一般口演
一般口演19
医療データ分析７（ DWH・臨床研究データベース）
2018年11月25日(日) 09:00 〜 10:30 F会場 (5F 502+503)

[4-F-1-5] 検査値間相関関係の網羅的解析による新たな診断方法の可能性
○久留宮 千賀[1], 菅田 夏央[1], 兵頭 勇己[2], 永田 桂太郎[2], 畠山 豊[2], 奥原 義保[2] （1.高知大学　医学部 医学科　先端医療学コース, 2.高知大学　医学部　附属医学情報センター）

背景現在の臨床では、鑑別疾患リストから特異度の高い検査の陽性結果による候補確率の上昇と、感度の高い検査の陰性結果による候補除外というプロセスを重ねて疾患を特定する手法が一般的である。つまり、検査結果を順次個別に評価することによって、一つの疾患が特定される仕組みとなっている。しかし、生体内では様々なメカニズムが互いに作用しあい恒常性を保とうとするゆえ、ある検査の値は他の検査の値と密接に関係していると考えられる。病気の状態においてはそのバランスが崩れ、検査値間の関係も健常状態とは異なったものになると考えられる。このため、検査結果を個別に見るだけではなく、検査間の関係性を網羅的に考えることにより、つまり複数の検査値間の関係を同時に考慮することにより疾患を判断する新たな材料となりうる可能性がある。目的疾患発症時における各検査項目間の関係変化を評価するため、検査項目間の相関係数を網羅的に計算し、非発症群(コントロール群)との相関係数の違いを評価した。方法高知大学医学部附属病院における匿名化病歴データベースにおいて、蓄積されている検査データから実施頻度の多いスクリーニング検査60種に対して、同時測定されている検査値を疾患ごとに抽出する。それぞれのスピアマンの相関係数を算出し、コントロール群との相関係数の違いを評価する。結果コントロール群と各疾患群の相関係数が変化することが確認できた。特にネフローゼ症候群患者群では、トータルコレステロールとアルブミンの相関係数がコントロール群と比較して0.73変動した。また、アルブミンと IgG抗体の相関係数が0.79変動した。考察これらの相関係数が大きく変動した検査項目間の関係は、医学知識と合致しており、この評価手法は十分妥当な結果を生成していると考える。そのため、他の関係を精査することで、新しい知見を得ることが期待できる。

検査値間相関関係の網羅的解析による新たな診断方法の可能性

久留宮千賀*1、菅田夏央*1、
兵頭勇己*2、永田桂太郎*2、畠山豊*2、奥原義保*2
*1 高知大学医学部医学科　先端医療学コース、*2 高知大学医学部　附属医学情報センター

Possibility of a new diagnostic method by comprehensive analysis of correlations between laboratory test values

Chika Kurumiya*1, Kao Sugata*1,
Yuki Hyohdoh*2, Keitaro Nagata*2, Yutaka Hatakeyama*2, Yoshiyasu Okuhara*2
*1 Center for Innovative and Translational Medicine, Kochi Medical School, Kochi University, *2 Center of Medical Information Science, Kochi Medical School, Kochi University

Abstract
In the clinical practice, physicians perform clinical reasoning by sequentially evaluating laboratory test results. Since various mechanisms interact with each other in the living body to try to maintain homeostasis, the value of a certain laboratory test is considered to be closely related to the value of other test. In the state of sickness, the balance will change and the relationship between the laboratory test values will be different from the healthy state. We hypothesize that changes in homeostasis due to the onset of disease alterate the relationship between each test. Of the laboratory tests from 1981 to 2016, 60 laboratory test items were extracted in descending order of recorded result frequencies. Patients were divided into disease groups (fatty liver, depression, pancreatic cancer) and non disease groups. In both groups, the correlation coefficients for 60 laboratory item data were calculated comprehensively, and the relationships between laboratory data ware drawn by network graphs and evaluated. The pattern of network graphs in the non disease groups were all similar, and laboratory tests with strong correlation were also almost the same. Common pattern of clusters were recognized for non disease groups in the network graphs. In contrast, clusters in the disease group were characteristic for each disease. Changes in the network graph suggest an alteration of specific homeostatic balance due to disease onset. The visualization of the laboratory tests relationships by the network graph may represent disease properties.

