toweranNA, a Novel, Prediction-Oriented R Package for Missing Values

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

R/FInance 2019, Chicago

Overview

Missing values (MVs):

- A perennial headache.
- Vast, VAST literature.
- Major R packages, e.g. **mice** and **Amelia**.
- New CRAN Task View, already quite extensive.

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*.

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR, e.g. multvar. normal.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR, e.g. multvar. normal.
- Time for a new paradigm!

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR, e.g. multvar. normal.
- Time for a new paradigm!
- We're interested in *prediction*.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR, e.g. multvar. normal.
- Time for a new paradigm!
- We're interested in *prediction*.
- We'll present a novel new technique we call the Tower Method.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR, e.g. multvar. normal.
- Time for a new paradigm!
- We're interested in *prediction*.
- We'll present a novel new technique we call the Tower Method.
- Non-imputational.

Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR, e.g. multvar. normal.
- Time for a new paradigm!
- We're interested in *prediction*.
- We'll present a novel new technique we call the Tower Method.
- Non-imputational.
- Available at `http://github.com/matloff/toweranNA`.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Theorem from Probability Theory

[Please be patient; R code and real-data examples soon. :-)]

Famous formula in probability theory:

$$EY = E[E(Y|X)] = E[g(X)]$$

Here $g()$ is regression function of Y on X.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Theoretical Background for Use in MVs

Theoretical Background for Use in MVs

- (Matloff, *Biometrika*, 1981)

Theoretical Background for Use in MVs

- (Matloff, *Biometrika*, 1981)
- My first published stat paper!

Theoretical Background for Use in MVs

- (Matloff, *Biometrika*, 1981)
- My first published stat paper!

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Theory Background (cont'd.)

Norm Matloff University of California, Davis

Pete Mohanty Stanford University

Theory Background (cont'd.)

- My context: Est. E(Y).

$$\widehat{EY} = \frac{1}{n} \sum_{i=1}^{n} \widehat{g}(X_i)$$

Here $\widehat{g}$ comes from linear model, logit, nonpar.

Theory Background (cont'd.)

- My context: Est. E(Y).

$$\widehat{EY} = \frac{1}{n} \sum_{i=1}^{n} \widehat{g}(X_i)$$

  Here $\widehat{g}$ comes from linear model, logit, nonpar.
  Maybe some Y_i missing; even if not, get smaller asympt. var.

- Steady stream of theory papers since then from various authors.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Theory Background (cont'd.)

- My context: Est. E(Y).

$$\widehat{EY} = \frac{1}{n} \sum_{i=1}^{n} \widehat{g}(X_i)$$

  Here $\widehat{g}$ comes from linear model, logit, nonpar.
  Maybe some Y_i missing; even if not, get smaller asympt. var.

- Steady stream of theory papers since then from various authors.
- E.g. (U. Müller, *Annals of Stat.*, 2009).

Norm Matloff University of California, Davis

Pete Mohanty Stanford University

Theory Background (cont'd.)

- My context: Est. E(Y).

$$\widehat{EY} = \frac{1}{n} \sum_{i=1}^{n} \widehat{g}(X_i)$$

  Here $\widehat{g}$ comes from linear model, logit, nonpar.
  Maybe some Y_i missing; even if not, get smaller asympt. var.
- Steady stream of theory papers since then from various authors.
- E.g. (U. Müller, *Annals of Stat.*, 2009).
- But all theoretical. Not used (or even known) by practitioners.

Tower Property

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Tower Property

More general version, known as the Tower Property:

$$E[E(Y|U,V)|U] = E(Y|U)$$

Tower Property

More general version, known as the Tower Property:

$$E[E(Y|U,V)|U] = E(Y|U)$$

Why is this relevant to us?

- Y: variable to be predicted
- U: vector of known predictor values
- V: vector of uknown predictor values

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (**prgeng** in pkg).

