Missing Data in Prediction Models: Pattern Mixture Kernel Submodel

Jeffrey D. Blume, PhD

joint work with Sarah Fletcher Mercaldo, PhD (NASA)

Departments of Biostatistics and Biomedical Informatics
Vanderbilt University

Netherlands, November 2019

The Problem

How do you find the 'best' prediction model when

- the data at hand has missingness
- future predictors may be missing
- the goal is to maximize predictive accuracy

Introduction

- Missingness occurs at the
  - Model construction step (in-sample)
  - Prediction step (out-of-sample)
- Multiple imputation (MI) methods
  - Well-studied for model construction
  - Hard to implement in prediction step
- Solution: Pattern Mixture Kernel Submodels (PMKS)
  - No imputation step (but related to MI)
  - Easy to construct and apply
  - Minimizes expected prediction error (EPE)
  - At least as predictive as standard methods
  - Unique model for every missing data pattern
  - Sacrifices (some) easy of interpretation

Background

How to predict for an out of sample individual with missing data?

	Imputation Requires
Zero Imputation	- Nothing
Mean Imputation	- Univariate means
MI	- Original data/Conditional distribution to make multiple draws - Computer/Imputation engine
CCS	- CCS to be fit in the model building phase
PMKS	- PMKS to be fit in the model building phase

Complete Case Submodels (CCS) - Fit submodels with all data
Pattern Mixture Kernel Submodel (PMKS) - Fit submodels by pattern

Pros and Cons

	Pros	Cons
Zero Imputation	- No computation time	- Zero may not be an appropriate value - Probably results in incorrect predictions
Mean Imputation	- No computation time	- May result in incorrect predictions for the non- average individual
MI	- Established method - Works when data are MAR	- Computation time - Not viable in the clinic
CCS	- No computation time - May be advantageous if data are MAR	- Large bias/variance tradeoff for MNAR
PMKS	- No computation time - Works for any missingness mechanism	- May be less efficient if data MAR - Patterns w/ few members may not fit well

Pattern Mixture Kernel Submodels (PMKS)

- Response (Y), Covariates (X_1, X_2), Missing Data Indicators (M_1, M_2)

$$\begin{aligned}
\text{Pattern 1: } & E[Y|X_1, X_2, M_1 = 0, M_2 = 0] = \gamma_{0,1} + \gamma_{1,1}X_1 + \gamma_{2,1}X_2 \\
\text{Pattern 2: } & E[Y|X_1, X_2, M_1 = 1, M_2 = 0] = \gamma_{0,2} + \gamma_{2,2}X_2 \\
\text{Pattern 3: } & E[Y|X_1, X_2, M_1 = 0, M_2 = 1] = \gamma_{0,3} + \gamma_{1,3}X_1 \\
\text{Pattern 4: } & E[Y|X_1, X_2, M_1 = 1, M_2 = 1] = \gamma_{0,4}
\end{aligned}$$

- CCS vs. PMKS

Minimizing the Expected Prediction Error

Minimizing the expected prediction error in each pattern will, in turn, minimize the overall expected prediction error. To see this, note that:

$$E_{Y|X}[L(Y, \hat{f}(\mathbf{X}))] = E_M \left[E_{Y|\mathbf{X},\mathbf{M}} \left[L\left(Y, \hat{f}_m\right) \right] \right]$$
$$= \sum_M P(M) E_{Y|\mathbf{X},\mathbf{M}} \left[L\left(Y, \hat{f}(\mathbf{X}, \mathbf{M})\right) \right]$$

where $\hat{f}_m = \hat{f}_m(\mathbf{X}, \mathbf{M})$. Hence, selecting $\hat{f}_m$ to minimize the pattern specific expected loss, $E_{Y|X,M} \left[L\left(Y, \hat{f}_m(\mathbf{X}, \mathbf{M})\right) \right]$, will in turn minimize the overall loss $E_{Y|X}[L(Y, \hat{f}(\mathbf{X}))]$.

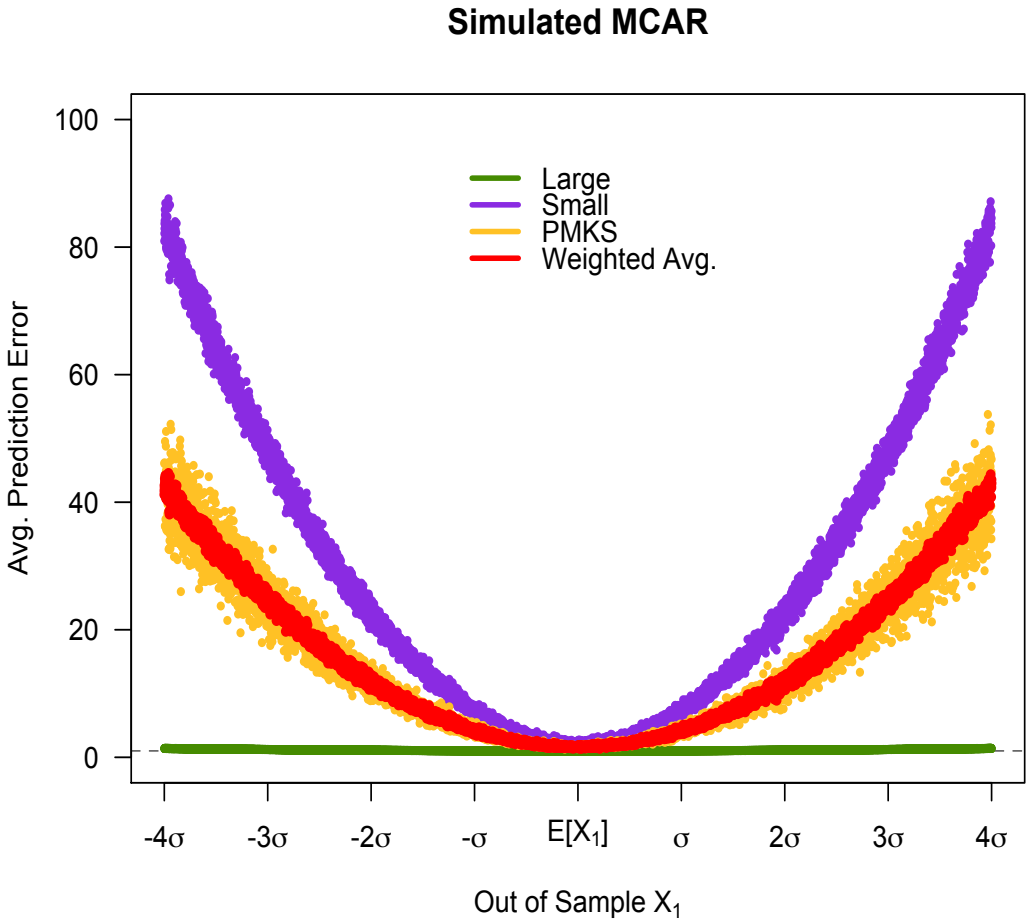
How does a missing predictor affect the EPE?

