Is it a computing algorithm or a statistical procedure:
Can you tell or should you care?

Xiao-Li Meng

Harvard University

Lee, Li and Meng (2019) ***Likelihood-free EM: Self Consistency as a Dual Principle for Incomplete or Irregular-Pattern Data***.

Machine Learning v.s. Statistics

- Shared a Grand Task: Separating signal from noise

Machine Learning v.s. Statistics

- Shared a 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 a 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 justification, to a point of being willing to deliver anything

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.

But can you tell which is which?

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

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

But can you tell which is which?

- 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

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

But can you tell which is which?

- 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

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

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

Is it possible?

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

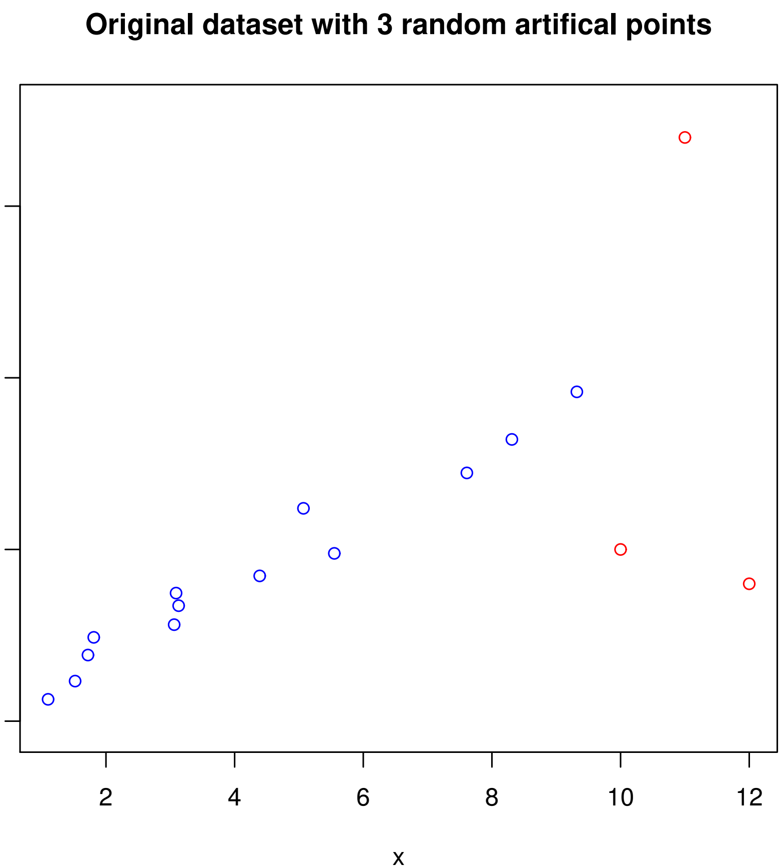
Is it possible?

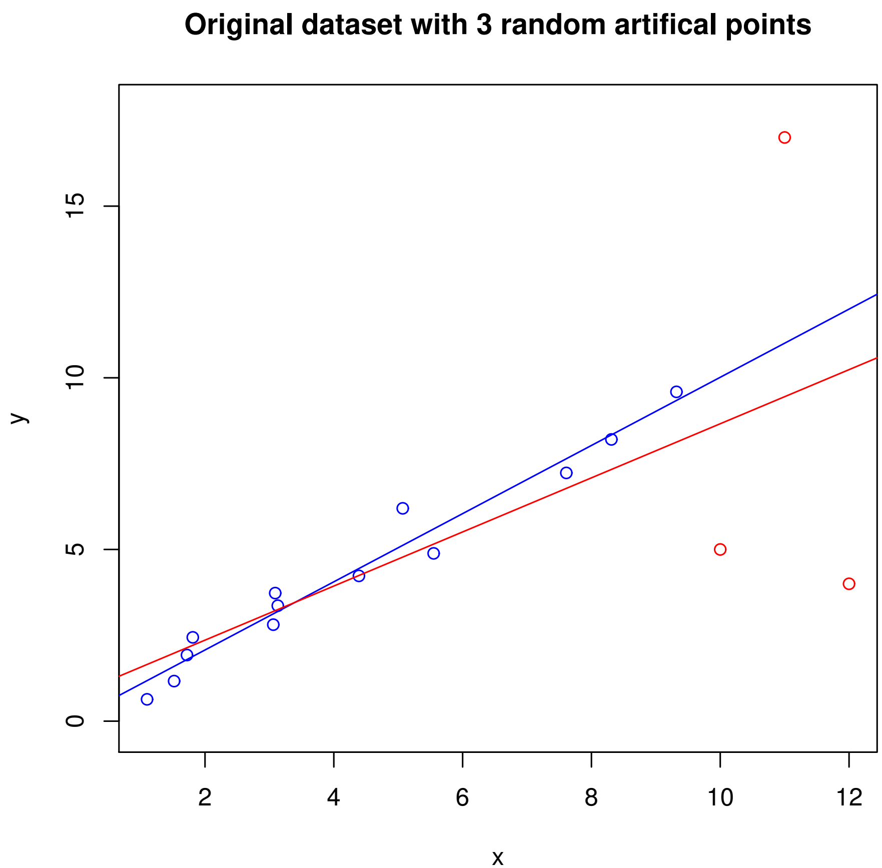
- Is it possible to use the **black box** designed for LSE with $n = 16$ to compute the LSE **exactly** with $n = 13$?
- The answer has to be **YES** because ...

Is it possible?