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (**prgeng** in pkg).
- Predict Y = wage income. In one particular case to be predicted, we might have

Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (**prgeng** in pkg).
- Predict Y = wage income. In one particular case to be predicted, we might have
    - U = (education,occupation,weeks worked)
    - V = (age,gender)

  In another case, maybe U = (age,gender,education,weeks worked) and V = (occupation). Etc.

Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (**prgeng** in pkg).
- Predict Y = wage income. In one particular case to be predicted, we might have
  - U = (education,occupation,weeks worked)
  - V = (age,gender)

  In another case, maybe U = (age,gender,education,weeks worked) and V = (occupation). Etc.
- Wish we had U,V, for prediction E(Y|U,V), but forced to use E(Y|U).
- But then must estimate many E(Y | U), since many different patterns for MVs (2^5 here).

Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (**prgeng** in pkg).
- Predict Y = wage income. In one particular case to be predicted, we might have
  - U = (education,occupation,weeks worked)
  - V = (age,gender)

  In another case, maybe U = (age,gender,education,weeks worked) and V = (occupation). Etc.
- Wish we had U,V, for prediction E(Y|U,V), but forced to use E(Y|U).
- But then must estimate many E(Y | U), since many different patterns for MVs (2^5 here).
- Hard enough to fit one good model, let alone dozens or more.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (**prgeng** in pkg).
- Predict Y = wage income. In one particular case to be predicted, we might have
  - U = (education,occupation,weeks worked)
  - V = (age,gender)

  In another case, maybe U = (age,gender,education,weeks worked) and V = (occupation). Etc.
- Wish we had U,V, for prediction E(Y|U,V), but forced to use E(Y|U).
- But then must estimate many E(Y | U), since many different patterns for MVs (2^5 here).
- Hard enough to fit one good model, let alone dozens or more.
- With Tower, need only one.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Tower (cont'd.)

Basic idea:

- Fit full regression model to the complete cases.
- Use Tower to get the marginal models from the full one:

$$\widehat{E}(Y \mid U = s) = \text{avg. } \underbrace{\widehat{E}(Y \mid U = s, V)}_{\text{full model}}$$

  over all complete cases with $U = s$
- In practice, use $U \approx s$ instead of $U = s$, using k nearest neighbors.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Tower (cont'd.)

Basic idea:

- Fit full regression model to the complete cases.
- Use Tower to get the marginal models from the full one:

$$\widehat{E}(Y \mid U = s) = \text{avg. } \underbrace{\widehat{E}(Y \mid U = s, V)}_{\text{full model}}$$

  over all complete cases with $U = s$
- In practice, use $U \approx s$ instead of $U = s$, using k nearest neighbors.
  In practice, $k = 1$ usually fine;

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Tower (cont'd.)

Basic idea:

- Fit full regression model to the complete cases.
- Use Tower to get the marginal models from the full one:

$$\widehat{E}(Y \mid U = s) = \text{avg. } \underbrace{\widehat{E}(Y \mid U = s, V)}_{\text{full model}}$$

  over all complete cases with $U = s$
- In practice, use $U \approx s$ instead of $U = s$, using k nearest neighbors.
  In practice, $k = 1$ usually fine; fitted values already smoothed, don't need more smoothing.

Census Example (cont'd.)

(a) Use, say, **lm()** on the complete cases, predicting wage income from (age,gender,education,occupation,weeks worked).

(b) Save the fitted values, e.g. **fitted.values** from **lm()** output.

(c) Say need to predict case with education = MS, occupation = 102, weeks worked = 52 but with age and gender missing.

(d) Find the complete cases for which (education,occupation,weeks worked) = (MS,102,52).

(e) Predicted value for this case is average of the fitted values for the cases in (d).

toweranNA Package API

- **toweranNA(x,fittedReg,k,newx,scaleX=TRUE)**
  - **x:** Data frame of complete cases.
  - **fittedReg:** Estimated values of full regress. ftn. at those cases (from **lm()**, **glm()**, random forests, neural nets, whatever).
  - **k:** Number of nearest neighbors.
  - **newx:** Data frame of new cases to be predicted.
  - Return value: Vector of predictions.

