Mini-curso

Tópicos sobre Sistema R

para Computação Estatística

Paulo Justiniano Ribeiro Junior

Laboratório de Estatística e Geoinformação (LEG)

Universidade Federal do Paraná

mailto:paulojus@ufpr.br

http://www.leg.ufpr.br/~paulojus

Programa

1. Introdução:

   Infos sobre o R e aspectos básicos sobre o uso

2. Miscelânia:

   Tópicos e exemplos de uso e funcionalidades do R

3. Tópicos adicionais:

   Interfaces, Sweave, montando pacotes, Tcl/Tk, etc

1 Introdução

Um primeiro contato com o sistema R e discussões sobre o uso

2 O projeto R

Histórico, características e informações diversas sobre o R

2.1 O que é o R?

R é um sistema para computação estatística e gráficos. Consiste de uma linguagem mais um ambiente de operação com gráficos, um *debbuger*, acesso à certas funções do sistema e capacidade de rodar comandos armazenados em arquivos (*script*.)

Influenciado por duas linguagens: S e Scheme, com aparência semelhante ao S e implementações e semânticas internas similares ao Scheme

Linguagem interpretada, programação modular via funções, interfaces com C, C++ e Fortran, implementando uma diversidade de métodos estatísticos

2.2 Sobre a linguagem S

Inicialmente trabalho de John M. Chambers do *Bell Laboratories* (antiga *AT&T*, atualmente *Lucent Technologies*).

Ganhador em 1998 do prestigioso prêmio *Association for Computing Machinery Award for Software Systems* por, nas palavras da citação:

> **pelo sistema S, que mudou para sempre a forma como pessoas analizam, visualizam e manipulam dados**

Durante a última década for o principal veículo para disponibilizar novas metodologias estatísticas aos usuários finais.

S tem uma longa história: o sistema gráfico remonta 1976

J. Chambers agora *Bell Labs Fellow*, membro do R *core team*

2.3 Breve histórico da linguagem S

Nome da linguagem oscilou e os “sabores” de S são conhecidos pelas cores das capas dos livros que tiveram J. Chambers como co-autor

- S1 1984 *brown book* linguagem baseada em macros
- S2 1988 *blue book* extensões por usuários como primeiras classes
- S3 1991 *white book* estrutura de classes, funcionalidade estatística
- S4 1998 *green book* sistema de classes mais rígido

No início eram programas Unix escritos em linguagem *C* e *Fortran*

Linguagem interpretada: facilidades e desvantagens

S-PLUS produzido em 1988 em Seattle (EUA) pela *Statistical Sciences* que em 1993 adquiriu direitos de exclusividade de mercado sobre S e fundiu-se com a *Mathsoft*. Em 2001 separaram-se e tornou-se *Insightful*.

Uma curiosodade: S não é (ou era) visto pelos desenvolvedores como um sistema estatístico, mas sim como um ambiente interativo para gráficos e análise de dados, um sistema para se fazer estatística dentro dele.

2.4 O sistema S-PLUS

Disponível para um limitado espectro de plataformas (Unix, DOS, Windows)

Versão para LINUX somente em 1998, e não disponível para Macintoshes.

Versão UNIX baseadas em S4 desde 1998. Para Windows a partir de 2001.

S-PLUS: muito usado para ensino de estatística, principalmente a nível de pós-graduação

Embora também usado para cursos de serviço, teve menor impacto para ensino a nível de graduação

Licenças acadêmicas caras

Atualmente tem feito muito sucesso em setores comerciais (finanças, indústria farmacêutica, etc)

2.5 O sistema R

R é um sistema originalmente escrito por Ross Ihaka and Robert Gentleman da *University of Auckland* no começo dos anos 90.

Ao usuário parece um dialeto da linguagem S mas internamente é baseado em idéias de *Scheme* (um membro da "família" LISP).

Muito parecido com S3

Provavelmente iniciado como um projeto de pesquisa, mas usado em Auckland para cursos básicos em Macintoshes com 2Mb de memória.

Artigo de R&R na Computer Sciences em 1996

Em 1997 outros se envolveram e criou-se um *core team* com acesso ao código fonte

Havia versão para Windows entretanto foram usuários de Linux que avalancaram o desenvolvimento – não havia versão de *S-PLUS* para Linux

Uma curiosidade: por que o nome R?

