopter View of The Self-Consistency Framework for Wavelets and Other Signal Extraction Methods In the Presence of Missing and Irregularly Spaced Data", *Wavelets XII, Proceedings of SPIE Vol. 6701 (Bellingham, WA, 2007).*
- Lee, Thomas C. M., Li, Zhan, and Meng, Xiao-Li (2010), "What can we do when EM is not applicable? Self Consistency: A general recipe for semi-parametric and non-parametric estimation with incomplete and irregularly spaced data." Revision for *Statistical Science*.

Missing at Random

degraded

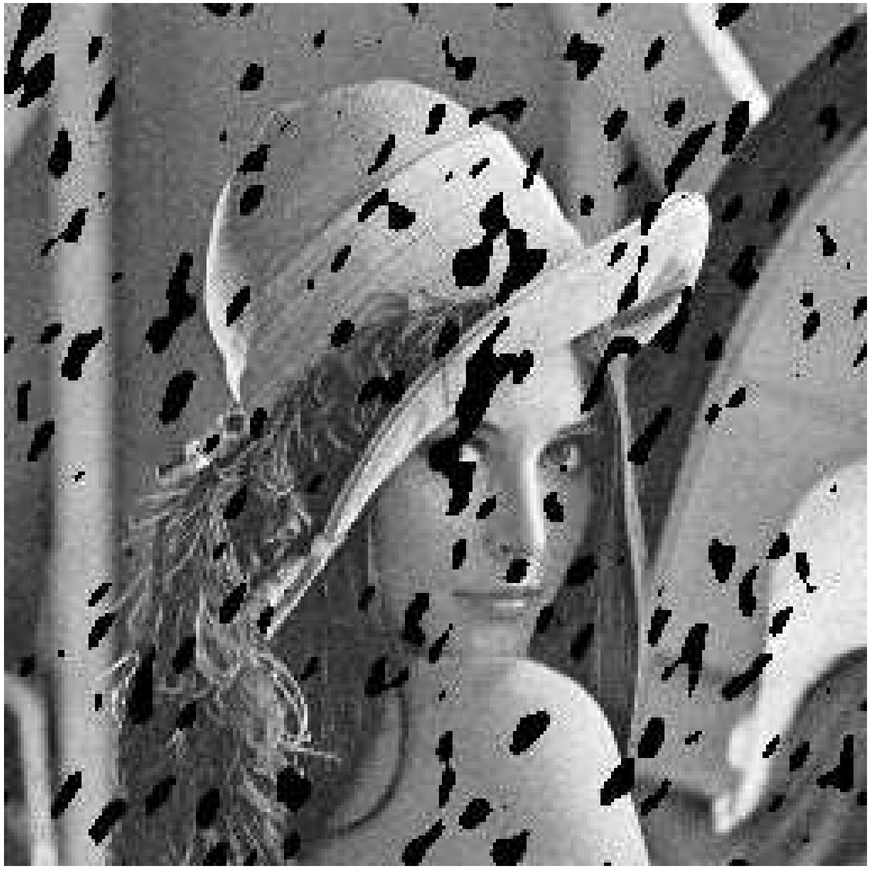
reconstructed

Missing at Random

degraded

reconstructed

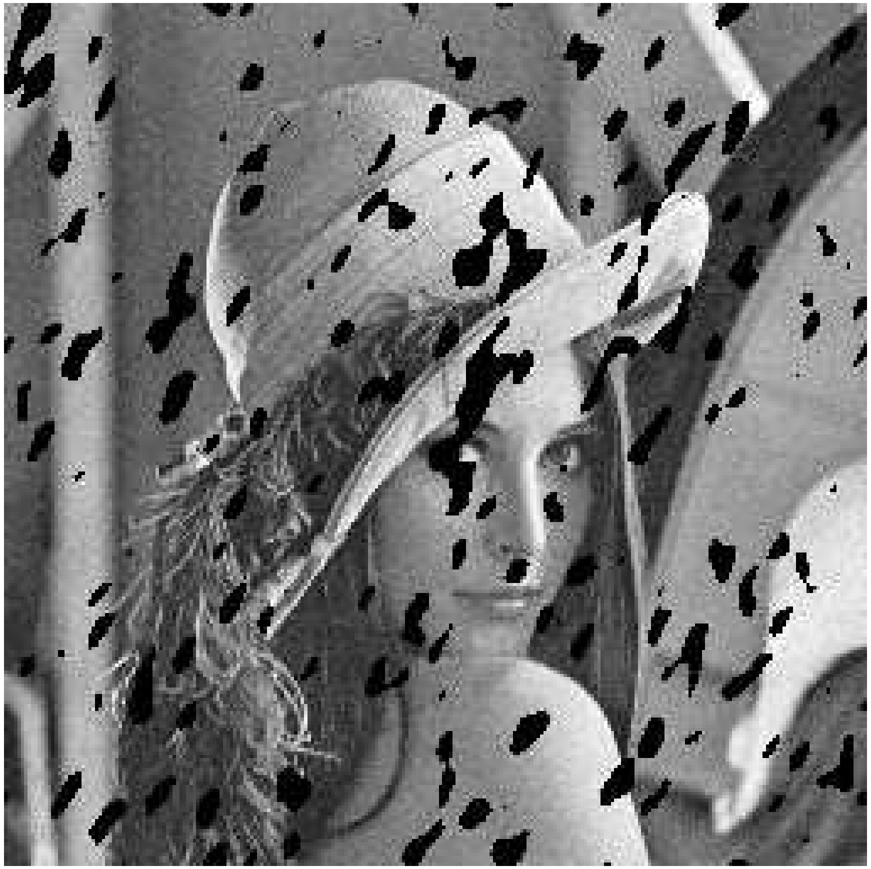
Clustered Missing Pixels – pushing beyond principles ...



degraded

reconstructed

Clustered Missing Pixels – pushing beyond principles ...



degraded

reconstructed

A Politically Correct Picture – Brad Efron in 1967-68