Large: $E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Small: $E[Y|X_2] = \delta_0 + \delta_2 X_2$

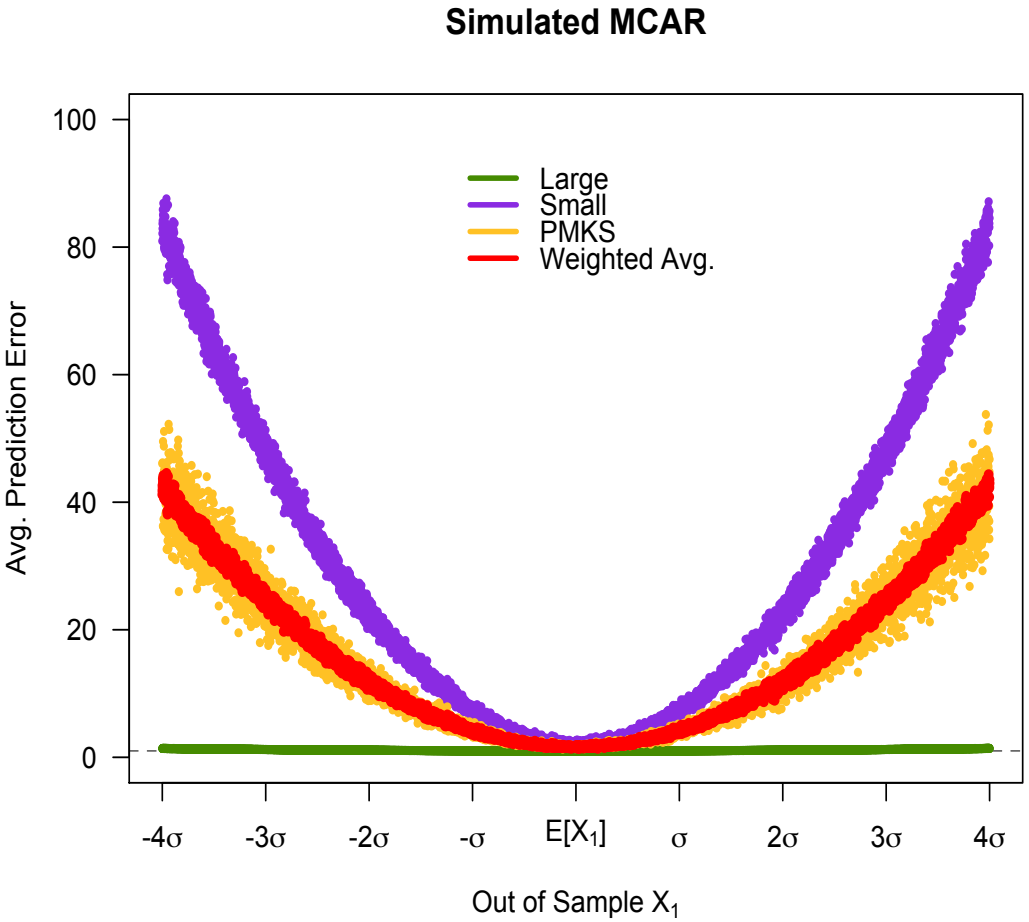
- X_1 missing: want to estimate $\hat{X}_1 = f(X_2) + \epsilon$ such that it yields optimal predictions from the large model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_1 + \hat{\beta}_2 X_2$
- Is there an $f(X_2)$ where the $\text{EPE}_{Large(\hat{X}_1)} \approx \text{EPE}_{Small}$

PMKS Minimizing Squared Prediction Error



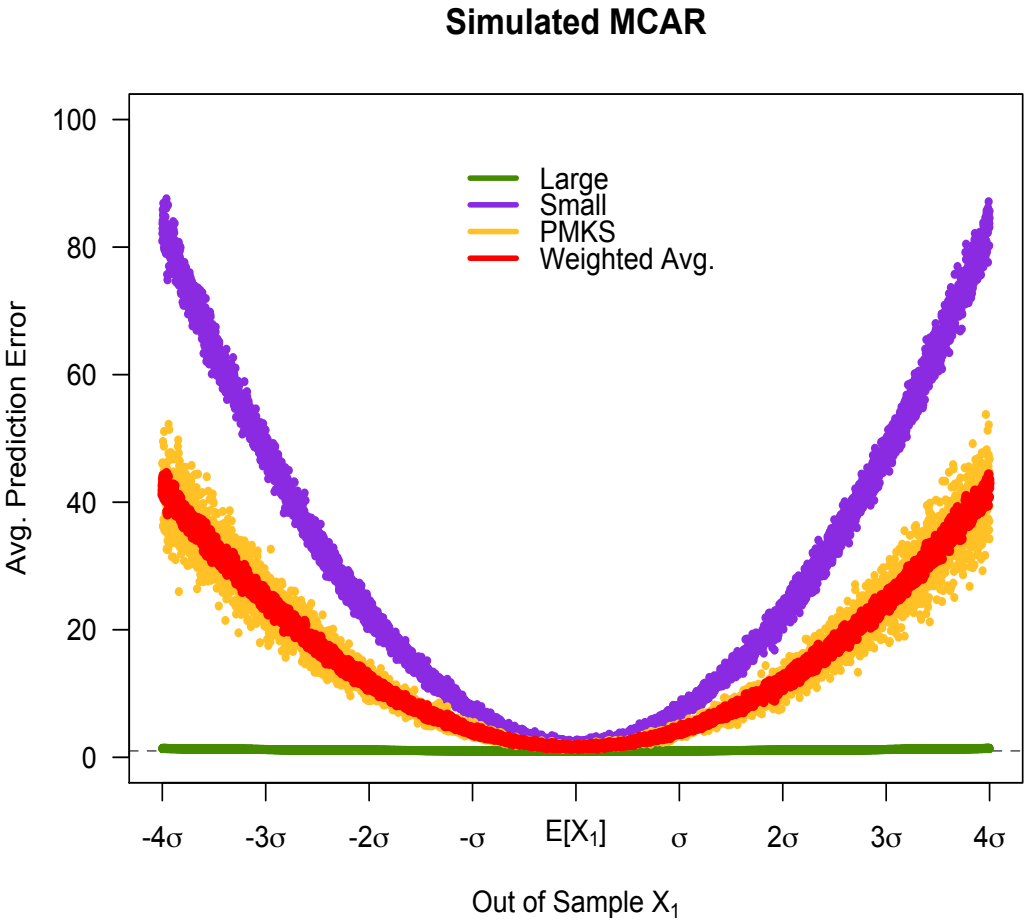
- Only X_1 missing, X_2 always observed

PMKS Minimizing Squared Prediction Error



- Only X_1 missing, X_2 always observed
- $EPE_{PMKS} = EPE_{Large}(1 - P(M)) + EPE_{Small}P(M)$

