Random Forestを用いた欠測データの補完とその応用

2010年11月16日

大学入試センター
研究開発部　石岡 恒憲

欠測データの取り扱い

- 欠測データの補完(impute):欠測値を埋める
- 欠測の確率をモデル化(Rubin,1976)
  - Missing completely at random, MCAR; 欠測するかどうかはモデリングに用いている変数に依存しない
  - Missing at random, MAR;欠測するかどうかは欠測値に依存せずに観測値に依存する
  - Not missing at random, NMAR;欠測値は観測していない他の変量にも依存する
- 欠測データを無視する
  - listwise deletion / pairwise deletion

欠測データの補完

- 単一代入法(single imputation)
  - 明示的(explicit)なモデリング
    - 予測分布に多変量正規などの統計的なモデル
    - 仮定が明示的なもの
  - 暗黙的(implicit)なモデリング
    - 補完の焦点がアルゴリズムにある
    - モデルの仮定を置かないわけではないが，明らかには示されていない

明示的なモデリング

- 平均値代入(mean imputation)
- 回帰代入(regression imputation)
- 確率的回帰代入(stochastic regression imputation)→モデルを想定し最尤推定

暗黙的なモデリング

- ホットデック(hot deck)
  - 最も似た個体の値を欠測値に代入する
- 代替個体の利用(substitution)
  - 標本に含まれない別の個体の値を欠測値として置き換える
- コールドデック(cold deck)
  - 観測の外部のソースからある一つの定数をもってきてそれで代入する
  - たとえば過去の時系列データから季節調整をした値を用いるなど

その他のバリエーション

- ホットデックと回帰代入法
  - 予測した値や最も似た個体の値に，データから得られた統計上の誤差をランダムに加える（Schieber, 1978; David, 1986）
  - 単一代入法の推定値の分散を過小に評価してしまうことを改良

多重代入法(multiple imputation)

- 単一代入法を複数回実施し，得られた複数の推定値の統合を行う(Rubin, 1987)
- マルコフ連鎖モンテカルロ法を用いたベイス推定の一種の近似 (星野, 2009)
- 標準誤差において良い推定量

多重代入法

- 幾つかの制約(Rubin, 1987; 1996)
  - ① データがMARの仮定を満たすこと
  - ② 行う補完がある意味で正しいこと
  - ③ 分析に用いているモデルが補完に用いられているモデルとある意味で合致していること
- 完全データ(観測データと欠測データ)の確率モデル
  - 多変量正規モデル(RではNORM)
  - 対数正規モデル(CAT)
  - 対数線形モデル+多変量正規モデルを結合させた一般位置モデル(MIX)など

欠測の確率モデル

- 適用上はMARの仮定
- NMARでのモデル化が難しい？
  - Multinomial mixture model (Marlin , AISTATS-2005)
  - Aspect model (Hofman, ECML-2001)
  - URP model (Marlin, NIPS-2003)

MARの仮定

- Y_{obs}とY_{mis}をそれぞれYの観測及び欠測データ
- 欠測識別行列をM； MARの定義により

$$f(M|Y,\phi) = f(M|Y_{\text{obs}},\phi) \quad \text{for all } Y_{\text{mis}},\phi$$

ただしφは未知パラメータ

- 観測されているデータと欠測識別変数についての同時分布：Yの周辺分布$f(Y|\theta)$を用いて

$$\begin{aligned} f(Y_{\text{obs}},M|\theta,\phi) &= \int f(Y|\theta)p(M|Y_{\text{obs}},\phi)dY_{\text{mis}} \\ &= f(Y_{\text{obs}}|\theta)p(M|Y_{\text{obs}},\phi) \end{aligned} \qquad (1)$$

- θの最尤推定を行う際には，MARを仮定すれば
欠測識別に関する部分$p(M|Y_{obs}, \varphi)$ を無視できる

完全尤度に基づく推測

- (1)式は完全尤度（full likelihood）
  - 観測されたデータを完全に利用しているという意味
  - 完全データの尤度とは違う
- Method of full-information maximum likelihood: FIML
  - よく知られるように収束しない場合がある
  - 欠測が多いと最尤法の良さがでない可能性が少なくない
  - 最尤法である以上，観測データに対して多変量正規などの統計的確率モデルの仮定

Random Forest(Breiman, 2001)

- 集団学習
- 分類や非線形回帰の方法
- 特別の統計的確率モデルの仮定なし
- 20%程度までの欠測を含む大量データに対し
  - 安定かつ精度よく推定
- Baggingを改良
  - RFでは変数をランダムサンプリングしたサブセット
  - 高次元の解析向き
  - 大量データに対して効率的に動作

RFによる欠測データの補完

- MARの仮定は不可避
- (現時点では)被説明変数(RFではpredictor)が必要
- RF自体についての説明
  - The Elements of Statisitical Learning
  - 金(2007)に少し;Webでもほとんどない
- 欠測データの補完に使えることの解説記事
  - Webにもない(2010年11月時点)
  - 学会のセミナー、ワークショップでも(たぶん)ない
  - (社会)統計学者にはほとんど知られていない

