

統計数理研究所で開発された R パッケージ R packages developed at the Institute of Statistical Mathematics

モデリング研究系 中野 純司 (Junji Nakano)

1. はじめに

統計数理研究所は計算機の出現以来、その時代の最新の統計計算環境を維持してきた。そしてその上で種々のソフトウェアが開発・公開されてきている。ただ計算機技術は急速に発展しており、過去のソフトウェアをソースコードでそのまま配布しても使える人は限られてしまう。そのため統計科学技術センターを中心として、それらのソフトウェアをその時代の計算機環境に合うように保守し、さらに利用しやすくすることが試みられてきた。われわれは、統計科学分野でデファクトスタンダードになっている統計解析ソフトウェア R からそれらのソフトウェアを利用できるように、R のパッケージを作成して公開している。本稿ではそれらを紹介するが、多くは CRAN(<https://cran.r-project.org/> またはそのミラーサイト、例えば <https://cran.ism.ac.jp/>) から入手できる。

2. timsac

TIMSAC(TIME Series Analysis and Control program) は、赤池弘次氏を中心として開発された時系列データの解析、予測、制御のための総合的プログラムパッケージである。オリジナル TIMSAC(TIMASAC-72) は 1972 年に発表され、その後、TIMSAC シリーズとして TIMSAC-74, TIMSAC-78, TIMSAC-84 が統計数理研発行の Computer Science Monograph に発表された。工業プロセスの最適制御、経済変動の分析など広い分野で現在でも利用されている。R パッケージ `timsac` は、FORTRAN で書かれているオリジナルプログラムの計算処理機能の多くをサブルーチン化し、R 関数を通して入出力を行い、必要であればその解析結果等を R でグラフィック表示することにより時系列データ解析を容易にしたものである。

3. SAPP

尾形良彦氏を中心として開発された SASeis (Statistical Analysis of Seismicity) は、地震活動などの統計的解析とモデリングのためのプログラムパッケージである。そこでは大森・宇津の公式と点過程 ETAS (Epidemic Type Aftershock Sequence) モデルを扱っているが、それは地震活動の標準化モデルとして世界各国で使用されている。そこで FORTRAN で書かれたこれらの計算処理機能を R パッケージにしたものが SAPP である。

4. TSSS

R パッケージ TSSS は、北川源四郎氏による書籍「FORTRAN 77 時系列解析プログラミング」(岩波書店、1993) に掲載されていたプログラムを基に作成された時系列データ解析のための関数群である。代表的な時系列のモデリングに必要な最小二乗法、最尤法、カルマンフィルタによる推定の方法、情報量規準 AIC を用いたモデルの評価・選択の方法が実現されてい

る。TSSS は本書のデータをデータセットとして組み込んでおり、関数のドキュメントにおける例題の一部ではこれらのデータセットを用いている。なお、時変係数 AR モデルの時変分散と時変 AR 係数を推定する関数 (tvvar, tvar) については、OpenMP を使った拡張パッケージ tvvarOMP を作成しており、それにより並列処理も行えるようになっている。このパッケージは <http://jasp.ism.ac.jp/ism/TSSS/> から入手できるが、近々、CRAN で公開予定である。

5. catdap

CATDAP (CATegorical Data Analysis Program) は、坂元慶行氏を中心に開発された最適な分割表 (クロス表) の探索のためのプログラムである。最適な説明変数の選択に AIC (Akaike Information Criterion) が使われている。R パッケージ `catdap` は、FORTRAN で書かれた CATDAP の計算処理機能をサブルーチン化することにより、R からこれらに関数として利用できるようにした。最近、主として石黒真木氏によりいくつかの拡張が行われた。例えば現在の関数 `catdap2` では連続値目的変数に適用できたり、目的変数、説明変数に欠測が含まれるデータに適用できたりするようになっている。

なお、関数 `catdap2` の機能をさらに使い易くするため、パッケージ R commander (`Rcmdr`) を使ったメニューインタフェース (`RcmdrPlugin.catdap`) も利用可能になっている。(<http://jasp.ism.ac.jp/ism/RcmdrPlugin.catdap/>)

6. NScluster

R パッケージ `NScluster` は、ネイマン・スコット型空間クラスターモデルのシミュレーションとパラメータ推定のための関数群である。これらの関数は U.Tanaka, Y. Ogata and K. Katsura, Simulation and estimation of the Neyman-Scott type spatial cluster models (Computer Science Monographs, No.34, 1-44, The Institute of Statistical Mathematics, 2008) の FORTRAN プログラムをもとに開発された。

パラメータ推定のためにシンプレックス法を用いているが、モデルによってはかなりの計算時間がかかる。そのため、この時間のかかる計算処理の部分を OpenMP を使って並列化している。OpenMP が利用可能な環境であれば、実行時間の大幅な短縮を図ることができる。

7. Rhpc

R パッケージ `Rhpc` は基本的な並列計算用パッケージ `snow` の流れをくむ R の並列化のためのパッケージである。その特徴は、並列化のために MPI ライブラリを用いるが 2GB 以上のデータ処理に対応している、多くの部分を C でプログラムして実行速度を上げている、R から MPI 外部プログラムを利用し易くなっている、などである。特に最近のスーパーコンピュータ上での利用を念頭において開発されている。

8. Rmpenv

R パッケージ `Rmpenv` は任意精度による実数と複素数の四則計算および基本的な数学関数、さらに行列積や逆行列を求める関数などを実現するパッケージである。現在機能拡張中であり、まだ公開にはいたっていない。

謝 辞

以上のパッケージの作成は嵯峨優美氏、中間栄治氏との共同研究によるところが大きい。また、オリジナルの FORTRAN プログラム作成者の方々に深く感謝したい。