PMKS Minimizing Squared Prediction Error



- Only X_1 missing, X_2 always observed
- $EPE_{PMKS} = EPE_{Large}(1 - P(M)) + EPE_{Small}P(M)$
- PMKS is optimal, fitting a weighted average of the large and small models

MIMI Model

Multiple Imputation with Missingness Indicator (MIMI) Model

- A general form of PMKS
- Same mean model / borrows strength to reduce variance
- Including indicators relaxes the MAR assumption

$$\begin{aligned} E[Y|X_1, X_2, M_1, M_2] = \beta_0 &+ \beta_1 X_1 + \beta_2 X_2 \\ &+ \delta_1 M_1 + \delta_2 M_2 \\ &+ \delta_3 M_1 X_1 + \delta_4 M_2 X_1 \\ &+ \delta_5 M_2 X_1 + \delta_6 M_2 X_2 \end{aligned}$$

- δ_k are 'auxiliary' parameters
- MIMI can only be fit with some flavor of MI

MIMI & PMKS have the same mean model

No missing predictors

$$E[Y|X_1, X_2, M_1 = 0, M_2 = 0] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Missing X_1, Plug in $E[X_1|X_2] = \alpha_0 + \alpha_2 X_2$

$$\begin{aligned} E[Y|X_1, X_2, M_1 = 1, M_2 = 0] &= (\beta_0 + \delta_1) + (\beta_1 + \delta_3)E[X_1|X_2] + (\beta_2 + \delta_6)X_2 \\ &= (\beta_0 + \delta_1) + (\beta_1 + \delta_3)(\alpha_0 + \alpha_2 X_2) + (\beta_2 + \delta_6)X_2 \\ &= (\beta_0 + \delta_1 + \beta_1\alpha_0 + \delta_3\alpha_0) + (\beta_2 + \delta_6 + \beta_1\alpha_2 + \delta_3\alpha_2)X_2 \\ &= \gamma_0 + \gamma_2 X_2 \end{aligned}$$

$E[Y|X_1, X_2, M_1 = 1, M_2 = 0] = \gamma_0 + \gamma_2 X_2$ is the submodel fit using only the group of individuals who are missing X_1.

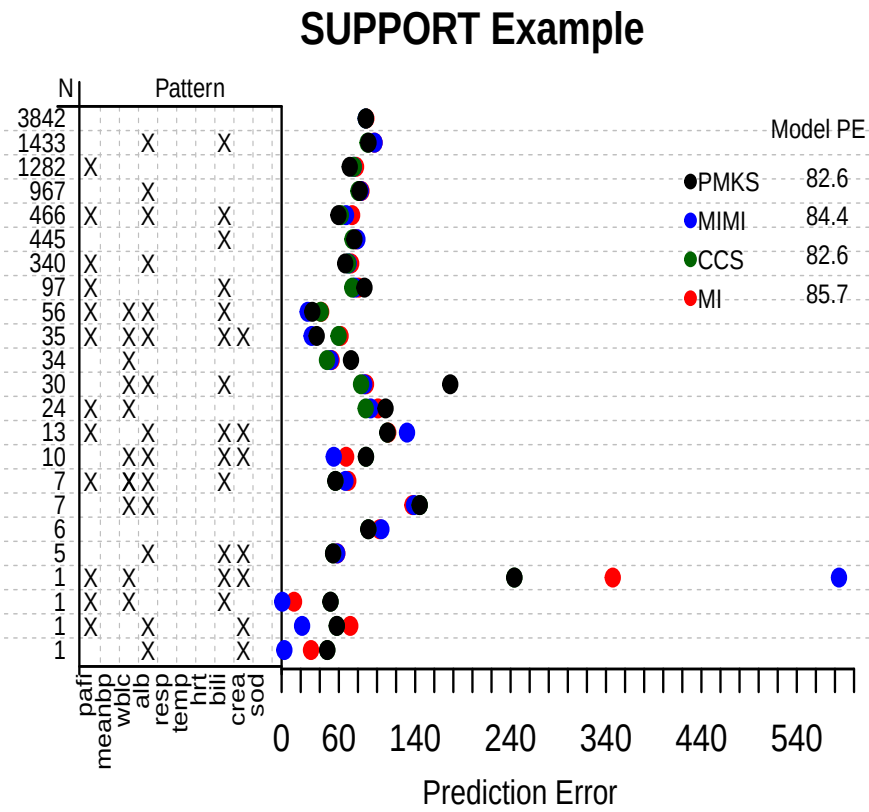
Why is the MIMI model important?

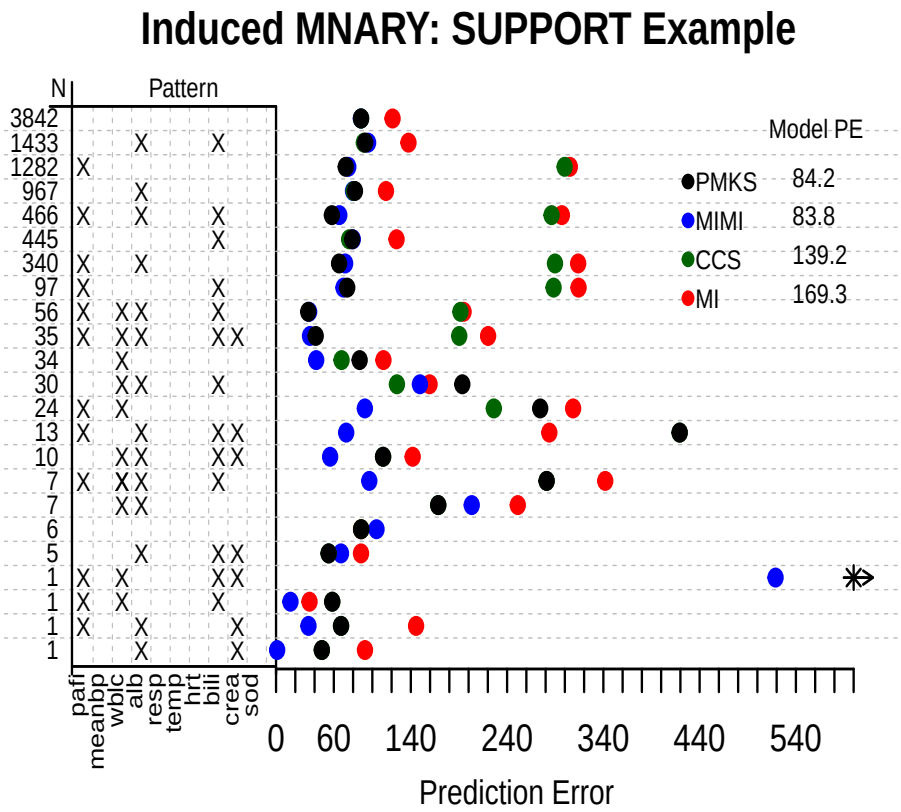
- Parameterizes the 'optimal' prediction model
- Makes assumptions about missingness transparent
- PMKS parameter interpretation is pattern specific
  - PMKS avoids complex computation to get predictions
- Imputation still required for out-of-sample missingness
- Solution to the 'missing indicator' model (computational)
- The 'classic' MI model can be recovered in two ways
  - Apply the SWEEP operator to the MIMI model over the auxiliary parameters
  - Fit the full pattern mixture model and weight the PMKS by the probability of the missing data pattern to get back to the marginal model.

