2B2-2

PSD推定の適用範囲拡大と精度向上に関する研究

- A study on matrix estimation -

グエン　ホン　ハ (1)　鷲尾　隆 (1)　宇野 毅明 (2)　桑島 洋 (1)
Nguyen Hong Ha　Washio Takashi　Uno Takeaki　Kuwajima Hiroshi

(1) 大阪大学産業科学研究所　(2) 国立情報学研究所

This paper assesses some approach to estimate missing elements of some PSD matrices. Its performance has been demonstrated through some numerical experiments.

1. はじめに

事例間の類似性評価は，クラスタリングなどに代表されるデータマイニングの中心的処理で用いられる．近年, このような処理において大量データを扱う必要性が増している．しかしながら, 科学的，工学的な実験や観測において, 直接の類似性測定に非常にコストがかかる場合, 大量事例間の類似性を測定することは現実的でない．そこで先行研究において, 事例間の類似性から構成される類似性尺度行列が Positive Semi-Definite (PSD) 行列になることに着目し，直接の計算または測定によって一部の事例間の類似性情報のみを収集し, それと PSD 行列の性質を基に, 他の全て事例間の類似性を高速に推定する PSD 推定手法を確立した [1].

しかし従来の PSD 推定手法では，幾つかの事例について他のすべての事例との類似性を知る必要があり，また，この条件を満たしても十分な推定精度が得られえない場合がある．そこで, これらの問題を克服するために, 我々は新たに PSD 推定の適用条件を拡張し, 推定精度を向上する手法を提案する.

2. 関連研究

提案手法では, 既存研究の PSD 推定手法のアルゴリズムである PSD Estimation by column Reduction based on incomplete CHolesky decomposition (PERCH) と，グラフ中の極大クリークを探索列挙するアルゴリズムである MAximal Cliques Enumeration (MACE) を適用して, PSD 推定の適用条件を拡張し, 推定精度を向上する．そこで, はじめに PERCH と MACE の原理について説明する.

2.1 PERCH[1]

全事例の集合を $OB(|OB| = n)$，OB に含まれる事例間の大きさ $n \times n$ の Positive Semi-Definite (PSD) 類似性尺度行列を A とする．また，$OA^k \subseteq OB(k = |OA^k|)$ に含まれる事例間の $k \times k$ 類似性尺度行列を A^k，OA^k に含まれる事例と $OB^{n-k} = OB - OA^k$ に含まれる事例間の $k \times (n-k)$ 類似性尺度行列を B^{n-k}, OB^{n-k} に含まれる事例間の $(n-k) \times (n-k)$ 類似性尺度行列を X^{n-k} とする．A はそれらの部分行列から, $A = \begin{pmatrix} A^k & B^{n-k} \\ B^{n-k\,T} & X^{n-k} \end{pmatrix}$ と構成される．PERCH では，不完全 Cholesky 分解 [3] と PSD 行列の全ての主小行列式が 0 以上である性質 [4] を利用して, 測定または計算によって得られた A の部分行列である A^k と B^{n-k} から，X^{n-k} を推定する．これより, 事例 p, q$\in OB^{n-k}$ の類似度 $x_{p,q}$ の値の存在許容区間 $\hat{x}^{(k)}_{p,q} - \Delta x^{(k)}_{p,q} \leq x_{p,q} \leq \hat{x}^{(k)}_{p,q} + \Delta x^{(k)}_{p,q}$ が得られる．これを我々は PSD 推定と呼ぶ.

2.2 MACE[2]

MACE は, 頂点集合 $V = \{v_1, v_2, \ldots, v_n\}$, 枝集合 $E = \{e_1, e_2 \ldots, e_m\}$ から成るグラフ $G = (V, E)$ から, その全ての極大クリークを探索する．任意の $X, Y \subseteq V$ について, $(X \setminus Y) \cup (Y \setminus X)$ に含まれる最小添え字 i を持つ頂点 v_i が X に含まれる時, X の辞書順は Y より大きいとする．この時, 図 1 により, 最初, G の辞書順で一番大きい極大サイズを持つクリーク K_0 を求めて, 極大クリークの親と子供の関係 [2] から, 重複なく K_0 の全ての子供のクリークを探索する．更に再帰を適用して, グラフの全ての極大クリークを探索する.