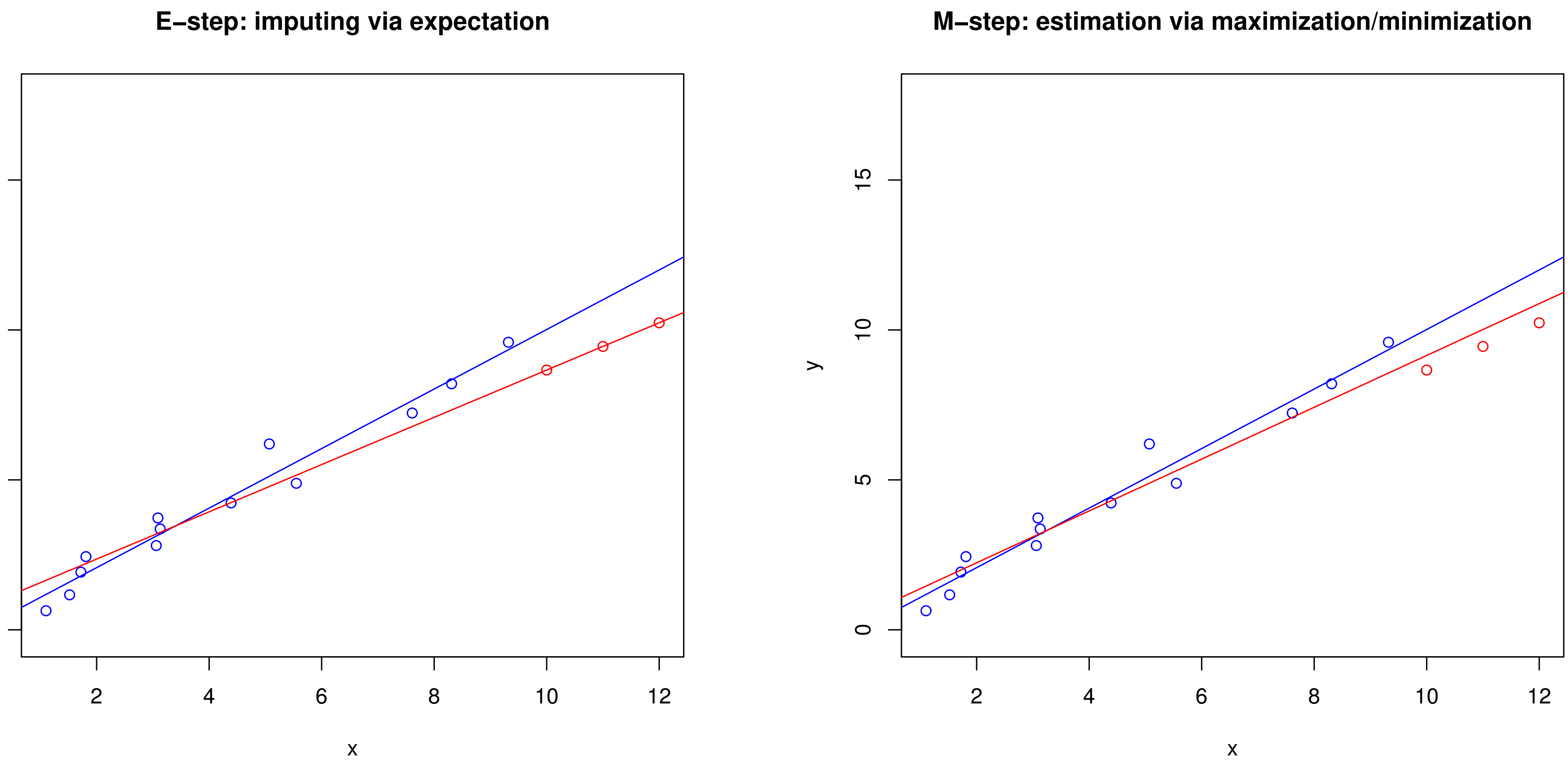
- Is it possible to use the **black box** designed for LSE with $n = 16$ to compute the LSE **exactly** with $n = 13$?
- The answer has to be **YES** because ...
- The ***Principle of Selection Bias***!

A Numerical Illustration



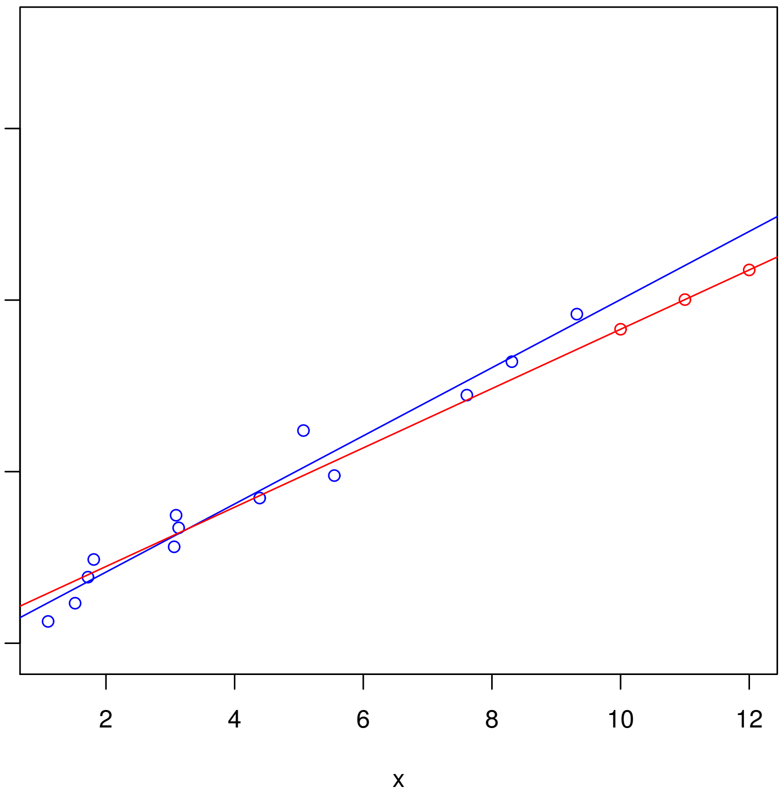


Iteration 1

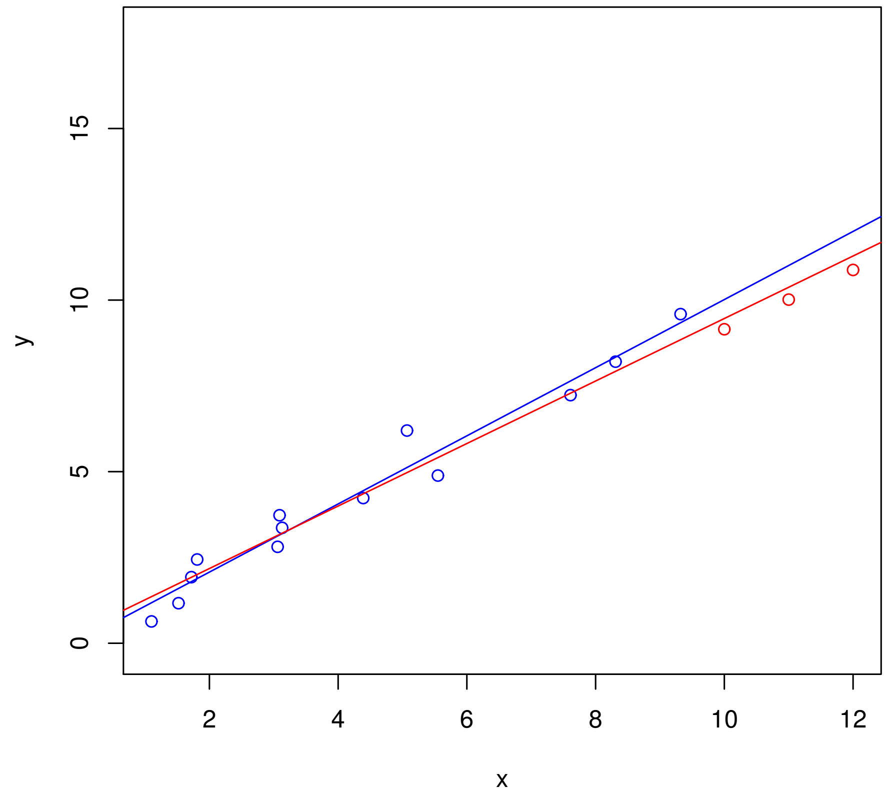

E−step: imputing via expectation
M−step: estimation via maximization/minimization
x
y
2
4
6
8
10
12
0
5
10
15

