4ZJ-01

カーネル行列モデルによる相互補完
Kernel matrix models for mutual completion

下山 愛祐美 * リベロ レイシェル † 加藤 毅 * ‡ §
Ayumi Shimoyama Rachelle Rivero Tsuyoshi Kato

1. はじめに

分子生物学分野においては，それぞれのタンパク質からアミノ酸配列，タンパク質間相互作用，発現データなど様々な表現が得られ，それぞれのデータタイプは機能を予測する上で有用な情報を保持している．これらの情報は，しばしば，細胞内の機能などの推定精度の向上のために，組み合わせて用いられている [2]．

カーネル行列は異種データ統合のための有用な表現方法である [2]．タンパク質の個数を ℓ とすると，カーネル行列の大きさは $\ell \times \ell$ の対称行列になる．データを $\ell \times \ell$ カーネル行列で表現しておくと，カーネル法と呼ばれる一連の方法論が利用可能になる．データの統合は，各々のデータタイプから得られたカーネル行列を単純に平均をとるだけで統合することができる．

分子生物学において複数のデータを統合するときに一部のタンパク質に対して，いくつかのデータタイプが欠けているような状況がある．このような状況においてカーネルの線形結合によるデータ統合を行うには，二つのアプローチがある．その一つは，そのタンパク質を解析対象から除外するか，そのデータタイプを除外する方法である．すると，解析対象が少なくなる，もしくは，有用なデータタイプを一部のデータの欠落だけのために放棄することになってしまう．もう一つのアプローチは，欠落しているカーネルの値を推定する方法である [3, 1]．

Rivero ら [3] は，**モデル行列** $\boldsymbol{M}$ を介して，各種データから得られる不完全なカーネル行列を相互に補完する方法を提案した (図 1)．しかし，そのモデル行列の**自由度**は大きく，過整合してしまう恐れがあった [3]．本論文では，モデル行列の自由度を適切に調整できる新しいモデルを提案し，実験により有効性を示す．

2. 問題設定

本研究で議論する**相互カーネル行列補完** (MKMC) というタスクを述べる．K 個の**不完全カーネル行列** $\boldsymbol{Q}^{(k)}$ $(k = 1, \ldots, K)$ が 図 1 のように与えられていたとする．$\boldsymbol{Q}^{(k)}$ の行列のサイズはいずれも $\ell \times \ell$ である．K 個の情報源のいずれかの対象に関する情報が欠損しているとする．すると，それぞれの情報源のカーネル行列 $\boldsymbol{Q}^{(k)}$ における対応する行と列は欠損していることになる．図 1 では欠損要素を白色，観測できる要素を灰色で示している．提案法は，$\boldsymbol{Q}^{(k)}$ 中の欠損した行と列の値を推定する．K 個の行列における欠損要素の集合を $\mathcal{H}$ と表すことにする．

*群馬大学大学院理工学府
†フィリピン大学数学研究所
‡群馬大学次世代モビリティ社会実装研究センター (CRANTS)
§早稲田大学規範科学総合研究所 (IIRS)

3. 既存の方法: FC-MKMC

先行研究 [3] において開発した方法論 **FC-MKMC** では，$\ell \times \ell$ のモデル行列 $\boldsymbol{M}$ を導入し，各不完全カーネル行列 $\boldsymbol{Q}^{(k)}$ とモデル行列 $\boldsymbol{M}$ との距離の和を最小化するように欠損部分 $\mathcal{H}$ とモデル行列 $\boldsymbol{M}$ を求めていた．行列間の距離として，カーネル行列を共分散行列に持つ正規分布の**カルバックライブラ距離** $\mathrm{KL}\left(\boldsymbol{Q}^{(k)}, \boldsymbol{M}\right)$ を用いていた．すなわち，先行研究の目的関数は

$$J_{\mathrm{FC}}\left(\mathcal{H}, \boldsymbol{M}\right) := \sum_{k=1}^{K} \mathrm{KL}\left(\boldsymbol{Q}^{(k)}, \boldsymbol{M}\right) \tag{1}$$

となる．この目的関数の最小化は，本質的には，正規分布モデルの最尤推定を行っていることを示すことができる [3]．最尤推定において**過整合**を防ぐためには，モデルの自由度を適切に設定する必要がある．FC-MKMC は，モデルとして対称行列を制約なしに用いているため，$(\ell + 1)\ell/2$ という大きな自由度に固定されていた．

4. 提案法その 1: PCA-MKMC

本研究では，モデル行列 $\boldsymbol{M}$ を行列 $\boldsymbol{W} \in \mathbb{R}^{\ell \times q}$ および正のスカラー $\sigma^2 \in \mathbb{R}$ を使って，次の形式に制限することにした：

$$\boldsymbol{M} = \boldsymbol{W}\boldsymbol{W}^{\top} + \sigma^2 \boldsymbol{I}. \tag{2}$$