Other Major Functions

Norm Matloff University of California, Davis

Pete Mohanty Stanford University

Other Major Functions

- **towerLM(x,y,k,newx,useGLM=FALSE)**
  Wrapper for **toweranNA()**.
- **towerTS(x,lag,k)**
  Adaptation of Tower Method for time series; see below.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Structure of Examples

- 3 real datasets.
- Break into random training and test sets.
- Predict all test-set cases with at least one MV.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Example: WordBank Data

- Kids' vocabulary growth trajectories.
- About 5500 cases, 6 variables. About 29% MVs.

Mean Absolute Prediction Errors:

Amelia	Tower
102.7	96.2
122.9	119.9
89.4	88.1
115.3	107.0
111.1	102.5

- Times about 6s each.
- The **mice** package crashed.

UCI Bank Data

- About 50K cases.
- Only about 2% MVs. Not much need for MV methods, but let's make sure Tower doesn't bring harm. :-)
- Tower run 8.3s, **mice** 442.2s.
- Too long to do multiple runs. About the same accuracy, 0.92 or 0.93.
- **Amelia** crashed.

World Values Study

- World political survey.
- 48 countries, sample 500-3500 from each.
- MVs artifically added.
- Tower outperformed **mice** in 39 of 48 countries.

	Tower	*Mice*
Mean Absolute Predictive Error	1.7603	1.8270
Elapsed Time (seconds)	0.1825	14.0822

Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.
- Precise def. of MAR tricky (Seaman, *Stat. Sci.*, 2013).

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.
- Precise def. of MAR tricky (Seaman, *Stat. Sci.*, 2013).
- Tower assumptions similar, but assumptions matter much less in prediction than in estimation.

Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.
- Precise def. of MAR tricky (Seaman, *Stat. Sci.*, 2013).
- Tower assumptions similar, but assumptions matter much less in prediction than in estimation.
- **Amelia**, **mice** assume X multvar. normal, very distorting.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

What about Time Series?

- How adapt toweranNA to time series?

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

What about Time Series?

- How adapt toweranNA to time series?
- Predict X_i from $X_{i-1}, X_{i-2}, ..., X_{i-m}$, lag m.
- E.g. lag 3:

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

What about Time Series?

- How adapt toweranNA to time series?
- Predict X_i from $X_{i-1}, X_{i-2}, ..., X_{i-m}$, lag m.
- E.g. lag 3:
  x_1, NA, NA, NA, x_5, x_6, x_7, x_8, x_9, x_{10}, NA, NA becomes

x_1	NA	NA	NA
x_5	x_6	x_7	x_8
x_9	x_{10}	NA	NA
...	...	...	...

Columns 1-3 are "X", col. 4 is "Y".
Then use Tower on this data frame.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Time Series (cont'd.)

Norm Matloff University of California, Davis

Pete Mohanty Stanford University

Time Series (cont'd.)

- A work in progress.

Time Series (cont'd.)

- A work in progress.
- Example: NH4 data in imputeTS package.

Time Series (cont'd.)

- A work in progress.
- Example: NH4 data in imputeTS package.
- Mean Absolute Prediction Error:
  na.ma (based on moving avg.): 1.51
  towerTS: 1.37

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Future Work

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Future Work

- Most pressing issue: May have too few (or no) complete cases.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Future Work

- Most pressing issue: May have too few (or no) complete cases.
- Solution: Relax our "one size fits all" structure.

Future Work

- Most pressing issue: May have too few (or no) complete cases.
- Solution: Relax our "one size fits all" structure.
- Instead of generating all marginal regression functions from one full one, have several "almost-full" ones.

Norm Matloff
University of California, Davis

Pete Mohanty
Stanford University

Future Work

- Most pressing issue: May have too few (or no) complete cases.
- Solution: Relax our "one size fits all" structure.
- Instead of generating all marginal regression functions from one full one, have several "almost-full" ones.
- E.g. have $p = 5$ predictors. Maybe fit four 4-predictor models. Each would be based on more complete cases than the 5-predictor models.