非負値行列因子分解の統一的な拡張と医学データ解析

阿部興[1] ・ 島村徹平[2]

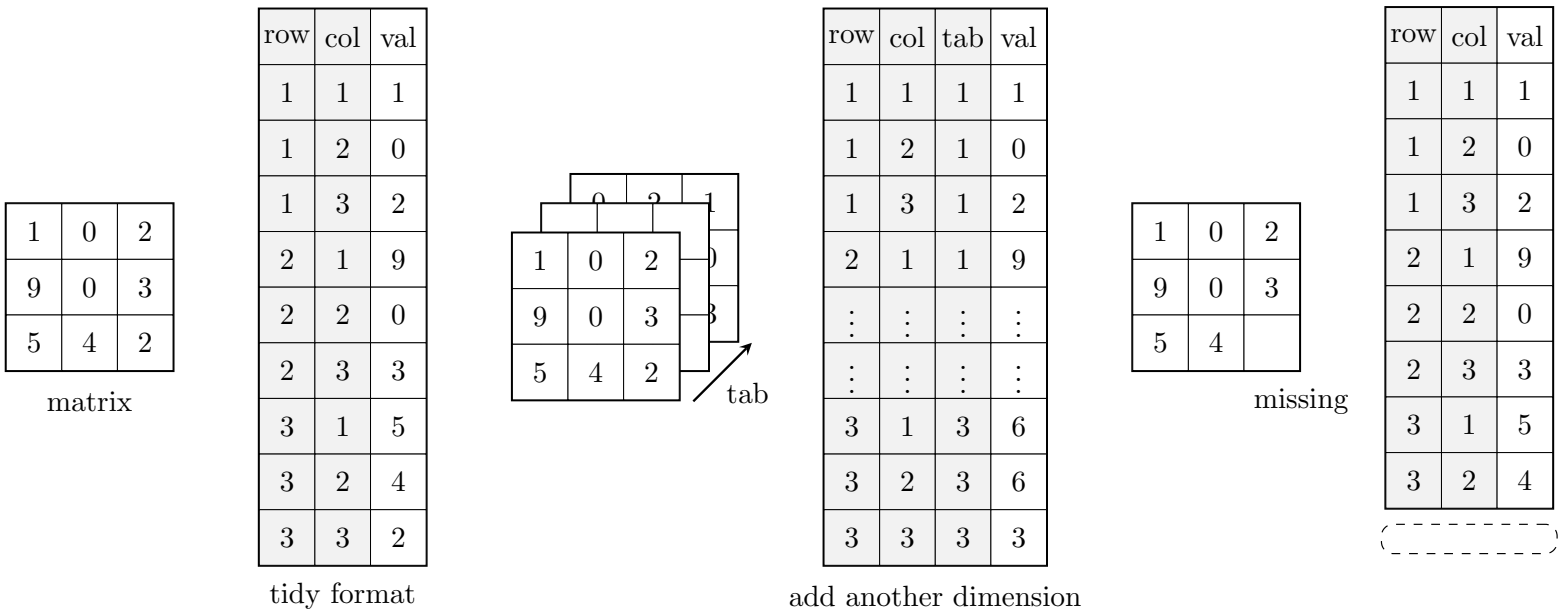
2023 年 6 月 3 日

[1]名古屋大学医学系研究科
[2]名古屋大学医学系研究科・東京医科歯科大学難治疾患研究所

背景

- 生命科学に関わる多数の要素の縦断的な観測が可能に
  - ゲノム, エピゲノム, トランスクリプトーム, プロテオーム, ...
- これらは非負の整数（カウント）データ
- 計算機上では多次元配列（テンソル）
- 欠損や重複の扱いが難しい

tidy data



1	0	2
9	0	3
5	4	2

row	col	val
1	1	1
1	2	0
1	3	2
2	1	9
2	2	0
2	3	3
3	1	5
3	2	4
3	3	2

row	col	tab	val
1	1	1	1
1	2	1	0
1	3	1	2
2	1	1	9
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
3	1	3	6
3	2	3	6
3	3	3	3

