(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2014-13601
(P2014-13601A)
(43)公開日　平成26年1月23日(2014.1.23)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/30	(2006.01)	G06F 17/30	210D	

審査請求 有　請求項の数 18　ＯＬ　（全 27 頁）

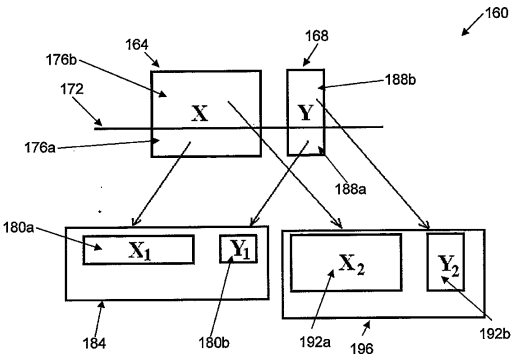
(21)出願番号　特願2013-190084 (P2013-190084)
(22)出願日　平成25年9月13日 (2013.9.13)
(62)分割の表示　特願2010-539868 (P2010-539868)の分割
原出願日　平成20年12月19日 (2008.12.19)
(31)優先権主張番号　61/015,973
(32)優先日　平成19年12月21日 (2007.12.21)
(33)優先権主張国　米国(US)

(71)出願人　592053963
エム　ケー　エス　インストルメンツ　インコーポレーテッド
ＭＫＳ　ＩＮＳＴＲＵＭＥＮＴＳ，ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国マサチューセッツ州０１８１０，アンドーバー，テック・ドライブ２，スイート　２０１
(74)代理人　100140109
弁理士　小野　新次郎
(74)代理人　100075270
弁理士　小林　泰
(74)代理人　100101373
弁理士　竹内　茂雄

最終頁に続く

(54)【発明の名称】部分的最小二乗分析（ＰＬＳ－ツリー）を用いたデータの階層編成

(57)【要約】　（修正有）
【課題】比較的短い処理時間で大量のデータを区分（クラスタリング）する。
【解決手段】方法は、第１データ・マトリクスおよび第２データ・マトリクスを供給することを伴い、第１および第２データ・マトリクスの各々は、１つ又は複数の変数と、複数のデータ点とを含む。また、本方法は、部分的最小二乗（ＰＬＳ）分析または直交ＰＬＳ（ＯＰＬＳ）分析を用いて第１データ・マトリクスから第１スコアを決定すること、ならびに第１および第２データ・マトリクスを（例えば、行単位で）第１グループおよび第２グループに区分することを伴い、ソートした第１スコア、第１データ・マトリクスの分散、ならびに第１および第２データ・マトリクスの分散に関連する第１および第２グループの分散に基づいて区分する。
【選択図】図１Ｂ



【特許請求の範囲】
【請求項１】
コンピュータにより実行される方法であって、
第１データ・マトリクスおよび第２データ・マトリクスを供給する、コンピュータにより実行されるステップであって、前記第１および第２データ・マトリクスの各々が、１つ又は複数の変数（マトリクス列）と、複数のデータ点（マトリクス行）とを含む、供給するステップと、
部分的最小二乗（ＰＬＳ）分析または直交ＰＬＳ（ＯＰＬＳ）分析を用いて、前記第１データ・マトリクスから第１スコアを決定する、コンピュータにより実行される、決定するステップと、
第１スコアの区分値に基づいて、前記第１データ・マトリクスおよび第２データ・マトリクスを第１グループおよび第２グループに行単位に区分する、コンピュータにより実行されるステップであって、該第１スコアの区分値は、（ｉ）前記第１および第２グループにおける前記第１データ・マトリクスの前記第１スコアの分散、（ｉｉ）前記第１および第２グループにおける前記第２データ・マトリクスの分散、及び（ｉｉｉ）区分後の前記第１グループのサイズと第２のグループのサイズとの差に関連する損失関数の合計を備える関数を最小化することによって求められる、区分するステップと、
を備えている、方法。
【請求項２】
請求項１記載の方法であって、区分するステップは、前記ＰＬＳ分析またはＯＰＬＳ分析を用いることにより生成される前記第１スコアの分散と、前記第２データ・マトリクスの分散との間の関係を表すパラメータｕを最小化するステップを含む、方法。
【請求項３】
請求項１記載の方法であって、前記第１データ・マトリクスは、プロセス・データを表すデータを収容する、方法。
【請求項４】
請求項１記載の方法であって、前記第２データ・マトリクスは、収量データ、品質データ、またはその組み合わせを表すデータを収容する、方法。
【請求項５】
請求項１記載の方法であって、前記第１データ・マトリクスは、対象の分子または高分子の構造的変動に関連のある測定データまたは計算データを表すデータを収容する、方法。
【請求項６】
請求項１記載の方法であって、前記第２データ・マトリクスは、同じ分子または高分子の生物的データを表すデータを収容する、方法。
【請求項７】
請求項１記載の方法であって、前記第１グループは、前記第１および第２データ・マトリクスを前記第１および第２グループに行単位に区分した結果各々得られた、第３データ・マトリクスおよび第４データ・マトリクスを含み、前記方法は、更に、
第２部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、前記第３データ・マトリクスから第２スコアを決定する、コンピュータにより実行されるステップと、
前記第２スコアの第２区分値に基づいて、前記第３データ・マトリクスおよび第４データ・マトリクスを行単位に前記第３グループおよび第４グループに区分する、コンピュータにより実行されるステップであって、該第２区分値は、前記第３データ・マトリクスおよび前記第４データ・マトリクスの前記第２スコアに関連する前記第３グループ、および第４グループ内の分散を備える第２関数を最適化することによって求められる、区分するステップと、
を備えている、方法。
【請求項８】
請求項７記載の方法であって、前記第２グループは、第５データ・マトリクスおよび第

６データ・マトリクスを含み、前記方法は、更に、
前記第２グループが閾値数よりも多いデータ点を含む場合、第３部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、前記第５マトリクスから第３スコアを決定する、コンピュータにより実行される、決定するステップと、
前記第３スコアの第３区分値に基づいて、前記第５データ・マトリクスおよび第６データ・マトリクスを行単位に前記第５グループおよび第６グループに区分する、コンピュータにより実行されるステップであって、該第３区分値は、前記第５データ・マトリクスおよび前記第６データ・マトリクスの前記第３スコアに関連する前記第５グループおよび第６グループ内の分散を備える第３関数を最適化することによって求められる、区分するステップと、を備えている、方法。

【請求項９】
請求項８記載の方法であって、更に、前記第１、第２、第３、第４、第５、または第６グループを階層的に表示するステップを備えている、方法。

【請求項１０】
請求項８記載の方法であって、更に、前記第２グループが閾値数未満のデータ点を含む場合、前記第２グループの区分を終了するステップを備えている、方法。

【請求項１１】
請求項８記載の方法であって、更に、前記第２スコアおよび前記第２データ・マトリクスの組み合わせた分散が、前記第２グループを前記第５および第６グループに区分したときに減少しない場合、前記第２グループの区分を終了するステップを備えている、方法。

【請求項１２】
請求項８記載の方法であって、更に、前記第１および第２データ・マトリクスと関連のある以前の区分数が所定の閾値に等しいかまたはこれを超える場合、前記第２グループの区分を終了するステップを備えている、方法。

【請求項１３】
請求項１２記載の方法であって、前記所定の閾値は、デンドログラムにおける階層レベルの最大数を表す制限値である、方法。

【請求項１４】
請求項１記載の方法であって、更に、前記第１データ・マトリクスおよび前記第２データ・マトリクスを表示するグラフ上において、前記第１グループおよび前記第２グループを表示するステップを備えている、方法。

【請求項１５】
請求項１記載の方法であって、四分位 (inter-quartiles)を用いて前記分散を計算する、方法。

【請求項１６】
請求項１記載の方法であって、更に、前記第１データ・マトリクスおよび前記第２データ・マトリクスを、第１セットのデータ行を備える第１グループおよび第２セットのデータ行を備える第２グループに行単位に区分するステップを備えている、方法。

【請求項１７】
データを階層的に編成するシステムであって、
（ａ）メモリであって、
（ａ１）第１データ・マトリクスおよび第２データ・マトリクスを含むデータ構造を含む、メモリと、
（ｂ）前記メモリに動作的に結合されているプロセッサであって、
（ｂ１）前記第１データ・マトリクスの部分的最小二乗分析またはＯＰＬＳ分析に基づいて第１スコアを決定するモジュールと、
（ｂ２）前記第１スコアの区分値に基づいて、第１グループおよび第２グループを発生するために前記第１および第２データ・マトリクスを区分するモジュールであって、該第１スコアの区分値は、（ｉ）前記第１グループおよび第２グループにおける前記第１データ・マトリクスの前記第１スコアの分散、（ｉｉ）前記第１グループおよび第２グルー

プにおける前記第２のデータ・マトリクスの分散、及び（ｉｉｉ）区分後の前記第１グループのサイズと第２グループのサイズとの差に関連する損失関数の合計を備える関数を最小化することによって求められる、モジュールと、
を備えている、プロセッサと、
（ｃ）前記第１および第２グループ、ならびに前記第１および第２グループの前記第１および第２データ・マトリクスに対する関連を表示するために、前記プロセッサに動作的に結合されているディスプレイと、
を備えている、システム。
【請求項１８】
データを分析するシステムであって、
メモリから第１データ・マトリクスおよび第２データ・マトリクスを読み出すデータ読み出し手段であって、前記第１および第２データ・マトリクスの各々が１つ又は複数のデータ点を含む、データ読み出し手段と、
部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、前記第１データ・マトリクスから第１スコアを決定するデータ分析手段と、
第１スコアの区分値に基づいて、前記第１データ・マトリクスおよび第２データ・マトリクスを第１グループおよび第２グループに分割するデータ区分手段であって、該区分値は、（ｉ）前記第１および第２グループにおける前記第１データ・マトリクスの前記第１スコアの分散、（ｉｉ）前記第１および第２グループにおける前記第２データ・マトリクスの分散、及び（ｉｉｉ）区分後の前記第１グループのサイズと第２のグループのサイズとの差に関連する損失関数の合計を備える関数を最小化することによって求められる、データ区分手段と、
を備えている、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
[0001] 本発明は、一般的には、大きなデータ集合の分析に関し、更に特定すれば、端的にはＰＬＳ－ツリーと呼ばれている、部分的最小二乗分析を用いたデータの階層的編成および分析に関する。
【背景技術】
【０００２】
[0002] 多くの産業において、製造ならびに研究および開発の双方において、非常に大きなデータ集合が収集される。
【０００３】
[0003] 半導体デバイス製造業界では、デバイス製造業者は、一層優れたおよび／または高速のプロセスおよびハードウェア構成を設計するためにプロセス・ツール製造業者を拠り所にすることによって、許容範囲を一層狭めたプロセスおよび材料仕様に移行するように管理している。しかしながら、デバイスの外形がナノメートル等級(nanometer scale)まで縮小しているため、製造プロセスが増々複雑になり、プロセスおよび材料の仕様を満たすのも一層難しくなっている。
【０００４】
[0004] 現在の半導体製造において用いられている典型的なプロセス・ツールは、数千個のプロセス変数の集合によって記述することができる。これらの変数は、一般に、製造プロセスの物理的パラメータ、および／または製造プロセスにおいて用いられるツールに関係付けられている。場合によっては、これら数千個の変数の内、数百個の変数が動的であることもある（例えば、製造プロセスにおいて、または製造プロセス間で時間的に変化する）。動的変数には、例えば、気体流量、気体圧力、配給電力、電流、電圧、および温度は、例えば、特定の処理方法(recipe)、処理ステップの全体的なシーケンスにおける特定のステップまたは一連のステップ、製造プロセスの間に起こる誤りおよび故障、または特定のツールまたはチェンバの使用に基づくパラメータ値の変化（例えば、「ドリフト」と

呼ばれる）に基づいて変化する。

【０００５】

[0005] プロセス変数は、収量(yield)変数または応答変数と関係付けられることが多い。プロセス変数は、予測子と考えることができる。即ち、変数間の基礎となる関係に基づいて収量変数を示すと考えることができる。プロセスおよび収量変数を示すデータは、製造プロセスの間に測定され、リアル・タイムの分析または後の分析のために格納される。

