本シリーズの編集にあたって

社会の進化に伴い，統計科学の環境が大きく変化している．その主な変化として次のような点があげられる．1) データの収集の方法が多様化されている．2) データの平均サイズがますます大きくなっている．3) データの流通が容易になっている．4) 統計計算やシミュレーションに必要となるコンピュータがますます安価になっている．5) 統計計算やシミュレーションの専用ソフトが無料で入手可能になった．6) 統計科学の役割の重要性の認知度が向上している．

このようなさまざまな変化は，統計的データ解析の新しい手法の開発と応用を促し，データマイニング (data mining) や統計的機械学習 (statistical machine learning) のような新しい研究分野が生まれるようになり，その応用が急速に広がっている．従来の統計学，近年のデータマイニングや機械学習（マシンラーニング）に関する定義はいろいろあるが，共通点はデータを対象としていることであるので，本シリーズではこれらを包含する用語として，狭義のデータサイエンス (data science) を用いることにする．

データサイエンスは，広義ではデータの収集，加工，蓄積，管理，流通，解析，マイニングなど，データの流れの上流から下流までを貫く科学である．昨今，データサイエンスは，工学，医学，薬学，生命科学，社会科学（社会，経済，マーケティングなど），心理学，教育学はもちろんのこと，文化学のような，従来は統計学やデータ解析があまり応用されていなかった分野でも，データサイエンスの手法による斬新な研究成果が多く報告されている．データサイエンスは，あらゆる分野において必要となる万人の科学と言っても過言ではない．

データ解析の手法のほとんどは数理的理論に基づいて開発されているので，データサイエンスに関する解説書では，数式を避けると厳密な説明ができなくなる．非理工系の研究者の中には数式が苦手である方が多いため，非理工系の研究分野におけるデータサイエンスの適用が遅れている．一方，データ解析のツールを用いると，数理的な理論が分からなくても，データを入力すると何らかの結果が出力され，形式上はデータ解析が可能な時代になっている．しかし，データ解析の理論に関する理解が不十分であると，統計手法の利用を間違えたり，出力された結果の解析を誤ったりする可能性がある．

データ解析を行うには，用いる手法の数理的理論の理解だけではなく，ツールを用いてデータを解析しなければならない．そのためには，データサイエンスの基礎理論を理解した上でツールを用いてデータを操作し，データ解析やデータマイニングを行うことが望ましい．データ解析やデータマイニングの手法は，データの構造と目的に依存する．万能なデータ解析やマイニングの

手法はない．データ解析やマイニングを行う際には，データの構造や目的に合う手法を用いることが必要である．そのためには，用いる手法の理論を正しく理解することが必要である．

データ解析の手軽なソフトとしては，表計算ソフト Excel や Calc がある．前者はマイクロソフト社の有料ソフトであり，後者はサン・マイクロシステムズ社が開発したフリーソフトである．最近，個人，法人を問わず，ほとんどのパソコンには Excel がインストールされていることもあり，広く利用されている．表計算ソフトは，データの整理や簡単な計算には便利なツールであるが，高度なデータ解析を行うためには，プログラムを作成するか追加ソフトを用いることが必要である．また，これらのソフトは列の数に制限があり，大量のデータ解析には向いていない．その一方，データ解析の専用ソフトとしては SAS，SPSS，S-PLUS などがあるが，これらは高価であるため，個人のポケットマネーでは購入しがたく，恵まれている環境でなければ使用できない．

このようなことから，1990 年代にニュージーランドのオークランド大学統計学科の Ross Ihaka とアメリカのハーバード大学の Robert Gentleman により R（R 環境，R 言語とも呼ぶ）というデータ解析ツールの開発が始められ，1997 年からは多くの賛同者が加わり，オープンソース方式で開発が続けられている．R はフリーソフトであり，インターネットが接続された環境であれば，誰でもどこでも自由にダウンロードできる．R は，基本的な統計計算の環境と専用パッケージの利用環境を提供している．2009 年の現在，公開された R 専用のフリーパッケージの数は 2 千を超えている．R では，表形式のように定型化されたデータ処理やモデリング，データマイニング，機械学習などはもちろん，定型化されていない遺伝子情報のデータ，画像データ，音声・音楽データ，テキストデータなどを解析することも可能である．R は，高度なデータ解析，繊細多様なグラフィックスの作成，データマイニング，機械学習，シミュレーションなどを行うツールであり，伝統的な統計計算やデータ解析の概念を超えたツールとして発展し続けている．従来は，研究者が考案したデータ解析の方法をエンドユーザが使用するまでには長い時間を要したが，R の普及のおかげで，研究者が考案した新しい方法をエンドユーザが使用できるようになるまでのサイクルが大幅に短縮されている．

R による統計学に関する単行本がわが国で初めて刊行されたのは 2003 年である．約 5 年の間に R に関する訳書・和書の数はすでに 30 冊を超えるようになった．これが R 普及の勢いを物語っている．しかし，その中には R による初級統計学や R のマニュアル形態のものが多く，高度なデータ解析やデータマイニングに関する理論を系統的に説明し，その方法を R で実践する，いわゆる理論と実践を両立したものが少ない．