RFによる欠測データ補完のメリット

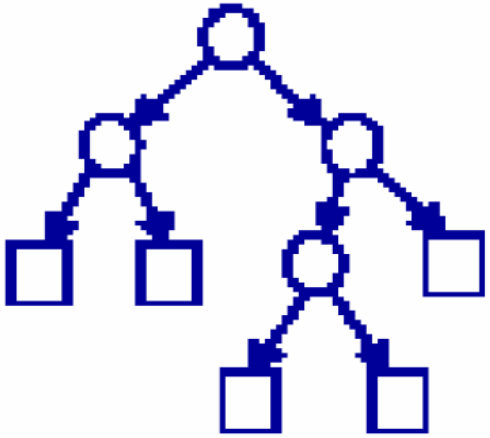
- 大量データ、多変数データに安定
  - 今後の標準の一つ
- データを有効に使える
  - 半教師学習(semi-supervised learning)
  - 「ラベルあり・なし混在データ (labeled and unlabeled data)」からの学習
  - ラベルありデータだけより予測精度が高い

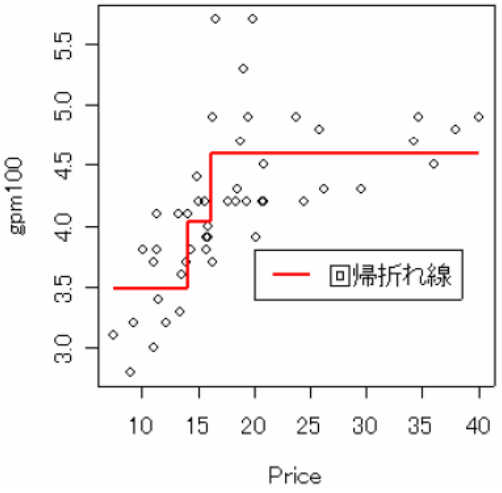
目次

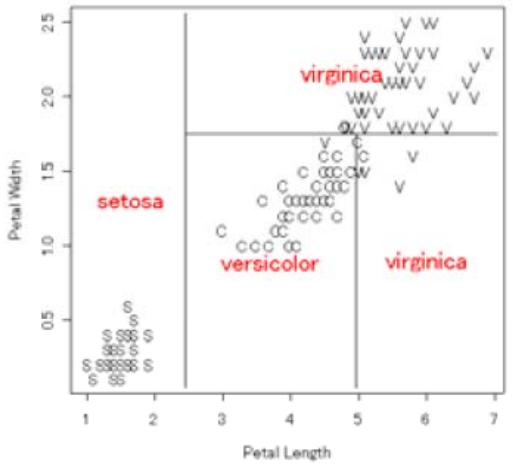
- RF理解の基礎知識
  - CART、Bagging
- RFによる欠測データの補完
- 応用例
  - 2009年のセンター試験
  - 「地理・歴史」「公民」「理科」
  - 科目間得点の差についての解析
- まとめと考察

CART

- Classfication And Regression Tress
  - スタンフォード大学のJ.H.Freedman
  - 1980年代初頭
  - 樹木モデル(tree-baesd model)の1つ
  - 非線形分類や非線形回帰に用いられる
- 説明変数を2進木に分岐