1	0	2
9	0	3
5	4	

row	col	val
1	1	1
1	2	0
1	3	2
2	1	9
2	2	0
2	3	3
3	1	5
3	2	4

- 各変数が 1 つの列を形成する
- 各観測値が 1 つの行を形成する
- 各値が 1 つのセルを構成する

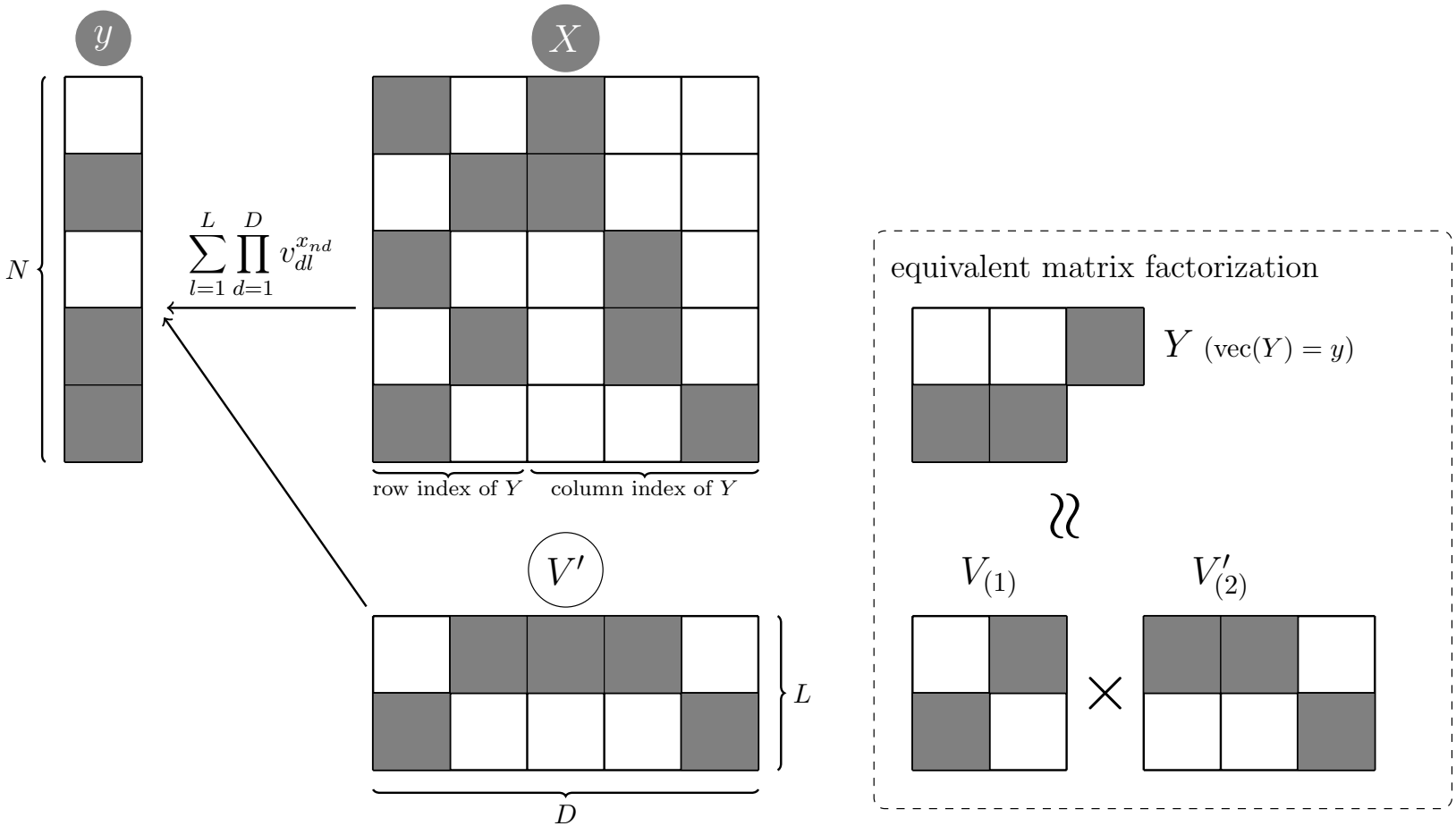


Fig: UNMF のコンセプト. 灰色：観測される変数, 白：潜在変数（推定の対象）

$$y_n \approx \sum_l \prod_d v_{dl}^{x_{nd}}$$

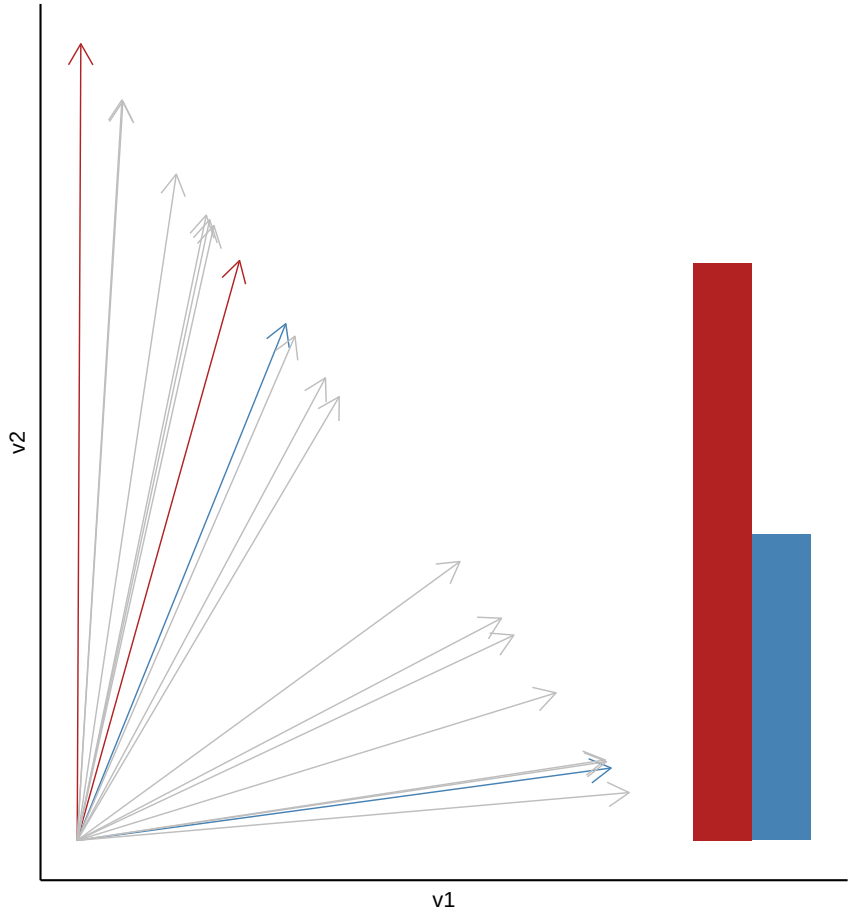


Fig: x で指定されるベクトルどうしの向きが近いほど右辺が大きくなる

表記法

- 確率分布（密度関数）
  - パラメータ λ のポアソン分布：$\mathrm{Poisson}(\lambda)$
  - 形状パラメータ a, レートパラメータ b のガンマ分布：$\mathrm{Gamma}(a, b)$
  - サイズパラメータ n, 確率パラメータ p の多項分布：$\mathrm{Multi}(n, p)$
- ベクトル
  - 添字付きの変数 x_{ij} $(j = 1, \ldots, m)$ に対して, $x_{i:} = (x_{i1}, \ldots, x_{im})'$

モデル

$$y_n = \sum_{d=1}^{D} \sum_{l=1}^{L} u_{nl},$$

$$u_{nl} \sim \text{Poisson}\left(\prod_{d=1}^{D} v_{dl}^{x_{nd}} \right)$$