Iteration 2

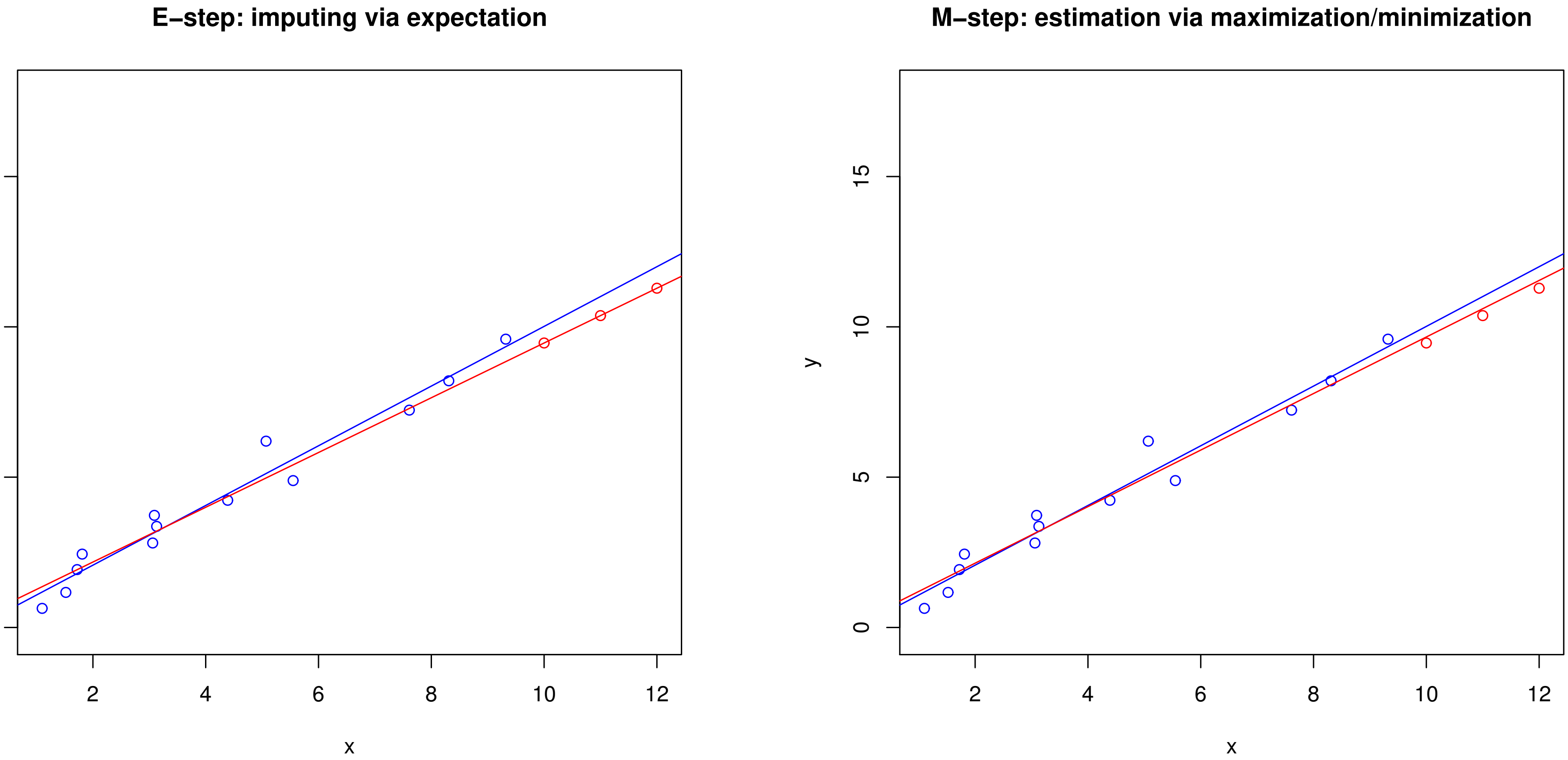
E–step: imputing via expectation



M–step: estimation via maximization/minimization

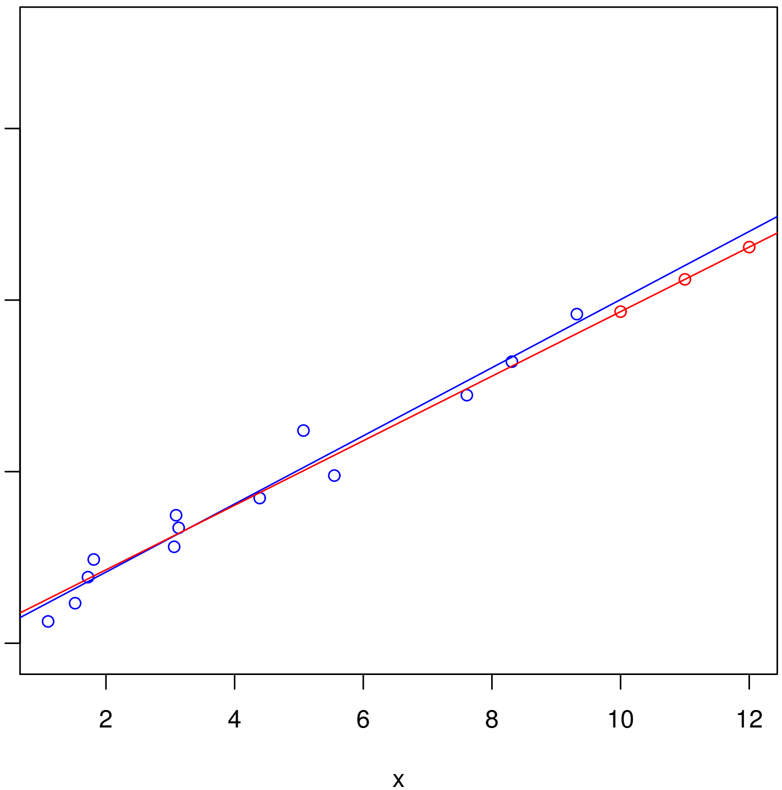


Iteration 3

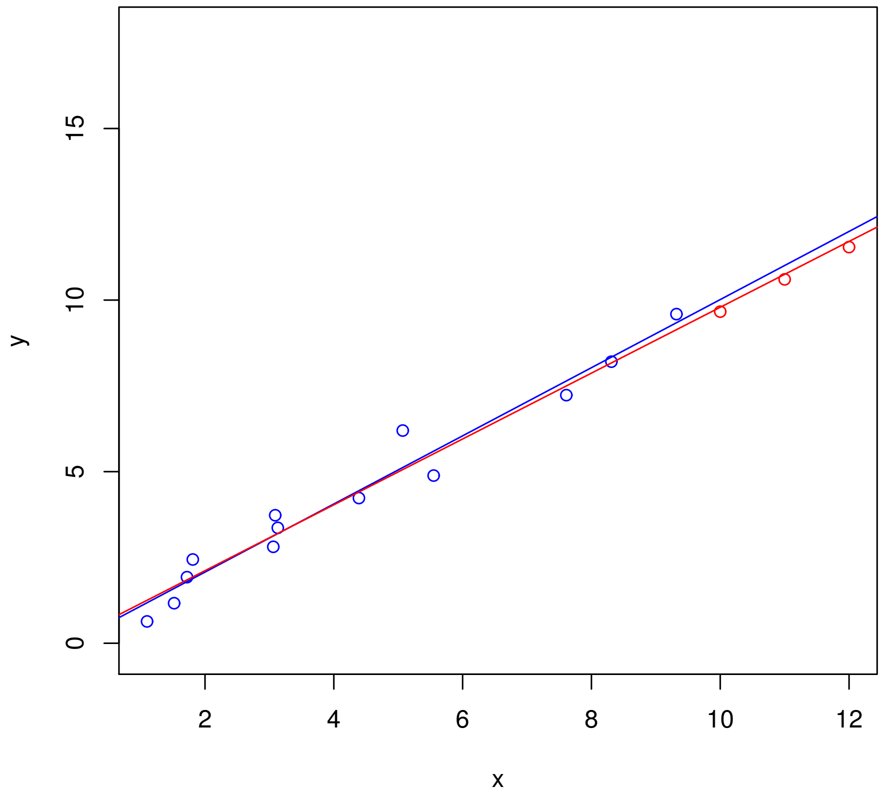

E–step: imputing via expectation
M–step: estimation via maximization/minimization
x
y
2
4
6
8
10
12
0
5
10
15