CARTの分岐基準

- クラス分類

$$\text{entropy} = -\sum_{i=1}^{c} p(i|t)\log_2 p(i|t)$$

$$\text{GI} = 1 - \sum_{i=1}^{c} [p(i|t)]$$

tはノード，iはクラス，pは分割された個体のクラスに属する比率

- 回帰

尤離度(deviance) $D = \sum_i (y_i - \mu(i))^2$

$\mu(i)$はy_iが属するセルの平均値

Bagging

- bootstrap aggregatingに由来する造語
  - 1996年　L.Breimanによって提案
  - 集団学習の方法の1つ
- 回帰モデル

  訓練データ $Z = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

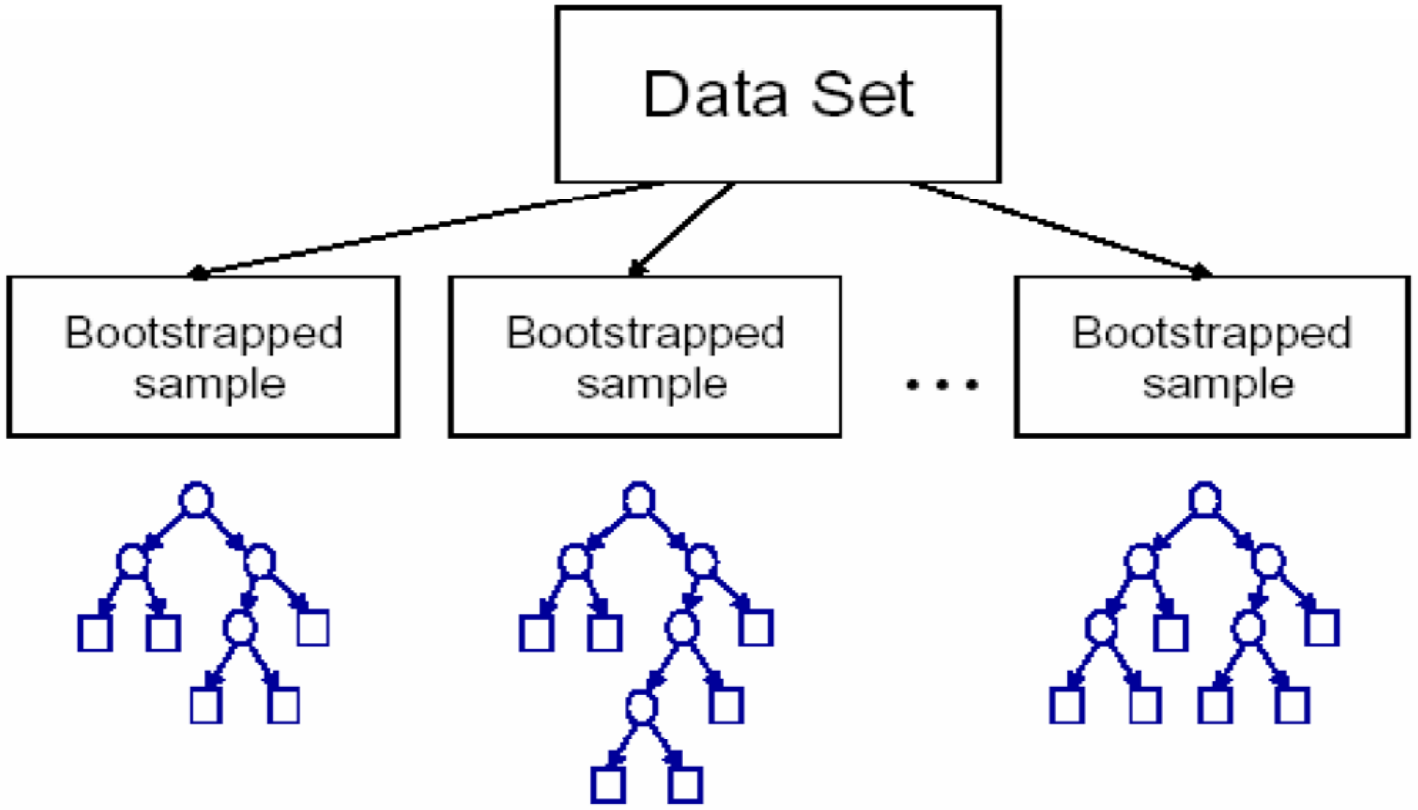
  入力データxにおける予測関数を $\widehat{f}(x)$

  それぞれのブートストラップ標本を $Z^{*b}, b = 1, 2, \ldots, B$

  そこからの予測モデルを $f^{*b}(x)$

  Bagging推定量は $$\widehat{f_{\mathrm{bag}}}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x)$$

Random Forest



出典 http://pegasus.cc.ucf.edu/~exsu/CLASS/STA5703

RFとBaggingの違い

- 変量をランダムサンプリングしたサブセットを用いる
  - 高次元データの解析に向く
  - 精度や計算資源の点で優れる
- ランダムサンプリングする変数の数m
  - 分類の場合 [$\sqrt{p}$], 回帰の場合 [$p/3$]
- 最小のノードサイズ $n_{\min}$
  - 分類の場合 1, 回帰の場合 5

RFのアルゴリズム

1. For b = 1 to B:
   a. ブートストラップ標本 Z* を訓練データから取り出す
   b. Z* からランダムフォレストの木 T_b を成長させる.

      そのために木のそれぞれの終端ノードに対して，木のノードに割り振られる訓練データの数が n_{min}に達するまで以下のステップを再帰的に繰り返す

      i. p変量からランダムに m変量を選ぶ
      ii. m変数の中から最良の変量と分割点を選ぶ
      iii. そのノードを2つの子ノードに分割する
2. できた木を統合させる:$\{T_b\}_1^B$

RFのアルゴリズム(続き)

- 新しいデータ x に対する予測値を出力する
- 回帰の場合:

$$\widehat{f}_{\mathrm{rf}}^{B}(x) = (1/B)\sum_{b=1}^{B} T_b(x)$$

- 分類の場合:
  $\widehat{C_b(x)}$ を b 番目のRFのクラス予測としその多数決をとる(一番多く予測されたクラスの値を選ぶ).

$$\widehat{C}_{\mathrm{rf}}^{B} = \arg\max \left| \{ (\widehat{C}_b(x) \}_1^B \right|$$

なぜRFがうまく働くのか

- もしB個のi.i.d.の確率変数が分散σ^2を持つとしたら平均の分散は$(1/B)\ \sigma^2$
- もし確率変数がi.d.(i.i.d.でない)であるとすれば，平均の分散は

$$\rho\sigma^2 + \frac{1-\rho^2}{B}\sigma^2$$

- もしBが十分に大きければ第2項→ゼロ；第1項だけが残る
- ρを小さくすることにより平均の分散を小さくできる
- RFではp個の変量の中からm《 pの変量を「ランダム」に選ぶことによりρの値を小さくしている.
- mを大きくすると個々の木の予測精度を高めるがその一方で木同士の相関も高めてしまう
- 最適な m の幅　→　mは 調整可能なパラメータ