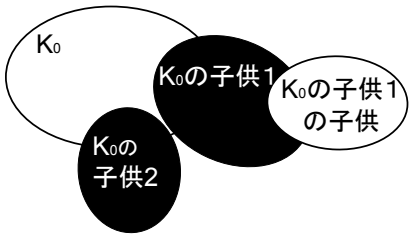

図 1: MACE のアルゴリズムの概要

3. 提案手法: 極大 PSD 推定手法

提案手法の極大 PSD 推定手法では, 任意の未知要素と既知要素を含む不完全 PSD 行列 A を入力して, PERCH[1] と MACE [3] を適用し, 既知要素部分から未知要素部分の真値を推定する．PERCH が適用可能な不完全 PSD 行列は, 既知部分に対応する要素が 1 で, 未知部分のそれが 0 である隣接行列が表すパターン・グラフが, Complete Split Graph with Multiple Independent Vertices (CSGMIV) でなければならない.

図 2 に示すように, CSGMIV はクリークと独立頂点集合から成り, クリークの頂点と独立頂点集合の全ての頂点が接続されたグラフである．これに対して, 提案手法は任意の不完全 PSD 行列に適用可能でなければならない．そこで，与えられた任意の不完全 PSD 行列から，最も精度の高い PSD 推定を可能にする極大な CSGMIV に対応する主小行列をすべて探索し，それぞれに PERCH を適用し，その結果を統合する.

連絡先: 氏名: Nguyen Hong Ha，所属: 大阪大学 産業研究所 鷲尾研究室，住所: 567-0047 大阪府茨木市美穂ヶ丘 8-1，Mail: hongha@ar.sanken.osaka-u.ac.jp

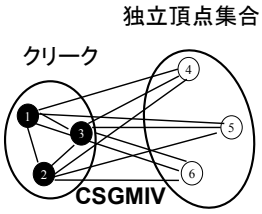

図 2: CSGMIV

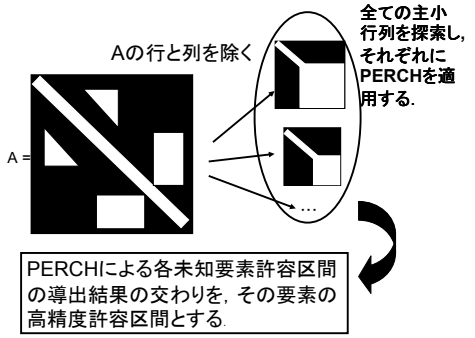

図 3: 提案手法のアルゴリズム

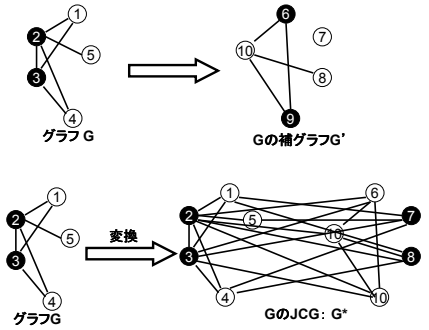

図 4: 各存在許容区間の交わり区間

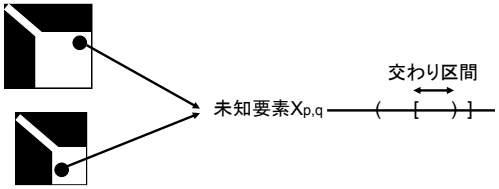

図 5: 各存在許容区間の交わり区間

提案手法の概要を図 3 に示す．左側の行列の黒い部分は既知要素部分で，白い部分は未知要素部分である入力行列 A から,PERCH が適用可能な CSGMIV を表す主小行列を全て探索する．入力行列 A のパターン・グラフ G の全ての CSGMIV の探索問題は，極大クリークの探索問題に帰着できる．図 4 のように，グラフ G の補グラフ G' を求めて，グラフ G と補グラフ G' を連結してグラフ G の Joined Complementary Graph（JCG）G^* を求める．即ち，グラフ $G = (V, E)$, $V = \{v_1, v_2, \ldots, v_n\}$, $E \subseteq V \times V$ の 補グラフを $G' = (V', \bar{E})$, $V' = \{v_{n+1}, v_{n+2}, \ldots, v_{2n}\}$,$(v_{n+i}, v_{n+j}) \in \bar{E}$, ただし $(v_i, v_j) \notin E, \forall i, j = 1, 2, \ldots, n$ とし，グラフ G の JCG $G^* = (V^*, E^*)$, $V^* = V \cup V'$, $E^* = E \cup \bar{E} \cup E'$ を導出する．ただし，各 $(v_i, v_{n+j}) \in E'$ は $(v_i, v_j) \in E, \forall i, j = 1, 2, \ldots, n$ を満たす．グラフ G に含まれる一つの CSGMIV と G の G^* の 1 つの極大クリークは，一対一対応する．例えば，図 4 においてグラフ G に含まれる頂点集合 $\{2, 3, 1, 4\}$ が構成する 1 つ CSGMIV に，G^* に含まれる頂点集合 $\{2, 3, 5, 8\}$ が構成する 1 つの