Iteration 4

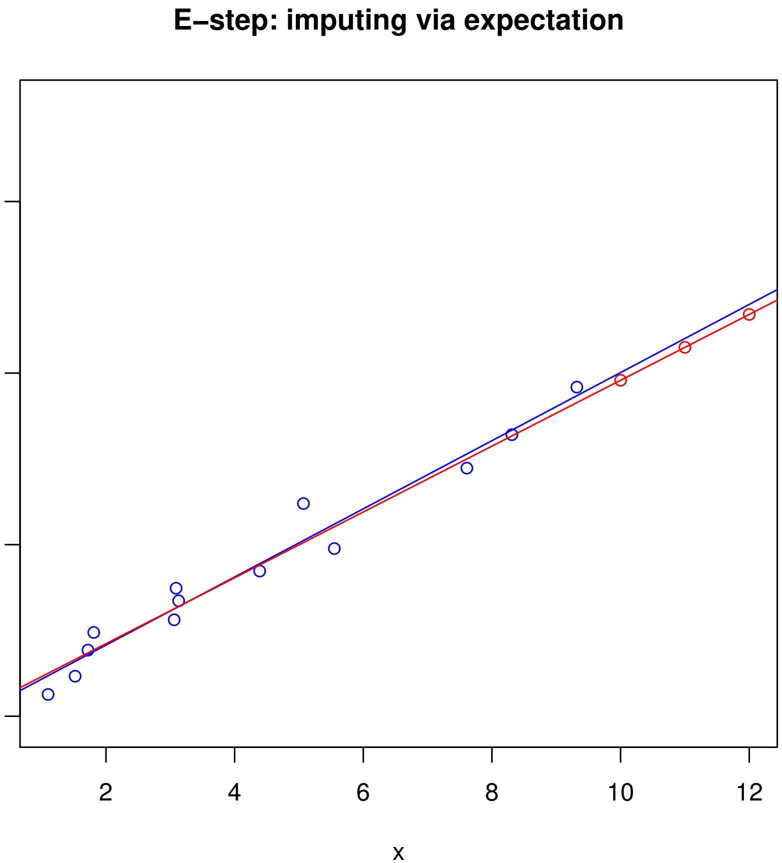
E−step: imputing via expectation

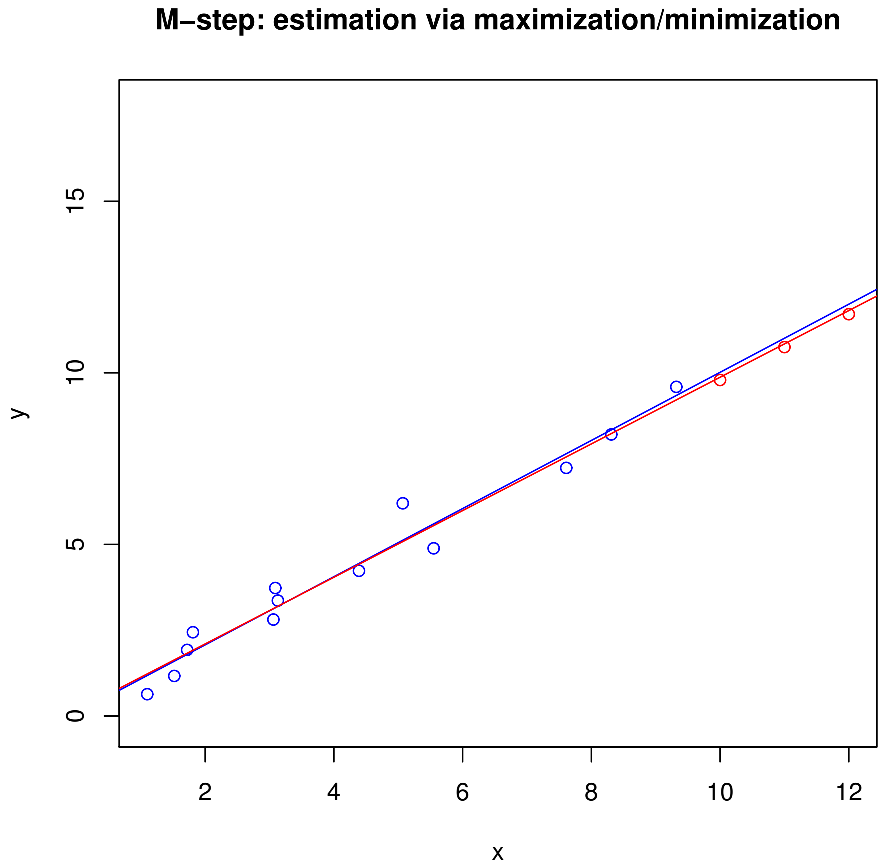


M−step: estimation via maximization/minimization

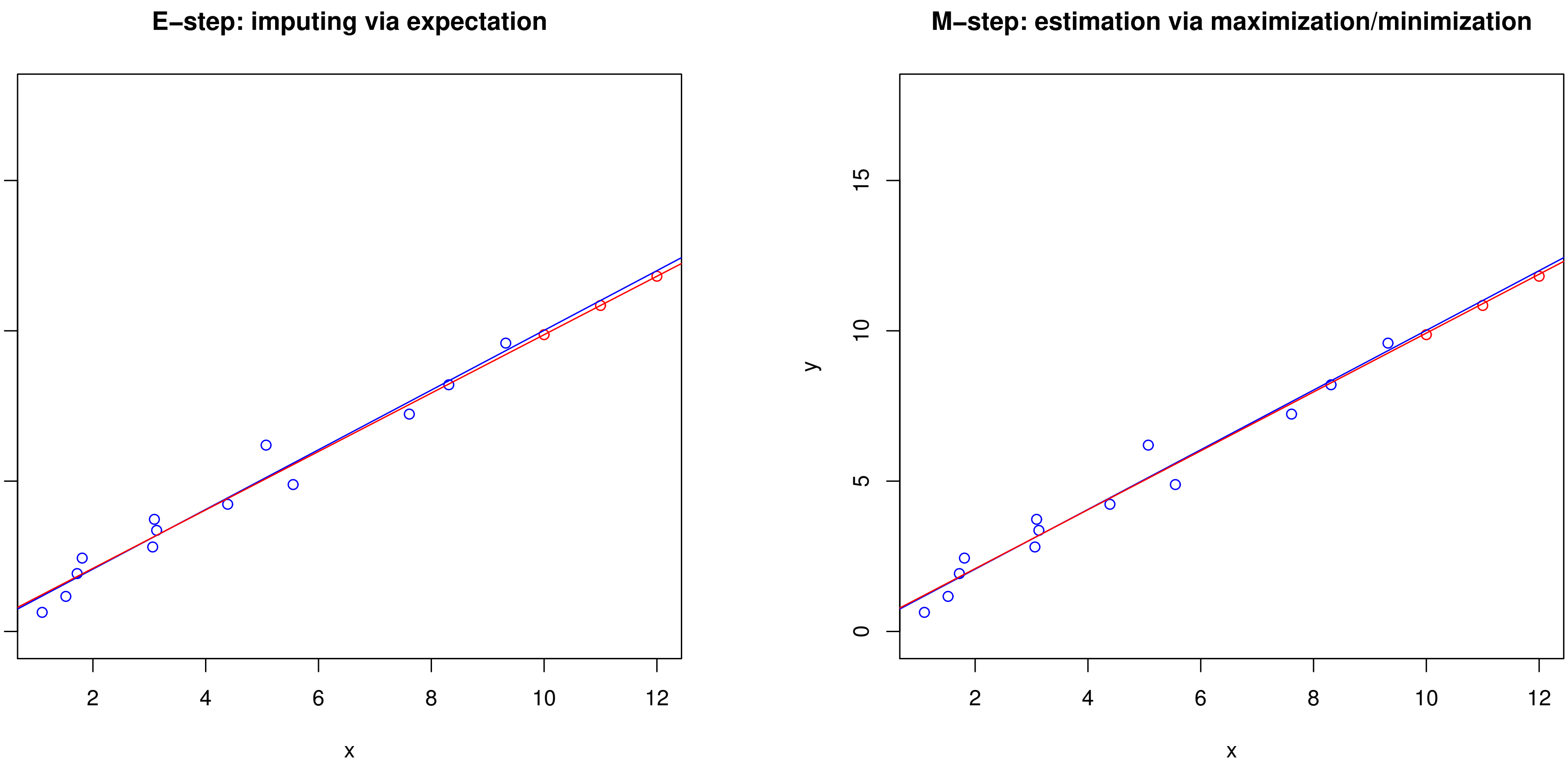


Iteration 5

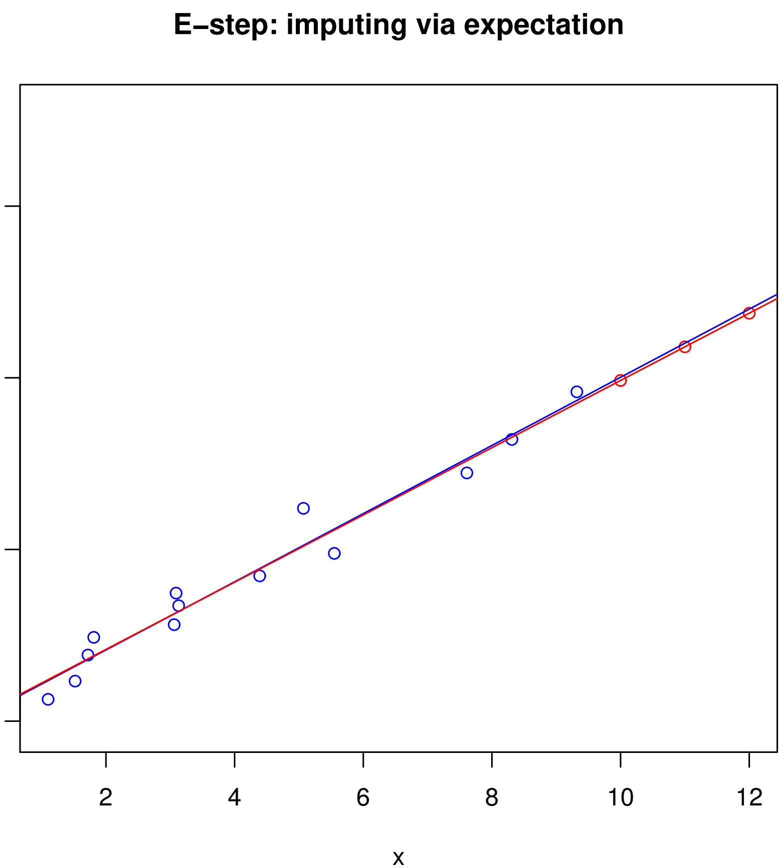

E–step: imputing via expectation