BreimanによるRFの実装

- Fortran77（g77）
  - Rでの実装はAndy Liaw (RandomForest V4.0)
  - rfImpute()もAndy Liaw
- out-of-bag; oob
  - 与えられたデータの2/3でモデルを作成
  - 残りの1/3データは交差検証用のデータ
  - 分類エラーの不偏推定量を得るために使用
  - 変量の重要度を得るためにも使用

RFの改良

- *m*をランダムに選ぶのではなく
  - 用いる変数間の独立性についての帰無仮説を設定
  - この仮説が棄却されないような変数を順次作成
  - 2分木を成長
  - 条件付き推論ツリーモデル(conditional inference tree model: cTree; Hothorn, 1996)
  - RFに適用した方法はcForest(Hothorn, 2007)
  - 変数の重要度(variable importance measure)の不偏推定ができる

RFを用いた欠測データの補完、その前に

- 用語
  - データ行列（個々のデータ×[被]説明変数）
  - 説明変数→クラス(class) ：列
  - 被説明変数→ ラベル（分類のとき）
  - 個々のデータ→observation / ケース(case)：行
  - 欠測データを埋め込むこと,その値→fill(s)

データ(ケース)間の近似度

- 近似度行列
  - $N \times N$ 行列;ゼロ初期化
  - ケースkとnが同じ終端ノードに落ちたならその近似度を1増やす
  - 終了後，木の数で行列全体を割り近似度を正規化
- 巨大データに対して
  - メモリー上に格納できない
  - MDS
  - たとえば$N \times T$(木の数)に縮約;変数nrnn

MDSによる近似度行列の計算

- ケース*n*と*k*の近似度をprox(*n*,*k*)
  - 正定；対称；上限が1
  - 1－ prox(*n*,*k*)　ユークリッド空間距離
- 距離の内積を示す行列

  cv (n,k)=.5 * (prox(*n*,*k*) - prox (*n*,-)
  - prox(-,*k*) + prox(-,-))
- cvの固有値$\lambda(j)$，固有ベクトル$v_j(n)$

$$x(n) = (\sqrt{\lambda(1)}v_1(n),\ \sqrt{\lambda(2)}v_2(n), \ldots)$$

x(*n*)と*x*(*k*)の間の2乗距離が1-prox(*n*,*k*)に一致

上位何個かの固有値を採用（エッカート-ヤング分解）

RFを用いた欠測データの補完

- ２段階：最初の段階
  - データ行列 $x(n,m)$
  - n:ケース，m:クラス（変数）
  - N:ケースの数（サンプルサイズ）
  - クラス j におけるメディアン（連続量）および最頻値（カテゴリカル変量）でfill

RFを用いた欠測データの補完（続き）

- 2段階目：
  - 前の粗い不正確なfillsを用いてRFの木を成長
  - データ間の近似度行列prox（$N \times N$)を作る
  - データ行列 $x(n,m)$,（n:ケース，m:変数)

→ 連続値をとるべき欠測値

  - m番目の変数の値を非欠測である他のケースとの近似度によって重み付けた平均で埋める

$$\widehat{x}(n,m) = \frac{1}{\#\text{ of non-missing } x(\cdot,m)} \sum_{i \neq n} \text{prox}(i,n) x(i,m)$$

RFを用いた欠測データの補完（続き）

- 2段階目（欠測値がカテゴリカル変量のとき）:
  - 非欠測値の最頻値を埋める
  - ただし頻度は近似度によって重みづけたもの

$$\widehat{x}(n,m) = \underset{C_m}{\operatorname{argmax}} \sum_{i \neq n} \operatorname{prox}(i,n)$$

  - C_mはm番目のクラスにおける値
  - 新しく埋めた値を用いて，もう一度forestを作る
  - また繰り返す
  - Breimanによれば経験上，4～6回の繰り返しで十分

センター試験データを用いた解析

- 得点調整
  - センター試験の本試験について次の科目間で，原則として20点以上の平均点差が生じ，これが試験問題の難易差に基づくものであると認められる場合
  - 地理歴史の「世界史B」「日本史B」「地理B」
  - 公民の「現代社会」「倫理」「政治・経済」
  - 理科の「物理I」「化学I」「生物I」「地学I」
- 分位点差縮小法
  - 科目間での最大幅20点以上→15点
  - 保守的かつ限定的

同じ学力層での比較

- 選択バイアスがあるか？
  - たとえば「物理I」の受験者群が他教科の受験者群に比べ相対的に総合学力が高いといった問題
  - 存在するのであれば，その分の補正や検討
- 大津（2009）による非線形因子分析
  - MARの仮定
  - 1次元の潜在学力分布に正規分布を仮定
  - 50個の等分位区間に対する科目間得点差についての比較研究

別のアプローチによる研究

- 共通学力
  - 大津は正規分布を仮定した（国語，数学，英語，英語リスニングの総和に基づく）潜在的学力
  - 著者らはこれら科目の総和を直接的に

別のアプローチによる研究（続き）

- 同じMARでも違う
  - 大津は欠測のメカニズムにMARを仮定し得られた科目得点のみを解析に用いる
  - 著者らはMARの仮定の下に欠測データを補完し，これらを解析に用いる．
  - 著者らの方法では例えば「物理I」における欠測データの補完にその人のとった「化学I」「生物I」「地学I」の得点にも影響する
  - 大津の方法では「物理I」の欠測データはMARの仮定のもとで欠測として取り扱われる