【０００６】

[0006] 同様に、薬品およびバイオテク生産では、U.S. Food and Drug Administration（米食品医薬品局）のような監督官庁が、指定された品質概要を中心としてばらつきが非常に少ない高品質の製品を維持するために、製造プロセスに対する厳格な仕様の遵守を要求している。これらの仕様は、プロセス変数のオンライン測定、ならびに、例えば、プロセス・ガス・クロマトグラフィ、近赤外線分光分析、および質量分光分析のような追加の多次元センサ技法を必要とする。理想的には、製造プロセスの間に測定されたデータがリアル・タイムの分析に利用することができ、プロセス状態がプロセス仕様にどの位近いかに関する指示または情報を提供することである。

【０００７】

[0007] 薬品およびバイオ技術の研究および開発では、数万以上のことも多い、多くの異なる分子が、新たな薬品を発見し最適化するプロセスの間に調べられる。多くの異なる物理的および生物的特性が、分子毎（例えば、潜在的な薬品候補）に測定および／または計算され、多くの理論的構造関係特性が分子毎に計算される。分子毎に決定された変数値の総数は数千を越える（例えば、２，０００個の変数値よりも多い）こともしばしばである。開発プロセスの一部には、一方では生物的特性と、他方では物理的、化学的、そして理論的に計算された構造関係特性との間における関係を発見することを含む。これらの関係を理解することによって、研究者は将来性のある分子の化学構造を変更し、生物的特性の概要を改善した新たな分子に向かって進むのに役立つ。

【０００８】

[0008] 大きなデータ集合では、多くの場合、データを互いにグループ化して、クラスタ化データを得る。データに対して有意な分析を行うために、同質のデータ即ちグループ化されないデータ間の比較が好ましい。したがって、グループ化したデータを同質のサブグループにクラスタ化するためのアルゴリズムが開発されている。

【０００９】

[0009] グループ化したデータを分析する１つの方法は、データに関する線形回帰分析の変形体（例えば、「回帰ツリー」または「分類および回帰ツリー」即ち「ＣＡＲＴ」と呼ばれることもある）を用いることである。回帰ツリー分析は、個々のＸ－変数またはＸ－変数の組み合わせに基づく一連のデータ分割を伴う。データを分割することができる可能性がある方法の数は、観察する変数の数と共に急激に増加する。この理由のために、回帰ツリーが一般に適しているのは、数個の変数だけを有するデータ集合であり、１０から２０個よりも多い変数を有するデータ集合の場合、回帰ツリー分析は、部分的に計算コストのために、データ集合を細分するのが一般的である。回帰ツリー分析の結果に基づいて、データをツリーまたは分岐編成にグループ化する。これは、デンドログラム(dendrogram)と呼ばれることもある。

【００１０】

[0010] 階層データ・クラスタリングの一種に、主成分分析（ＰＣＡ）に基づくものがある。このような技法は、階層レベル毎に、データ集合をＰＣＡ分析の第１主成分軸上に投影することを伴う。こうして、投影されたデータは、第１主成分軸に沿って一次元的に整列され、第１主成分軸上の中央位置付近においてデータは区分(partition)される。この種の区分またはクラスタリングは、クラスタ・メンバ間の最大距離が所定の（例えば、ユーザが定めた）閾値を超過するまで、再帰的に反復される。ＣＡＲＴ分析と同様、ＰＣＡに基づく分析も、大きなデータ集合には比較的遅い。更に別の欠点は、ＰＣＡに基づく分析が通常Ｘ－変数のみを考慮し、得られたデータ関係に対するＹ－変数の影響を無視する

ことである。
【００１１】
[0011] 別の技法に、Ｙ－変数を２つのランダム・グループに分割する、ランダム、二進（０または１）Ｙ－ベクトル値を伴うものがある。部分的最小二乗（ＰＬＳ）アルゴリズムを用いて、１－成分モデルを用いて新たなＹ－変数を予測し、予測したＹ－変数がランダムＹ－変数値に取って代わる。分析が収束した後、予測したＹ－変数を最も近い整数（例えば、０または１のいずれか）に丸め、この丸めたＹ－変数を用いて、データをグループに区分する。ＰＣＡに基づく分析およびＣＡＲＴ分析と同様、この手法は、内部計算用ＰＬＳを用いるにも拘わらず、Ｘ－変数のみについて動作する傾向がある。この技法の一例では、二進区分（０または１）の代わりに多数（例えば、３、４、またはそれ以上）の区分についてフレームワークを確立することによって、２つよりも多いクラスタを可能にする。
【００１２】
[0012] ニューラル・ネットワーク型分析は、データ分析の別の手法である。しかしながら、ニューラル・ネットワーク型分析は、多くの用途に適するだけ十分に計算が高速にはなっておらず、しかも変数の数が１０から２０を超えると困難が生ずる。
【発明の概要】
【発明が解決しようとする課題】
【００１３】
[0013] 以前からの手法の欠点には、多数の変数および変数の組み合わせにおいて多くの潜在的な分割を調査する際の集約的計算およびコストが含まれる。回帰ツリーおよびニューラル・ネットワーク型分析は、変数の数が普通または多い（例えば、約２０を超える）場合、困難に直面し困難を生ずる。
【課題を解決するための手段】
【００１４】
[0014] 本明細書に記載する概念は、部分的最小二乗（「ＰＬＳ」）手法を用いたデータ分析およびデータのクラスタリングまたはグループ化を伴う。部分的最小二乗手法を用いてデータを分析することによって、反復プロセスにおいて比較的大きなデータ集合を小さな部分集合（グループまたはクラスタとも呼ぶ）に区分する。データを区分する毎に得られるグループは、内部同質性(internal homogeneity)（例えば、クラスタ内における変動が少ない）および最大外部異質性(external heterogeneity)（例えば、他のクラスタ（群）に対して相対的に多い変動）のレベルが高くなる。データ分析および編成に対する部分的最小二乗手法は、大きなデータ集合を同様の観察またはデータ点（例えば、プロセス変数および収量変数を関係付けるデータ点）のクラスタまたはグループに分離でき、以前の手法に伴う計算の集中やコストが不要であるという利点がある。また、部分的最小二乗手法は、分析を補助するサブグループ（クラスタ）におけるプロセス変数と収量変数との間の関係を保存する。
【００１５】
[0015] 部分的最小二乗手法は、１０，０００個よりも多い変数を有するデータ集合を含む、比較的大量の変数を有するデータ集合に対処することができる。更に、部分的最小二乗手法は、データ集合が、多数の共線変数またはプロセス変数と収量変数との間に多数の関係を含む場合でも動作することができ、および／または、例えば、検出器の異常またはデータ格納の問題によってデータ集合からデータが部分的に失われた場合でも動作することができる。部分的最小二乗手法の別の利点は、コンピュータ処理時間が比較的速く、比較的速い計算および／またはグラフ即ちプロット上に階層的に編成したデータの提示が容易であることである。
【００１６】
[0016] 部分的最小二乗手法を用いることの別の利点は、Ｙ－変数（例えば、応答）がクラスタリングおよびクラスタリングする決定に影響を及ぼすことである。例えば、Ｙ－変数は、明示的に、「分割判断基準」の一部、またはデータをサブグループに区分すべきか

そしてどこで区分すべきか判断するパラメータとして用いることができる。別の利点として、部分的最小二乗手法は、二進および連続Ｙ－変数ならびに１つまたは多数のＹ－変数でも動作可能である。部分的最小二乗分析は、分析はデータ集合全体から始まり、連続的にデータをより小さいグループに分割していくということから、「トップ－ダウン」手法である。トップ－ダウン手法は、１つのデータ点を有するグループから始まって、グループを組み合わせ（２データ点のグループにする）、全ての観察点が組み合わされて１つのデータ・クラスタになるまで続けられるボトム－アップ手法と対照をなす。ボトム－アップ手法は計算コストがかかる傾向がある。何故なら、データをクラスタリングするか否かの判断が、観察点間距離（例えば、Ｘ－変数間の類似度）、ならびに観察点－クラスタおよびクラスター－クラスタ距離（例えば、同質性および異質性における相違）を監視しなければならないからである。本明細書に記載する手法の別の利点は、本方法が、データ集合または観察点における逸失またはノイズの多いデータにも拘わらず、有用な結果を生成することである。本発明の実施態様には、四分位時間を用いた分散の計算を特徴とするものもある。

【００１７】

[0017] 部分的最小二乗分析に基づくグループ化、区分、またはクラスタリングの利点は、Ｘ－変数自体の値ではなく、ＰＬＳ回帰モデルのＸ－スコアを、分割またはグループ化判断基準の一部として用いることである。ＰＬＳ手法を１対のマトリクス、Ｘ－変数のＸ－マトリクスおよびＹ－変数のＹ－マトリクスに適用すると、その結果、一連のデータ分割、グループ化、または区分が行われる。データは、行単位（例えば、観察点単位）で区分され、ＰＬＳモデルによって表されるツリー構造またはデンドログラムとなる。このデンドログラムにおける各ノードは、特定のグループまたはクラスタにおけるデータのＰＬＳモデルを表す。

【００１８】

[0018] 一般に、１つのデータ集合またはクラスタを２つ（以上）に分割するには、何らかの区分値または位置を決定する。例えば、第１マトリクス（Ｘ－マトリクス）におけるＸ－変数について第１スコアｔ１を計算し、クラスタの観察点をこのスコアｔ１に沿ってソートする。次いで、（ａ）Ｘ－マトリクスの分散、（ｂ）Ｙ－マトリクスの分散、および（ｃ）後続の各データ集合における観察点の数と関連のある関数（例えば、損失関数）を含む、数個のファクタ(factor)の加重組み合わせの改良に基づいて区分の位置を決定する。この関数は、データ集合を、実質的に等しくないまたは均衡が取れていないデータ量を有する２つのサブグループに分割し難くするファクタと考えることができる。実施形態によっては、サブグループのクロス確認を用いて、デンドログラムの分岐を終了することもある（例えば、更なるサブグループ化が不要であると判断するために）。実施形態によっては、ユーザがＰＬＳ－ツリーにおける最大レイヤ数を指定することもあり、典型的な値は４または５である。

【００１９】

[0019] 部分的最小二乗手法は、種々のデータ集合に適用することができる。例えば、本明細書において記載する概念の検査は、プロセス・データ、定量的構造－活動関係（ＱＳＡＲ）データ集合、およびハイパー・スペクトラル画像データ(hyper-spectral image data)に対して実行されたことがある。

【００２０】

[0020] 概して言えば、一態様において、比較的短い処理時間で大量のデータを区分するコンピュータ実装方法およびシステムを提供する。本方法は、第１データ・マトリクスおよび第２データ・マトリクスを供給することを伴う。第１および第２データ・マトリクスの各々は、１つ又は複数の変数（例えば、マトリクス列）と、複数のデータ点（例えば、マトリクス行）とを含む。また、本方法は、部分的最小二乗（ＰＬＳ）分析または直交ＰＬＳ（ＯＰＬＳ）分析を用いて第１データ・マトリクスから第１スコアを決定すること、ならびに第１および第２データ・マトリクスを第１データ・グループおよび第２データ・グループに区分することを伴い、第１データ・マトリクスの第１スコア、第１データ・マ

トリクスの分散、ならびに第１および第２データ・マトリクスの分散に関連する第１および第２グループの分散に基づいて区分する。一実施形態では、第１および第２データ・マトリクスの各々は、１つ異常のマトリクス列および複数のマトリクス行を含む。

【００２１】

[0021] 実施形態によっては、区分するステップは、第１および第２データ・マトリクスを行単位に区分することを伴う。また、区分するステップは、第１ＰＬＳまたはＯＰＬＳスコアの分散と、第２データ・マトリクスの変動との間の関係を表すパラメータを最小化するステップを伴うことができる。区分するステップは、第１および第２データ・グループ間における統計的差異を最大化するステップを伴うことができ、第１データ・マトリクスの第１ＰＬＳまたはＯＰＬＳスコアの分散、各グループの第２データ・マトリクスの分散、ならびに区分後において第１および第２グループに残っているデータ・マトリクスのサイズ（例えば、サイズの均衡）に関する関数に基づいて、統計的差異を計算する。実施形態によっては、区分するステップは、第１データ・マトリクスの第１スコアの変動、第１および第２データ・グループの各々における第２データ・マトリクスの分散、ならびに区分後に第１および第２データ・グループに残っているデータ（例えば、サイズの均衡）に関する関数を最小化することを伴うこともある。実施形態によっては、ＰＬＬ－ツリーにおけるレイヤ数がユーザ指定の最大値に達したときに、区分が終了する。

