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. multivar. normal.
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. multivar. 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. multivar. normal.
- Time for a new paradigm!
- We’re interested in *prediction*. 
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. multivar. normal.
- Time for a new paradigm!
- We’re interested in *prediction*.
- We’ll present a novel new technique we call the Tower Method.
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. multivar. 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. multivar. 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.
Theorem from Probability Theory

Famous formula in probability theory:

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

Here \( g() \) is regression function of \( Y \) on \( X \).
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!
Theory Background (cont’d.)
Theory Background (cont’d.)

- My context: Est. $E(Y)$.

$$\hat{E}Y = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(X_i)$$

Here $\hat{g}$ comes from linear model, logit, nonpar.
Theory Background (cont’d.)

- My context: Est. \( E(Y) \).

\[
\hat{E}Y = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(X_i)
\]

Here \( \hat{g} \) 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.
Theory Background (cont’d.)

- My context: Est. E(Y).

\[ \hat{E}_Y = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(X_i) \]

Here \( \hat{g} \) 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.

Theory Background (cont’d.)

• My context: Est. E(Y).

\[ \hat{E}Y = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(X_i) \]

Here \( \hat{g} \) 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.


• But all theoretical. Not used (or even known) by practitioners.
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 unknown predictor values
More general version, known as the Tower Property:

\[ E[E(Y|U, V)|U] = E(Y|U) \]
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 unknown predictor values
Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 ($\texttt{prgeng}$ in pkg).
Example: Census Data

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

- \(U = (\text{education, occupation, weeks worked})\)
- \(V = (\text{age, gender})\)

In another case, maybe \(U = (\text{age, gender, education, weeks worked})\) and \(V = (\text{occupation})\). 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.
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
  - $U = (\text{education}, \text{occupation}, \text{weeks worked})$
  - $V = (\text{age}, \text{gender})$

In another case, maybe $U = (\text{age}, \text{gender}, \text{education}, \text{weeks worked})$ and $V = (\text{occupation})$. 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
  - $U = (\text{education}, \text{occupation}, \text{weeks worked})$
  - $V = (\text{age}, \text{gender})$

In another case, maybe $U = (\text{age}, \text{gender}, \text{education}, \text{weeks worked})$ and $V = (\text{occupation})$. 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
  - $U = (\text{education, occupation, weeks worked})$
  - $V = (\text{age, gender})$

In another case, maybe $U = (\text{age, gender, education, weeks worked})$ and $V = (\text{occupation})$. 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.
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
  - \( U = (\text{education,occupation,weeks worked}) \)
  - \( V = (\text{age,gender}) \)

In another case, maybe \( U = (\text{age,gender,education,weeks worked}) \) and \( V = (\text{occupation}) \). 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.
Tower (cont’d.)

Basic idea:

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

\[
\hat{E}(Y \mid U = s) = \text{avg. } \hat{E}(Y \mid U = s, V) \quad \text{full model}
\]

over all complete cases with \( U = s \)

- In practice, use \( U \approx s \) instead of \( U = s \), using \( k \) nearest neighbors.
Basic idea:

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

\[ \hat{E}(Y \mid U = s) = \text{avg.} \left( \hat{E}(Y \mid U = s, V) \right) \]

over all complete cases with \( U = s \)

3. In practice, use \( U \approx s \) instead of \( U = s \), using \( k \) nearest neighbors.
   In practice, \( k = 1 \) usually fine;
Basic idea:

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

\[ \hat{E}(Y \mid U = s) = \text{avg.} \hat{E}(Y \mid U = s, V) \]

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, \texttt{lm()} on the complete cases, predicting wage income from (age,gender,education,occupation,weeks worked).

(b) Save the fitted values, e.g. \texttt{fitted.values} from \texttt{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
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.
Structure of Examples

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

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

Mean Absolute Prediction Errors:

<table>
<thead>
<tr>
<th></th>
<th>Amelia</th>
<th>Tower</th>
</tr>
</thead>
<tbody>
<tr>
<td>01.2</td>
<td>96.2</td>
<td></td>
</tr>
<tr>
<td>22.9</td>
<td>119.9</td>
<td></td>
</tr>
<tr>
<td>89.4</td>
<td>88.1</td>
<td></td>
</tr>
<tr>
<td>115.3</td>
<td>107.0</td>
<td></td>
</tr>
<tr>
<td>111.1</td>
<td>102.5</td>
<td></td>
</tr>
</tbody>
</table>

- 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 artificially added.
- Tower outperformed mice in 39 of 48 countries.

<table>
<thead>
<tr>
<th></th>
<th>Tower</th>
<th>Mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Predictive Error</td>
<td>1.7603</td>
<td>1.8270</td>
</tr>
<tr>
<td>Elapsed Time (seconds)</td>
<td>0.1825</td>
<td>14.0822</td>
</tr>
</tbody>
</table>
Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.
Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.
Concerning Assumptions

- Most MV methods assume MAR, Missing at Random.
- Tower assumptions similar, but assumptions matter much less in prediction than in estimation.
Concerning Assumptions

• Most MV methods assume MAR, Missing at Random.
• Tower assumptions similar, but assumptions matter much less in prediction than in estimation.
• *Amelia*, *mice* assume $X$ multvar. normal, very distorting.
What about Time Series?

- How adapt toweranNA to time series?
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:
What about Time Series?

- How adapt toweranNA to time series?
- Predict \( X_i \) from \( X_{i-1}, X_{i-2}, \ldots, X_{i-m} \), 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

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>NA</th>
<th>NA</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_5 )</td>
<td>( x_6 )</td>
<td>( x_7 )</td>
<td>( x_8 )</td>
</tr>
<tr>
<td>( x_9 )</td>
<td>( x_{10} )</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Columns 1-3 are “X”, col. 4 is “Y”. Then use Tower on this data frame.
Time Series (cont’d.)
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:
  \texttt{na.ma} (based on moving avg.): 1.51
  \texttt{towerTS}: 1.37
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.
Future Work

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