  → 「物理I」の解析においては，その人のとった「化学I」「生物I」「地学I」の得点は影響を与えない

センター試験データへの適用

- あわや得点調整となった2009年のデータ
  - 同じ総合学力における各科目間の得点比較
  - 総合学力の定義：
    「国語(200点)＋数学I・数学A(100点)＋英語筆記(200点)＋英語リスニング(50点)」
  - 各科目の偏差値の総和は主成分分析における第1主成分に相当；大津の方法も適切
  - 取り扱う対象が各科目間の得点差(素点の差)
  - 総合学力も素点である方がわかりやすい？
  - 数学的/統計的妥当性よりも一般の人に向けた理解の容易性

欠測は少ない方がいい？

- 従来法
  - 欠測データが少なくなる工夫
  - 理科系科目の解析：「数学I・数学A」と「英語（筆記とリスニング）」を受験した受験生を対象に理科の科目の得点比較
  - 文科系科目の解析：「国語」と「英語（筆記とリスニング）」を受験した受験生を対象に地理・歴史や公民の科目の得点比較
  - その上で欠測したデータについてはMARの仮定

アラカルト方式

- 任意の科目を好きなだけ選択することが可能
  - 理科系希望者の中には「国語」を受けないが公民の科目を受けている者はいる
  - 従来、解析の対象から外れた中に，解析すべき科目を受験している者は当然存在する
  - そもそも得点調整では全受験者を対象とした科目得点の平均差を論じている
  - 可能な限り捨てるデータを少なくしたいという要請
  - センター試験約55万受験者のうち，国立大学を志望する者は約30万人
  - 多科目受験者だけを解析の対象とすると誤った結論

MARに基づくデータ補完

- そのための一つの方法
  - MARの仮定が成立している保障のないところにMARの仮定をおいて欠測データを補完
  - 捨てるデータは少なくなる一方で疑わしいデータも多くなる
  - データ補完の方が優れているかどうかは必ずしも明白ではない
  - 問題の性質による
- 一般論として
  - 欠測率が高くなければ従来捨てていたデータを取り入れた方が全体としての推測の精度はあがる