【００２２】

[0022] 第１データ・マトリクスは、例えば、半導体または薬品および／またはバイオ技術製造プロセスからのプロセス・データを表すデータを収容することができる。更に、第１データ・マトリクスは、例えば、薬品またはバイオ技術研究開発における薬品開発プロジェクトにおいて研究される分子または高分子というような、対象の分子または高分子の構造的変動に関連のあるあるいはこれらを記述する測定データまたは計算データを表すデータを収容することができる。第２データ・マトリクスは、プロセス収量データ、プロセス品質データ、またはその組み合わせを表すデータを収容することができる。別の例では、第２データ・マトリクスは、同じ分子または高分子の生物的データを表すデータを収容することができる。

【００２３】

[0023] 実施形態によっては、第１データ・グループは、第１および第２データ・マトリクスを第１および第２グループに行単位に区分した結果各々得られた、第３データ・マトリクスおよび第４データ・マトリクスを含むことがある。このような実施形態では、第３および第４データ・マトリクスの第２部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、第３データ・マトリクスから第２スコアを決定するステップと、第３データ・マトリクスの第２スコア、第３データ・マトリクスの分散、ならびに第３および第４データ・マトリクスの分散に関連する第３および第４グループにおける分散に基づいて、第３および第４データ・マトリクスを区分する（例えば、行単位の区分）ステップとを伴う。第２データ・グループは、第５データ・マトリクスおよび第６データ・マトリクスを含み、このような実施形態では、本方法は、更に、第２データ・グループが閾値数よりも多いデータ点を含む場合、第３部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、第５マトリクスから第３スコアを決定するステップと、第５データ・マトリクスの第３スコア、第５データ・マトリクスの分散、ならびに第５および第６データ・マトリクスにおける分散に関連する第５および第６グループにおける分散に基づいて、第５および第６データ・マトリクスを行単位に区分するステップとを伴う。

【００２４】

[0024] 一部の実施形態は、第１、第２、第３、第４、第５、または第６グループを階層的に表示するステップを伴う。更に、本方法は、更に、グループが閾値数未満のデータ点を含む場合、データ・グループの区分を終了するステップを伴うことができる。また、本方法は、更に、第２スコアおよび第２データ・マトリクスの組み合わせた分散が、データ・グループをサブグループに区分したときに減少しない場合、データ・グループの区分を終了するステップを伴うこともできる。実施形態によっては、本方法は、第１および第２

データ・マトリクスと関連のある以前の区分数が所定の閾値に等しいかまたはこれを超える場合、第２グループの区分を終了するステップを伴う場合もある。所定の閾値は、デンドログラムにおける階層レベルの最大数を表す制限値である。

【００２５】

[0025] 実施形態によっては、本方法は、第１データ・マトリクスおよび第２データ・マトリクスを表示するグラフ上において、第１データ・グループまたは第２データ・グループを識別するステップを伴う場合もある。実施形態によっては、ユーザがＰＬＳ－ツリーにおけるレイヤの最大数を指定することもあり、典型的な値は４または５である。

【００２６】

[0026] 概して、別の態様では、情報担体に有形的に具現化されているコンピュータ・プログラム・プロダクトであって、コンピュータ・プログラム・プロダクトが、データ処理装置に、多数のステップを実行させるように動作可能な命令を含む。例えば、これらのステップは、第１データ・マトリクスおよび第２データ・マトリクスを受け取るステップであって、第１および第２データ・マトリクスの各々が、１つ又は複数のデータ点を含む、ステップと、第１および第２データ・マトリクスの部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、第１データ・マトリクスから第１スコアを決定するステップと、第１データ・マトリクスの第１スコア、第１データ・マトリクスの分散、ならびに第１および第２データ・マトリクスの分散に関連する第１および第２データ・グループにおける分散に基づいて、第１および第２データ・マトリクスを行単位に区分するステップとを含むことができる。これらの分散は、実施態様によっては、最初の数成分のそれぞれのＰＬＳ－スコアの分散によって表される場合もある。

【００２７】

[0027] 更に別の態様では、データを階層的に編成するシステムがある。このシステムはメモリを含む。このメモリは、第１データ・マトリクスおよび第２データ・マトリクスを有するデータ構造を含む。また、本システムはメモリに動作的に結合されているプロセッサも含む。このプロセッサは、部分的に第１データ・マトリクスの部分的最小二乗分析またはＯＰＬＳ分析に基づいて第１スコアを決定するモジュールと、第１グループおよび第２グループを発生するために第１および第２データ・マトリクスを（例えば、行単位で）区分するモジュールを含む。区分は、第１データ・マトリクスの第１スコア、第１データ・マトリクスの分散、ならびに第１および第２データ・マトリクスの分散に関連する第１および第２グループにおける分散に部分的に基づく。また、本システムは、第１および第２データ・グループ、ならびに第１および第２データ・グループの第１および第２データ・マトリクスに対する関連を表示するために、プロセッサに動作的に結合されているディスプレイも含む。

【００２８】

[0028] 別の態様では、データを分析するシステムである。このシステムは、メモリから第１データ・マトリクスおよび第２データ・マトリクス（例えば、データ構造）を読み出すデータ読み出し手段を含む。第１および第２データ・マトリクスの各々は、１つ又は複数のデータ点を含む。本システムは、部分的最小二乗（ＰＬＳ）分析またはＯＰＬＳ分析を用いて、第１データ・マトリクスから第１スコアを決定するデータ分析手段を含む。また、本システムは、第１および第２データ・マトリクスを第１データ・グループおよび第２データ・グループに分割するデータ区分手段であって、第１データ・マトリクスの第１スコア、第１マトリクスの分散、ならびに第１および第２データ・マトリクスの分散に関連する第１および第２グループにおける分散に基づいて区分する、データ区分手段を含む。

【００２９】

[0029] 実施態様の中には、前述の態様のいずれかを含み、以上の実施形態またはその効果を特徴とするものもある。

【００３０】

[0030] これらおよびその他の特徴は、以下の説明および図面を参照することによって

、一層深く理解されよう。図面は、例示であって、必ずしも同じ拡縮率で描かれている訳ではない。本明細書では、製造プロセス、特に、半導体、薬品、またはバイオ技術製造プロセスに関してその概念を記載するが、この概念には追加の用途、例えば、データ・マイニング用途、財務データ分析用途、あるいは多数のデータ点または観察を伴うその他の用途もあることは、当業者には明白であろう。

【図面の簡単な説明】

【００３１】

[0031] 以上のおよびその他の目的、特徴、および利点は、添付図面に示す実施形態の、以下の更に特定的な説明から明白となろう。図面において、同様の参照符号は、異なる図全てを通じて同じ部分を指す。図面は、必ずしも同じ拡縮率で描かれているのではなく、むしろ、実施形態の原理を図示する際には強調が加えられている。

【図１Ａ】図１Ａは、測定データを示すグラフである。

【図１Ｂ】図１Ｂは、データ区分の前および後における、図１Ａのグラフ上に表されたデータを示すブロック図である。

【図２】図２は、階層的にデータを編成し表示するデータ処理システムのブロック図である。

【図３】図３は、部分的最小二乗分析を用いてデータを分析する方法を示すフロー・チャートである。

【図４】図４は、部分的最小二乗ツリー分析の後における階層的編成データを示す分類ツリーである。

【図５】図５は、データを表示するためのユーザ・インターフェースの一例である。

【図６】図６は、近似検索を用いるためのアルゴリズムの一例を示すフロー・チャートである。

【発明を実施するための形態】

【００３２】

[0039] 図１Ａは、測定データ１０５を示すグラフ１００である。データ１０５は、グラフ１００上において複数のデータ点１１０として表されている。データ点１１０の各々は、製造プロセスまたはその他の何らかの測定または監視プロセス中に収集または測定されたデータを表す。データ点１１０は、「観察点」(observation)と呼ばれることもある。グラフ１００は、第１軸１１５と、この第１軸１１５に垂直な第２軸１２０とを含む。軸１１５および１２０は、プロセス変数（観察可能または予測可能変数と呼ばれることもある）または収量変数（結果または予測変数と呼ばれることもある）を表すことができる。実施形態によっては、これらの軸１１５および１２０をＸ－軸と呼ぶ場合もある。また、軸１１５および１２０は、Ｙ－軸と呼ぶこともできる。実施形態によっては、第１軸１１５および第２軸１２０の単位は、無次元であるか、または目盛りが振られている。実施形態によっては、グラフ１００はＸ－Ｘ空間またはＹ－Ｙ空間においてデータ１０５を図示し、グラフ１００は、１つ又は複数のデータ・マトリクスの平面（または低次元面）への投影を図示することもできる。これらの軸は、データ・マトリクスにおける変数によって定めることができる。

【００３３】

[0040] 実施形態によっては、データ点１１０が、プロセス・データおよび対応する収量データ（例えば、プロセス・データを測定したバッチについての収量データ）を表すデータを順序付けた対の一部である場合もある。実施形態によっては、データ点１１０は、１つ又は複数のデータ・マトリクスにおけるエントリを表すこともある。例えば、プロセス・データは、第１データ・マトリクスにおけるエントリであることもできる。第１データ・マトリクスをＸ－マトリクスとも呼ぶ。Ｘ－マトリクスは、Ｎ行（観察点とも呼ぶ）およびＫ列（変数とも呼ぶ）を含む、Ｎ×Ｋマトリクスであることができる。収量データは、第２データ・マトリクスにおけるエントリであることができる。第２データ・マトリクスをＹ－マトリクスとも呼ぶ。Ｙ－マトリクスは、Ｎ行およびＭ列を含むＮ×Ｍマトリクスであることができる。

【００３４】
[0041] グラフ１００は、少なくとも１０，０００個のデータ点、そして場合によっては、１０，０００個よりも遥かに多いデータ点を含むことができる。実施形態によっては、第１データ・マトリクスおよび／または第２データ・マトリクスにおけるデータは、グラフ１００上に表示する前に、前処理を受ける。例えば、グラフ１００を作成または表示する前に、第１データ・マトリクスおよび第２データ・マトリクスをストレージから読み出し、前処理アルゴリズム（図示せず）によってマトリクスにおけるデータの変換、中心合わせ、および／または倍率変換を行う。
【００３５】
[0042] 実施形態によっては、前述の前処理を、第１または第２データ・マトリクスにおけるデータの統計的分析と関連付ける。例えば、ユーザ（例えば、コンピュータまたは人）が、グラフ１００を発生する前に、１組の倍率パラメータを指定して、データに適用することができる。適した倍率パラメータの特定の値を指定するために、倍率調整ファイル(scaling file)を用いることができる。倍率調整は、後続の処理またはモデル発生にデータを用いる前における、データの一種の前処置(pre-treatment)または前処理と呼ばれることもある。データ・マトリクスにおける観察点および変数の測定値は、非常に異なる数値範囲を有することが多く、このためにデータにおいて大きな統計的分散が生ずる。部分的最小二乗分析は、一般に、最大共分散投影方法と見なされる。その結果、大きな分散がある変数またはデータは、比較的分散が低い変数よりも、グラフ１００上で大きく表現される可能性が高い。比較的分散が大きな変数を第１軸１１５（例えば、Ｘ－軸）に沿って散乱プロット(scatter plot)にプロットし、比較的分散が小さな変数を同じ目盛りの第２軸１２０に沿って散乱プロットにプロットすると、散乱が大きな変数における拡散(spread)が分散が小さな変数における拡散を支配することがあり得る。矯正手段として、双方の変数のデータ（および軸）の倍率を調整することができる。これらの変数を倍率調整することにより、双方の変数が特定のデータ・モデルに寄与することができる。
【００３６】
[0043] 軸１１５および１２０の双方に相対的にまたは近似的に等しい重みを与えるために、データ値を標準化、倍率調整、または重み付けする。これによって、Ｘ－マトリクスのエントリおよびＹ－マトリクスのエントリ（または変数）がモデルにほぼ等しく寄与することを促進する。倍率調整プロセスは、所定の判断基準にしたがって変数空間における座標軸の長さを規制することを伴う（例えば、各座標軸の長さを同じ分散に設定する）。データを倍率調整する共通の技法に「単位分散」、「ＵＶ」倍率調整、または「自動倍率調整」と呼ばれるものがある。単位分散倍率調整は、データ集合からの特定の変数に対して標準偏差（例えば、σ）を計算することを伴う。倍率調整重みは、標準偏差の逆数（例えば、ｗ＝１／σ）として計算する。変数の各値に倍率調整重みを乗算して、倍率調整変数を決定する。データ・マトリクスにおける変数の全ての倍率を調整した後、座標軸１１５および１２０の各々は単位分散を有する。
【００３７】
[0044] 実施形態によっては、ユーザが特定の変数（例えば、ノイズが多い変数または関連のない変数）の価値を減じたり、またはある種の変数のグラフ１００に対する寄与を増大させたいこともあり得る。ユーザは、特定のデータ集合についてこの目的を達成するために、倍率調整重みを修正することができる（例えば、このために分散も修正することができる）。また、変数の変換を用いて、その変数に与える分布を対称に近付けることも多い。例えば、対数変換、負対数倍率調整(negative logarithm scaling)、ロジット倍率調整(log-it scaling)、二乗根倍率調整、第４根倍率調整、逆倍率調整、またはべき変換倍率調整(power transformation scaling)を用いることができる。
【００３８】
[0045] 同様に、マトリクス内にあるデータは、倍率調整した座標系の原点Ｏ以外のある点（図示せず）を中心にして配することもできる。これを行う場合、必要に応じて、マトリクス要素を他の点を中心にして配するために、中央値をマトリクス列の各々に加算する