$$v_{dl} \sim \text{Gamma}(a, b).$$

次と等価：

$$y_n \sim \text{Poisson}\left(\sum_{l=1}^{L} \prod_{d=1}^{D} v_{dl}^{x_{nd}} \right)$$

変分ベイズ推定

- $u_{n:}$ の変分事後分布 $q(u_{n:}) = \mathrm{Multi}(y_n, r_{n:})$

$$r_{nl} = \frac{\exp(\widetilde{\log v_{dl}})}{\sum_{l=1}^{L} \exp(x_{nd}\widetilde{\log v_{dl}})}$$

- v_{dl} の変分事後分布 $q(v_{dl}) = \mathrm{Gamma}(\hat{a}_{dl}, \hat{b}_{dl})$

$$\hat{a}_{dl} = \sum_{n=1}^{N}\sum_{n=1}^{N} x_{nd}\widetilde{u_{nl}} + a,$$

$$\hat{b}_{dl} = \sum_{n=1}^{N} x_{nd} \left(\prod_{d' \neq d} \widetilde{v_{d'l}}^{x_{nd'}} \right) + b.$$

$\widetilde{x}$: 変分事後分布による平均

推定量の平均と標準誤差

$N = 100 \times 100 \times 3$. 真値は $v_{dl} \sim \mathrm{Gamma}(1, 0.01)$ で設定.