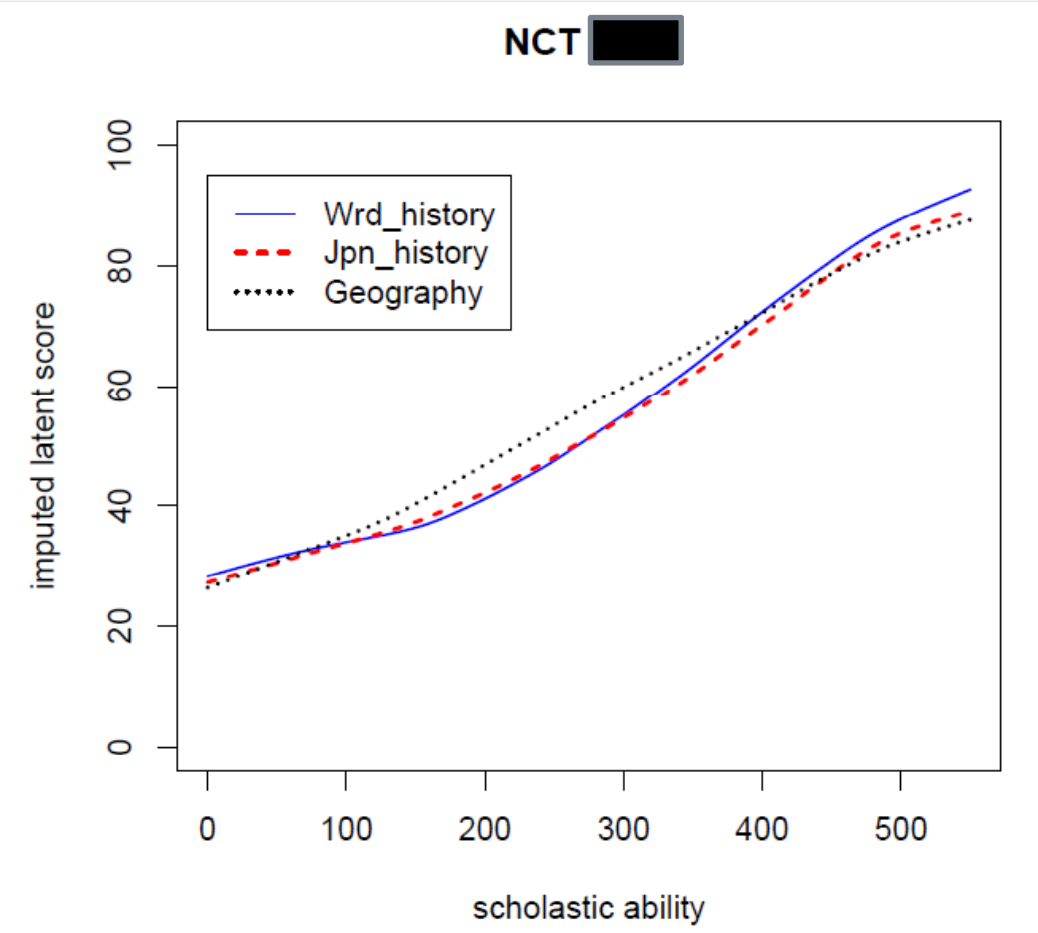
2段階でのデータ補完

- 欠測率の高いデータ対象に補完を行なわないようにするため
  - 最初に(「科目」ではなく)「教科(試験枠)」単位でデータ収集し，欠測値に対してデータ補完
  - 「教科(試験枠)」とは
    「国語」,「地理・歴史」,「公民」,「数学①」,
    「数学②」,「理科①」,「理科②」,「理科③」,
    「外国語」,「英語リスニング」→説明変数
  - 「総合学力(国語＋数学I・数学A＋英語筆記＋英語リスニング)」→被説明変数
  - 回帰モデルをRFでモデル化し，欠測値を埋める

データ補完2段階目(着目している教科)

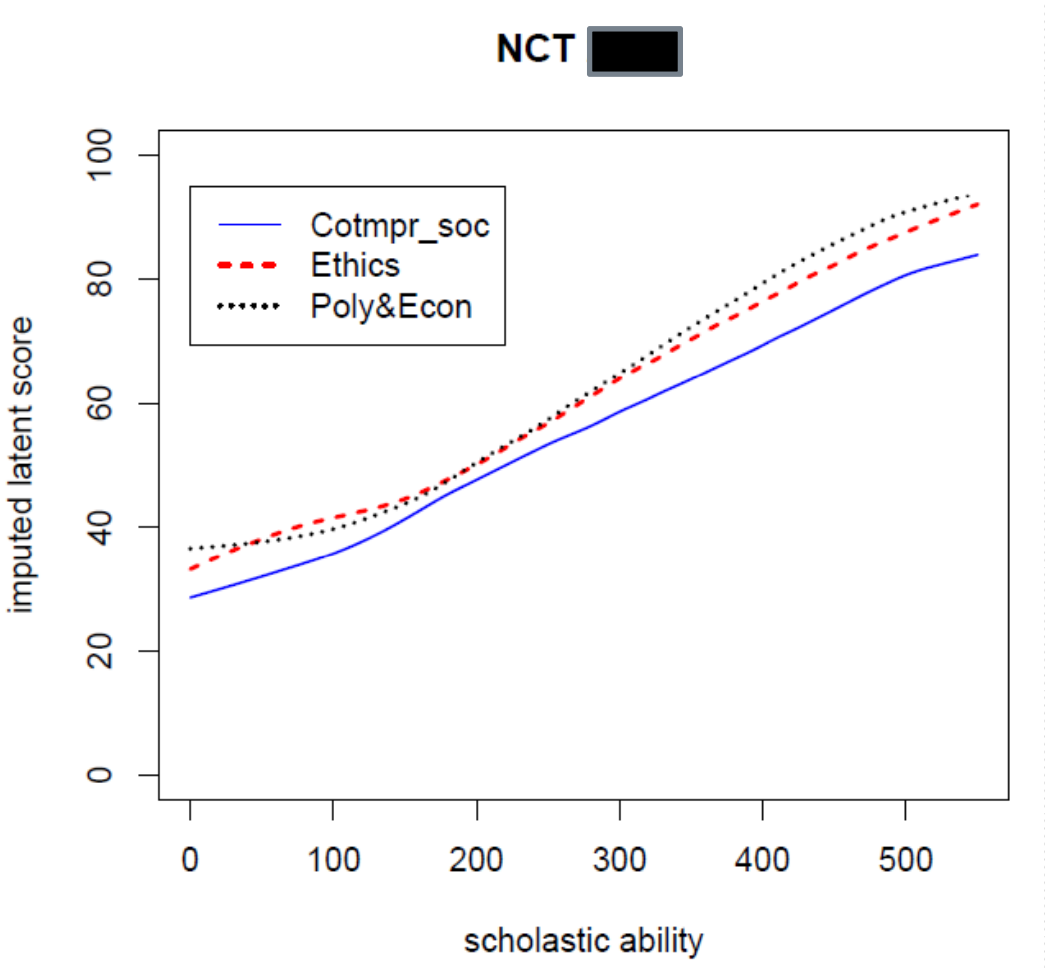
- (「地理・歴史」,「公民」,「理科」)において
  - 得点調整の対象となる科目ごとの得点に分け欠測値を埋める
  - たとえば「地理・歴史」において「世界史B」「日本史B」「地理B」の3つに分け欠測値を埋める
  - このとき「公民」や「理科」については教科を分けない
  - データ全体に対する欠測の割合の低下を抑えるため
  - RFにおける回帰モデルでは「世界史B」「日本史B」「地理B」を被説明変数
  - この結果に基づき「世界史B」「日本史B」「地理B」の得点差を検討する.