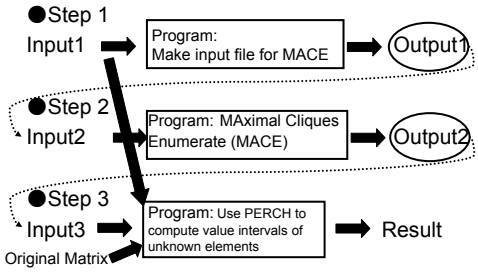

図 6: プログラムの構成

極大クリークが対応する．従って，グラフ G の全ての CSGMIV を探索する問題は，G^* の極大クリークを探索する問題に帰着できる．そこで，前述の MACE を適用し，G^* の全極大クリーク探索によって G の全 CSGMIV を探索する．そして，各 CSGMIV をパターン・グラフとして持つ部分不完全 PSD 行列に PERCH を適用して，未知要素の値の存在許容区間を求める．更に，図 5 に示すように，各未知要素について各部分不完全 PSD 行列から推定される存在許容区間の交わり区間を計算する．即ち，$x_{p,q}$ を含む CSGMIV を表す主小行列が ℓ 個ある場合，PERCH が推定する区間を $[\hat{x}_{p,q(1)} - \Delta x_{p,q(1)}, \hat{x}_{p,q(1)} + \Delta x_{p,q(1)}]$, $[\hat{x}_{p,q(2)} - \Delta x_{p,q(2)}, \hat{x}_{p,q(2)} + \Delta x_{p,q(2)}]$,..., $[\hat{x}_{p,q(\ell)} - \Delta x_{p,q(\ell)}, \hat{x}_{p,q(\ell)} + \Delta x_{p,q(\ell)}]$ とすると，$x_{p,q}$ の区間は $[max_{i=1,\ldots,\ell}(\hat{x}_{p,q(i)} - \Delta x_{p,q(i)}), min_{i=1,\ldots,\ell}(\hat{x}_{p,q(i)} + \Delta x_{p,q(i)})]$ とする．上記の各 PERCH による推定区間と比べ，交わり区間の幅はより狭く，より高精度である．従来の PERCH と比べ，極大 PSD 推定手法の適用可能な不完全 PSD 行列の種類は CSGMIV に限らず幅広く，推定精度もより高いと期待される．

プログラムは，図 6 に示す 3 つのステップより構成される．Input 1 は推定対象とする不完全 PSD 行列 A である．Step1 では，不完全 PSD 行列 A のパターン・グラフ G をその JCG である G^* に変換する．Step2 では，G^* の全ての極大クリークを探索し，G の全 CSGMIV を探索する．Step3 では，各 CSGMIV に対応する部分不完全 PSD 行列に PERCH を適用して，不完全 PSD 行列の未知要素部分を推定し，更に推定した区間の交わり区間を導出する．

4. 性能評価実験

4.1 人工データによる基本性能評価

人工データとして，乱数で n 個の平均値が 0 の m 次元のベクトル $v_i (i = 1, ..., n)$ を生成し，$x_{i,j} =< v_i, v_j > /(|v_i||v_j|)$ を要素とする $n \times n$ の相関係数行列 R を計算し，更に 行列 R から一様ランダムに要素を除去し，不完全 PSD 行列 A を得た．