2
4
6
8
10
12
x


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

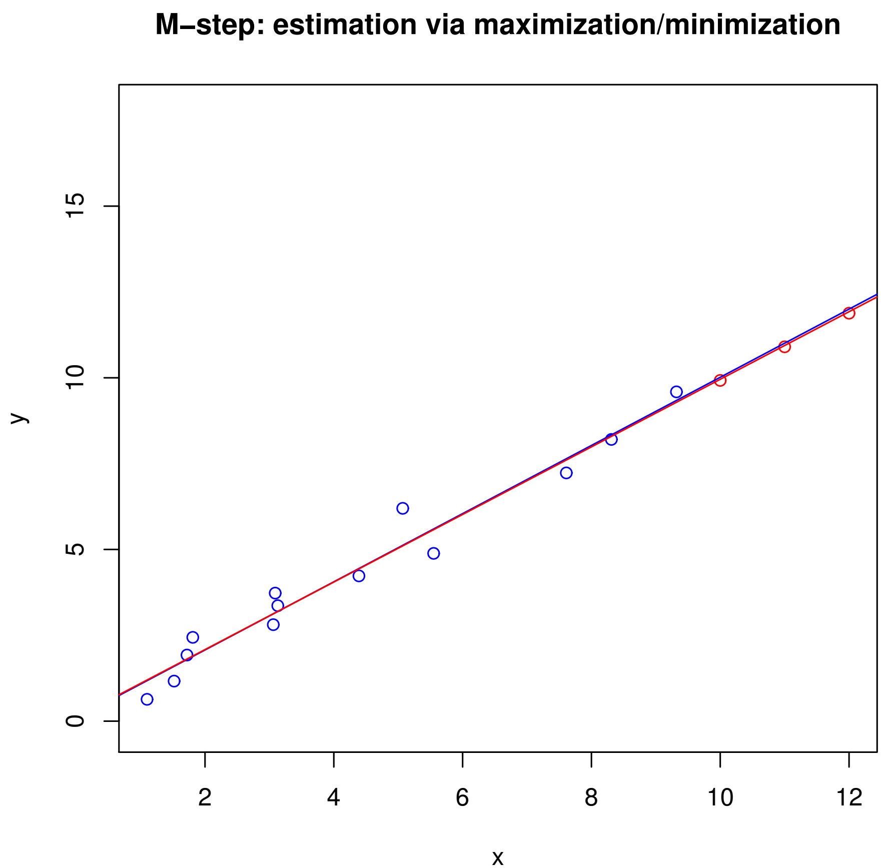
Iteration 6


E−step: imputing via expectation
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
2
4
6
8
10
12
x



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

OK, what is it?