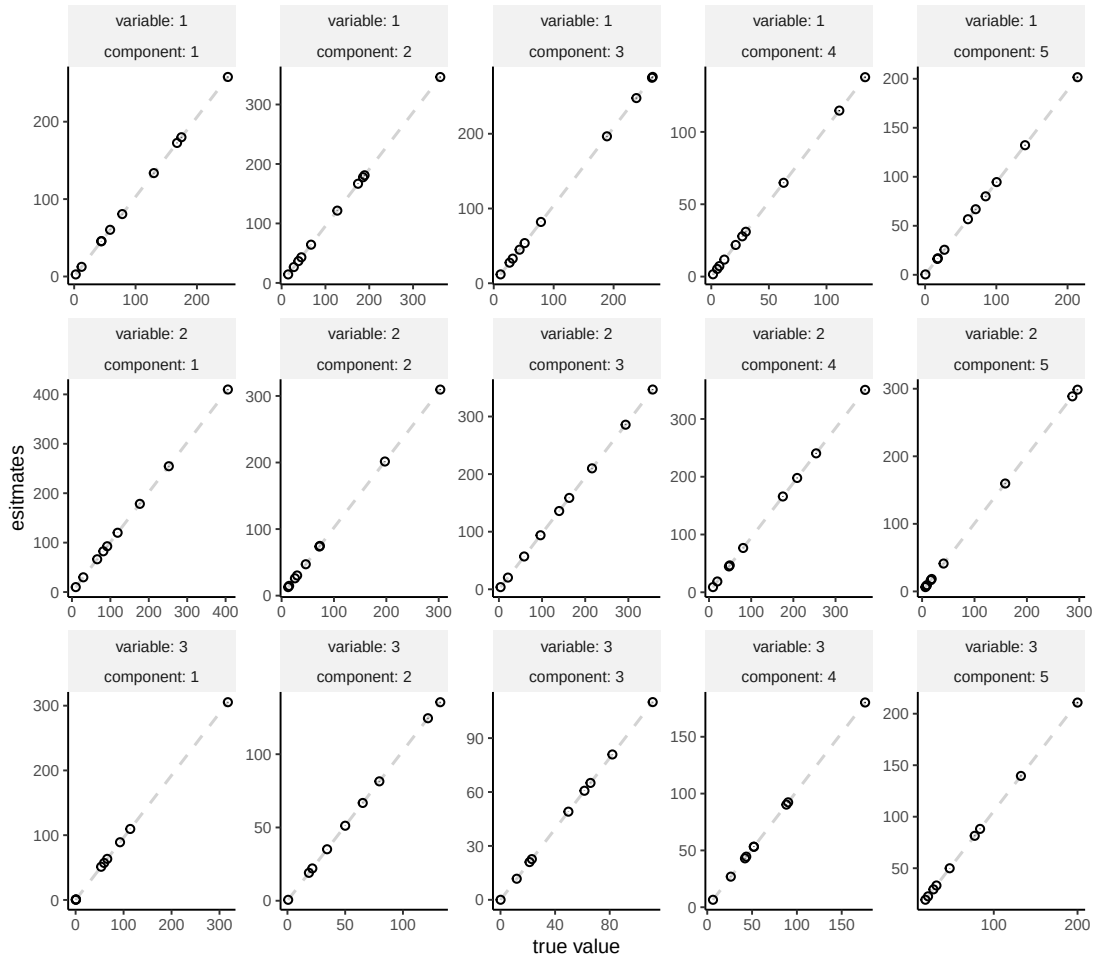


Fig: シミュレーションで設定した真値と 100 回のシミュレーションから得られた推定値の平均の比較. 縦棒: 標準誤差.