提案手法をこの人工データに適用した．図 7 と図 8 は事例数 n に対する PERCH との比較結果，図 9 と図 10 は辺密度 β に対する PERCH との比較結果である．ここで辺密度 β とは，グラフ G のノード数を n とした時，$\beta = \frac{G \text{ の辺の数}}{n(n-1)/2}$ である．Interval Length Δ は推定区間の幅，δ は未知要素の真値と推定区間中央値との誤差である．図 7, 図 9 の縦軸は推定区間の幅を表す．図 8, 図 10 の縦軸は誤差を表す．図 7, 図 8 より，n は大きいほど推定精度はより高い．一般に n が大きいほど入力行列 A の CSGMIV を表す主小行列の数は多く，極大 PSD 推定手法の推定では交わり区間を取る操作が増えるため，推定精度が高くなるためと考えられる．図 9, 図 10 より，β が大きいほど推定精度はより高い．これは β が大きいほど入力行列 A の既知要素部分がより多く，PERCH の推定精度が高まるためである．図 7, 図 9 より PERCH と比べると極大 PSD 推定手法の推定区間幅は 1/3 程度に狭くなる．同様に図 8, 図 10 よ

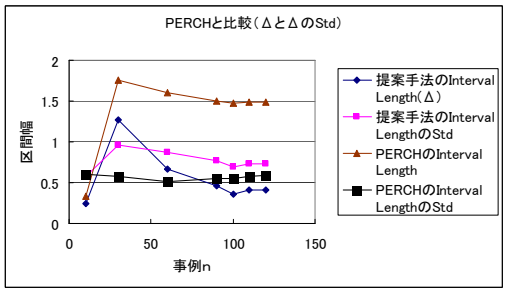

図 7: 事例数 n に関する Δ の精度比較

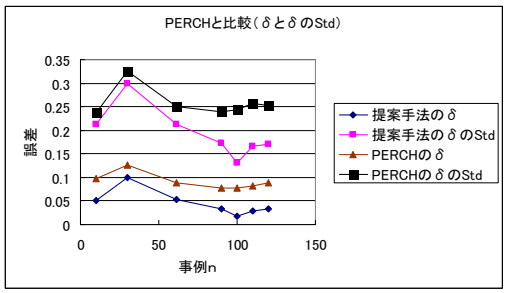

図 8: 事例数 n に関する δ の精度比較

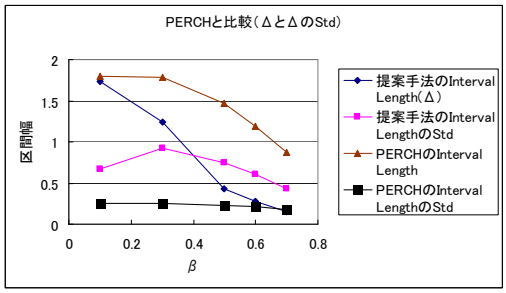

図 9: 辺密度 β に関する Δ の精度比較

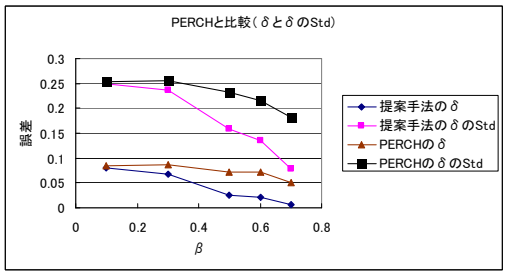

図 10: 事例数 β に関する δ の精度比較

り, PERCH と比べると極大 PSD 推定手法での中央推定値の誤差 δ も同様に 1/3 程度に小さくなる.

一方, 図 7, 図 9 より, 極大 PSD 推定手法の Δ の標準偏差は PERCH のより高い. これは, 極大 PSD 推定手法は各推定区間の交わりを取るので, Δ のバラッキが大きいからである. それに対して, 図 8, 図 10 より, 極大 PSD 推定手法の δ の標準偏差は PERCH より小さい. 極大 PSD 推定手法の方が PERCH より高精度であるので, 誤差 δ のバラッキも小さいためである.