または中央値をマトリクス列の各々から減算することができる。中心合わせおよび倍率調整は双方共、グラフ１００を発生する際の計算上の要求および／またはデータの部分的最小二乗分析の計算上の要求を減少させることができる。または、中心合わせおよび倍率調整は、データの解釈、および結果的に得られるパラメータまたは解釈モデル(interpretive model)を使用し易くする。

【００３９】

[0046] 第１および第２マトリクスの中にあるデータがインポート、および／または中心合わせ、変形、あるいは倍率調整された場合、部分的最小二乗アルゴリズムをそのデータに適用して、ｔ１スコアを判定する。実施形態によっては、部分的最小二乗アルゴリズムは、データの直交部分的最小二乗分析に基づき、ｔ１スコアはデータのこのＯＰＬＳ分析に基づくこともある。ｔ１スコアは、データのクラスタを近似し、第２マトリクスにおけるデータに相関付けられた、Ｘ－空間（例えば、グラフ１００上）にあるラインに対応する。部分的最小二乗成分に沿った座標は、個々のデータ点即ち観察点についてのｔ１スコアを定義または決定する。Ｘ－空間において蓄積した観察点についてのｔ１スコアは、ｔ１スコア・ベクトルを定義または決定する。ｔ１スコア・ベクトルは新たな変数と見なすことができる。

【００４０】

[0047] ｔ１スコアは、グラフ１００上におけるライン１２５を表す（例えば、個々のｔ１スコアの累積、またはｔ１スコア・ベクトルを表す）。ｔ１スコア（例えば、ライン１２５）に対して垂直なライン１４５を用いて、グラフ１００を２つのセクション１３０および１３５に分割または区分する。セクション１３５は、ライン１４５よりも下にある観察点即ちデータ点を表し、セクション１３０は、ライン１４５よりも上にある観察点即ちデータ点を表す。グラフ１００上にあるデータ１０５を、ｔ１スコア・ベクトル（例えば、ライン１２５）に沿ってソートする。ライン１２５に沿ったｔ１スコア毎に計算を行う。

【００４１】

[0048] ｔ１スコア・ベクトルを決定した後、ライン１２５上にある点（例えば、ライン１４５に沿った）の各分割値を式１によって評価する。

$$u = (1-b)*\{a[V(t1_1)+V(t1_2)]/V(t1)+(1-a)[V(y1)+V(y2)]/V(y)\}+b*F(n_1, n_2) \quad \text{式１}$$

ここで、

ｕ＝最小化すべきパラメータ、

ａ＝ユーザが調節可能なパラメータであり、通例、０と１との間、

ｂ＝ユーザが調節可能なパラメータであり、通例、０と１との間、

Ｖ＝特定のマトリクスまたはベクトル内における分散

$t1_i$＝ライン１２５上におけるｉ番目の座標、例えば、ｉ番目の観察点についてのｔ１スコア値、

y_i＝Ｙマトリクスにおけるｉ番目の行、例えば、ｉ番目の観察点のＹ－ベクトル、

n_1＝サブグループ１（例えば、セクション１３０内にある）の中にあるデータ点即ち観察点の数、

n_2＝サブグループ２（例えば、セクション１３１内にある）の中にあるデータ点即ち観察点の数、

Ｆ＝n_1およびn_2を関係付ける関数。データを、n_1およびn_2についてほぼ同様の値を有するサブグループに区分し易くするために用いられる。

【００４２】

[0049] 式１は、定性的には、Ｘマトリクス（例えば、ｔ１スコア）における分散と、Ｙ－マトリクスにおける分散と、各潜在的なサブグループまたはサブパーティションにおけるデータ量との間の関係と考えることができる。「ｕ」の値は、通例、Ｘ－スコアｔ１の分散、Ｙ－マトリクスの分散、および連続する各サブグループにおけるデータ量と関連付けられている関数（Ｆ（n_1，n_2））の組み合わせの全体的な改良によって、最小化される（つまり、ｔ１スコアに沿った区分が最適化される）。例えば、関数Ｆ（n_1，n_2

）は、結果的に得られる各サブグループにおいてほぼ等しい数の観察点（例えば、Ｘ－変数）が得られ易くする損失関数と考えることができる。実施形態によっては、この関数Ｆ（n_1，n_2）は式２によって与えられる。

【００４３】

【数１】

$$F=\frac{(n_1-n_2)^2}{(n_1+n_2)^2} \qquad \text{式2}$$

【００４４】

[0050] 当業者には、他の損失関数も明白であろう。実施形態によっては、ユーザ調節可能パラメータａが、スコアｔ１およびＹ－変数を関係付ける場合もある。例えば、ａの値が０に近づく程、スコア１に起因するウェイトが大きくなる。ａの値が１に近づく程、Ｙ－変数に起因するウェイトが大きくなる。ユーザ調節可能パラメータｂは、サブグループのサイズに関係する。例えば、ｂの値が０に近づく程、区分によって得られるサブグループは、区分後にほぼ等しいサイズとなる可能性が低くなる。ｂの値が１に近づく程、区分によって得られるサブグループは、区分後にほぼ等しいサイズとなる可能性が高くなる。実施形態によっては、パラメータａのデフォルト値は０．３であり、パラメータｂのデフォルト値も０．３である。パラメータａおよびｂについて、他のデフォルト値も可能である。

【００４５】

[0051] 実施形態によっては、パラメータｂの値は０にすることができる。このような実施形態では、損失関数Ｆは、最小化すべきパラメータｕには影響を及ぼさない。具体的には、損失関数Ｆは、それぞれのサブグループにおけるデータ量の増大を促すためまたはデータ量に影響を与えるためには用いられない。実施形態によっては、パラメータａの値は０にすることができる。このような実施形態では、式１は、区分化がＸ－変数自体の値の代わりにＰＬＳ－スコアに基づくことを除いて、分類および回帰ツリー（ＣＡＲＴ）分析に類似する。

【００４６】

[0052] 実施形態によっては、パラメータn_{min}を、例えば、パラメータｂの値が０に近いかまたは比較的小さいときに、各々比較的少量のデータを収容する比較的多数のクラスタまたはグループが生ずる式１に対する解を防止する境界条件またはパラメータとして指定することができる。パラメータn_{min}は、関数的に、$n_{min}=min(n_1, n_2)$と表すことができる。n_{min}の値の一例は５である。n_{min}には他の値も可能であり、ユーザが選択することができる。実施形態によっては、階層レベルの数（暗示的に、サブグループまたはクラスタの数）をユーザが決定または選択することができる。例えば、ユーザは４つまたは５つの階層レベルを選択することができる。ユーザが階層レベルの数を選択していない場合、デフォルトを指定することができる（例えば、４階層レベル）。

【００４７】

[0053] 式１におけるパラメータ「ｕ」の値を最小化するライン１２５上の座標１４０が決定されるおよび／または突き止められる。パラメータ「ｕ」を最小化することによって、ｔ１スコア（またはｔ１スコア・ベクトル）および第２マトリクス（Ｙ－マトリクス）における変動にしたがって、データ１０５が区分される。交差座標１４０においてライン１２５に垂直なライン１４５が決定され、グラフ１００上に図示される。ライン１４５はこのグラフをセクション１３０および１３５に分割する。セクション１３０は、ライン１４５よりも上にあるデータ１０５を含み、セクション１３５は、ライン１４５よりも下にあるデータ１０５を含む。セクション１３０は、第１データ・グループを含み、セクショ

ン１３５は第２データ・グループを含む。式１を最小化した結果、（ｉ）第１データ・グループないにおけるスコアｔ１の分散および第１グループの第２マトリクスの分散、ならびに（ｉｉ）第２データ・グループ内におけるスコアｔ１の分散および第２グループの第２マトリクスの分散の組み合わせが、ｔ１の特定の値の選択によって最小化される。この組み合わせを最小化することは、変数ｔ１およびＹに関して、第１データ・グループと第２データ・グループとの間において組み合わせ分散(combinded variance)を最大化することと同等である。

【００４８】

[0054] グラフ１００をセクション１３０および１３５に分割した後、同様の手順を用いてセクション１３０および１３５の各々におけるデータを分析することができる。例えば、セクション１３５におけるデータ点１１０は、第３データ・マトリクスＸ１（例えば、セクション１３５におけるデータ１０５のＸ－マトリクス値を含む）、および第４データ・マトリクスＹ１（例えば、セクション１３５におけるデータ１０５のＹ－マトリクス値を含む）と見なすことができる。先に論じたのと同様にして、第３データ・マトリクスから第２ｔ１スコアを決定することができる（しかし、セクション１３５の中にあるデータのみに基づく）。セクション１３５は、第２ｔ１スコアに基づいて第２ライン（図示せず）に沿って分割または区分することができる。第２ｔ１スコアを決定した後、式１は、第２ｔ１スコアおよび第４データ・マトリクス（例えば、Ｙ－マトリクス）における変動に関して最小化し、更にセクション１３５を第２垂直ライン（図示せず）に沿って第１および第２サブグループ（図示せず）に細分することができる。

【００４９】

[0055] 次いで、セクション１３０（グループ２）の類似した分析が続き、セクション１３０の中にあるデータをサブグループに区分することができる。

【００５０】

[0056] 以上で説明した手順は、グラフ１００上のデータ１０５全てを分析し、増々小さなクラスタ（サブグループ）の階層構造にグループ化し終えるまで、連続するサブグループ毎に継続することができる。実施形態によっては、この区分プロセスは、サブグループが収容するデータ点が閾値データ点数よりも少なくなったとき、またはデータを更に区分してもｔ１スコア・ベクトルにおける相対的分散またはＹ－マトリクスにおける変動が小さくならないときに、特定のサブグループに対して終了する。データ点の閾値数は、ユーザが選択可能であり、例えば、５データ点とすることができる。

【００５１】

[0057] 実施形態によっては、クラスタ外形(geometry)の検査から、クラスタまたはサブグループが、第１スコア・ベクトルｔ１に平行でない方向に沿って方位付けられていることが示唆されることがある。サブグループが第１スコア・ベクトルｔ１に平行に方位付けられていない場合、スコア・ベクトルの組み合わせ（例えば、２つ、３つ、またはそれ以上のスコア・ベクトルの組み合わせ）を用いることができる。スコア・ベクトルを組み合わせるには、実施形態によっては、第３パラメータｃを導入することもある。第３パラメータｃは、通常－１と＋１との間の値を有する。第３パラメータｃは、第１スコアｔ１を第２スコアｔ２に関係付ける。パラメータｃを用いた、スコア・ベクトルｔ１とｔ２との間における適した関係の一例は、次の通りである。{c*t2+(1-|c|)*t1}。他の関係も当業者には明白であろう。パラメータｃとスコア・ベクトルｔ１およびｔ２との間における関係から、第１スコア・ベクトルｔ１または第２スコア・ベクトルｔ２のみに沿った分析ではなく、スコア・ベクトルｔ１およびｔ２によって定義される平面において表されるデータの分析ができる。