- Is this a computing algorithm or a statistical procedure?

OK, what is it?

- Is this a computing algorithm or a statistical procedure?
- Who cares? Does it really matter?

OK, what is it?

- Is this a computing algorithm or a statistical procedure?
- Who cares? Does it really matter?
- Well, do you want to know how general it is?

If you do, then ...

- From a statistical estimation perspective: What’s the statistical principle behind it? Is it consistent? Is it (asymptotically) efficient in some sense? What assumptions on missing-data mechanism are needed to justify its validity?

If you do, then ...

- From a statistical estimation perspective: What's the statistical principle behind it? Is it consistent? 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

- $\hat{f}_{\rm com}$: estimator for f given complete data $\boldsymbol{y}_{\rm com}$

The Self-Consistency Principle

- $\hat{f}_{\text{com}}$: estimator for f given complete data $\boldsymbol{y}_{\text{com}}$
- But we only observed data $\boldsymbol{y}_{\text{obs}}$.

The Self-Consistency Principle

- $\hat{f}_{\mathrm{com}}$: estimator for f given complete data $\boldsymbol{y}_{\mathrm{com}}$
- But we only observed data $\boldsymbol{y}_{\mathrm{obs}}$.
- Intuitively, the “best" estimate of f given the procedure $\hat{\boldsymbol{f}}_{\mathrm{com}}$ and the imputation model $p(\boldsymbol{y}_{\mathrm{com}}|\boldsymbol{y}_{\mathrm{obs}}, f)$, $\hat{\boldsymbol{f}}_{\mathrm{obs}}$, should satisfy the fixed-point equation

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

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 F is an NPMLE.
- Efron (1967) shown K-M estimator $\hat{\boldsymbol{F}}_{\text{obs}}$ is *self-consistent*:

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

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 $\boldsymbol{F}$ is an NPMLE.
- Efron (1967) shown K-M estimator $\hat{\boldsymbol{F}}_{\rm obs}$ is *self-consistent*:

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

  where $\hat{\boldsymbol{F}}_{\rm com}$ 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

- Seek $\hat{\beta}_{13}$:

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

Least Squares Estimator is Self-consistent

- Seek $\hat{\beta}_{13}$:

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

- Starting with $\beta_{13}^{(0)}$, (1) impute the missing y_i by $y_i^{(t)} = \beta_{13}^{(t)} x_i$ and (2) compute

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

Least Squares Estimator is Self-consistent

- Seek $\hat{\beta}_{13}$:

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

- Starting with $\beta_{13}^{(0)}$, (1) impute the missing y_i by $y_i^{(t)}=\beta_{13}^{(t)}x_i$ and (2) compute

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

- The limit $\hat{\beta}_{13}$ satisfies

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

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

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

... 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 $I(\theta)$

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

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

... 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 $I(\theta)$

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

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

- Because of the Fisher's identity (fundamental for EM)

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

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

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

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

   (nuisance part of f can be handled differently.)

MISC: No corner cutting, but ...

- Advantages:
  1. A generic algorithm: can be applied with any complete-data *procedure*;
  2. 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. 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.

So What Are The Theoretical Guarantees?

- Use L^p norm

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

So What Are The Theoretical Guarantees?

- Use L^p norm

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

- Project $\hat{f}_{\text{com}}$ under the *conditionally expected* norm:

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

So What Are The Theoretical Guarantees?

- Use L^p norm
$$||\hat{f}_{\mathrm{com}} - f||_p = \left[E\left(\int |\hat{f}_{\mathrm{com}}(t) - f(t)|^p dt \right) \right]^{1/p}$$
- Project $\hat{f}_{\mathrm{com}}$ under the *conditionally expected* norm:
$$M(f; \boldsymbol{y}_{\mathrm{obs}}) = \mathrm{argmin}_g E\left[\int |\hat{f}_{\mathrm{com}}(t) - g(t)|^p dt \middle| \boldsymbol{y}_{\mathrm{obs}}; f \right]$$
- For $p = 2$, $M(f; \boldsymbol{y}_{\mathrm{obs}})(t) = E[\hat{f}_{\mathrm{com}}(t)|\boldsymbol{y}_{\mathrm{obs}}; f]$

So What Are The Theoretical Guarantees?

- Use L^p norm

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

- Project $\hat{f}_{\rm com}$ under the *conditionally expected* norm:

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

- For $p = 2$, $M(f; \boldsymbol{y}_{\rm obs})(t) = E[\hat{f}_{\rm com}(t) | \boldsymbol{y}_{\rm obs}; f]$
- $M(\hat{f}) \equiv M(f = \hat{f}; \boldsymbol{y}_{\rm obs})$ a map from $\mathcal{F}_{\rm obs}$—a sub-space of L^p that contains the true f_0—into itself.

The Power of Contraction Mapping

- Define $|f|_p = \left[\int |f(t)|^p dt\right]^{1/p}$. If $M(f)$ contracts on $\mathcal{F}_{\text{obs}}$ wrt $|f|_p$, then there exists a unique solution to $M(\hat{f}_{\text{obs}}) = \hat{f}_{\text{obs}}$.

The Power of Contraction Mapping

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

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

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

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

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

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 $\hat{f}_{\mathrm{obs}}$ and $\hat{f}_{\mathrm{com}}$ have the same order of *rate of convergence*, as long as we can show $M(f)$ is a contraction mapping.