そこで，数理的な基礎が一定程度ある方は，関連手法の数理的理論を理解し，R による実践を通じてその方法の理論と応用を学び，数理的基礎が弱い方々は，R を用いて実践的に入門し，数理的理論を徐々に理解するようにと，数理に強い，弱いに関係なく幅広く使用できる本を提供することが本シリーズの主な目的である．ただし，企画した時点ですでに上記の理念と一致する本が刊行されている分野もある．それらの内容に関しては，重複を避けている．本シリーズでは，可能であれば社会的ニーズに応じて新たな内容の巻を追加していく予定である．

各巻の著者は，それぞれの分野で教育と研究にご活躍されている専門家である．ご多忙にもかかわらずご執筆をお引き受けいただいたことに感謝する．

本シリーズがデータサイエンスの発展に少しでも寄与できれば幸いである．

編者　金 明哲

まえがき

画像は，日常生活から最先端の医療や科学研究にいたるまで，さまざまな分野で広く用いられているデータ形式である．近年では，画像データの大部分がデジタルデータとして取得されるため，即時にコンピュータによる処理・解析が可能になった．デジタルカメラが自動で顔を認識し，ピントや明るさを調節するといった機能は多くの人が経験しているだろう．医療や科学研究の専門分野においては，MRI，CT，共焦点顕微鏡，電波望遠鏡といったさまざまな観察技術が開発され，それらによって記録されたデジタル画像データが診断や理論の実証に不可欠な情報をもたらしている．これらの技術を支えているのが画像処理および画像解析である． デジタル画像からパターンやシグナル強度などの情報を抽出し，定量的に解析することにより，アナログ画像からでは得られなかった機能や知識が得られるようになっている．今後，画像処理は，日常生活においても科学技術の分野においても，ますます身近で重要なスキルになると考えられる．

R[1] は，S 言語という統計解析ソフトウェアを参考に開発された，GNU に準拠するソフトウェアである．ソースが公開されており誰でも無料で利用することができるだけでなく，パッケージという形で世界中から新たな機能（ライブラリ）が追加されている．R は，その高度な統計ライブラリを活かして統計解析や検定などの目的に用いられることが多いが，実は画像処理にも威力を発揮する．画像は，画素が 2 次元に配置された行列型のデータであり，ベクトルや行列操作を得意とする R とは相性がよいうえに，近年，R には画像を入出力・処理するためのライブラリが充実してきている．これらを利用すると，単純な処理であれば，画像の入力から処理までをほんの数個のコマンドで実現できる．さらに，目的に応じたプログラムを組むことで，定型的な画像処理だけでなく，より柔軟で複雑な処理を行うことも可能である．このため，R の統計処理パッケージと画像処理パッケージを併用すれば，画像データの入力から，画像処理，統計解析，そして結果のグラフィカルな出力までを一つのソフトウェア上で行うことができる．これにより，それぞれの処理を別々のソフトウェアで行う場合に比べて大幅な省力化が達成され，大規模な画像データを効率よく扱えるようになる．また，R がフリーウェアであり，かつさまざまなプラットフォームで機能するため，開発した画像処理プログラムをほかのユーザや共同研究者と共有することも容易である．

このようにさまざまな利点をもつ R であるが，これまで R による画像処理を系統立てて解説した書籍は洋書和書を含めて見あたらなかった．本書では，R を使って画像処理の基礎を概説するとともに，R の特徴を活かした実用的な画像処理の手法を紹介する．R はシンプルで学びやす

い言語体系であり，高度な画像処理が驚くほど簡潔に記述できる．このため，本書を通じて，R のユーザだけでなく，プログラミングそのものになじみのなかった読者にも画像処理をより身近に感じてもらえるのではないかと思う．また，並列処理による高速化や他言語との連携方法などを取り挙げ，画像処理をほかのシステムで行ってきた経験者にも興味をもってもらえるような内容を目指した．本書を通じて一人でも多くの読者に画像処理に興味をもってもらい，R による画像処理を研究・勉学に役立ててもらえれば幸いである．そして，読者のフィードバックによって，まだ発展段階である R の画像処理分野の開発コミュニティのさらなる原動力にもつながっていくことを期待したい．

本書の執筆にあたっては多くの方々の協力をいただいた．同志社大学の金明哲氏には，企画から脱稿にいたるまで常に励ましと助言をいただいた．共立出版株式会社の横田穂波氏には，執筆のスケジュールや本書の体裁などに関してさまざまな便宜を図っていただいた．北爪（山本）美和子氏と勝木公雄氏には本文とコードを丁寧に読んで有益なコメントをいただいた．カリフォルニア大学サンディエゴ校 Kavli Institute for Brain and Mind の Ryan Shultzaberger 氏，Minh Tong 氏，Ralph Greenspan 氏には本書のために未発表のデータを快く提供していただいた．RNiftyReg パッケージの利用方法は UCL Institute of Child Health の Jonathan Clayden 氏に，EBImage の利用方法は EMBL の Gregoire Pau 氏にお世話になった．また，世界中の R ユーザの築き上げた共有知なくしては，本書の執筆はかなわなかった．この場を借りて深く感謝したい．

2011 年 10 月

著　者