モデル選択

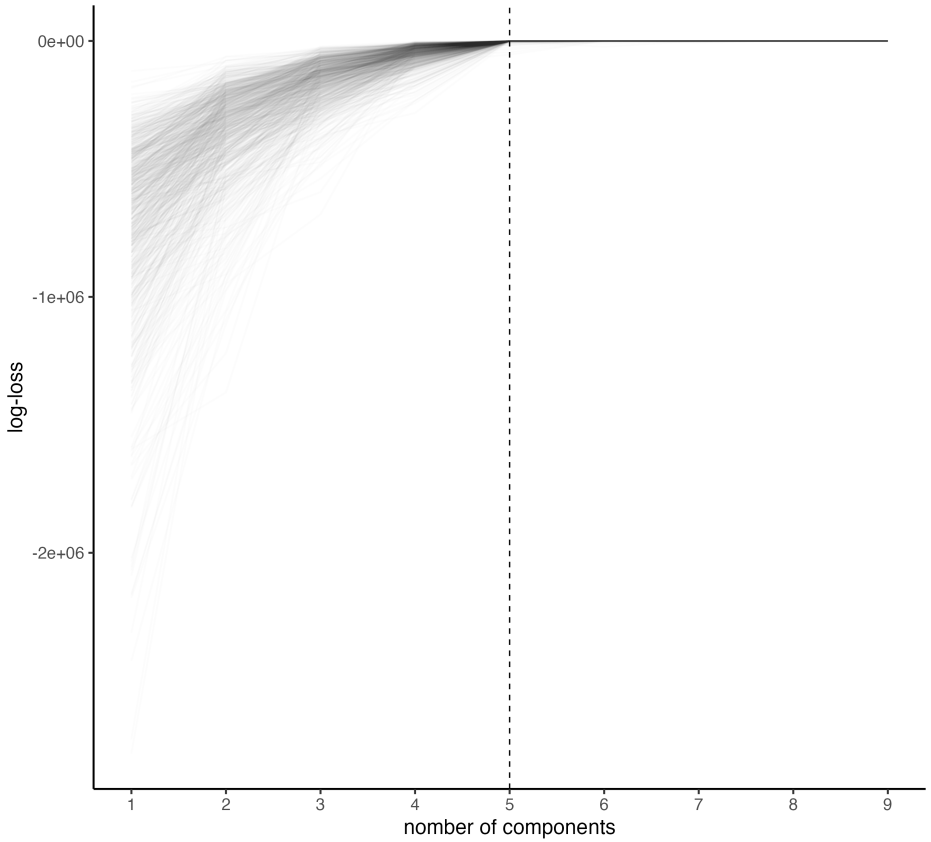


Fig: データの 10%をホールドアウトしてテストデータとし, 対数尤度を評価.
1000 回の試行中, $L = 5$ （正解）が 644 回選択

簡単な例：予測

Fig: 左：データとした画像, 右：欠損値を補完した画像.

画像の 8 割をランダムに欠損させ, $L = 15$ とした UNMF で補完. 左図は欠損値に 1 を入れて表示.

```
R> mNMF_vb(Y~row+col+rgb, data=X, L=15)
```

簡単な例：タイタニックデータ

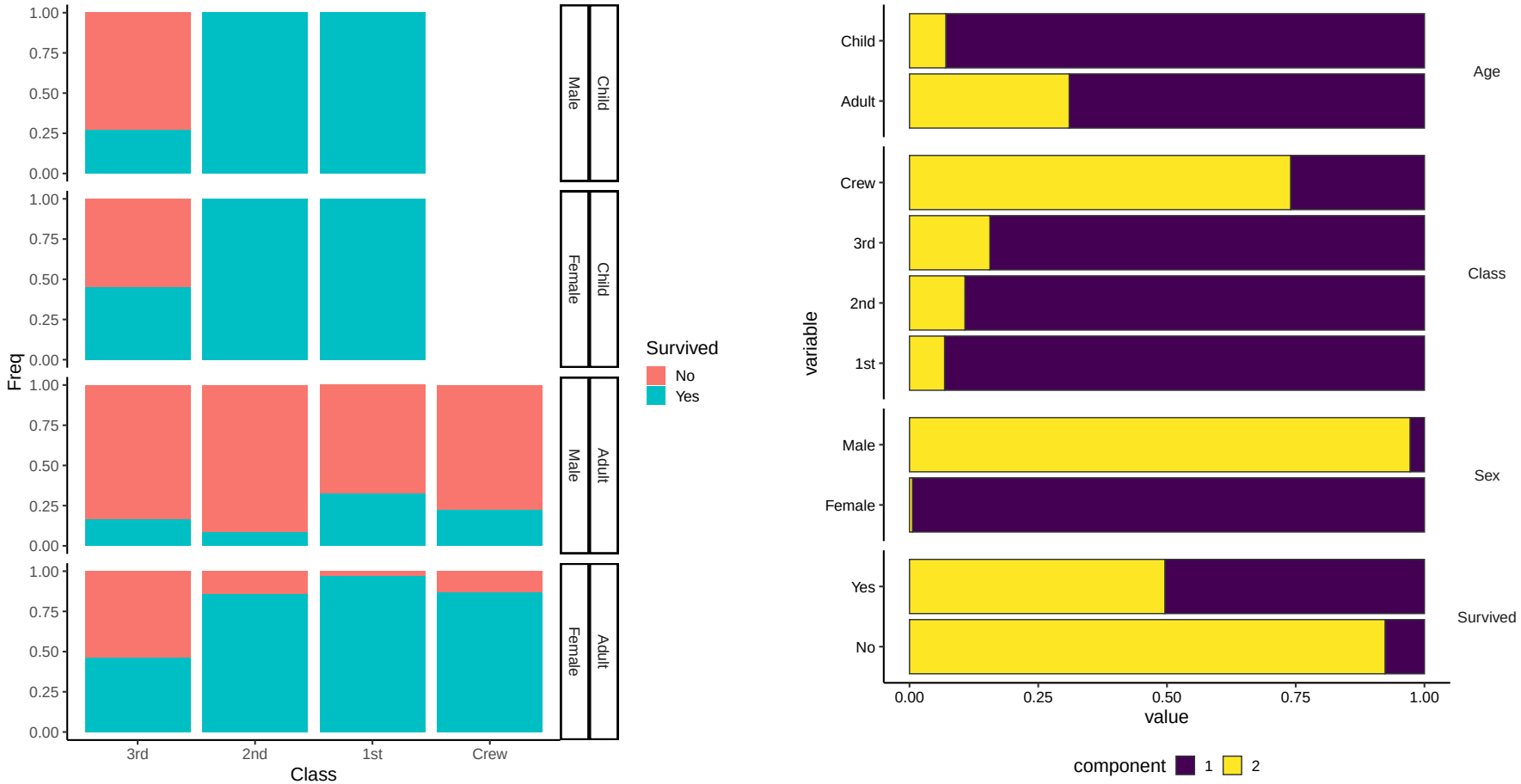


Fig: 左：集計されたデータ, 右：V の推定値

```
R> mNMF_vb(Freq ~ Survived+Class+Sex+Age, data=Titanic, L=2)
```