SUPPORT Data Example

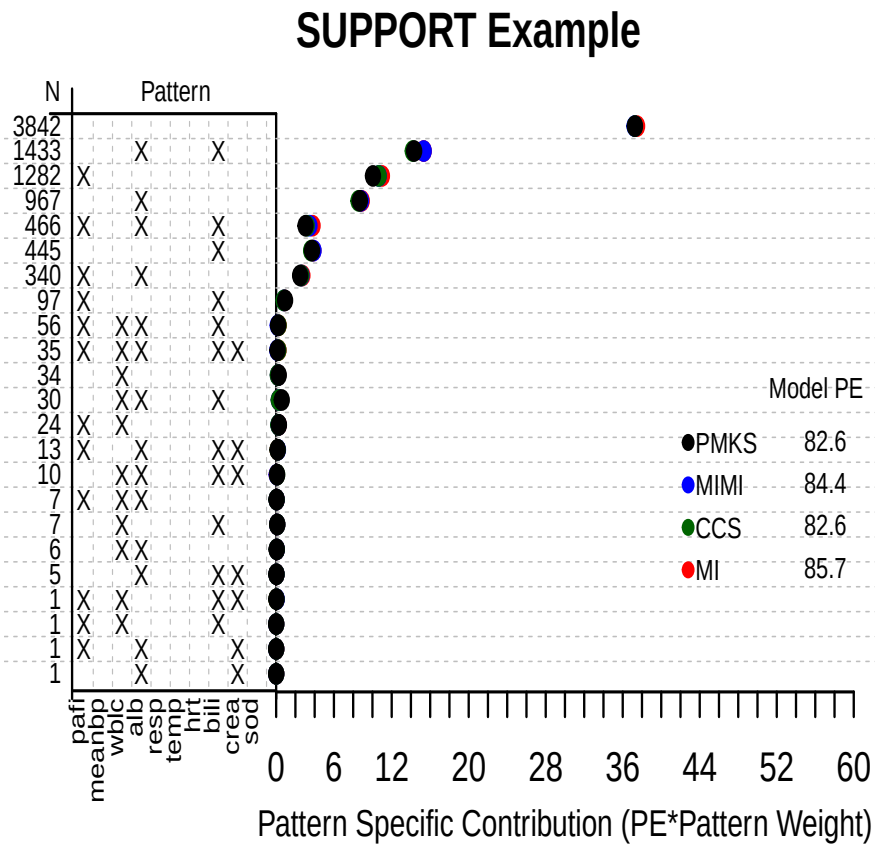
- Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT Data)
- Continuous outcome: SUPPORT day 3 physiology score (SPS Model)
  - `sps` $\sim$ `pafi + meanbp + wblc + alb + resp + temp + hrt + bili + crea + sod`
- Compared PMKS, CCS, MIMI, MI
  - 9105 patients, 10 fold cross-validation of all models
  - MAR assumption met – All methods perform similarly
  - Induce MNAR mechanism – Pattern mixture model and MIMI outperform other methods

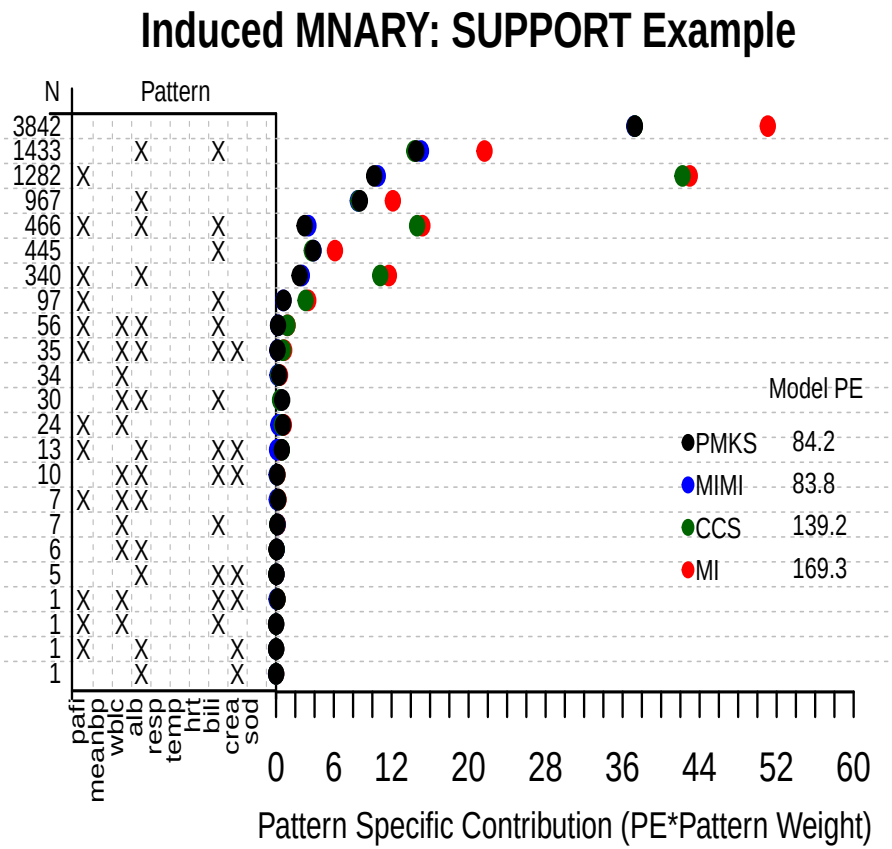
SUPPORT Results (Prediction Error)