【００５２】

[0058] 実施態様によっては、ユーザが分析対象変数の数を減少させることができる場合もある（例えば、データ選択または前処理とも呼ばれている）。例えば、ある種の変数は、モデルにおける最良の予測変数と強く相関付けることができ、あるいはＹ－変数（例えば、結果変数）とは相関付けられない。データ選択の一例では、データ（例えば、Ｘ－変

数）のパラメータを所定値と比較することを伴う。例えば、Ｙとの相関が所定の百分率（例えば、７５％）よりも低いことを表示する変数を、分析の前に、データ集合から排除する。

【００５３】

[0059] 図１Ｂは、データ区分の前および後における、図１Ａのグラフ１００上に表されたデータを示すブロック図１６０である。ブロック図１６０は、第１データ・マトリクス１６４（Ｘで示す）および第２データ・マトリクス１６８（Ｙで示す）を含む。データ・マトリクス１６４、１６８の各々は、１つ又は複数の列（変数とも呼ぶ）および複数の行（観察点とも呼ぶ）を含むことができる。実施形態によっては、第１データ・マトリクス１６４がプロセス・データを含み、第２データ・マトリクス１６８が収量データおよび／または製品品質データを含む場合もある。別の実施形態では、第１データ・マトリクス１６４が測定および計算した物理化学的データおよび／または構造関係データを含み、第２データ・マトリクス１６８は、例えば、１組の分子または高分子と関連のある生物データを含む。

【００５４】

[0060] また、ブロック図１６０は、第１および第２データ・マトリクス１６４、１６８を分割する区分ライン１７２も含む。第１データ・マトリクス１６４の部分１７６ａは、第１データ・マトリクス１６４の区分時に、第１データ・グループ１８４の部分１８０ａ（X_1で示す）となる。同様に、第２データ・マトリクス１６８の部分１８８ａは、区分時に、第１データ・グループ１８４の部分１８０ｂ（Y_1で示す）となる。第１データ・マトリクス１６４の部分１７６ｂは、第１データ・マトリクス１６４の区分時に、第２データ・グループ１９６の部分１９２ａ（X_2で示す）となる。同様に、第２データ・マトリクス１６８の第２部分１８８ｂは、区分時に、第２データ・グループ１９６の部分１９２ｂ（Y_2で示す）となる。したがって、後続の区分は、第１データ・マトリクス１６４および第２データ・マトリクス１６８を表す第１データ・グループ１８４の第１部分１８０ａ（X_1）および部分１８０ｂ（Y_1）について同様に続けられる。第２データ・グループ１９６も同様に区分することができる。

【００５５】

[0061] 実施形態によっては、区分は、第１データ・マトリクス１６４における観察点のスコア、および第２データ・マトリクス１６８の分散に応じた行単位の区分である。他の区分技法も用いることができる。実施形態によっては、区分が中止または終了するのは、特定のデータ・グループ（例えば、データ・グループ１８４）がデータ点の閾値数を超えないとき、例えば、観察点の数が少なすぎる場合である。データ点の閾値数は、ユーザによって設定または決定することができる。実施形態によっては、データ・グループの区分を終了するのは、Ｘ－マトリクス（ｔ１）およびＹ－マトリクスの第１スコア・ベクトルにおける分散が、区分の結果減少しなくなったときである。これらの分散を比較することができる（例えば、［Ｖ１＋Ｖ２］／Ｖの比または分数として比較することができる。これは、式１からの変数ｕと呼ばれることもある）。ｕが１以上である場合、区分プロセスの結果得られたマトリクスにおける分散は、直前のデータ・グループから得られた分散以上となり、区分を終了する。ｕが１未満である場合、区分プロセスの結果得られたマトリクスにおける分散は、直前のデータ・グループの分散よりも小さく、ｕが１以上になるまで区分を継続する。

【００５６】

[0062] 図２は、データを階層的に編成し表示するデータ処理システム２００のブロック図である。データ処理システム２００は、プロセッサ２１０に結合されているメモリ２０５を含む。また、データ処理システム２００は、プロセッサ２１０に結合されているディスプレイ２１５も含む。更に、データ処理システム２００は、図示しないその他のコンポーネントまたはモジュール、例えば、データを測定し、収集し、メモリ２０５に格納するデータ取込モジュール、または多変量統計分析にしたがって収集したデータに基づいてモデルを発生するモデル発生モジュールも含むことができる。データ処理システム２００は

、現場分析および／またはリアル・タイム分析のために製造設備に設置することができ、あるいは処理後の分析またはデータ・マイニングの用途のために他の場所に設置することもできる。

【００５７】

[0063] メモリ２０５は、例えば、処理変数に関するデータ（Ｘ－マトリクス・データ）および収量変数に関するデータ（Ｙ－マトリクス・データ）のような、例えば、製造プロセスを表すデータを含む。データは、生データとして、データ・モデルまたはテンプレートとして、あるいは前処理済みのデータ（例えば、倍率調整、中心合わせ、および／または変換後）として格納することができる。

【００５８】

[0064] プロセッサ２１０は、メモリ２０５と交信するデータ読み出しモジュール２２０を含む。データ読み出しモジュール２２０は、分析のためにメモリ２０５からデータを読み出す。また、プロセッサ２１０はデータ分析モジュール２２５およびデータ区分モジュール２３０も含む。データ分析モジュール２２５は、データ読み出しモジュール２２０およびデータ区分モジュール２３０と交信する。データ区分モジュール２３０は、データ読み出しモジュール２０５と交信する。データ読み出しモジュール２０５は、ディスプレイ２１５と交信し、読み出したデータのユーザへの表示をし易くする（例えば、図１のグラフ１００上にあるデータ点１１０）。データ区分モジュール２３０およびデータ分析モジュール２２５もディスプレイ２１５と交信し、データをユーザに表示し易くする。

【００５９】

[0065] データ読み出しモジュール２２０がメモリ２０５から特定のデータ集合を読み出した後、データ分析モジュールは、読み出したデータ（たとえば、第１Ｘ－マトリクスおよびＹ－マトリクス）の部分的最小二乗分析（ＰＬＳ）または直交ＰＬＳ分析（ＯＰＬＳ）を実行して、第１ｔ１スコアを決定する。先に論じたように、ｔ１スコアは、読み出したデータを細分するための基準を形成するＸ－空間内のラインを表す。データ分析モジュール２２５は、ディスプレイ２１５と交信し、分析したデータのユーザへの表示をし易くする（例えば、図１のグラフ１００上にあるライン１２５のように）。

【００６０】

[0066] データ分析モジュール２２５がｔ１スコアを決定し終えると、データ区分モジュール２３０がｔ１スコアをＸ－マトリクスおよびＹ－マトリクス（例えば、第１および第２データ・マトリクス）に関して分析し、区分されたグループ内における分散を最小化し、区分されたグループ間の分散を最大化する値を、ｔ１について決定する。データ区分モジュール２３０は、ディスプレイ２１５と交信し、区分したグループのユーザへの表示をし易くする（例えば、図１のグラフ１００上におけるライン１４５およびセクション１３０、１３５として）。

【００６１】

[0067] 実施形態によっては、データ読み出しモジュール２２０、データ分析モジュール２２５、およびデータ区分モジュール２３０の内１つ又は複数が、同じアプリケーション、プロセス、またはプログラムのサブルーチンあるいはサブアルゴリズムである場合もある。実施形態によっては、データ分析モジュール２２５およびデータ区分モジュール２３０が同じサブルーチンまたはアルゴリズムの一部であることもある。

【００６２】

[0068] ディスプレイ２１５は、ユーザ入力デバイス（図示せず）、例えば、ユーザにパラメータを指定させるまたは命令をプロセッサ２１０に発行させるキーボードまたはマウスを含むことができる。実施形態によっては、ディスプレイがユーザ・インターフェースを含み、ユーザとプロセッサとの間における交信をやり易くする場合もある。例えば、ユーザは、ユーザ・インターフェースを通じて、前述の式１からパラメータ「ａ」および「ｂ」の値を指定することができ、またはユーザは命令をプロセッサ２１０に発行し、データ読み出しモジュール２２０に、分析のために、メモリ２０５から指定のデータ集合を読み出すように指令することができる。加えて、ユーザは、データをメモリ２０５から読み

出す前または後のいずれでも、前処理モジュール（図示せず）によってデータを前処理するために、データの倍率調整、変換、または中心合わせを指定することができる。ユーザが別のコンピュータ・システムまたはプロセッサ（図示せず）である実施形態では、ユーザ・インターフェースは、メモリ２０５内にあるデータに関して、そしてプロセッサ２１０がデータをどのように処理するかに関して、システム２００にパラメータを指定することができるマシン－マシン・インターフェースとすることができる。

【００６３】

[0069] 本発明の特徴および態様を具現化した市販製品の一例に、スエーデン、UmeaのUmetrics, Inc.が販売するSIMCA-P+（商標）ソフトウェア製品がある。

【００６４】

[0070] 図３は、部分的最小二乗分析を用いてデータを文政する方法を示すフロー・チャート３００である。ステップ３０４は、分析対象データを読み出すことを伴う。読み出されたデータは、マトリクス形状あるいはその他の何らかの形態または構造とすることができる（例えば、図１Ａおよび図１Ｂに関して先に論じた第１および第２データ・マトリクス）。実施形態によっては、データはメモリ（例えば、コンピュータ化されたメモリ）から読み出される。また、ステップ３０４は、フロー・チャート３００に示した方法が完了した後に行われる。例えば、ステップ３０４は、図１Ａおよび図１Ｂに関して先に論じたように第１および第２グループに分割されたデータの後続処理または一連の繰り返し処理における最初のステップを表すことができる。ステップ３０４は、フロー・チャート３００における方法が行われている間に、アレイまたは一時的メモリに格納されているデータからデータを読み出すことを伴うことができる。

【００６５】

[0071] ステップ３０４においてデータを読み出した後、任意のステップ３０８においてデータを前処理することができる。このデータの前処理は、前述のように、読み出されたデータ・マトリクスにおけるデータの変換、中心合わせ、および／または倍率調整を伴う。実施形態によっては、前処理はユーザの指定に応答して行われる（例えば、倍率調整重み、閾値のような前処理パラメータを含む、または特定の繰り返しに応答して）。実施形態によっては、ユーザは、デフォルト値または予め選択してある値を用いて前処理を行うべきことを指定する。データを前処理するか否かは、本方法の初期セットアップの間にユーザによって指定されるデフォルト設定とすることができる。ステップ３０６は、前処理ステップ３０８が完了したか否かに関連する判断ステップを示す。

【００６６】

[0072] データを前処理した後（ステップ３０８）、ステップ３１２において部分的最小二乗分析または直交部分的最小二乗分析を選択することができる。部分的最小二乗分析の種類は、ユーザによって指定することができ、またはデフォルト設定とすることもできる。部分的最小二乗分析が選択された場合（ステップ３１６）、ＰＬＳ分析を用いて第１ｔ１スコアを決定する（ステップ３２４）。直交部分的最小二乗（ＯＰＬＳ）分析が選択された場合（ステップ３２０）、直交部分的最小二乗分析を用いて第１ｔ１スコアを決定する（ステップ３２４）。第１ｔ１スコアは、第１および第２データ・マトリクスによって定義される空間または座標系（例えば、Ｘ－Ｙ空間）におけるラインまたはその他の何らかの曲線によって表すことができる。実施形態によっては、ＰＬＳおよびＯＰＬＳ分析の双方を用いることもある。