「地理・歴史」における科目間比較



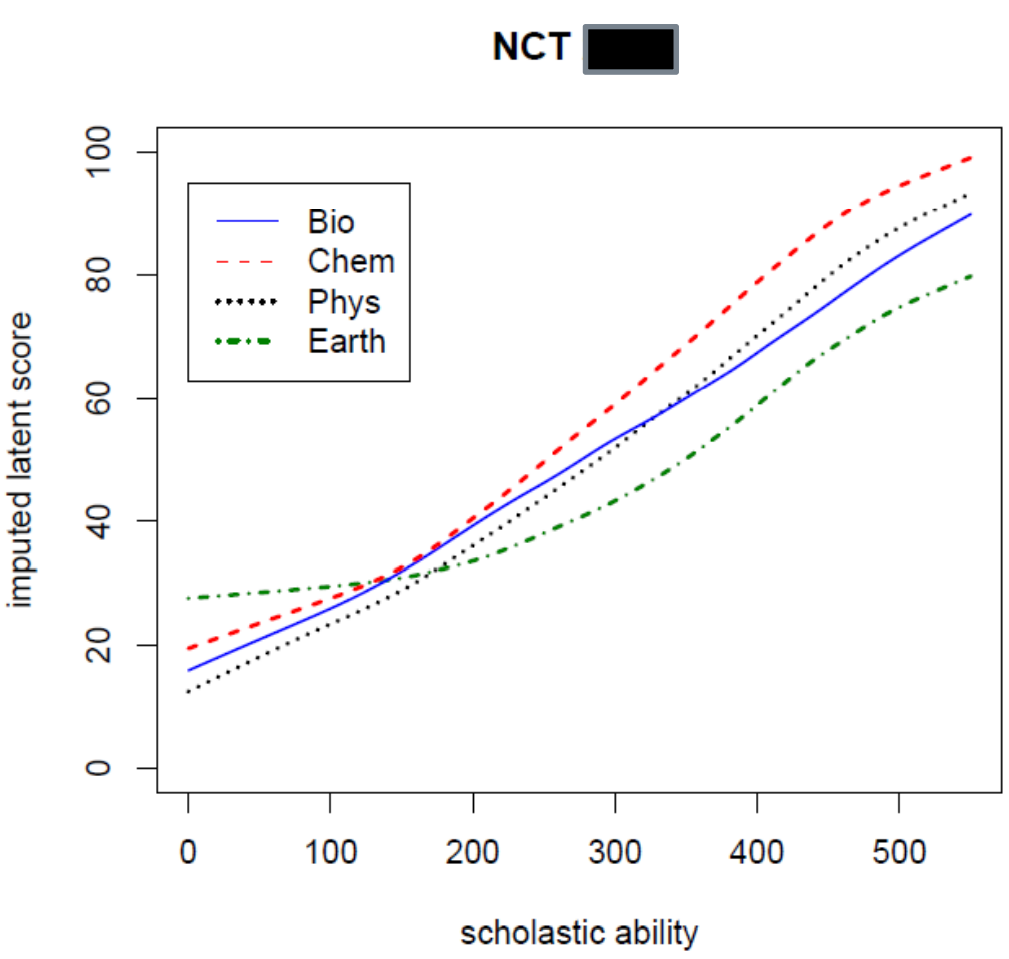
- 総合学力300点の地点で「地理B」のスコアが他の2科目に比べ5点ほど高い
- 総合学力400点で3者はほぼ一致
- 総合学力500点では「世界史B」が最も高いスコア
- 中学力層においては「地理B」が良いスコアを取りやすく
- 高学力層においては「世界史B」が良いスコアを取りやすい