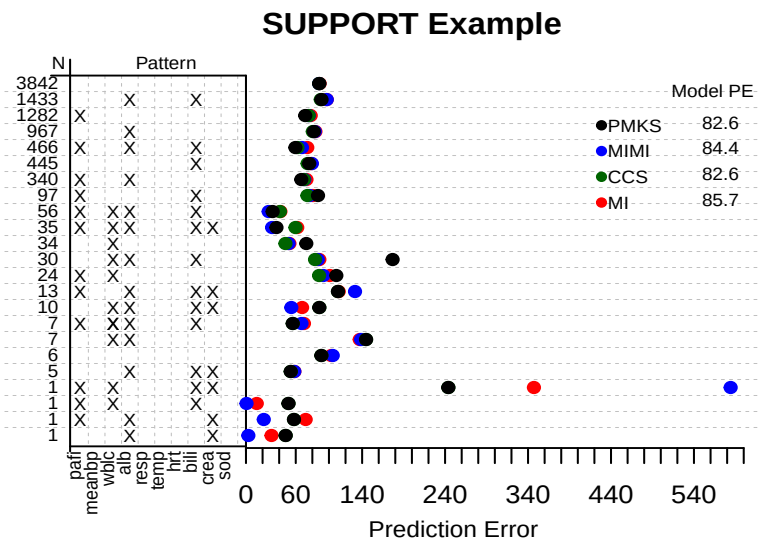


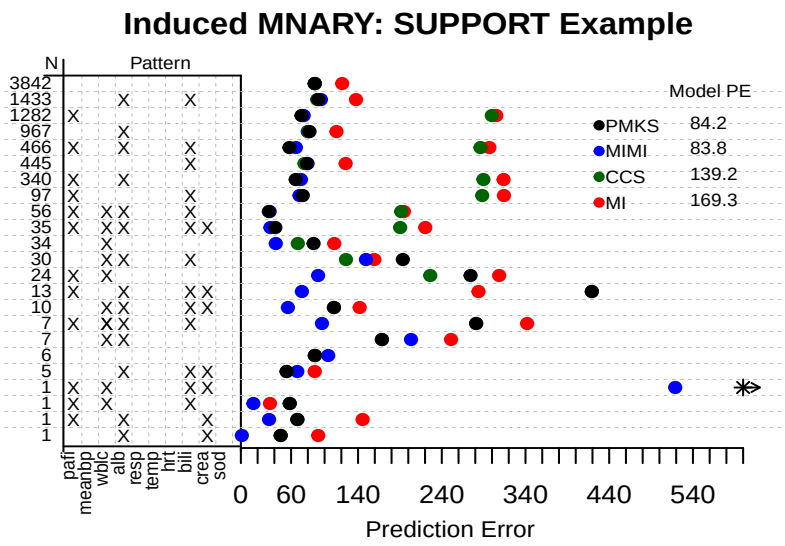
SUPPORT Results (Contributed Prediction Error)