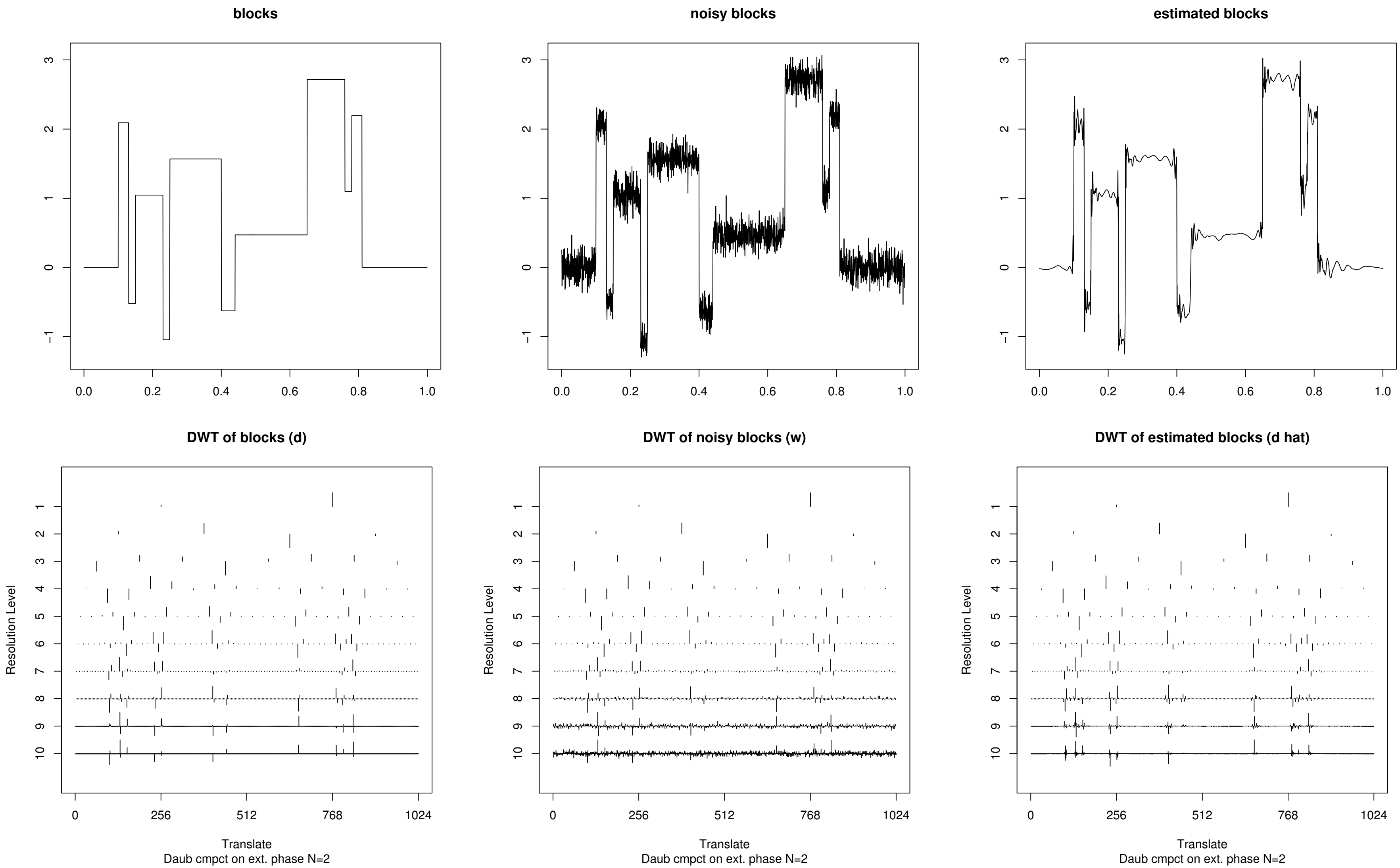
Generality and Implications

- 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, $M(f)$ is a contracting map

Generality and Implications

- The result holds for any $p \geq 1$. Important for LASSO, L^1 regressions, etc.
- Potentially a useful theoretical tool, ensuring $\hat{f}_{\rm obs}$ and $\hat{f}_{\rm com}$ 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, $M(f)$ is a contracting map
- $M(f)$ is *not* a contraction map for hard thresholding.

Wavelet Denoising (e.g., Donoho and Johnstone, 1994)



Incomplete Designs

- We observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^n$:

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

Incomplete Designs

- We observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$:

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

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

Incomplete Designs

- We observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$:

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

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^{n}$ 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

- We observe $\boldsymbol{y}_{\text{obs}} = \{x_i, y_i\}_{i=1}^{n}$:

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

- $\boldsymbol{X}_{\text{obs}} = \{x_i\}_{i=1}^{n}$ 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

- We observe $\boldsymbol{y}_{\rm obs} = \{x_i, y_i\}_{i=1}^n$:

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

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

airplane

murder in progress

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
  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

- 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 to account for the effect of the imputed $y_i^{(t)}$'s on estimating 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 to account for the effect of the imputed $y_i^{(t)}$'s on estimating 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 to account for the effect of the imputed $y_i^{(t)}$'s on estimating σ^2.
- Derived by assuming the conditional expectation

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

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

A Refined (REF) Algorithm: Much More Principled 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:

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

A Refined (REF) Algorithm: Much More Principled 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:

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

- Estimate σ by

$$
\hat{\sigma}^{(t)} = \sqrt{\frac{n_{\text{com}}}{n_{\text{obs}}}} \hat{\sigma}_{\text{com}}
$$

A Refined (REF) Algorithm: Much More Principled Corner Cutting

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

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

- Estimate σ by

$$
\hat{\sigma}^{(t)} = \sqrt{\frac{n_{\rm com}}{n_{\rm obs}}} \hat{\sigma}_{\rm com}
$$

- $\eta_l \approx C_m = 1 - \frac{n}{N}$, and $c = g(\hat{\sigma}^{(t)})$

Contracting Properties for Soft-Thresholding

- Proved: $\forall\ y_{\rm obs}$ and $\hat{w}^{(0)}$, $\exists\ \hat{\rho}_n = \rho(y_{\rm obs}, w^{(0)}) < 1$,

$$|\hat{w}^{(t+1)} - \hat{w}^{(t)}|_n \leq \hat{\rho}|\hat{w}^{(t)} - \hat{w}^{(t-1)}|_n, \quad t = 1, \ldots, \ldots \qquad (1)$$

Contracting Properties for Soft-Thresholding

- Proved: $\forall\, y_{\text{obs}}$ and $\hat{w}^{(0)}$, $\exists\, \hat{\rho}_n = \rho(y_{\text{obs}}, w^{(0)}) < 1$,

$$|\hat{w}^{(t+1)} - \hat{w}^{(t)}|_n \leq \hat{\rho}|\hat{w}^{(t)} - \hat{w}^{(t-1)}|_n, \quad t = 1, \ldots, \ldots \tag{2}$$