4.2 実データによる性能評価

極大 PSD 推定手法の実適用性を評価するために, UCI Machine Learning Repository[5] の 4 つのデータの musk, isolet, spambase, ionosphere に適用した. musk, isolet, spambase, ionosphere の事例数はそれぞれ 6598, 6238, 4601, 351 である. また, 属性数 m はそれぞれ 167, 618, 58, 34 である. これらのデータからランダムに n 個の事例を選んで, 事例ベクトル $v_i(i = 1, ..., n)$ を得た. そして, 人工データと同様に, 相関係数行列 R を計算し, その各要素を一様ランダムに除去して, 不完全 PSD 行列を得た. 以下, これらの結果のうち, musk の結果のみを示す. 図 11, 図 12 は提案手法のステップ 2, ステップ 3 の計算時間と入力不完全 PSD 行列のサイズ n との関係である. この時, 辺密度 β =0.5 とした. 図 11, 図 12 より, n が大きくなると各ステップ 2, ステップ 3 の計算時間もより多くなる. これは, n が大きいほど, グラフ G 中の CSGMIV の数も多くなり, G の全ての CSGMIV を探索する時間もかかるためである. G の CSGMIV 数が多いので, ステップ 3 では PERCH の適用回数が増え, その計算時間も大きくなる. 図 13, 図 14 は提案手法のステップ 2, ステップ 3 の計算時間と辺密度 β との関係である. この時, 入力不完全 PSD 行列のサイズ n =90 とした. 図 13 より, β =0.5 の時, ステップ 2 の計算時間は最小である. これは, β =0.5 の時には, グラフ G もその補グラフ G' も辺密度があまり高くないためである. そのため, グラフ G の JCG G^* の辺密度も余り高くならず, グラフ G^* のクリーク数も余り多くならず, 計算時間が少なくて済む. 一方, β が小さい時, 補グラフ G' は密度が高いので計算時間がかかる. 同様に, β が大きい時には, グラフ G の密度が高いので, 計算時間がかかる. 図 14 より, β が大きいほどステップ 3 の計算時間は大きくなる. これは, β が大きいとき, グラフ G の CSGMIV の数が多くなり, PERCH の適用回数が増えるためである. 図 15 と図 16 は事例 n に対する PERCH との比較結果, 図 17 と図 18 は辺密度 β に対する PERCH との比較結果である. これらより, PERCH と比べて極大 PSD 推定手法の精度は非常に良い. ただし, 図 15, 図 17 より, 極大 PSD 推定手法と PERCH の Δ の標準偏差は同じくらいである. 一方, 図 16, 図 18 により, 極大 PSD 推定手法の δ の標準偏差は PERCH より小さい. これは, 極大 PSD 推定手法が PERCH より高精度なので, 誤差 δ のバラッキも小さいからである. 他のデータ isolet,spambase, ionosphere についても, 同様の傾向が見られた

