

確率モデルを用いたテンソル因子化法 Tensor Factorization with Stochastic Models

池田 和司

KAZUSHI IKEDA

奈良先端科学技術大学院大学 情報科学研究科

NARA INSTITUTE OF SCIENCE AND TECHNOLOGY

林 浩平

KOHEI HAYASHI

東京大学 大学院情報理工学系研究科

UNIVERSITY OF TOKYO

Abstract

主成分分析や特異値分解などの行列因子化法と同様のアイデアで、一般の次元のテンソルをより自由度の小さいテンソル/行列に分解することができる。これはテンソル因子化法と呼ばれる。誤差を許し近似を伴うテンソル因子化法では、通常、各要素の誤差の二乗和を最小化する。これは誤差がガウス分布に従うと仮定して最尤推定をしているとみなせるので、より一般的な確率モデルを導入したり最尤推定の代わりに MAP 推定あるいはベイズ推定を行うなどの拡張が可能である。本稿ではテンソルの性質および従来のテンソル因子化法を簡単に述べ、その後に確率モデルに基づくテンソル因子化法を紹介する。

Abstract

A tensor can be decomposed to smaller tensors/matrices as the principal component analysis or the singular value decomposition for matrices. This is called the tensor factorization. The conventional tensor factorization with approximation minimizes the Frobenius norm (the sum of the squared elements) of the error tensor. Since this is regarded as the maximum likelihood estimation assuming Gaussian noise, we can extend it to more general distributions or to the MAP estimation or the Bayes estimation with a prior distribution. In this paper, we show some properties of tensors and the conventional tensor factorization and explain our tensor factorization method based on stochastic models.

1 はじめに

主成分分析や特異値分解などの行列因子化法は、機械学習や信号処理の分野において次元削減のために利用されている。行列は2次のテンソルであるので、同様のアイデアで一般の次元のテンソルを扱うことができる。すなわち、多少の誤差を許すことで“大きなテンソル”はより自由度の小さいテンソルに分解される。これがテンソル因子化法である。

近似を伴うテンソル因子化法では、通常、各要素の誤差の二乗和を最小化する。これは誤差がガウス分布に従うと仮定して最尤推定をしているとみなすことができる。したがって、より一般的な確率モデルを導入したり最尤推定の代わりに MAP 推定あるいはベイズ推定を行うなどの拡張が可能である。

本発表では [1] に従ってテンソルの性質および従来のテンソル因子化法を簡単に述べ、その後に確率モデルに基づくテンソル因子化法を紹介する。

2 テンソルの性質

行列 $x = \{x_{ij}\}$ を特異値分解すると

$$x_{ij} = \sum_q \lambda_q u_{iq} v_{jq} \quad (1)$$

となる。ここで、 u, v は正規直交行列である。ある q に対する $\lambda_q u_{iq} v_{jq}$ はランク 1 の行列であることから、特異値分解は行列をランク 1 の行列の和に分解しているとみなせる。

行列に欠損値がある場合や、より小さなランクの行列で近似したい場合には、

$$x_{ij} = \sum_q \lambda_q u_{iq} v_{jq} + \varepsilon_{ij} \quad (2)$$

と誤差項 ε を導入する。拘束条件を満たし、かつ二乗誤差 $\|\varepsilon\|$ を最小化するように λ, u および v を定めることで欠損値の推定や外れ値検知が可能となるため、推薦システムなどに利用されている。

行列は2次のテンソルであるので、同様の考え方を3次以上のテンソルにも応用することができる。ここでは簡単のため、3次のテンソル $x = \{x_{ijk}\}$ だけを考える。特異値分解と同様に、 x は

$$x_{ijk} = \sum_q u_{iq} v_{jq} z_{kq} \quad (3)$$

という形でテンソルの和に分解することができる。ただし、ここでは u, v, z は正規直交でなくてもよい。これは PARAFAC と呼ばれるテンソル因子化法である [2]。ここで、ある q に対する $u_{iq} v_{jq} z_{kq}$ をランク 1 テンソルと呼び、 x を表すのに必要なランク 1 テンソルの数をテンソル x のランクと呼ぶ。この定義は行列におけるランクの定義の自然な拡張

張であるが、テンソルの次数が3以上の時には、ランクの性質は行列のそれとは大きく異なる。例えば、一般にランクはテンソルの次元より小さいとは限らず、また、

$$x_{ijk} = u_i v_j z_k + \varepsilon_{ij} \quad (4)$$

において $\|\varepsilon\|^2$ が最小になるように u_i, v_j, z_k を定めても、 ε のランクが x のランクより小さいとは限らない [5].