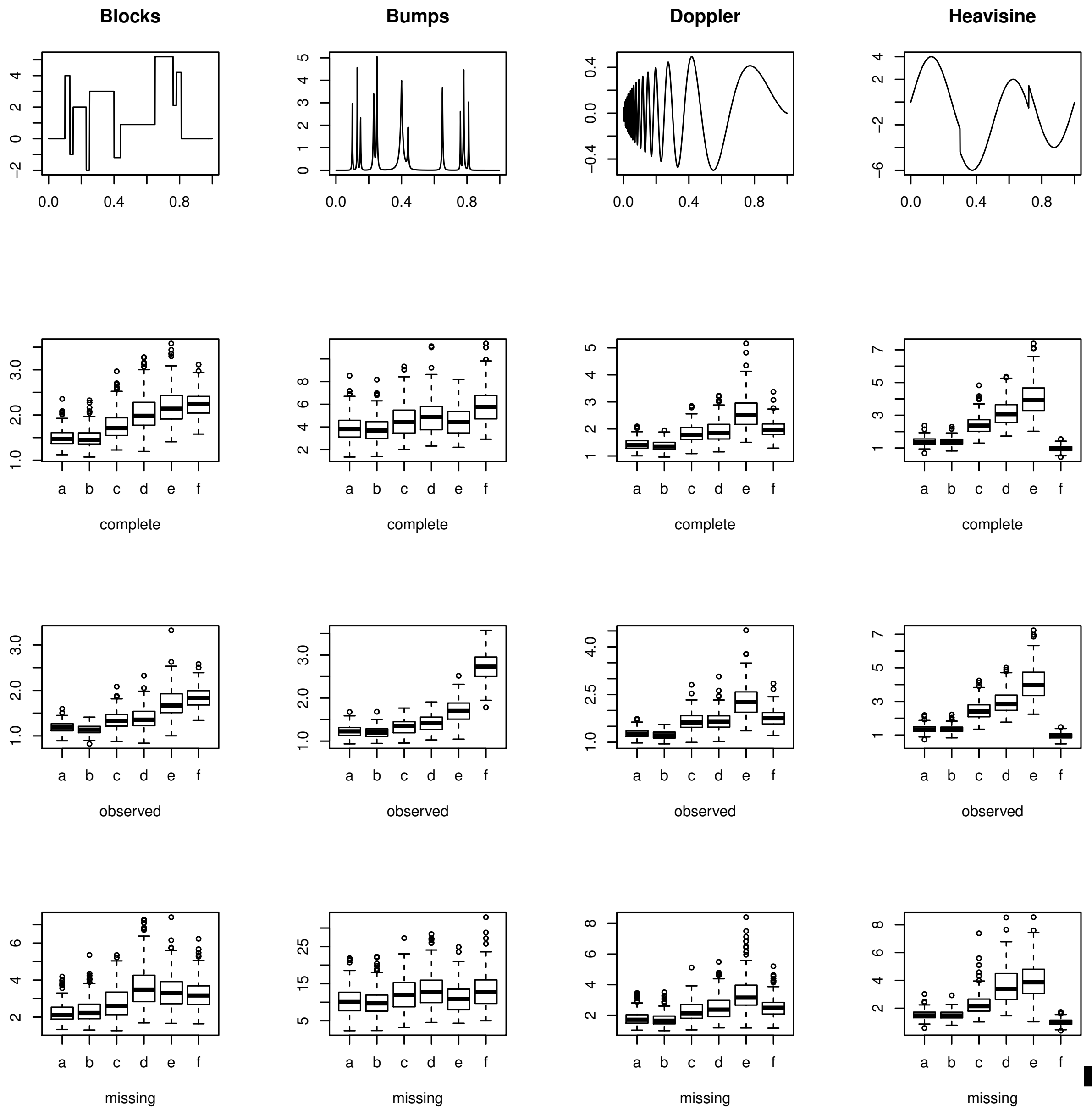
- Conjecture: $\exists\, \rho < 1$ independent of n,

$$||M(\hat{w}^{(t)}) - M(w)||_{2,n} \leq \sqrt{\rho}||\hat{w}^{(t)} - w||_{2,n}$$

Simulation Experiments

- compared 6 procedures:
  - a: MISC, L_1 norm, EBayes of Johnstone & Silverman (2005) for $\hat{f}_{\mathrm{com}}$
  - b: MISC, L_2 norm, EBayes of Johnstone & Silverman (2005) for $\hat{f}_{\mathrm{com}}$
  - c: MISC, L_1 norm, Universal Thresholding for $\hat{f}_{\mathrm{com}}$
  - d: MISC, L_2 norm, Universal Thresholding for $\hat{f}_{\mathrm{com}}$
  - e: REF with hard thresholding
  - f: REF with soft thresholding

Boxplots of MSE: snr=3, 30% missing (results similar for MAE)



Let's see how it works - Airplane

degraded

reconstructed

Murder in Progress

degraded

reconstructed

Variable Selection with Missing Data

- applied MISC to adaptive lasso

Variable Selection with Missing Data

- applied MISC to adaptive lasso
- "Stacking" method by treating the m data sets as a big one with size $N = n \times m$

Variable Selection with Missing Data

- applied MISC to adaptive lasso
- "Stacking" method by treating the m data sets as a big one with size $N = n \times m$
- conducted experiments to compare PPV and NPV:

$$\text{PPV} = \frac{\text{number of selected significant variables}}{\text{number of true significant variables}}$$

$$\text{NPV} = \frac{\text{number of removed non-significant variables}}{\text{number of true non-significant variables}}$$

Experimental Results

method		LCLM	HCLM	LCHM	HCHM	mean	weighted mean
1	PPV	85.4	67.3	85.4	67.3	76.4	87.9
	NPV	93.5	92.2	93.5	92.2	92.8	
2	PPV	80.6	56.5	72.7	46.6	64.1	86.4
	NPV	95.9	94.9	97.2	95.9	96.0	
3	PPV	88.1	70.1	83.7	64.4	76.6	84.2
	NPV	88.6	87.1	88.7	85.3	87.4	
4	PPV	79.8	55.9	71.5	45.1	63.1	86.1
	NPV	95.6	95.1	97.1	96.2	96	

- Tested 4 methods:
  1. complete data available (for benchmark comparison)
  2. MISC with L_1 norm
  3. MISC with L_2 norm
  4. Stacking

Experimental Results

method		LCLM	HCLM	LCHM	HCHM	mean	weighted mean
1	PPV	85.4	67.3	85.4	67.3	76.4	87.9
	NPV	93.5	92.2	93.5	92.2	92.8	
2	PPV	80.6	56.5	72.7	46.6	64.1	86.4
	NPV	95.9	94.9	97.2	95.9	96.0	
3	PPV	88.1	70.1	83.7	64.4	76.6	84.2
	NPV	88.6	87.1	88.7	85.3	87.4	
4	PPV	79.8	55.9	71.5	45.1	63.1	86.1
	NPV	95.6	95.1	97.1	96.2	96	

- Tested 4 methods:
  1. complete data available (for benchmark comparison)
  2. MISC with L_1 norm
  3. MISC with L_2 norm
  4. Stacking
- LC/HC: low/high correlation in X; LM/HM: low/high missing %

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.

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.
- Provided an initial unified theory via contraction mapping.

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.
- Provided an initial unified theory via contraction mapping.
- Obtained Refined Algorithm as a 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.
- Provided an initial unified theory via contraction mapping.
- Obtained Refined Algorithm as a good compromise between statistical and computational efficiency for wavelet applications.
- BUT, there are a lot more to be done ...

A Telescopic, Microscopic, and Kaleidoscopic View of Data Science

HDSR

HARVARD DATA SCIENCE REVIEW