【００６７】

[0073] ステップ３２４においてｔ１スコアを決定した後、ソーティング・プロセスを行う（ステップ３２８）。その後、分割プロセス（ステップ３３２）を行い、第１および第２データ・マトリクスを２つの部分（パート１および２）に分割する。パート１および２は、ｔ１スコア値を第１および第２データ・マトリクスにおける分散に関係付けるパラメータを、パート１およびパート２におけるデータ値が最小化するか否か判断するための更に別の処理および分析のための、データの一時的細分化とすることができる。第１および第２データ・マトリクスをパート１および２に分割した後（ステップ３３２）、この区分

を先の式１によって評価する（ステップ３３６）。ステップ３４０において、式１からのパラメータ「ａ」および「ｂ」を読み出す。実施形態によっては、パラメータ「ａ」および「ｂ」をユーザが選択することもある。ステップ３４４において、プロセッサは、ｔ１スコア（ステップ３４０）によって表されるラインに沿った値ならびにおよびパラメータ「ａ」および「ｂ」を用いて、式１を分析し、パート１および２がパラメータ「ｕ」の値を最小化するか否か評価する。パート１および２が「ｕ」の値を最小化しない場合、プロセッサは、ｔ１ラインに沿った値を評価し続ける（例えば、スコア・ベクトル）。プロセッサは、パラメータ「ｕ」の値を最小化する１つまたは複数の値を特定する（ステップ３４８）。式１からの「ｕ」の値を最小化するｔ１スコアの値を選択し、パート１および２（ステップ３３２）をグループ１および２と置き換える（ステップ３４８）。

【００６８】

[0074] ステップ３４４から３４８においてパラメータ「ｕ」を最小化した結果、グループ１および２は、各々、式１に基づくパラメータ「ａ」および「ｂ」の特定の値について、ｔ１スコアの最小グループ内分散(intra-Group variance)ならびにＸ－マトリクス・データおよびＹ－マトリクス・データ、そしてｔ１スコアの最大グループ間分散(inter-Group variance)ならびにＸ－マトリクス・データおよびＹ－マトリクス・データを有する。グループ１および２は、ステップ３０４において読み出した第１および第２データ・マトリクスからのデータの部分集合を収容する。グループ１は、例えば、図１Ｂに示すような、第３データ・マトリクス（例えば、グループ１に分割された第１Ｘ－マトリクスからの値を含むＸ－マトリクス）および第４データ・マトリクス（例えば、グループ１に分割された第１Ｙ－マトリクスからの値を含むＹ－マトリクス）を含むというように考えることができる。同様に、グループ２は、例えば、図１Ｂに示すような、ステップ３０４において読み出した値を含むＸ－マトリクスおよびＹ－マトリクスを含む２つのデータ・マトリクスを含むことができる。

【００６９】

[0075] ステップ３５２において、グループ１およびグループ２におけるデータ点の数、ならびにパラメータ「ｕ」の値を評価する。グループ１（またはグループ２）の中にあるデータ点の数ｎが、データ点の閾値数$n_{threshold}$よりも小さい場合、または「ｕ」の値が所定値（例えば、１）を超える場合、グループ１（またはグループ２）を終了する（ステップ３６０）（例えば、これ以上グループの区分を行わない）。ｎがグループ１（またはグループ２）について$n_{threshold}$に等しいかまたはこれを超え、更に「ｕ」の値が所定値（例えば、１）よりも小さい場合、本方法は問い合わせ(query)ステップ３５６に進む。ステップ３５６において、階層レイヤまたはレベルの数を評価し、指定された限度と比較する。階層レベルの数がこの限度未満である場合、本方法はステップ３０４に戻り、グループ１におけるデータ・マトリクスをデータ（例えば、ステップ３１６、３２０において部分的最小二乗または直交部分的最小二乗分析を実行したデータ・マトリクスとして）として用いて、処理を開始する。階層レベルの数が限度に等しい場合、プロセスを終了する（ステップ３６０）。その後、グループ１と同様にグループ２を処理する。フロー・チャート３００における方法は、ＰＬＳ－ツリー・レイヤの指定最大数に達するまで、またはグループの区分が終了しており、データ点の閾値数$n_{threshold}$よりも多いデータ点を含むグループがなくなるまで繰り返される。実施形態によっては、ツリー・レイヤの数$n_{threshold}$および／または「ｕ」の値の限度は、ユーザによって指定されることもある。

【００７０】

[0076] 図４は、部分的最小二乗（またはＯＰＬＳ）ツリー分析後における階層的編成データを示す分類ツリー４００である。分類ツリー４００は、第１データ集合４０８を含む第１階層レベル４０４を含む。第１階層レベル４０４は、後続の処理が行われる前にメモリ（または最上位の階層レベル）から読み出したデータを示すことができる。実施形態によっては、第１階層レベル４０４が、既に処理されているデータを示すこともある（例えば、第１データ集合４０８は、区分されているサブグループを表す）。第１データ集合４

０８は、例えば、図１Ｂに示すような、第１データ・マトリクス（例えば、Ｘ－マトリクス）および第２データ・マトリクス（例えば、Ｙ－マトリクス）を含む。

【００７１】

[0077] また、分類ツリー４００は、第２階層レベル４１２も含む。第２階層レベル４１２は、第１データ・グループ４１６および第２データ・グループ４２０を含む。第１データ・グループ４１６および第２データ・グループ４２０は、図３のフロー・チャート３００に図示した方法にしたがって、そして図１Ｂに例示したように決定される。例えば、第１グループ４１６および第２グループ４２０は、式１においてパラメータ「ｕ」を最小化することによって特定され、パラメータ「ｕ」は、第１集合４０８の第１データ・マトリクスの第１ｔ１スコア、および第１集合４０８の第２マトリクスの分散に基づいて決定することができる。第１ｔ１スコアおよびＹ－マトリクスに関して、第１グループ４１６は最小の内部変動を保有し、第２グループ４２０に関してまたは関係して最大化された変動を保有する。第１グループ４１６は、例えば、図１Ｂに示すように、第１データ・マトリクスからのＸ－マトリクス値およびＹ－マトリクス値を含む。第２グループ４２０は、例えば、図１Ｂに示すような区分に基づいて、第１データ・マトリクスからの異なるＸ－マトリクス値およびＹ－マトリクス値を含む。

【００７２】

[0078] 分類ツリー４００は、第３階層レベル４２４を含む。第３階層レベル４２４は、第１データ・サブグループ４２８、第２データ・サブグループ４３２、第３データ・サブグループ４３６、および第４データ・サブグループ４４０を含む。第１サブグループ４２８および第２サブグループ４３２は、第１データ集合４０８から第１グループ４１６を決定したのと同様に決定される（例えば、図１Ｂおよび図３に関して先に論じたように）。更に特定すれば、第２ｔ１スコアは、部分的最小二乗（またはＯＰＬＳ）分析に基づいて、第１データ・グループ４１６の中にあるデータから計算される。第２ｔ１スコアは、Ｘ－マトリクスおよびＹ－マトリクスにおける分散と共に、第１グループ４１６におけるデータに関して式１からのパラメータ「ｕ」を最小化するために用いられる。パラメータ「ｕ」が最小化されたときに、第１サブグループ４２８および第２サブグループ４３２が発生する。第３サブグループ４３６および第４サブグループ４４０は、第２グループ４２０の中にあるデータに基づいて、第２グループ４２０の部分的最小二乗分析に基づいて計算された第３ｔ１スコアを用いて、同様に決定される。

【００７３】

[0079] 分類ツリー４００は、第４階層レベル４４４を含む。第４階層レベル４４４は、第３階層レベル４２４の第１サブグループ４２８から決定される、第１サブサブグループ４４８および第２サブサブグループ４５２を含む。第２サブグループ４３２は、第４階層レベル４４４においては表されていない。これは、第２サブグループ４３２から作ろうとしたサブグループにおけるデータ点数が、分析継続に対するデータ点の閾値レベルを超えなかったため（分岐とも呼ぶ）、またはパラメータｕについて含まれる値が１．０を超えているからである。また、第４階層レベル４４４は、更に別のサブサブグループ４５６ａ、４５６ｂ、４５６ｃ、４５６ｄも含む。これらは、第１サブサブグループ４４８および第２サブサブグループ４５２と同様に決定される。

【００７４】

[0080] 尚、階層レベル４０４、４１２、４２４、および４４４の各々は、直上にある階層レベルからのデータ（例えば、第１データ・マトリクスおよび第２データ・マトリクス）を表すことは言うまでもない。例えば、 第２階層レベル４１２によって表されるデータ・マトリクスは、第１階層レベル４０４では（例えば、データ集合４０８では）存在したが、ソートされておらず、分類もされていない形態であった。分類ツリー４００における各「分岐」４６０は、部分的最小二乗（またはＯｐｌｓ）分析を用いて（例えば、図３のフロー・チャートにしたがって）データを分類または区分するプロセスを表す。分岐４６０は、図１Ｂからのブロック図１６０を表す。図示のように、各データ・グループと直前の階層レベルとの間の図４におけるｙ－軸に沿った距離は、ほぼ同一である（例えば、

ｙ－軸に沿ったサブサブグループ４４８とサブグループ４２８との間の距離は、サブサブグループ４５６ｃとサブグループ４４０との間の距離にほぼ等しい）。実施形態によっては、異なる階層レベル上におけるデータ・グループ間の距離が異なる場合もある（例えば、グループ４１６とサブグループ４２８および４３２との間のｙ－軸に沿った距離は、グループ４２０とサブグループ４３６および４４０との間のｙ－軸に沿った距離とは異なることも可能である）。ｙ－軸は、実施態様によっては、サブグループへの区分と関連付けられたｔ１スコアを表すこともできる。

【００７５】

[0081] 図５は、データを表示するユーザ・インターフェース５００の一例である。ユーザ・インターフェース５００は、第１表示部５０４、第２表示部５０８、および第３表示部５１２を含む。第１表示部５０４は、図１のグラフ１００を含む。第２表示部５０８は、図４の分類ツリー４００を含む。第１表示部５０４および第２表示部５０８は、ユーザが、分類ツリー４００からのデータを、グラフ１００上における対応するデータと素早くおよび／または視覚的に関連付けることを可能にする。例えば、ユーザが分類ツリー４００のサブサブグループ４４８に含まれるデータに関心がある場合、ユーザは、第２表示部４００において分類ツリー４００からサブサブグループ４４８を選択することができる。選択されたサブサブグループ４４８は、第２表示部５０８において、楕円５１６またはその他グラフィック指示手段によって強調される。対応する楕円５２０、またはその他のグラフィック指示（例えば、強調）が、第１表示部５０４におけるグラフ１００の上に現れる。楕円５２０またはその他のグラフィック指示は、第１サブサブグループ４４８において表されているデータを示す。実施形態によっては、サブサブグループ４４８の中にあるデータが密接にクラスタ化されていない場合や、楕円５２０またはその他のグラフィック指示によって識別が容易にできない場合もある。このような実施形態では、グラフ１００におけるデータを異なる方法で（例えば、色または強調によって、あるいはラインまたは曲線の当てはめによって）表すことができる。

【００７６】

[0082] 同様に、図４の分類ツリー４００のその他の階層レベルもユーザが選択し、第１表示部５０４に表示することができる。尚、追加の情報は、グラフ１００上では第１表示部５０４において提示できることは言うまでもない。例えば、図示のように、グラフ１００は、ライン１２５（第１階層レベル４０４におけるｔ１スコアに対応する）と、データ１０５を第１グループまたはセクション１３０および第２グループまたはセクション１３５に分離するライン１４５とを含む。第１グループ１３０は、第１階層レベル４０４における第１グループ４１６に対応し、第２グループ１３５は、第１階層レベル４０４における第２グループ４２０に対応する。第１グループ４１６を更に第１サブグループ４２８および第２サブグループ４３２に分類するとき、追加のライン（追加のｔ１スコアに対応するライン、およびそれを通過する垂直ライン）を第１表示部５０４におけるグラフ１００に追加することができる。