なお、テンソルのランクは他にも定義があり、例えばある1成分に着目して残りの成分を列挙することで行列表示し、その行列のランクをテンソルのランクとすることも可能である。すなわち、

$$x_{i,(jk)} = u_{iq}(v_{jq}z_{kq}) \quad (5)$$

という行列のランクが x のランクである。

3 テンソル因子化法

ランク1のテンソルの和ではなく、コアテンソルと呼ばれる“小さなテンソル” $t = \{t_{qrs}\}$ と基底となる行列 $u = \{u_{iq}\}$, $v = \{v_{jr}\}$, $z = \{z_{ks}\}$ の積に分解するのが Tucker 分解と呼ばれる因子化法である [3]. すなわち、残差項 $\varepsilon = \{\varepsilon_{ijk}\}$ を許して

$$x_{ijk} = \sum_q t_{qrs} u_{iq} v_{jr} z_{ks} + \varepsilon_{ijk}. \quad (6)$$

と分解する。分解は残差項の Frobenius ノルム

$$\sum_{ijk} \left(x_{ijk} - \sum_q t_{qrs} u_{iq} v_{jr} z_{ks} \right)^2 \quad (7)$$

が小さくなるように t, u, v, z を選ぶ。上式は各変数については2次式であるので、他を固定して順番に最適化する Alternating Least Squares (ALS) によって局所最適化できる。

しかし ALS は局所解が多く適切な解を得られないため、Higher-Order SVD (HOSVD) という方法が提案されている [4]. HOSVD は前説の“もう一つのランクの定義”のように、テンソルを

$$x_{i,(jk)} = \sum_q u_{iq} \sum_{r,s} t_{qrs} (v_{jr} z_{ks}) \quad (8)$$

と行列とみなし、3つの行列の積に分解する。ここで u, v, z は直交行列、 t は all-orthogonal と呼ばれる条件 (u を求める上の例では、 $t_{q..}$ が互いに直交) を満たすように選ぶと、これは SVD の自然な拡張となる。この時、 $\|t_{q..}\|$ が特異値に相当する。

4 テンソル因子化法への確率モデルの導入

Tucker は二乗誤差を最小化するので、誤差項 ε の各要素が独立にガウス分布に従うと仮定して最尤推定をしているとみなせる。すなわち、 $\varepsilon \sim N(0, \sigma^2)$, $y_{ijk} = \sum_q t_{qrs} u_{iq} v_{jr} z_{ks}$ として

$$\|x - y\|^2 = -\log p(x|t, u, v, z) \rightarrow \min \quad (9)$$

である。そこで、他の要素も確率変数とみなして事前分布を導入すると、ベイズ統計に基づく推定が可能となる。

pTucker は Chu & Ghahramani によって提案されたテンソル因子化法の一つであり、コアテンソル t を隠れ変数とみなしてガウス事前分布を導入した上で周辺化し、ガウス事前分布を導入した u, v, z を MAP 推定する方法である [6]。すなわち、 $t \sim N(0, \sigma^2)$ として

$$p(y|u, v, z) = \int p(y|t, u, v, z) p(t) dt \rightarrow \max \quad (10)$$

とする。

彼らの実験により、pTucker はコアテンソルの要素数が大きすぎる場合でも過学習しにくいことが示されており、これは解が一意に定まらない特異モデルにおいては、ベイズ推論により過学習が抑えられ、高い汎化性能を示すという Watanabe の理論 [7] と関係している可能性もある。

5 分布のクラスの拡張

pTucker では t, u, v, z および ε はガウス分布に従うと仮定していた。しかし明らかにガウス分布には従わないデータが数多くある。例えば個数データなどは非負整数であり、関係のあるなしを示すデータは 0 または 1 のバイナリである。Hayashi et al. はこのような不均一なデータを扱えるように Tucker および pTucker を拡張した Exponential-family Tensor Factorization (ETF) および Bayesian ETF (BETF) を提案した [8]。これらでは x の従う分布をガウス分布から一般の指数分布族に拡張し、

$$x_{ijk} \sim \exp [x_{ijk} w_{ijk}^T t - \psi_{ijk}(w_{ijk}^T t) + f_{ijk}(x_{ijk})] \quad (11)$$

としている。ただし

$$w_{ijk} = u_i \cdot v_j \cdot z_i \quad (12)$$

である。

BETF においては、すべてガウス分布を仮定する pTucker とは異なり、 t による周辺化を陽に解くことができない。そのため、EM アルゴリズムによる反復解法を利用してい