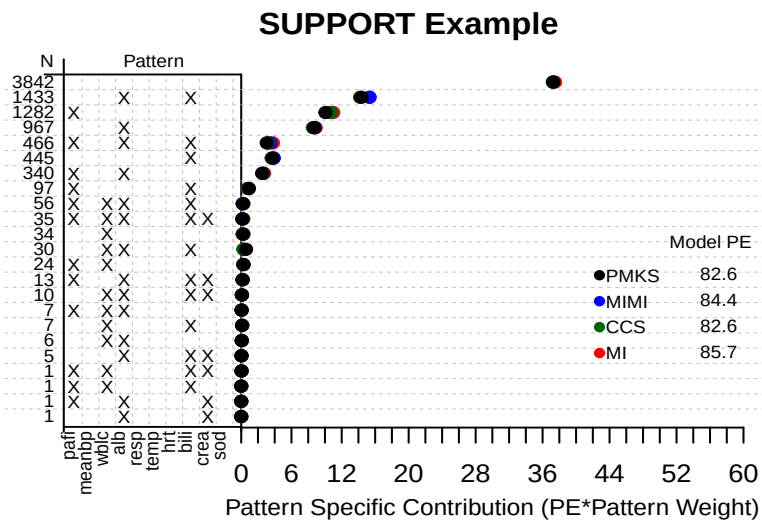


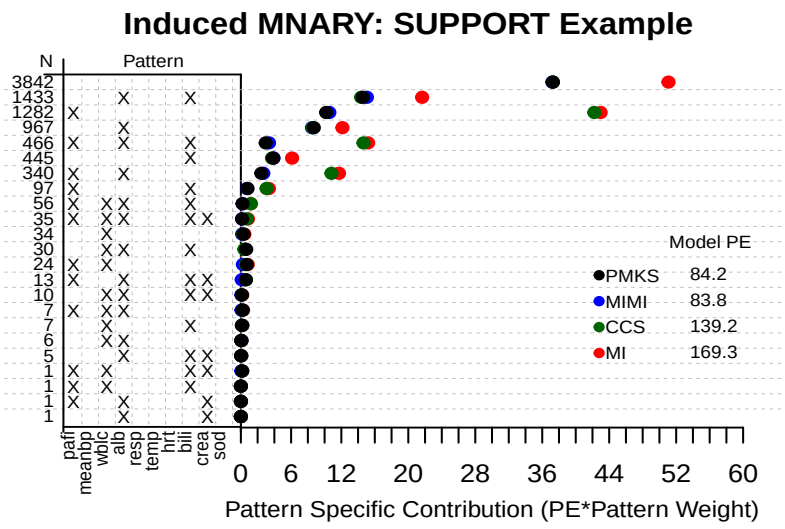


SUPPORT Results









Simulations

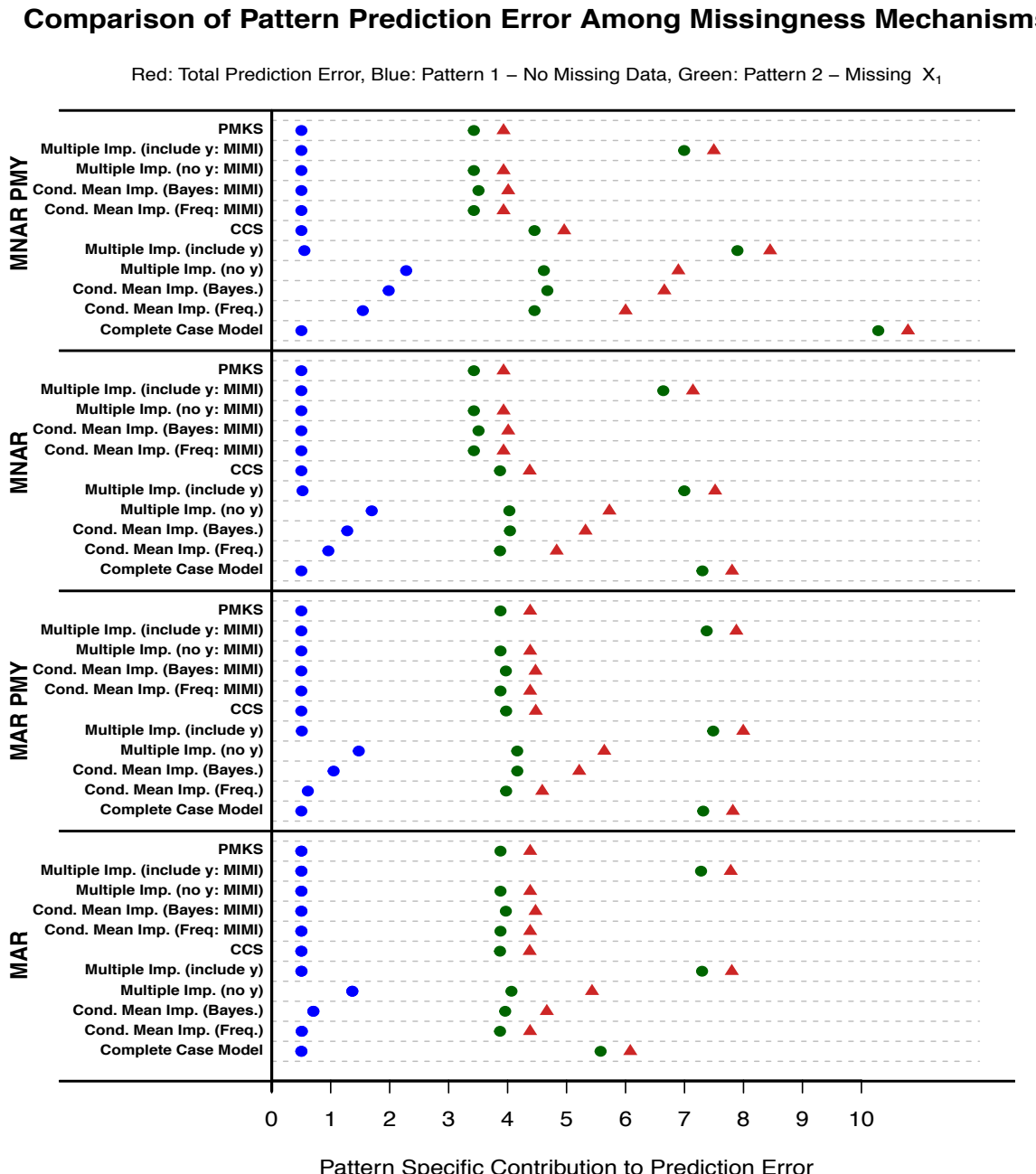
- Generate $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$
- Selection Model ($P(\mathbf{Y}, \mathbf{M}|\mathbf{X}) = P(\mathbf{Y}|\mathbf{X})P(\mathbf{M}|\mathbf{Y}, \mathbf{X})$)
  - Generate X
  - Generate $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - Generate Missing Data
- Pattern Mixture Model ($P(\mathbf{Y}, \mathbf{M}|\mathbf{X}) = P(\mathbf{Y}|\mathbf{X}, \mathbf{M})P(\mathbf{M}|\mathbf{X})$)
  - Generate X
  - Generate Missing Data (MCAR, MAR, MNAR only)
  - Generate $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \delta_1 M_1 + \delta_2 M_2 + \delta_3 M_1 X_1 + \delta_4 M_2 X_1 + \delta_5 M_2 X_1 + \delta_6 M_2 X_2 + \epsilon$

Missingness Mechanisms

- MCAR : $\mathbf{M} \sim c$
- MAR : $\mathbf{M} \sim \mathbf{X_{obs}}$
- MNAR : $\mathbf{M} \sim \mathbf{X_M}$
- MARY : $\mathbf{M} \sim \mathbf{X_{obs}} + Y$
- MNARY: $\mathbf{M} \sim \mathbf{X_M} + Y$

Expected Prediction Error

Parameters: $\beta_0 = 1,\ \beta_1 = 3,\ \beta_2 = 1,\ \delta_1 = \delta_3 = \delta_5 = 1,\ p(M_1 = 1) = 0.5$