「公民」における科目間比較



- 全ての学力層において
- 「現在社会」が他の2科目より5～8点ほど少ない→難しい

「理科」における科目間比較

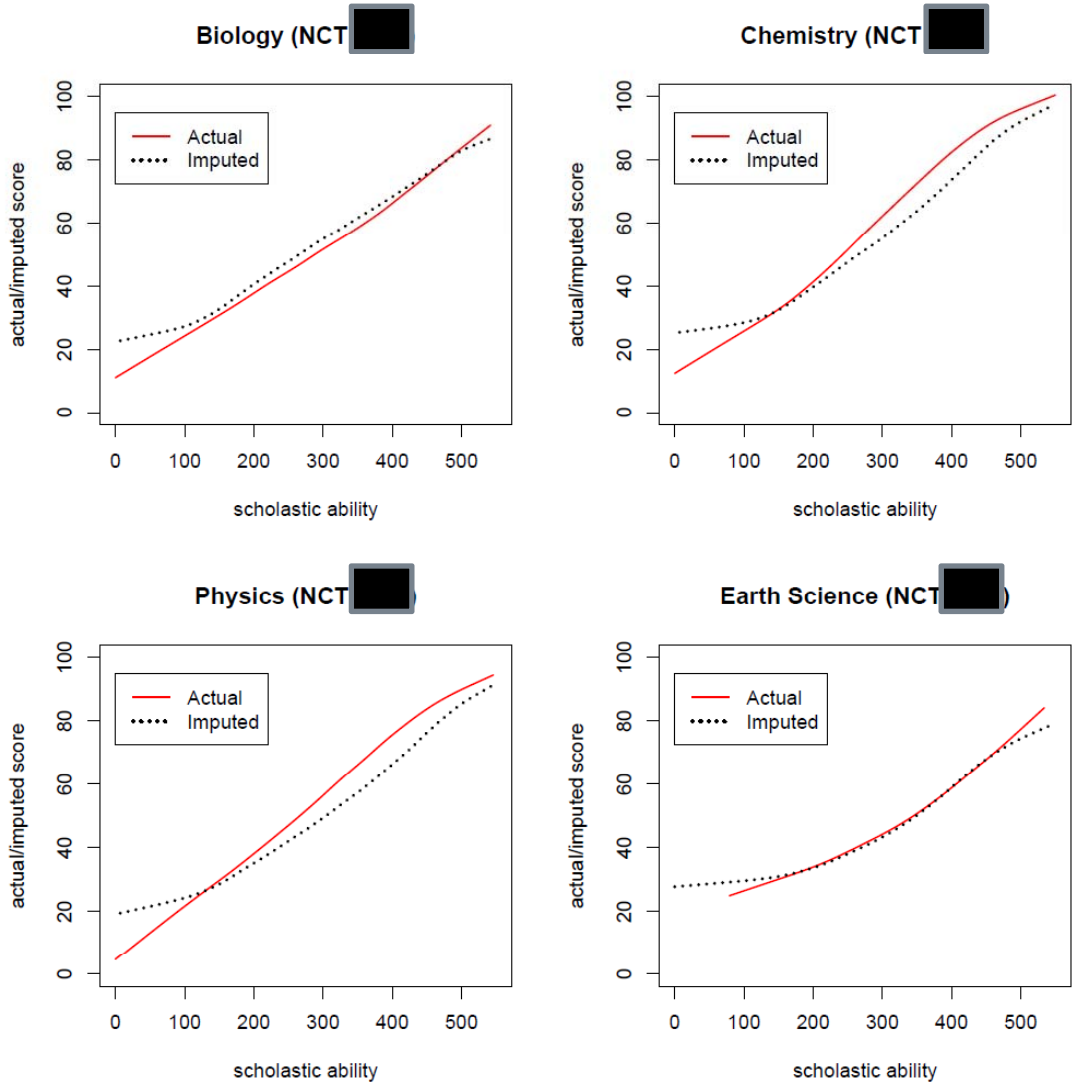


- 総合学力300点の中位のあたりで「化学I」と「地学I」の間には20点近くの違い
- その差は総合学力500点の高学力層においても変わらない
- 中上位層において「化学I」が易しく「地学I」が難しかった
- 「化学I」と「地学I」の差は試験の難易差の違いにほぼ負うもの
- 大津（2009）の解析とほぼ同じ結果

補完データと実データとの差異

- 欠測を補完したデータと実測値との同じ総合学力に対する差の検討
  - 横軸に総合学力
  - 縦軸に欠測データ(Imputed)のスコアと実測データ(Actual)のスコア
- 実測データが補完データよりも（同じ総合学力において）高スコアなら
  - 本来その学力から期待されるスコアよりも実際に受験したスコアの方が高い
  - その人にとっての得意科目

補完データと実データとの差異（理科）



- 大学受験を目指すほぼ全ての受験生において，「物理I」と「化学I」の受験者はそれを得意科目にしている（あるいは得意とする人が受験している）
- 「生物I」においては，その逆で受験者はそれを不得意科目にしている（あるいは不得意とする人が受験している）
- 「地学I」においては得意でも不得意でもない

Rプログラム

- RFの使用法
  - Package ‘randomForest’
    http://stat-www.berkley.edu/users/breiman/RandomForests
  - 金(2007)他
  - 判別・回帰における変数の重要度としてのジニ係数の表示の仕方についての言及

Rプログラム（続き）

- RFによる欠測データの補完
  - CRAN; RandomForest ライブラリにある rfImpute()
  - Rライブラリー　yaImpute() なる*k*NNによるデータ補完アルゴリズム(Crookston, 2008).
  - 著者はこのライブラリによるデータ補完(*k*=1)を使用→妥当な推定値を得ることができなかった
  - 変数が増えるといわゆる「次元の呪い(curse of dimensionality)」→どのデータも互いに似なくなる
  - 多次元かつ欠測率が大きいデータに単に*k*NNを適用しただけでは，データのもつ本質を捉えることが難しい。少なくとも本事例ではうまくいかない

おわりに

- RFによるデータ補完rfImpute()
  - 高次元でかつ欠測が少なくないデータに対して比較的頑健
  - データ構造の理解に有効に機能する
- 完全データを補完により作ってしまう
  - MARの前提はあるものの
  - 既存の手法を特別な工夫無く利用でき
  - 実用的な意義は大きい

欠測データの取り扱いにおける位置づけ

- データ補完；多重代入法と似たアプローチ
- 両者の違い
  - rfImpute()がデータモデルに恣意的な確率モデルをおく必要がない
  - データ間の近さを分類木における最終ノードの位置を問題としている
  - 統計的なアプローチとはかなり様相が異なる
  - 両者の相違性や近似性についての数学的な検証を期待
- 半教師学習との関連