る。EM アルゴリズムの E ステップにおいては事後分布の十分統計量を計算するが、ここではラプラス近似を用いている。一方 M ステップにおいてはその積分を最大化する u, v, z を求めるが、ここではガウス過程による近似計算を行っている。

Hayashi et al. は BETF を人工データ、オフィスログデータ、cross-national statistics データのそれぞれにおける欠損値予測および異常検出に応用し、その優位性を実験的に示している [8]。

6 ダイナミクスの導入

テンソル因子化法は強力なツールであるが、成分の順序についての情報は全く考慮されていない。たとえば適当な列を置換すれば、分解結果も単に置換したものが得られる。このことは、テンソルの次元の一つが時間のような情報の場合には、重要な情報を失っていることを意味している。

そこで林は、行列因子化の状態空間モデルを考え、3 次のテンソルを時間発展する行列因子によって構成されたものとみなしてその因子を推定する方法を提案した [9]。すなわち、

$$u(t) = u(t-1) + \gamma(t) \quad \gamma_{ki}(t) \sim N(0, \gamma_0^2), \quad (13)$$

$$v(t) = v(t-1) + \xi(t) \quad \xi_{lj}(t) \sim N(0, \xi_0^2), \quad (14)$$

とモデル化した。

このモデルにおけるベイズ推論は、観測値を $x(t)$ 、推定値を $y(t) = u(t)v(t)$ とすれば

$$p(y(t), u(t), v(t)|x(t)) = p(y(t)|u(t), v(t)) p(u(t)|x(t-1)) p(v(t)|x(t-1)) \quad (15)$$

が成り立つことを利用すると、逐次的に推定することができる。ただしモデルがガウス分布でない時には $p(u(t)|x(t-1))$ および $p(v(t)|x(t-1))$ が陽には求められないため、ラプラス近似を用いて計算している。

本モデルの有効性を人工データで検証した結果、 u および v のランクが小さい時には単純な SVD などと比べてよい汎化性能を示すが、ランクが大きすぎる時には汎化性能が劣化することがわかった。これはモデルのパラメータが多いためにモデルの複雑さが容易に増大し、過学習を起こしていると推察される。

7 まとめ

本稿では、行列の拡張としてテンソルの性質を紹介し、特異値分解の拡張としてテンソル因子化法を紹介した。さらに、従来のテンソル因子化法を統計的推定とみなすことで、より一般的な確率モデルを導入することで MAP 推定や周辺化 MAP 推定問題としてテンソル因子化を定式化できることを示した。また、行列の因子の状態空間モデルを考えると、より適切に時系列を扱うことができることを紹介した。

テンソルの性質は行列と異なる点が多く興味深い研究対象であるが、アルゴリズム導出においても数値計算においてもテンソル特有の性質はあまり利用されておらず、多くの場合、高次元行列として扱われている。今後、テンソルそのものを扱う方法が開発されると、テンソルの本質的な性質を利用することが可能になることが期待される。

- [1] 林, 池田 (2009): 確率モデルを用いたテンソル因子化法の拡張に関するサーベイ, 人工知能学会研究会資料, SIG-DMSM-A902-06.
- [2] R. Harshman (1970): Foundations of the PARAFAC procedure: Models and conditions for an “exploratory” multimodal factor analysis, *UCLA Working Papers in Phonetics*, **16**, 1–84.
- [3] L. R. Tucker (1966): Some mathematical notes on three-mode factor analysis, *Psychometrika*, **31**/3, 279–311.
- [4] L. De Lathauwer, B. De Moor, J. Vandervalle (2000): A multilinear singular decomposition, *SIAM J. Matrix Analysis and Applications*, **21**/4, 1253–1278.
- [5] P. Comon (2009): Tensors vs matrices — usefulness and unexpected properties, *Proc. SSP*.
- [6] W. Chu, Z. Ghahramani (2009): Probabilistic models for incomplete multidimensional arrays, *Proc. AISTAT*.
- [7] S. Watanabe (2001): Algebraic analysis for non-identifiable learning machines, *Neural Computation*, **13**/4, 899–933.
- [8] K. Hayashi et al. (2012): Exponential family tensor factorization: an online extension and applications, *KAIS*, **33**/1, 57–88.
- [9] 林, 平山, 石井 (2009): 指数族行列因子化の状態空間モデルへの拡張と時系列関係データ解析への応用, 信学技報, NC2008-114.