Keywords: data mining, network analysis, laboratory test value

1　緒論

1.1　研究背景

現在の臨床では、検査情報を用いた診断過程において、鑑別疾患リストから特異度の高い検査の陽性結果による候補確率の上昇と、感度の高い検査の陰性結果による候補除外というプロセスを重ねて疾患を特定する手法が一般的である。つまり、検査結果を順次個別に評価することによって、一つの疾患が特定される仕組みとなっている。

しかし、生体内では様々なメカニズムが互いに作用しあい恒常性を保とうとするゆえ、ある検査の値は他の検査の値と密接に関係していると考えられる。病気の状態においては、そのバランスが崩れ、検査値間の関係も健常状態とは異なったものになると考えられる。このため、検査結果を個別に見るだけでなく、検査値間の関係性を網羅的に考えることにより、つまり複数の検査値間の関係を同時に考慮することにより疾患を判断する新たな材料となりうる可能性がある。

関係性を同時に考慮する解析手法の一つとして、ネットワーク理論を応用した解析および可視化の手法がある。特に、複雑なネットワーク理論に基づく研究が、金融工学分野や生物学分野、遺伝医学分野においては盛んに行われている[1][2]。Mantegna は各株式をノード、相関係数をエッジとして市場のネットワークを構築し、その階層構造を明らかにした[3]。また、Szymanski らは、異なる環境ストレス条件に曝露された代謝産物の相関ネットワークの特性変化に着目し、ストレス応答の基礎となるメカニズムを探索した[4]。このように、ネットワーク理論に基づいた解析は、個々の構成要素だけでなく、それらの相互作用および構成要素全体を考慮した全体のシステムとして理解するための一つの方法となり得る。しかしながら、検査値間の関係と健康状態について、このようなネットワーク理論に基づいた解析を行った研究は少ない。

1.2　目的

我々は、検査値を一種の構成要素とし、健康状態を身体のシステムの結果とみなした場合、疾患の発症によるシステムの変化は複数の検査値間の関係性が変化し、その変化はネットワークによって表現されるという仮説のもと、以下の 2 つを本研究の目的とした。

第一に、特定の疾患発症群と、その疾患を発症していない群の 2 群における各検査値間の関係変化を評価するために、検査項目間の相関係数を網羅的に計算し、その相関係数の違いを評価した。

第二に、検査間の関係を、全体として俯瞰的にとらえるために、相関ネットワーク解析による関係性の可視化を行った。

2 方法

2.1 データセット

使用するデータは、高知大学医学部附属病院の匿名化医療データウェアハウス(the Retrieval sYstem for Open Medical Analysis 2 warehouse: RYOMA2)から抽出した。なお、RYOMA2 には 2018 年現在、患者数約 34 万人、臨床検査値数約 1 億 6 千万件のデータをはじめ、病名、薬剤オーダ歴などの外来および入院患者データが格納されている。

2.2 抽出データ

1981 年から 2016 年までに行われた生化学スクリーニング検査の項目のうち、検査結果の記録数が多い順に 60 項目抽出した。また、検査方法の変更がある項目に関しては、補正を行い、比較可能な検査値となるようデータクレンジングを実施した。その他の抽出項目として、年齢、性別、該当患者の病名(病歴)、病名登録日、検査日を抽出した。

2.3 相関係数算出

全患者を、特定の疾患の病名登録が行われた群(以下、発症群)と、行われなかった群(以下、非発症群)の 2 群に分割し、それぞれの群で、60 の検査項目に対し、それぞれ 2 項目間の相関係数を計算することを網羅的に行った。病名登録時(登録前 30 日から登録後 10 日まで)の検査値が対象2検査項目に対し存在している患者データ、非発症群では対象2検査項目における初回検査時の間隔が 90 日以内で存在している患者データを抽出し、スピアマンの順位相関係数をもって各群における対象相関係数として定義した。ただし、20 歳未満または 60 歳以上での測定データは除外した。

なお、今回対象とした疾患は従来の生化学検査だけでは診断が難しい、脂肪肝、うつ病、膵臓癌とした。

2.4 相関ネットワーク解析

それぞれの群における検査項目間の関係をネットワークグラフによって描画し評価を行った。ネットワークは、各検査項目をノード、それらの検査間の対象相関係数をエッジとして構築した。ネットワーク解析における検査項目は、相関係数の絶対値が 0.4 以上の検査項目間を関係ありと定義し、描画の対象とした。