このモデル (以後，**PCA-MKMC** と呼ぶ) の自由度は $\ell q + 1 - (q-1)q/2$ である．$\boldsymbol{W}$ の列数 q を変更することでモデルの自由度を任意に調整することができる．目的関数 $J(\mathcal{H}, \boldsymbol{W}\boldsymbol{W}^{\top} + \sigma^2 \boldsymbol{I})$ を最小化する $\mathcal{H}, \boldsymbol{W}, \sigma^2$ は次の 2 ステップを反復することで見つける：

補完ステップ

$$\mathcal{H}^{(t)} := \operatorname*{argmin}_{\mathcal{H}} J(\mathcal{H}, \boldsymbol{W}^{(t-1)}(\boldsymbol{W}^{(t-1)})^{\top} + \sigma_{t-1}^2 \boldsymbol{I});$$

モデル更新ステップ

$$(\boldsymbol{W}^{(t)}, \sigma_t^2) := \operatorname*{argmin}_{\boldsymbol{W}, \sigma^2} J(\mathcal{H}^{(t)}, \boldsymbol{W}\boldsymbol{W}^{\top} + \sigma^2 \boldsymbol{I}); \tag{3}$$

ただし，添え字の t および $(t-1)$ は反復回数を表す．補完ステップは FC-MKMC と同様な方法で実現できる．本研究の主要な理論的成果は，モデル更新ステップを閉形式で与えられることを発見したことである．

Theorem 1. 行列 $\boldsymbol{S} := \sum_{k=1}^{K} \boldsymbol{Q}^{(k)}/K$ の q 個の主要な固有値を $\lambda_1, \ldots, \lambda_q$ とし，また，対応する固有ベクトルを $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$ とする．$\boldsymbol{\lambda}_q :=$

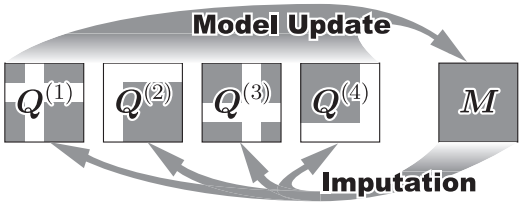

図 1: カーネル行列相互補完法の概要．$\boldsymbol{Q}^{(1)},\ldots,\boldsymbol{Q}^{(4)}$ の白い部分が欠損している要素である．

$[\lambda_1,\ldots,\lambda_q]^\top$, $\boldsymbol{U}_q := [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_q]$ とおく．この時，$(\boldsymbol{W}^{(t)},\sigma_t^2) = \operatorname*{argmin}_{\boldsymbol{W},\sigma^2} J(\mathcal{H}^{(t)}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I})$ となる最適な $(\boldsymbol{W},\sigma^2)$ は

$$\sigma_t^2 = \langle \boldsymbol{\lambda}_q, \mathbf{1}\rangle / (\ell - q), \quad \boldsymbol{W}^{(t)} = \boldsymbol{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\boldsymbol{I})^{1/2}$$

で表される．ただし，$\boldsymbol{\Lambda}_q := \mathrm{diag}(\boldsymbol{\lambda}_q)$ とする．

5. 提案法その２：FA-MKMC

PCA-MKMC のモデル (2) に代わる定式化として次のようなモデル行列も考えることができる：

$$\boldsymbol{M} = \boldsymbol{W}\boldsymbol{W}^\top + \mathrm{diag}(\boldsymbol{\psi}). \tag{4}$$

式 (2) では第 2 項を単位行列の定数倍に限定していた．このモデル (以後，**FA-MKMC** と呼ぶ) の自由度は $\ell q + \ell - (q-1)q/2$ である．FA-MKMC モデルは (2) の第 2 項を対角行列に一般化したものである．この一般化によって，モデル更新ステップにおいて，最適なモデルパラメータを閉形式で表現できなくなる．本研究では，最小化に代替する更新方法として，目的関数の非増加

$$\begin{aligned} &J(\mathcal{H}^{(t)}, \boldsymbol{W}^{(t)}(\boldsymbol{W}^{(t)})^\top + \mathrm{diag}(\boldsymbol{\psi}^{(t)})) \\ &\quad \le J(\mathcal{H}^{(t)}, \boldsymbol{W}^{(t-1)}(\boldsymbol{W}^{(t-1)})^\top + \mathrm{diag}(\boldsymbol{\psi}^{(t-1)})) \end{aligned} \tag{5}$$