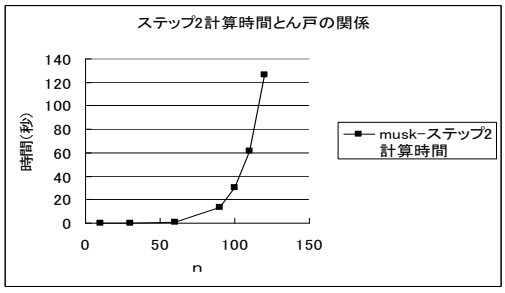

図 11: ステップ 2 計算時間と n との関係

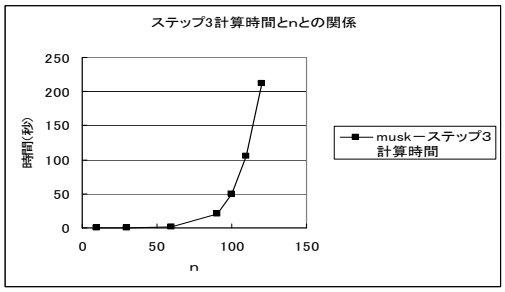

図 12: ステップ 3 計算時間と n との関係

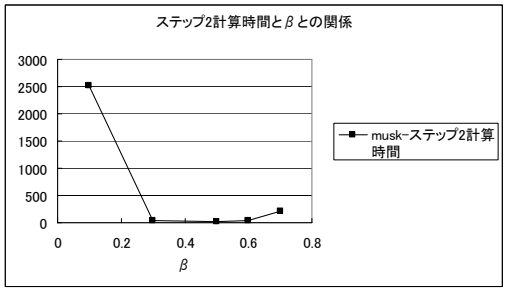

図 13: ステップ 2 計算時間と β との関係

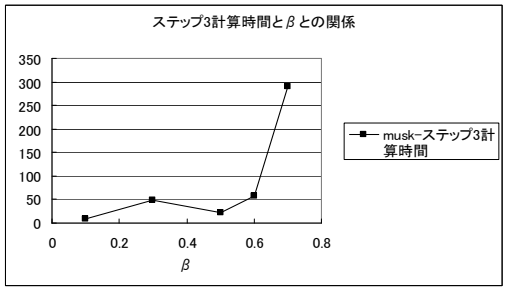

図 14: ステップ 3 計算時間と β との関係

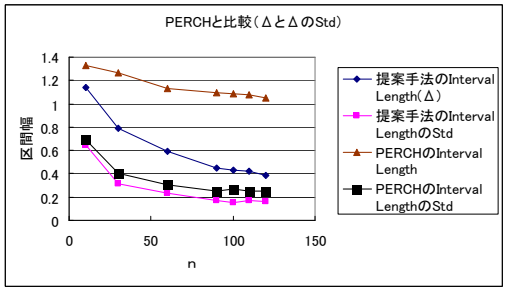

図 15: musk における事例数 n に関する Δ の精度比較係

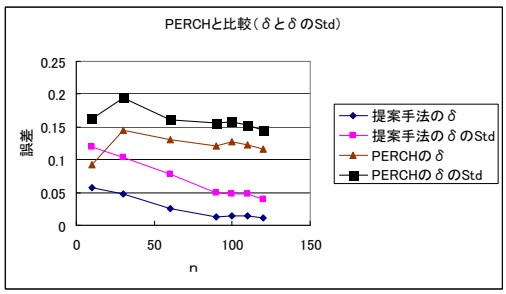

図 16: musk における事例数 n に関する δ の精度比較

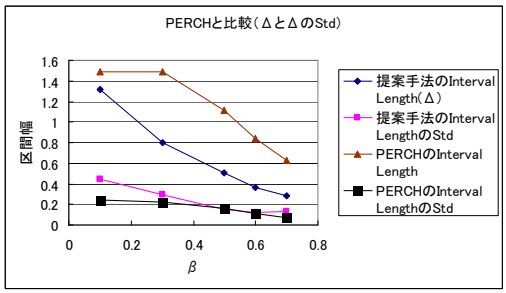

図 17: musk における辺密度 β に関する Δ の精度比較

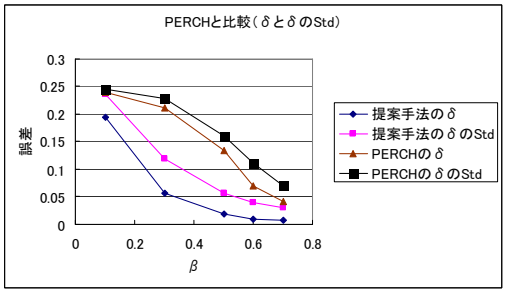

図 18: musk における辺密度 β に関する δ の精度比較

5. まとめ

以上の結果より, 我々が提案した極大 PSD 推定手法は, 人工データ, 実データのいずれにおいても, 推定区間幅 Δ, その中央値誤差 δ ともに, 推定精度が従来の PERCH に比べて非常に高い. 従って, 事例間の類似性を直接測定・計算するとコストがかかり現実的でない場合には, 極大 PSD 推定手法によって高効率, 高精度に類似性を推定できることが判った.

本研究では, 広範な適用性を有する高精度な PSD 推定手法を提案した. 提案手法をプログラム実装し, 人工データと実データに適用, 評価実験を行い, 期待された結果を得た. 今後取り組むべき課題としては, より大規模な不完全 PSD 行列の推定が挙げられる.

参考文献

[1] H. Kuwajima and T. Washio. Large PSD Matrix Estimation from Partial Elements. *Workingnotes of Seventh IEEE International Conference on Data Mining - Workshops: 0-7695-3019-2/07, (2007) IEEE DOI 10.1109/ICDMW.2007.24, pp.337-342.*

[2] K. Makino and T. Uno. New Algorithms for Enumerating All Maximal Cliques. *Proc. of SWAT2004, Scandinavia Workshop on Algorithm Theory, LNCS 3111 (2007), Springer-Verlag, pp.260-272*

[3] R. A. Horn and C. R. Johnson. *Matrix Analysis, Section7.2.* Cambridge University Press, Cambridge, UK, 1985

[4] R. Bhatia. *Positive definite matrices.* Princeton series in Applied Mathematics, Princeton University Press, Princeton, New Jersey, 2007.

[5] C.B. D. J. Newman, S. Hettich and C. Merz. Uci repository of machine learning database. *University of California, Irvine, Dept. of Information and Computer sciences, http://mlearn.ics.uci.edu/MLRepository.html, 1998*