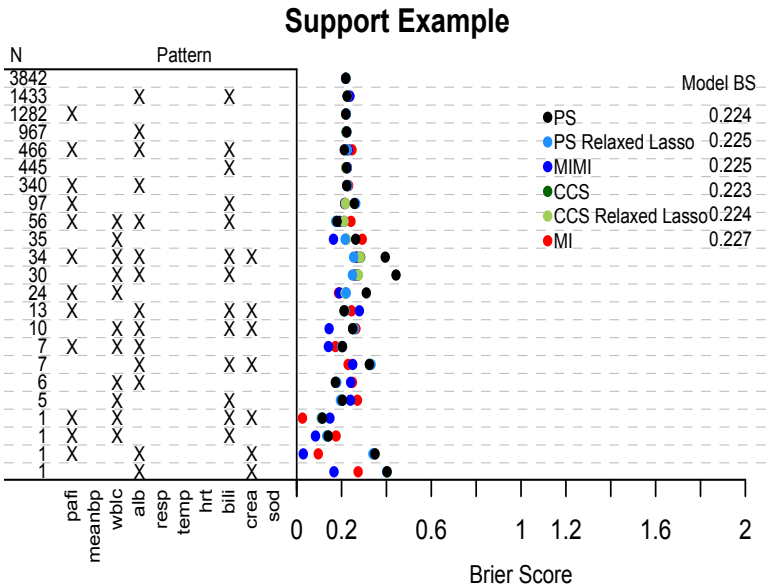
Graph Summary

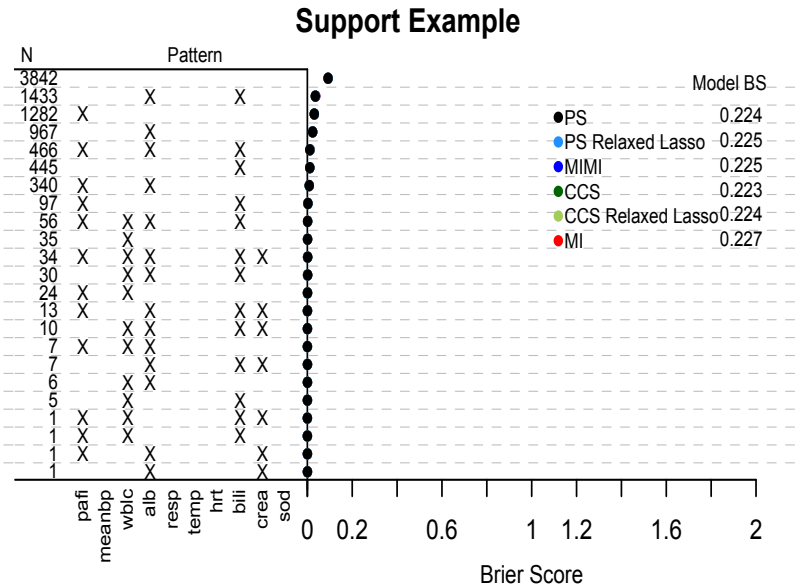
- PMKS has the lowest total prediction error, and requires the least computation
- The magnitude of the differences is highly dependent on the missing data mechanism
- General ordering holds
- PMKS has an advantage when the missing data mechanism depends on the response, otherwise CCS does well

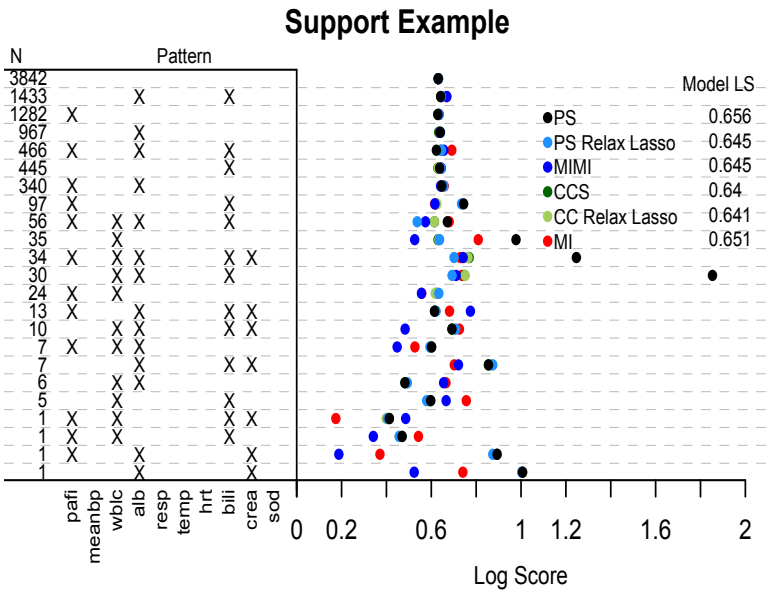
Connections to Machine Learning

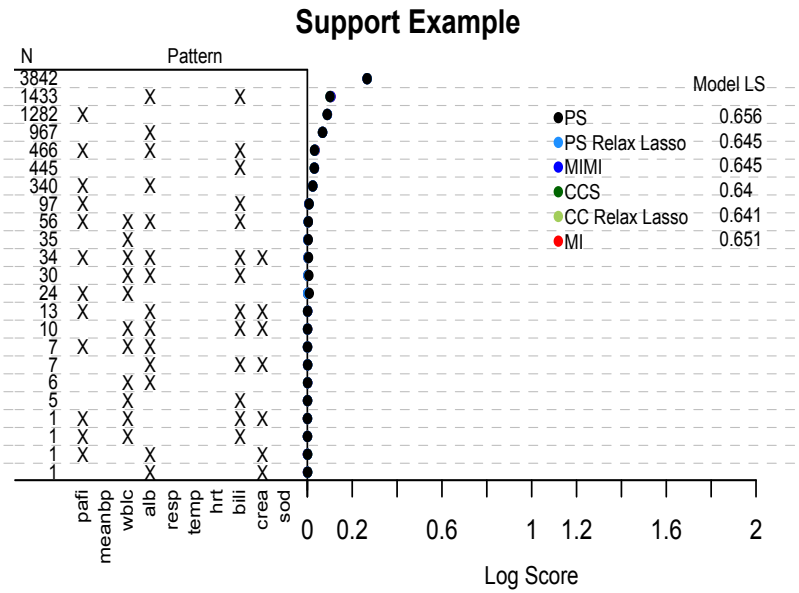
- Machine learners are fancy prediction models
- Not concerned with parameter interpretation
- Feature space usually includes missingness indicators
- Can outperform statistical model that assumes MAR
- PMKS should be applied to prediction algorithms instead of including missingess indicators in the feature space