が保証される更新式を発見した．その更新式とは次のようなものである：

$$\begin{aligned} \boldsymbol{W}^{(t)} &:= \boldsymbol{S}_{xz}^{(t)} \left(\boldsymbol{S}_{zz}^{(t)}\right)^{-1}, \\ \boldsymbol{\psi}^{(t)} &:= \mathrm{diag}\left(\boldsymbol{S}^{(t)} - \boldsymbol{S}_{xz}^{(t)}(\boldsymbol{S}_{zz}^{(t)})^{-1}(\boldsymbol{S}_{xz}^{(t)})^\top\right), \end{aligned} \tag{6}$$

ただし，

$$\begin{aligned} &\boldsymbol{F}^{(t)} := \left(\boldsymbol{W}^{(t-1)}\right)^\top \mathrm{diag}\left(\boldsymbol{\psi}^{(t-1)}\right)^{-1}, \\ &\boldsymbol{C}^{(t)} := \boldsymbol{I} + \boldsymbol{F}^{(t)}\boldsymbol{W}^{(t-1)}, \\ &\left(\boldsymbol{M}^{(t)}\right)^{-1} := \mathrm{diag}\left(\boldsymbol{\psi}^{(t-1)}\right)^{-1} - \left(\boldsymbol{F}^{(t)}\right)^\top \left(\boldsymbol{C}^{(t)}\right)^{-1} \boldsymbol{F}^{(t)}, \\ &\boldsymbol{B}^{(t)} := \left(\boldsymbol{W}^{(t-1)}\right)^\top \left(\boldsymbol{M}^{(t)}\right)^{-1}, \\ &\boldsymbol{S}_{xz}^{(t)} := \boldsymbol{S}^{(t)}\left(\boldsymbol{B}^{(t)}\right)^\top,\ \boldsymbol{S}_{zz}^{(t)} := \boldsymbol{I} - \boldsymbol{B}^{(t)}\boldsymbol{W}^{(t)} + \boldsymbol{B}^{(t)}\boldsymbol{S}_{xz}^{(t)}. \end{aligned}$$

紙面の制約から証明は割愛するが，実際に次の定理を示すことができる．

表 1: タンパク質機能予測における ROC スコア．

	従来法 [3]			提案法	
機能	ZI	MI	FC-MKMC	PCA-MKMC	FA-MKMC
1	0.791	0.792	0.800	**0.802**	0.801
2	0.792	0.793	0.798	**0.803**	0.801
3	0.794	0.793	0.800	**0.805**	0.803
4	0.842	0.843	0.850	**0.853**	0.852
5	0.884	0.884	0.896	**0.897**	0.896
6	0.767	0.767	0.775	**0.778**	0.777
7	0.832	0.833	0.841	**0.844**	0.843
8	0.734	0.734	0.735	**0.741**	0.739
9	0.762	0.763	0.765	**0.771**	0.769
10	0.744	0.745	0.749	**0.755**	0.753
11	0.577	0.576	**0.583**	0.579	0.579
12	0.938	0.935	0.944	**0.945**	0.944
13	0.682	0.685	0.679	**0.691**	0.684

Theorem 2. 不等式 (5) は各反復で常に成り立つ．

6. 実験結果

出芽酵母におけるタンパク質の機能予測問題にカーネル行列の補完法を適用した．用いたデータセットは文献 [2] と同一で，$\ell = 3{,}588$ 個のタンパク質を含み，各タンパク質は 13 種類の生物学的機能の有無がアノテーションされている．これより，13 種類の 2 クラス分類タスクを構成した．このデータセットは，配列データのカーネル行列や発現データのカーネル行列など $K = 6$ 種類のカーネル行列を含んでいる．各データタイプで，無作為に選んだ 20% のタンパク質に対してデータが得られていないと仮定して，カーネル行列を欠損させた．その上で，提案法で欠損部分を推定した．因子数 q は **Guttman-Kaiser 基準**を用いて選択した．その上で，無作為に選びなおした 20% のタンパク質を訓練用，残りを評価用に用いた．予測性能の評価には ROC スコアを用いた．これを 10 回繰り返した．

13 タスクそれぞれに対する ROC スコアの平均を表 1 に示す．**太字**は，各タスクにおける最高性能，下線は最高性能と統計的有意差がないものである．いずれのタスクにおいても提案法が最高性能を示した．以上の結果は，高い自由度に固定された従来モデルよりも，Guttman-Kaiser 基準によって調整した自由度の提案モデルの方が予測性能が向上することを示唆している．

参考文献

[1] Kato, T., Tsuda, K. and Asai, K.: Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, Vol. 21, pp. 2488–2495 (2005).

[2] Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I. and Noble, W. S.: Kernel-based data fusion and its application to protein function prediction in yeast., *Pac Symp Biocomput*, pp. 300–11 (2004).

[3] Rivero, R., Lemence, R. and Kato, T.: Mutual kernel matrix completion, *IEICE Transactions on Information & Systems*, Vol. E101-D, No. 8, pp. 1844–1851 (2017).