【００７７】

[0083] 第３表示部５１２は、選択したデータを処理するためにプロセッサ（例えば、図２のプロセッサ２１０）に命令を供給する、複数のユーザ選択可能ボタン５５０ａ～５５０ｈを含む。図示のように、サブサブグループ４４８が第２表示部５０８において選択されている。更に、ユーザは、ボタン５５０ａ～５５０ｈを通じて、サブサブグループ４４８の中にあるデータの特性を調査または評価することができる。例えば、ボタン５５０ａは、サブサブグループ４４８内におけるｔ１値の分散を計算するモデルと関連付けられており、ユーザがボタン５５０ａを選択したときに、ｔ１値の分散が計算される。本明細書において言及する場合、分類ツリー内の特定のグループについての統計計算は、データ・マトリクスについて行われる計算を指し、その結果はマトリクスまたは１つの値（例えば、二乗の和）とすることができる。

【００７８】

[0084] ボタン５５０ｂは、サブサブグループ４４８内におけるｔ１値の標準偏差を計算

するモジュールと関連付けられており、ユーザがボタン５５０ｂを選択すると、ｔ１値の標準偏差が計算される。ボタン５５０ｃは、サブサブグループ４４８内におけるＹ値の分散（例えば、サブサブグループ４４８の第２データ・マトリクスまたはＹ－マトリクスの分散）を計算するモジュールと関連付けられている。ボタン５５０ｃは、サブサブグループ４４８内にある異なる値（例えば、Ｙ値に対してｔ１値）が分析されることを除いて、ボタン５５０ａと同様である。ボタン５５０ｄは、サブサブグループ４４８内におけるＹ値の標準偏差を計算するモジュールと関連付けられている。ボタン５５０ｄは、サブサブグループ４４８内における異なる値（例えば、Ｙ値に対してｔ１値）を分析することを除いて、ボタン５５０ｄと同様である。

【００７９】

[0085] ボタン５５０ｅおよび５５０ｆは、それぞれ、サブサブグループ４４８内におけるｔ１値およびＹ－値の平均を計算するモジュールに関連付けられている。様々な計算技法のいずれかにしたがって、平均の代替物を計算することもでき、ｔ１値またはＹ－値またはマトリクスの中央値またはモードを決定することを含む。実施形態によっては、計算技法はユーザが選択する。ユーザが選択したボタン５５０ｅ～５５０ｆに応答して計算されたｔ１値および／またはＹ－値の平均は、更に別の評価（例えば、収量データ）または分析のためのデータを示すことができる。

【００８０】

[0086] ボタン５５０ｇは、R^2を計算するモジュールと関連付けられている。R^2は、Ｙ－値または収量値の変動を示し、ＰＬＳまたはＯＰＬＳ分析の正確さを判断するため、例えば、ｔ１スコア・ラインがどれ位正確に対応するＹ－データに当てはまるかを判断するために用いることができる。実施形態によっては、R^2が多重相関係数として分かっている場合もある。ボタン５５０ｈは、Q^2を計算するモジュールと関連付けられている。Q^2は、クロス確認手順を用いる特定のＰＬＳまたはＯＰＬＳモデルによって予測される、サブサブグループ４４８（またはいずれかのＹ－マトリクス）における全変動の端数(fraction)を示す。

【００８１】

[0087] 実施形態によっては、ボタン５５０ａ～５５０ｈは、階層レベル内にある個々の分岐やグループではなく、階層レベル４０４、４１２、４２４、および４４４と関連付けることができる場合もある。

【００８２】

[0088] 実施形態によっては、近似検索(approximate search)を用いて最適な区分を判断する速度を増大させることができる。例えば、近似検索は、データの検索概要の多項式近似に基づくことができる。検索概要の近似の一例は、区分的二次多項式近似である。図６は、近似検索を用いるためのアルゴリズムの一例を示すフロー・チャート６００である。

【００８３】

[0089] ステップ６０４において、各多項式近似に用いられる点（n_{pol}）の数を選択する。点（n_{pol}）の数は、ユーザが選択することができる。ユーザが点（n_{pol}）の数に対して値を選択していない場合、点（n_{pol}）の数のデフォルト値を用いる（ステップ６０８）。例えば、点（n_{pol}）の数のデフォルト値は、関数によって$n_{pol}=\min(11,\mathrm{sqrt}(N))$と表すことができる。ここで、Ｎはデータ集合における点の総数を表す。

【００８４】

[0090] 点（n_{pol}）の数を決定した後、データを当てはめるために用いられる多項式区間(polynomial piece)の数を決定する（ステップ６１２）。多項式区間の数N_{poly}を決定するために用いられる関数関係の一例は、$N_{poly}=\min(7,\text{整数}[2N/n_{pol}]-1)$である。

【００８５】

[0091] 多項式区間の数N_{poly}を決定した後、近似の初期刻み長を計算する（ステップ６１６）。刻み長は、最初の観察点および最後の観察点を除いた観察点（例えば、Ｘ－変数）の範囲が包含されるように、そして各多項式区間が当該多項式区間の中点の各側において観察点の半分と重複するように決定される。

【００８６】
[0092] 「切断表現」(cut expression)、即ち、式１におけるパラメータ「ｕ」の値を、各多項式区間における点毎に計算する（ステップ６２０）。各多項式区間における点毎に式１を計算した後（ステップ６２０）、例えば、最小二乗当てはめ方法を用いて、二次多項式を各多項式区間に当てはめる（ステップ６２４）。ステップ６２４に関して、別の当てはめ技法を用いることもできる。当てはめられた多項式区間に対して最小値が生ずる「ｕ」の値を計算する（ステップ６２８）。
【００８７】
[0093] 最小値を計算した後（ステップ６２８）、刻み長を評価する（ステップ６３２）。当てはめプロセスの刻み長は、多項式における２点間の観察点の数を表す。刻み長が所定値（例えば、１）を超えていない場合、プロセスは終了する（ステップ６３６）。逆に、刻み長が所定値（例えば、１）を超過している場合、刻み長を短縮する（ステップ６４０）。例えば、刻み長を１／４に分割する（例えば、刻み長の値を４で除算する）ことができる。刻み長を短縮した後、新たな多項式区間を生成する（ステップ６４４）。新たな多項式区間は、パラメータ「ｕ」の最小値を中心として位置付けられており、データのほぼ半分がこの多項式区間の中心のいずれの側にも来るようにしている。この新たな多項式区間について、この多項式区間における点毎に「ｕ」の値を決定する（ステップ６４８）。ステップ６４８は、刻み長をステップ６３２において評価する前に行われるステップ６２０と動作が同様である。
【００８８】
[0094] ステップ６４８においてパラメータ「ｕ」の値を最小化した後、多項式を新たな多項式区間に当てはめ（ステップ６５２）、この新たな多項式区間におけるパラメータ「ｕ」の最小値を計算する（ステップ６５６）。ステップ６４８の後、再度刻み長を評価し（ステップ６３２）、刻み長が所定数（例えば、１）を超えない場合、プロセスは終了する（ステップ６３６）。そうでない場合、刻み長が所定の閾値未満に減少するまで、プロセスを繰り返す。
【００８９】
[0095] 以上で説明した技法は、ディジタル電子回路において、またはコンピュータ・ハードウェア、ファームウェア、ソフトウェア、あるいはその組み合わせにおいて実現することができる。実施態様は、コンピュータ・プログラム・プロダクト、例えば、データ処理装置、例えば、プログラマブル・プロセッサ、コンピュータ、または多数のコンピュータによって実行するため、あるいはその動作を制御するために、情報担体、例えば、機械読み取り可能記憶デバイスに有体的に具現化されているコンピュータ・プログラムとすることができる。コンピュータ・プログラムは、コンパイル型言語またはインタプリタ型言語を含む、あらゆる形態のプログラミング言語でも書くことができ、単体プログラム、あるいはモジュール、コンポーネント、サブルーチン、あるいは計算環境において用いるのに適したその他のユニットを含む、あらゆる形態で展開することができる。コンピュータ・プログラムは、１カ所にある１つのコンピュータ、あるいは多数の箇所を跨って分散し通信ネットワークによって相互接続されている多数のコンピュータ上で実行するように展開することができる。
【００９０】
[0096] 方法ステップは、１つ又は複数のプログラマブル・プロセッサによって実行することができ、入力データに作用し出力を発生することによって本技術の機能を実行するコンピュータ・プログラムを実行する。また、方法ステップは、特殊目的論理回路、例えば、ＦＰＧＡ（フィールド・プログラマブル・ゲート・アレイ）またはＡＳＩＣ（特定用途集積回路）によって実行することができ、そして装置はこの特殊目的論理回路として実現することができる。モジュールは、コンピュータ・プログラムの一部、および／またはその機能を実現するプロセッサ／特殊回路に言及することができる。
【００９１】
[0097] コンピュータ・プログラムの実行に適したプロセッサは、一例として、汎用およ

び特殊目的用マイクロプロセッサの双方、ならびにあらゆる種類のディジタル・コンピュータの１つ又は複数のあらゆるプロセッサを含む。一般に、プロセッサは命令およびデータをリード・オンリ・メモリまたはランダム・アクセス・メモリあるいはその双方から受け取る。コンピュータの必須要素は、命令を実行するプロセッサと、命令およびデータを格納する１つ又は複数のメモリ・デバイスである。一般に、コンピュータは、データを格納する１つ又は複数の大容量記憶デバイス、例えば、磁気、光磁気ディスク、または光ディスクも含むか、またはこれらからデータを受け取りこれらにデータを転送する、あるいはこれらの双方を行うように動作的に結合される。データ送信および命令は、通信ネットワーク上においても生ずることができる。コンピュータ・プログラム命令およびデータを具現化するのに適した情報担体は、あらゆる形態の不揮発性メモリを含み、一例として、半導体メモリ・デバイス、例えば、ＥＰＲＯＭ、ＥＥＰＲＯＭ、およびフラッシュ・メモリ・デバイス、磁気ディスク、例えば、内部ハード・ディスクまたはリムーバブル・ディスク、光磁気ディスク、ならびにＣＤ－ＲＯＭおよびＤＶＤ－ＲＯＭディスクが含まれる。プロセッサおよびメモリは、特殊目的論理回路によって補強すること、または特殊目的論理回路の中に組み込むことができる。

【００９２】

[0098] 「モジュール」および「機能」という用語は、本明細書において用いる場合、ある種のタスクを実行するソフトウェアまたはハードウェア・コンポーネントを意味するが、これに限定されるのではない。モジュールは、アドレス可能な記憶媒体上に常駐するように構成し、１つ又は複数のプロセッサにおいて実行するように構成することができる利点がある。モジュールは、汎用集積回路（「ＩＣ」）、ＦＰＧＡ、またはＡＳＩＣによって全体的または部分的に実現することができる。つまり、モジュールは、一例として、ソフトウェア・コンポーネント、オブジェクト指向ソフトウェア・コンポーネント、クラス・コンポーネントおよびタスク・コンポーネントというようなコンポーネント、プロセス、関数、属性、手順、サブルーチン、プログラム・コードのセグメント、ドライバ、ファームウェア、マイクロコード、回路、データ、データベース、データ構造、表、アレイ、ならびに変数を含むことができる。コンポーネントおよびモジュールにおいて設けられる機能は、もっと少ない数のコンポーネントおよびモジュールに組み合わせることができ、あるいは追加のコンポーネントおよびモジュールに更に分離することもできる。加えて、コンポーネントおよびモジュールは、多くの異なるプラットフォーム上で実現することができるという利点があり、これらのプラットフォームには、コンピュータ、コンピュータ・サーバ、アプリケーション対応(application-enabled)スイッチまたはルータのようなデータ通信インフラストラクチャ機器、あるいは公衆または個人電話スイッチまたは個人分岐交換機（「ＰＢＸ」(private branch exchange)）のような電気通信インフラストラクチャ機器が含まれる。これらの場合のいずれにおいても、選択したプラットフォームにネーティブなアプリケーションを書くことによって、あるいはプラットフォームを１つ又は複数の外部アプリケーション・エンジンにインターフェースすることのいずれかによって、実現を達成することができる。

【００９３】

[0099] ユーザとの双方向処理に備えるために、前述した技法は、ユーザに情報を表示するディスプレイ・デバイス、例えば、ＣＲＴ（陰極線管）またはＬＣＤ（液晶ディスプレイ）モニタ、ならびにキーボードおよびポインティング・デバイス、例えば、マウスまたはトラックボールを有するコンピュータ上で実現することができる。ユーザは、ポインティング・デバイスによってコンピュータに入力を供給することができる（例えば、ユーザ・インターフェース・エレメントと双方向処理を行う）。ユーザとの双方向処理に備えるためには、他の種類のデバイスも用いることができる。例えば、ユーザに宛てるフィードバックは、あらゆる形態の感覚的フィードバック、例えば、視覚フィードバック、聴覚フィードバック、または触覚フィードバックとすることができ、ユーザからの入力は、音響入力、音声入力、または接触入力を含む、あらゆる形態で受け取ることができる。