メタゲノムデータ

- David *et al.*(2014) * は糞便から採取したメタゲノムを縦断的に調査
- 今回はドナー B についてのみ分析
- ドナー B については, 530 種の細菌（不明な種は “unknown” としてまとめて集計）が 176 の時点で記録
- ドナー B は 151 日目から 159 日目の間に腸管感染症に罹患

*David LA, Materna AC, Friedman J, *et al.* (2014). Host lifestyle affects humanmicrobiota on daily timescales. *Genome Biology*, 15(R89). https://doi.org/10.1186/gb2014-157-r89.

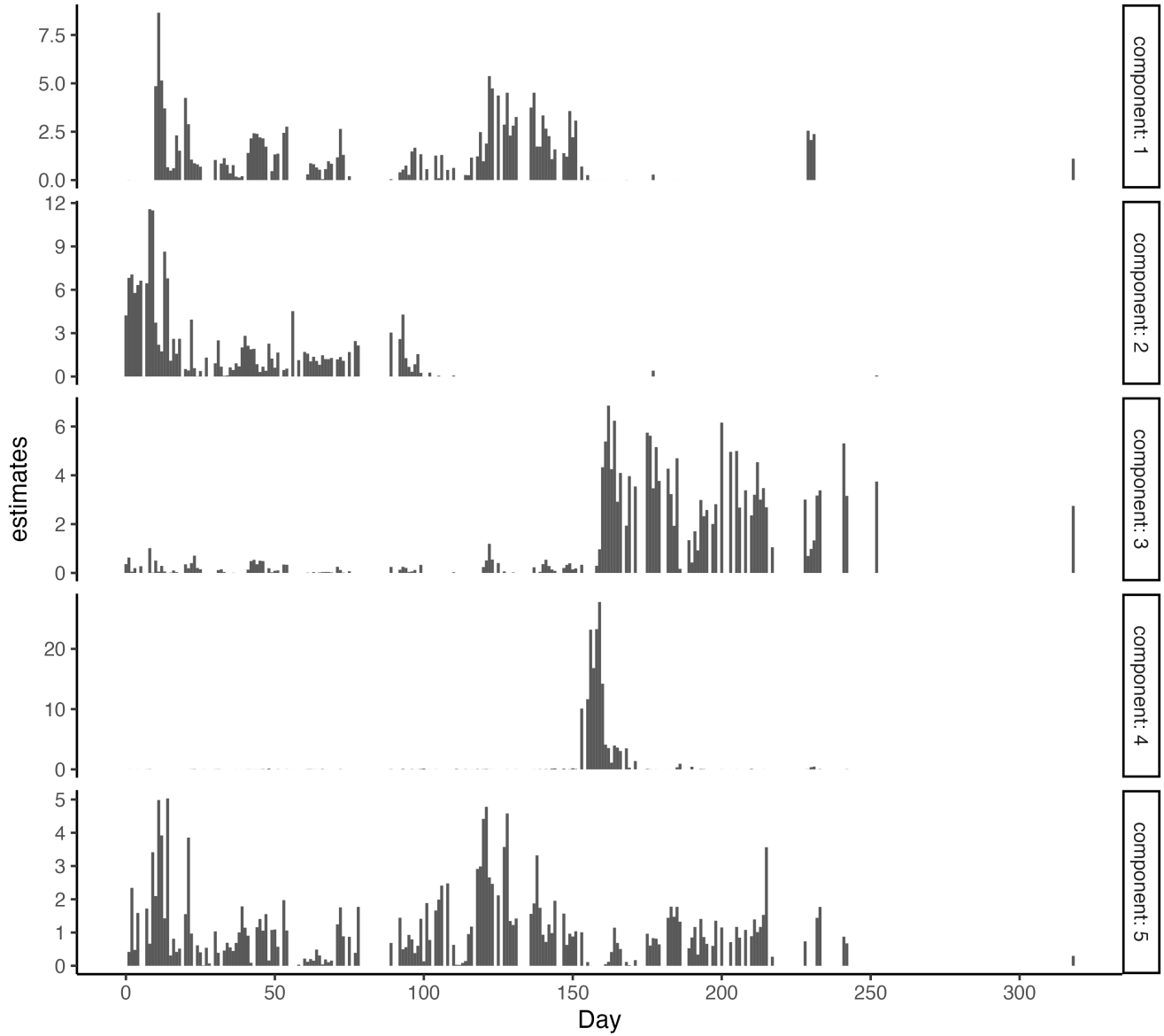


Fig: 推定された V の時間ダミーに対応する部分

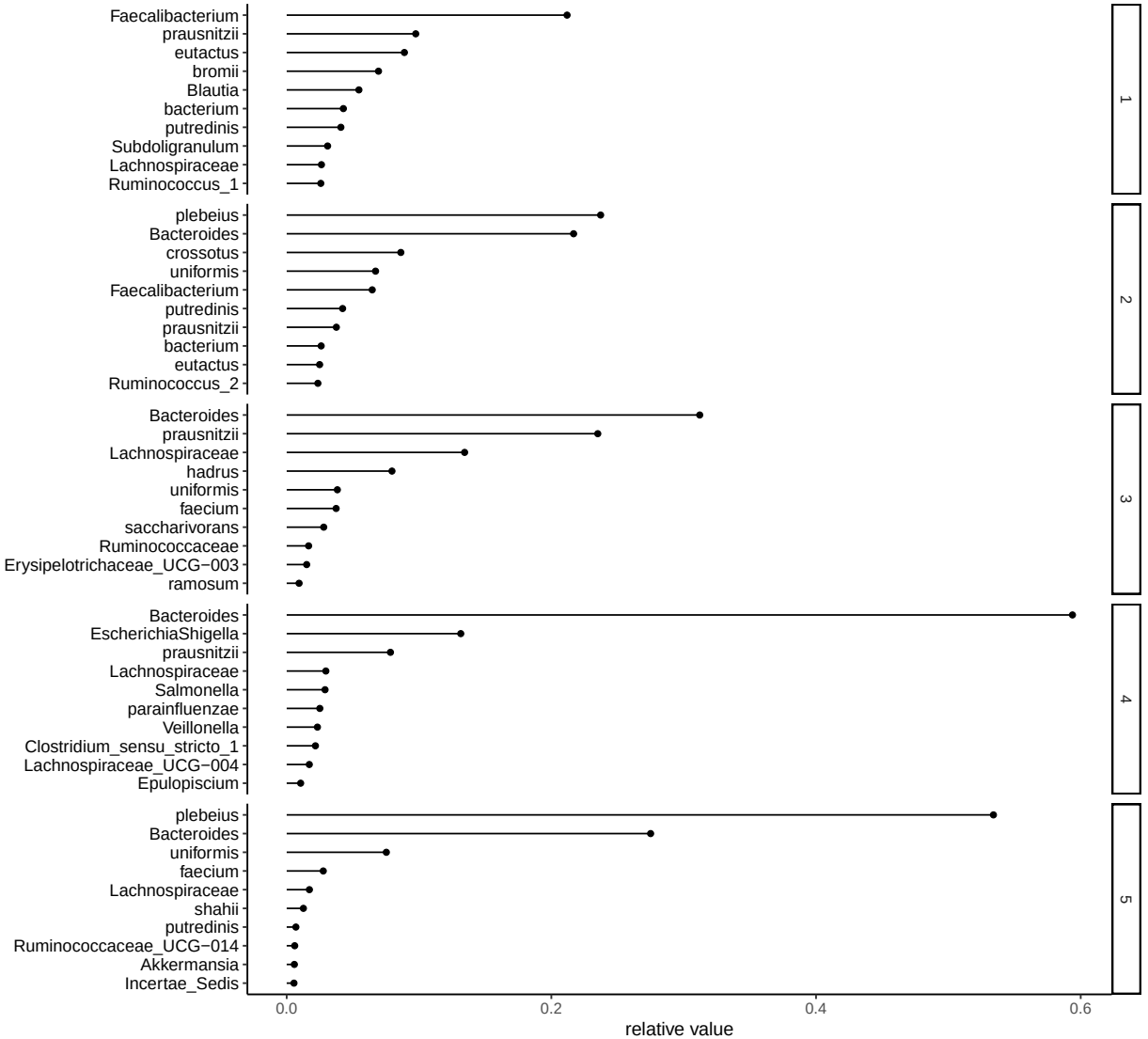


Fig: 推定された V の菌種ダミーに対応する部分. 代表的な細菌上位 10 種のみ.