Including a Relaxed Lasso

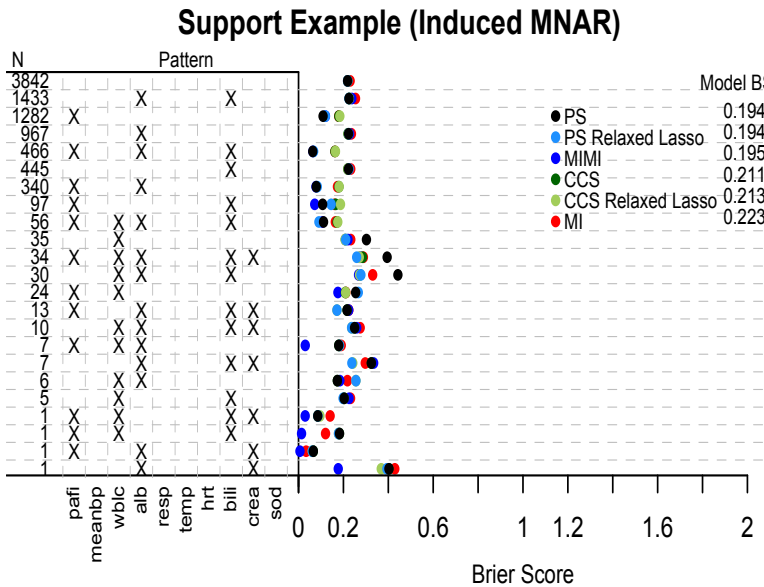


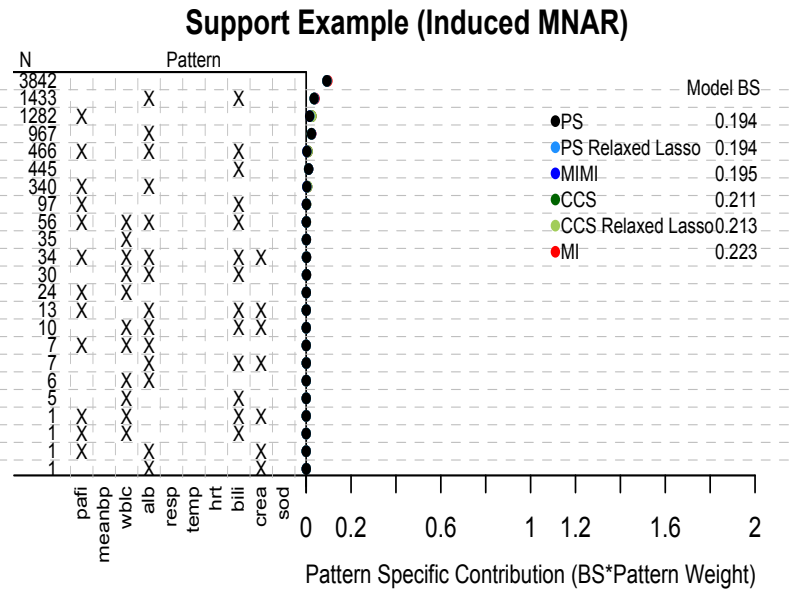


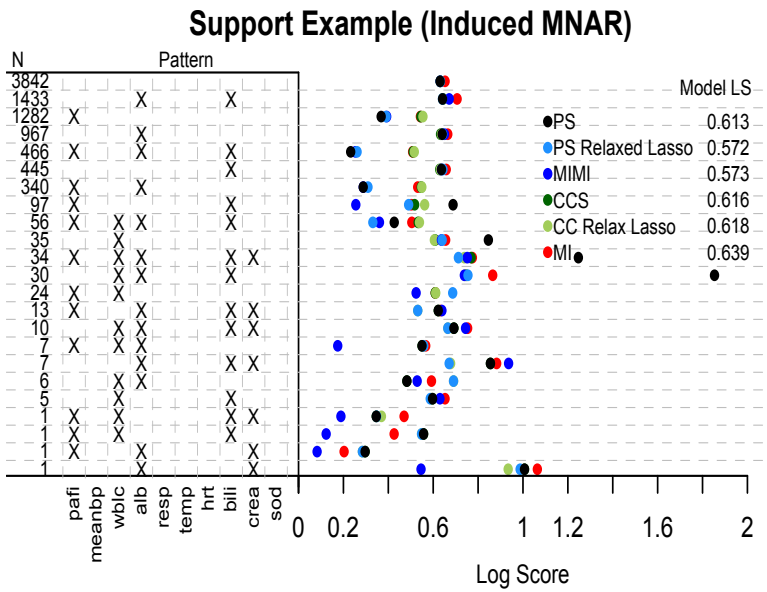


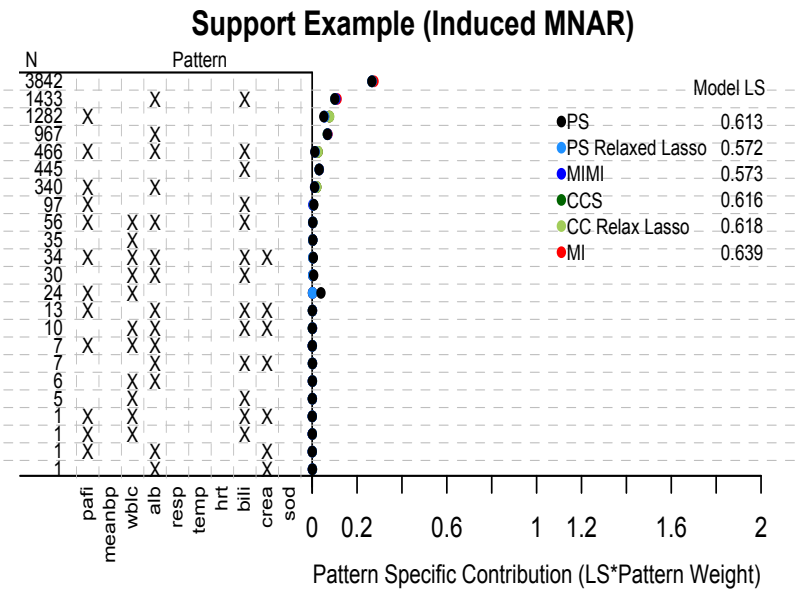


Including a Relaxed Lasso with MNAR









Final Thoughts

- Summary
  - PMKS provides optimal predictions with missing data.
  - PMKS Computationally feasible
  - PMKS and CCS are similar when your missingness is not dependent on the outcome
  - WARNING: Complete Case Analysis not the same as CCS (Complete case performed too poorly to be included)
- Future Extensions of Work
  - Generalizable to more complicated models
  - The number of auxiliary parameters can be reduced by assumption or by penalization

Questions?

Mercaldo, S., J. Blume. Missing Data and Prediction: The Pattern Submodel. September 2018, kxy040, https://doi.org/10.1093/biostatistics/kxy040

Imputation Error of X_1

Squared Imputation Error of the true X_1 compared to the imputed X_1 under different imputation methods and missing data mechanisms: Imputation Error of $X_1 = \sum_i (X_{1i} - \hat{X}_{1i})^2$. The Mean (SD) are presented for each type of imputation and missingness mechanism.

	MAR	MAR PMY	MNAR	MNAR PMY
Unconditional Mean	0.56 (0.03)	0.56 (0.03)	0.76 (0.03)	0.76 (0.03)
Cond. Mean	0.38 (0.02)	0.38 (0.02)	0.53 (0.03)	0.53 (0.03)
Cond. Mean (Bayes)	0.49 (0.03)	0.49 (0.03)	0.69 (0.03)	0.69 (0.03)
MI (No Y)	0.47 (0.03)	0.47 (0.03)	0.61 (0.03)	0.61 (0.03)
MI (Y)	0.76 (0.06)	0.74 (0.06)	0.75 (0.6)	0.71 (0.08)

TREAT Model

- The TREAT model for predicting lung cancer was developed in 492 individuals who were evaluated for surgery with known or suspected lung cancer
- The TREAT model depends on a physician having access to the following predictors: age (`age`), BMI (`bmi`), gender (`gender`), pack years (`pack_years`), pre-operative lesion maximum diameter (`ct_size`), spiculation (`spicul`), lesion growth (`growthcat`), previous cancer (`prev_cancer`), any symptoms (`anysympt`), FEV1 predicted (`fev1_pred`), and FDG-PET avidity (`petpos34`).
- 264 individuals had complete data
- FDG-PET avidity (most impactful variable) was missing for 109 individuals

Logarithmic Score TREAT Model

AUC TREAT Model