【００９４】

[0100] 前述した技法は、分散型計算システムにおいて実現することができる。分散型計算システムは、例えば、データ・サーバのようなバック・エンド・コンポーネント、および／またはミドルウェア・コンポーネント、例えば、アプリケーション・サーバ、および／またはフロント・エンド・コンポーネント、例えば、グラフィカル・ユーザ・インターフェースおよび／またはウェブ・ブラウザを有し、これを通じてユーザが実施態様例と双方向処理することができるクライアント・コンピュータ、あるいは、このようなバック・エンド、ミドルウェア、またはフロント・エンド・コンポーネントのあらゆる組み合わせを含む。本システムのコンポーネントは、あらゆる形態または媒体のディジタル・データ通信、例えば、通信ネットワークによって相互接続することができる。通信チャネルとも呼ばれる通信ネットワークの例には、ローカル・エリア・ネットワーク（「ＬＡＮ」）およびワイド・エリア・ネットワーク（「ＷＡＮ」）、例えば、インターネットが含まれ、そして有線ネットワークおよびワイヤレス・ネットワークの双方が含まれる。例の中には、通信ネットワークが仮想ローカル・エリア・ネットワーク（「ＶＬＡＮ」）のような仮想ネットワークまたはサブネットワークを特徴とすることができるものもある。特に明確に示さない限り、通信ネットワークは、ＰＳＴＮの全部または一部、例えば、特定の電気通信事業者が所有する一部も含むことができる。

【００９５】

[0101] 前述の計算システムは、クライアントおよびサーバを含むことができる。クライアントおよびサーバは、一般に、互いに離れており、通信ネットワークを通じて双方向処理するのが通例である。クライアントおよびサーバの関係は、それぞれのコンピュータ上において実行し、互いにクライアントーサーバ関係を有するコンピュータ・プログラムによって生ずる。

【００９６】

[0102] 種々の実施形態は、交信状態にあるものまたは１系統以上の通信経路によって接続されたものとして図示されている。通信経路は、特定のデータ転送媒体に限定されることはない。情報は、電気信号、光信号、音響信号、物理信号、熱信号、またはそのあらゆる組み合わせを用いて、通信経路上で送信することができる。通信経路は、多数の通信チャネル、例えば、データ・フローの容量が同一であるまたは異なる、多重化チャネルを含むことができる。

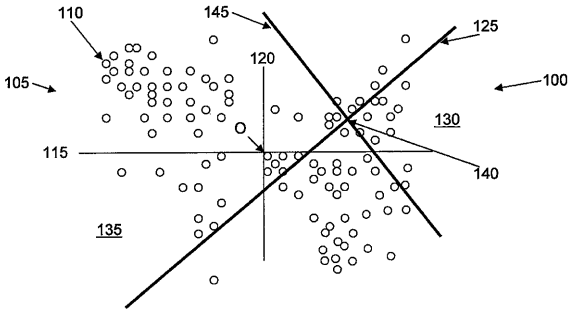
【００９７】

[0103] 図示したユーザ・インターフェース機構(user interface feature)のパラメータを設定するためには、多数のユーザ入力を用いることができる。このような入力の例には、ボタン、ラジオ・ボタン、アイコン、チェック・ボックス、コンボ・ボックス、メニュー、テキスト・ボックス、ツールチップ、トグル・スイッチ、ボタン、スクロール・バー、ツールバー、ステータス・バー、ウィンドウ、あるいはユーザが、本明細書に記載したモジュールまたはシステムのいずれとでも通信すること、および／またはデータを提供することを可能にするユーザ・インターフェースと関連付けられている、その他の適したアイコンまたはウィジェット(widget)が含まれる。

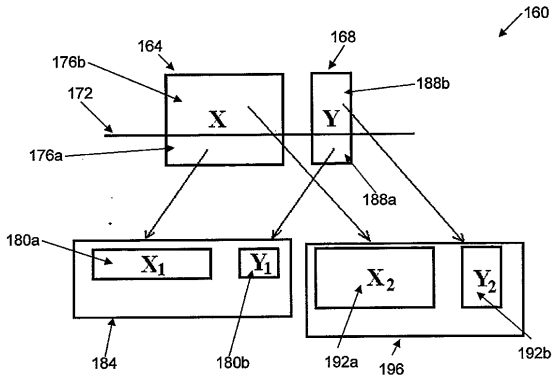
【００９８】

[0104] 以上、具体的な実施形態を参照しながら、本発明について特定的に示し説明したが、添付した特許請求の範囲によって定められる発明の主旨および範囲から逸脱することなく、本発明の形態および詳細には種々の変更が可能であることは、当業者には理解されてしかるべきである。

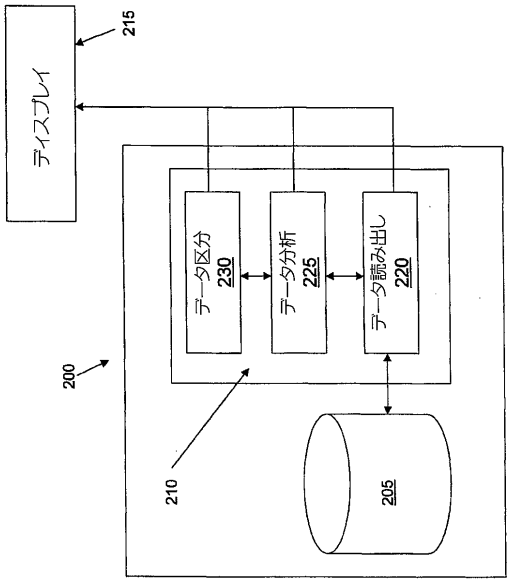
【図１Ａ】



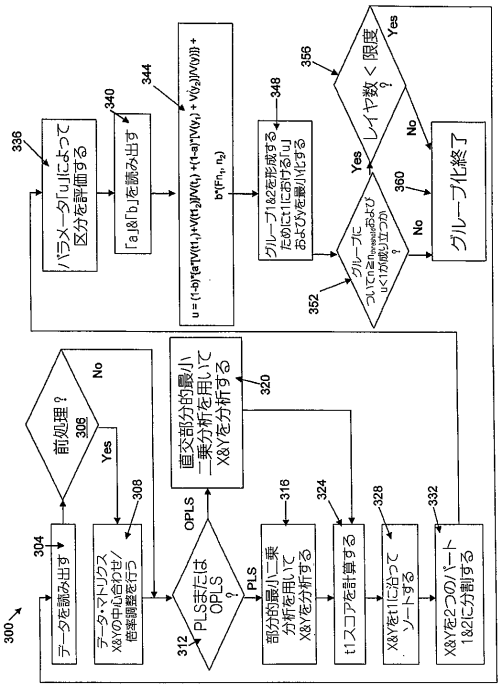
【図１Ｂ】



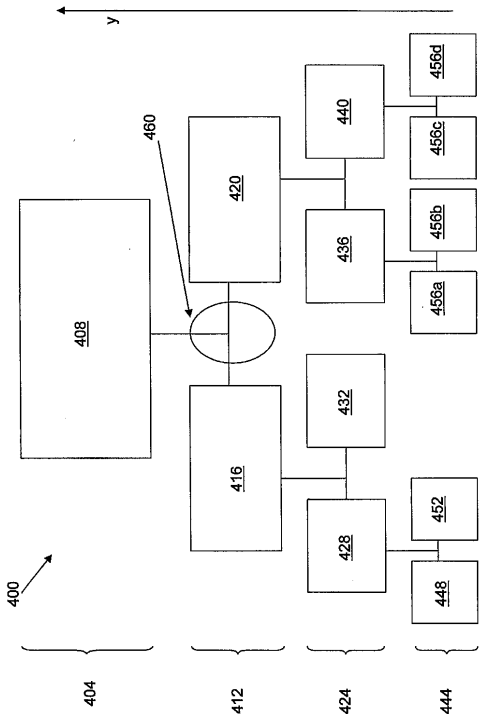
【図２】



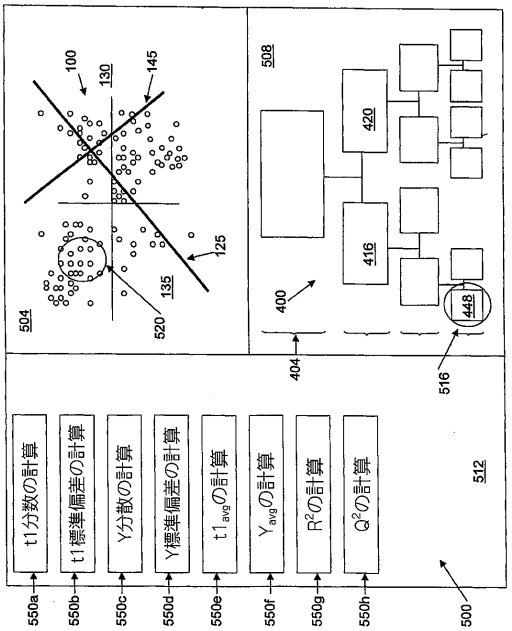
【図３】



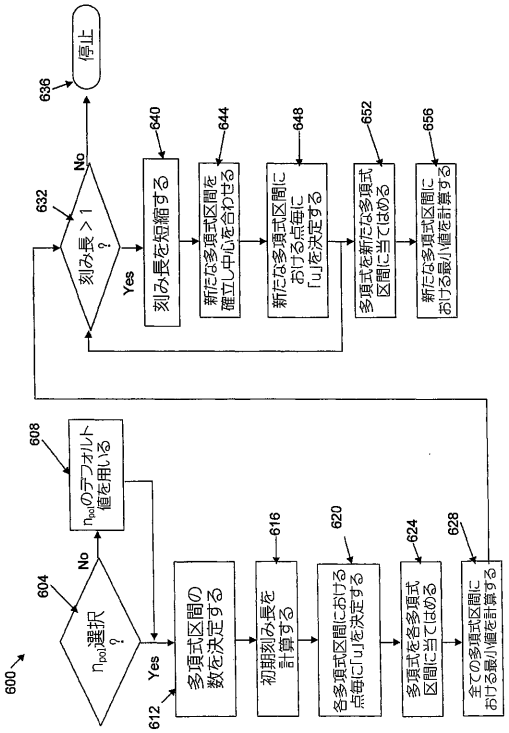
【図４】



【図５】


t1分数の計算
t1標準偏差の計算
Y分散の計算
Y標準偏差の計算
t1avgの計算
Yavgの計算
R2の計算
Q2の計算
550a
550b
550c
550d
550e
550f
550g
550h
500
512
504
508
100
130
145
125
135
520
400
404
416
420
448
516

【図６】


600
604
npol選択 ?
No
Yes
608
npolのデフォルト値を用いる
612
多項式区間の数を決定する
616
初期刻み長を計算する
620
各多項式区間における点毎に「u」を決定する
624
多項式を各多項式区間に当てはめる
628
全ての多項式区間における最小値を計算する
632
刻み長 > 1 ?
No
Yes
636
停止
640
刻み長を短縮する
644
新たな多項式区間を確立し中心を合わせる
648
新たな多項式区間における点毎に「u」を決定する
652
多項式を新たな多項式区間に当てはめる
656
新たな多項式区間における最小値を計算する

フロントページの続き

(74)代理人 100118902
弁理士 山本 修
(74)代理人 100162846
弁理士 大牧 綾子
(72)発明者 ウォルド，スヴァント・ブジャルヌ
アメリカ合衆国ニューハンプシャー州０３０４９，ホリス，パイン・ヒル・ロード ４２
(72)発明者 トリグ，ユーハン
スウェーデン国 エスイー－９０４ ３５，スウェーデン，ウメア，オクスベルスヴェーゲン ２０
(72)発明者 エリクソン，レンナート
スウェーデン国 エスイー－９０７５１，スウェーデン，ウメア，スヨフルヴァーゲン ８５