遺伝子発現解析

- 薬物に対する遺伝子発現応答は創薬の知識につながる
- 細胞, 遺伝子, 薬剤投与量といった条件の組み合わせは多数になり, 薬剤反応の遺伝子発現データには欠損値が多数存在
- 薬物反応遺伝子発現データにおける欠損値の補完が求められる
- 遺伝子発現オムニバス データベース (GEO) から取得した遺伝子発現プロファイル (GSE92742) を分析
- 遺伝子 × 時間 × 細胞タイプ × 投与量
- 補完の精度を評価するために, データのレコードをランダムに削除し, 学習用に

予測の比較

Tab: RMSE の比較. GSE データを 10^6 行にリサンプリングし, 10^5 をテストデータとして、残りを検証用とした

hyperparameter				(XGBoost)
eta	max_depth	lambda	alpha	RMSE
0.30	6.00	1.00	0.00	1314.17
0.30	5.00	0.50	0.00	1340.58
0.50	15.00	0.50	0.10	1032.77
1.50	30.00	1.00	0.00	800.95
				(UNMF)
			L	RMSE
			2	816.67
			3	780.11
			4	762.93
			5	763.50

予測の比較（デフォルト）

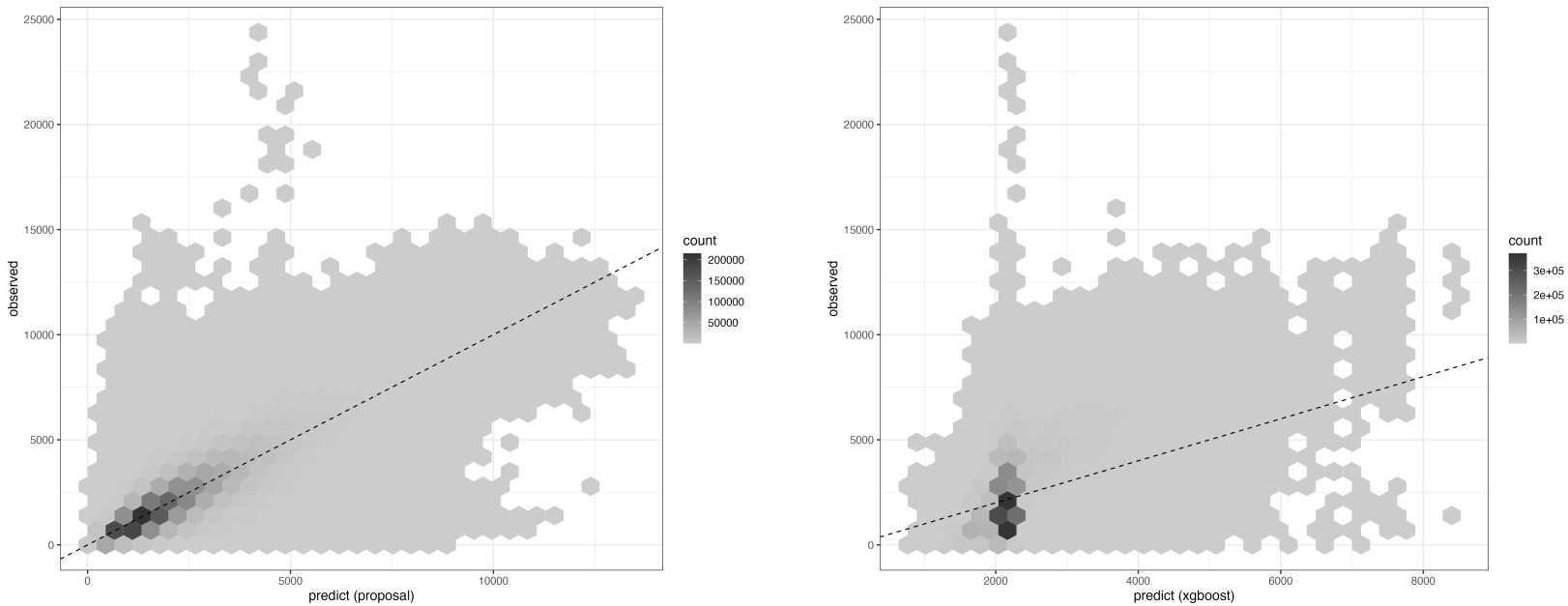


Fig: x 軸: 予測, y 軸: 実測値 (正解). 左: UNMF, 右: xgboost

まとめ

- UNMF は欠測値や反復測定を含む幅広いデータ形式を扱える
- 解釈性を維持したまま予測精度を実現
- 「前処理」が正規化, 平滑化, 欠測値の補完といった複数のステップからなる場合, 分析全体としての自由度は多重に大きくなり, 検証はより複雑になる
- そのため一つのモデルで予測と解釈, 検証を行えることは重要
- 今後の研究：マルチタスク・マルチモーダル学習