ネットワークグラフの描画アルゴリズムは、力学モデルである Kamada-Kawai アルゴリズム[5]を使用した。このアルゴリズムは、各ノード間に重みに応じた力を割り当て、最終的に力学的平衡状態を作り出した状態を描画する。本研究の場合は、相関係数の絶対値を重みと定義しており、その挙動はエッジにより接続されていないノード、すなわち関係性が低い検査項目は、ネットワークグラフ上は離れて描画される傾向となる。また、各ノードの大きさを該当検査項目の媒介中心性の大きさに比例するよう描画した。媒介中心性は該当ノードに対して、その他の 2 点を結ぶ最短経路がどの程度該当ノードを通過しているかをもとに、中心性を評価する指標である[6]。

相関係数の算出、ネットワークグラフの描画および評価は R (3.0.0) 上にて行った。ネットワークグラフの描画は、igraph パッケージを使用した。

3 結果

評価対象となった検査項目およびその略称を表 1 に示す。概観すると、スクリーニング検査項目の中でも、特に日常診療で頻繁に行われる検査項目が抽出された。

ネットワークグラフにより可視化した結果を図 1–6 に示す。脂肪肝を対象とした際の非発症群、発症群のグラフがそれぞれ図 1,2 であり、うつ病を対象とした際の非発症群、発症群が図 3,4 であり、膵癌を対象とした際の非発症群、発症群が図 5,6 に示す。各ネットワークのノード名は検査項目の略称であり、表 1 で示した検査項目名と対応している。

ネットワークグラフにおいて、相関係数が 0.4 以上の検査項目同士がエッジ(線)で繋がっている。発症群および非発症群におけるエッジ数は、脂肪肝にて 56/50（非発症群/発症群）、うつ病にて 57/45（非発症群/発症群）、膵臓癌にて 57/50（非発症群/発症群）であった。

表1 検査略称と検査項目名

検査略称	検査項目名
A/G	A/G比
ALB	ｱﾙﾌﾞﾐﾝ
ALD	ｱﾙﾄﾞﾗｰｾﾞ（ALD）
ALP	ｱﾙｶﾘ性ﾌｫｽﾌｧﾀｰｾﾞ（ALP）
ChE	ｺﾘﾝｴｽﾃﾗｰｾﾞ
CL	ｸﾛｰﾙ
CRE	ｸﾚｱﾁﾆﾝ
hs-CRP	高感度CRP
DB	直接ﾋﾞﾘﾙﾋﾞﾝ
eGFRcreat	推算糸球体濾過量
F-CHO	遊離ｺﾚｽﾃﾛｰﾙ（F-CHO）
GLB	ｸﾞﾛﾌﾞﾘﾝ
GLU	ｸﾞﾙｺｰｽ
AST	GOT,AST
ALT	GPT,ALT
HB	ﾍﾓｸﾞﾛﾋﾞﾝ
HbA1c	HbA1c
HDL-C	HDL-ｺﾚｽﾃﾛｰﾙ
HT	ﾍﾏﾄｸﾘｯﾄ
gammaGTP	γ-GTP
Ig-A	免疫ｸﾞﾛﾌﾞﾘﾝ A
Ig-G	免疫ｸﾞﾛﾌﾞﾘﾝ G
LAP	ﾛｲｼﾝｱﾐﾉﾍﾟﾌﾟﾁﾀﾞｰｾﾞ（LAP）
LD	乳酸脱水素酵素（LDH）
LYMPHO	ﾘﾝﾊﾟ球
MACRO	大型赤血球率
MCV	平均赤血球容積（MCV）
METAMYEL	後骨髄球
MICRO	小型赤血球率
MPV	平均血小板容積
Na	ﾅﾄﾘｳﾑ
NORMO	正常赤血球率
PDW	血小板分布幅
P-LCR	大型血小板比率
RBC	赤血球数
RDW-CV	赤血球分布幅-CV
RDW-SD	赤血球分布幅-SD
Ca	ｶﾙｼｳﾑ（S-Ca,Ca）
SEGMENT	分葉核球
TB	総ﾋﾞﾘﾙﾋﾞﾝ
TC	総ｺﾚｽﾃﾛｰﾙ
TG	中性脂肪
TP	総蛋白
UA	尿酸
UN	尿素窒素
WBC	白血球数

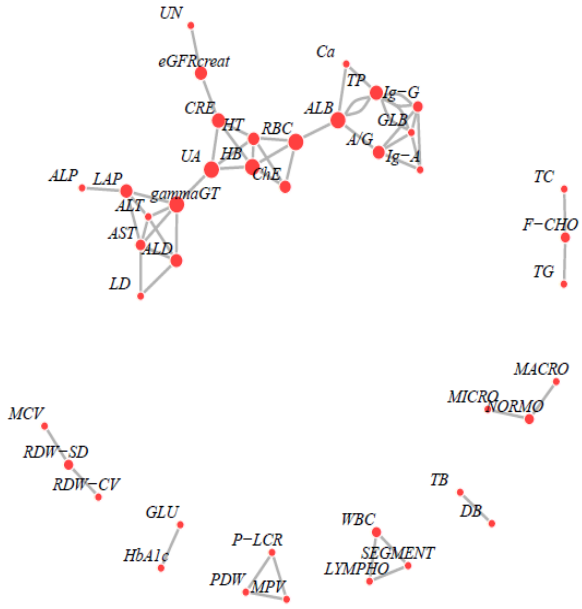

図１　脂肪肝非発症群におけるネットワークグラフ

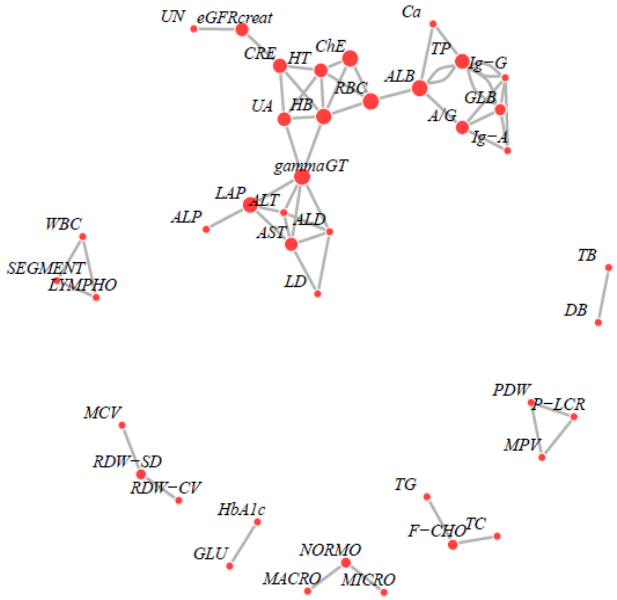

図３　うつ病非発症群におけるネットワークグラフ

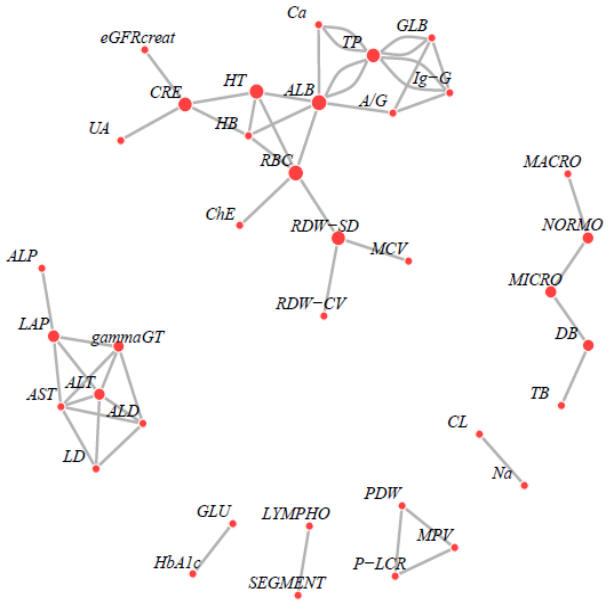

図２　脂肪肝発症群におけるネットワークグラフ

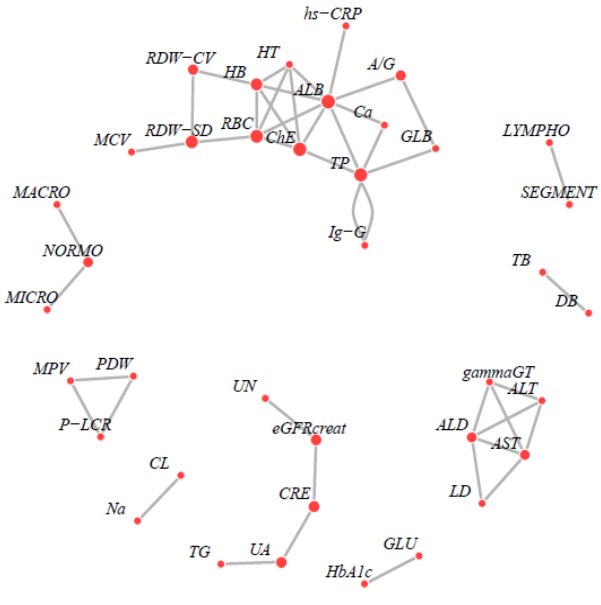

図４　うつ病発症群におけるネットワークグラフ

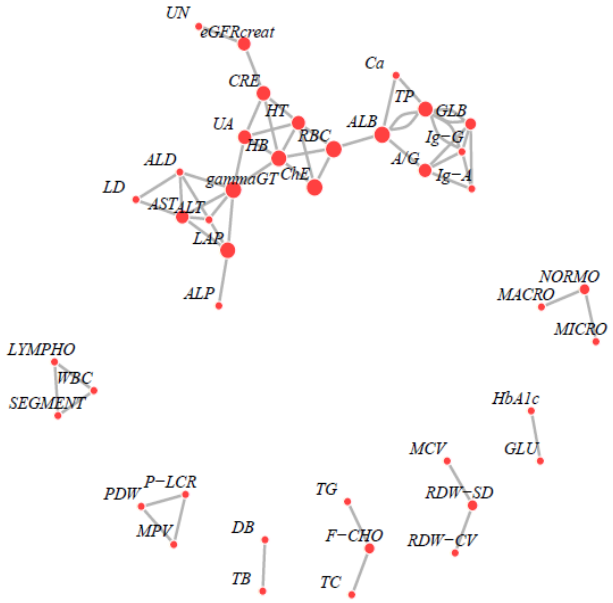

図５　膵臓癌非発症群におけるネットワークグラフ

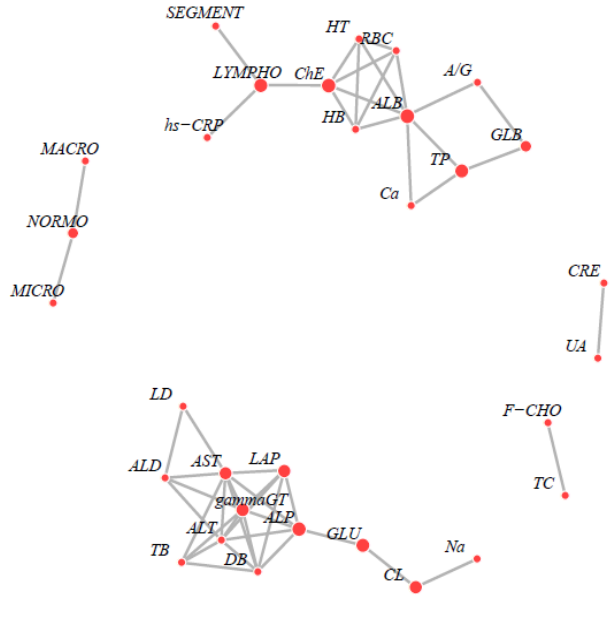

図６　膵臓癌発症群におけるネットワークグラフ

4　考察

結果の図から、非発症群の分布の形態はどの疾患についてもよく似ており、検査値間の相関の強いものなどもおおよそ一致していることがわかる。非発症群は発症群の病名を全く含まない群という定義なので、健常者ではなく、発症群の病名以外の様々な病名は含んでいるはずである。にもかかわらず、ほぼ共通のパターンが見られたのは、様々な疾患に罹患はしていても、それらを全て含めたことによって各疾患の特徴が平均化され、共通のパターンになったと考えられる。

非発症群の共通パターンとして、肝機能、腎機能、血球、蛋白などの検査項目間が、常に連結しており、かつそれらの検査項目が、一つの島になっている結果が得られた。相関があるもの同士の集まりを島と表現する。それ以外にも、独立した島がいくつか見られた。例えば、HbA1c とグルコース、脂質系の検査項目、胆道系などが相関をもつのは妥当な結果である。

発症群では、非発症群にある大きな島の一部が分裂したようになっている。その原因は疾患を発症したことによる特定部位の生体機能に生じた障害にあると考えられる。

また、疾患を個別にみると、脂肪肝では、γ GT, ALD,AST,ALT,LD,LAP,ALP の島が元の大きな島から分離され、独立な島を形成している。脂肪肝が発症している患者は、γ GT 値が異常値を示すことがよく知られている。そのため非発症群で結合していたγ GT と UA が発症群では分離したと考える。逆に、非発症群では独立した島であった RDW-CV,RDW-SD,MCV は RBC と結びついて、一つの大きな島の一部を形成している。

うつ病でも同様に、γ GT、ALD, AST, ALT, LD など肝機能に関連した検査と連結して独立した島として分離している。うつ病の治療薬はほとんど肝臓で代謝されるため、その影響を反映している可能性がある。また、UN, eGFRcreat, CRE UA など腎機能に関係した検査も独立した島として分離している。逆に、独立した島であった MCV,RDW-SD,RDW-CV からなる島が、RBC,HB に連結して大きな島の一部になっている。

膵臓癌では、γ GT や LAP などの肝機能検査が分離し DB,TB,GLU,Na と連結して比較的大きな島を形成している。これは発症によって、肝機能や胆道系の検査異常を示すと言われており、これらの異常を示す検査が互いに関係する結果となった可能性があると考える。腎機能検査の CRE,UA は大きな島からは完全に分離して独立した島になった。

これらの発症群及び非発症群の連結性の変化が示された結果は、特定の疾患を発症することによって恒常性が崩れた状況を反映していると考える。このことは同時に、検査項目それぞれの値が基準値に対して外れているかどうかの評価だけではなく、項目間の値の関係性についても考慮することによって、より精度のよい診断手法の確立が期待できることを示している。本研究では発症による関係性変化の有無の確認を網羅的な探索によって行ったが、今後変化項目の意味についてより詳細なデータに基づき解析を行っていく必要がある。

本研究では検査項目間の関係を相関係数によって評価したが、因果関係を必ずしも示していないことが本研究の限界と考える。つまり、検査項目間の関係が崩れたことによって対象疾患が発症したのか、発症したことによって関係が崩れたのかを識別することは困難である。また、うつ病の考察で示したように薬剤などの治療行為によって引き起こされた関係性変化も識別することは難しい。患者の背景情報などの詳細な評価が必要であると考える。

また、検査値データ分布に基づく解析であるため、本研究

の結果を、直接的に診断方法として活用することは難しい。ただし、疾患発症時の検査値データ解析研究を行う際、本研究で示した結果は重要な検査項目として期待できると考える。

5 結論

我々は、蓄積された診療情報を利用し、日常診療でよく行われる生化学スクリーニング検査項目間の関係性をネットワークグラフによって可視化し、疾患の罹患状態により関係性が変化する可能性を見出した。本研究の結果は、検査間の関係性を網羅的に考える、すなわち複数の検査値間の関係を同時に考慮することにより、疾患を判断する一つの材料となりうることを示している。疾患ごとに異なる島を形成した検査項目に着目した更なる分析は、より複雑な関係性を明らかにし、診断の一助となることが期待される。

参考文献

[1] GUO, Nancy Lan; WAN, Ying-Wooi. Network-based identification of biomarkers coexpressed with multiple pathways. *Caner informatics*, 2014, 13: CIN. S14054.

[2] ROSATO, Antonio, et al. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics*, 2018, 14.4: 37.

[3] MANTEGNA, Rosario N. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 1999, 11.1: 193-197.

[4] SZYMANSKI, Jedrzej, et al. Stability of metabolic correlations under changing environmental conditions in Escherichia coli–a systems approach. *PLoS One*, 2009, 4.10: e7441.

[5] KAMADA, Tomihisa, et al. An algorithm for drawing general undirected graphs. *Information processing letters*, 1989, 31.1: 7-15.

[6] 金明哲. *R によるデータサイエンス: データ解析の基礎から最新手法まで*. 森北出版, 2007.pp228–252