2.6 R *vs* **S-PLUS - I**

Co-existência dos sistemas: nem sempre pacífica mas atualmente de franca colaboração

S-PLUS é comercial R é gratuito

R é mais leve, requer menos *hardware*: S-PLUS é monolítico e R tem um pequeno núcleo e extensões

S-PLUS distribuído com GUI “oficial”

Performance comparável, embora R seja mais tolerante a código “mal escrito” que podem fazer o S-PLUS “travar”

No início R apresentava mais *bugs*, porém são mais rapidamente corrigidos

Ambos tem excelente qualidade gráfica, limitações em 3D e gráficos dinâmicos

Pesquisadores com ênfase em computação estatística migraram do S para o R — John Chambers é atualmente membro do R *core team*

Atualmente possuem público alvo diferente. R possui mais recursos em pacotes

2.7 R *vs* **S-PLUS - II**

S (como C) usa *static scoping*

R (como Scheme) usa *lexical scoping*

Consequências práticas:

1. incompatibilidades entre códigos
2. tratamento de variáveis livres em funções
3. objetos em vários arquivos (S) *vs* arquivo único (*workspace* - R)
4. velocidade
5. riscos de perda de trabalho (*crashes*)
6. outras diferenças

2.8 Para que o R é usado?

Obviamente para análise estatística de dados, mas . . .

. . . impossível dizer pois é livremente disponível

Listas e páginas na WEB dão uma idéia

Palavras de um influente membro do R *core team*:

> *One of my main motivations for being involved is a (perhaps the) major use, to provide a first-class statistical system to students and researchers in the third world.*

Atualmente usado para análises estatísticas de larga escala

Aplicações em *micro-arrays* - THE BIOCONDUCTOR PROJECT

Pesquisadores em várias companhias estão desenvolvendo seus sistemas a partir do R

Ambiente de desenvolvimento, implementação e teste de novas metodologias estatísticas através dos *pacotes*

2.9 Como está o projeto R no momento

Primeira versão não-beta (1.0) lançada em 29 de Fevereiro de 2000.

Versão atual: 2.6.1 (outubro de 2007)

Sistema disponível com código aberto

Distribuído segundo termos da GNU—GPL2

Disponível no formato compilado (binários) e/ou fontes + *scripts* de compilação

Compila em Windows, Linux, Mac, Unix, FreeDBS, etc

Sítio oficial: `http://www.r-project.org`

CRAN: Área de *download* e espelhos: `http://cran.r-project.org`

2.10 R FOUNDATION

R Foundation: criada em 2003

Citando o R em publicações:

```
R Development Core Team (2007). R: A language and
  environment for statistical computing. R Foundation
  for Statistical Computing, Vienna, Austria.
  ISBN 3-900051-00-3, URL http://www.R-project.org.
```

Use o comando: `citation()`

2.11 Alguns recursos - I

Uso típico - linha de comando

Mas há muito além disto . . .

- Rcgi, Rweb
- interfaces TCL/TK
- Rsciview, Rcmdr
- . . . projetos R-GUI's em franca atividade
- Tinn-R é uma excelente opção de interface para o windows (se voce tem estômago e paciência para usar este sistema operacional . . .)

2.12 Alguns recursos - II

Pacotes : + de 1000, atualizações frequentes

Metodologias recentes/em pesquisa

“Patch” e versão “devel” diáriamente disponíveis

Disponível como biblioteca compartilhada *(shared library)* e/ou estática *(static library)*

Interfaces com programas, linguagens e SGBD: possibilidades diversas via integração com outros recursos

Embedding reserva ao R o que ele tem de melhor: capacidade de produzir análises estatísticas e gráficos

2.13 R é um projeto atípico

R não tem um líder e se baseia no consenso entre membros do R *core team*

Há áreas de *expertise* dentre estes

Deferência especial com os “fundadores”

R Core team: *modus operandi* e diretrizes , encontros regulares (*DSC* e *useR meetings*) e aparentemente excelente relacionamento

2.14 Alguns tópicos “difícies”

Ross Ihaka disputa com sua Universidade para “liberar” seu trabalho com o R.

O direito de se construir um sistema comercial baseado no R não é claro

A propriedade do código fonte não é bem definida

por ex: R usa algorítmos estatísticos da *RSS*, com licença sob o entendimento de que o projeto não é comercial

Projetos livres são incrivelmente trabalhosos:

usuários demandam: funcionamento como esperam e reparos

tem o hábito de reportar/perguntar antes de ler manuais

usuários que mais demandam provavelmente usam para ganhos comerciais.

Possível solução (como em LINUX) é prover suporte para produto gratuito.

Compatibilidade entre versões e dificuldades com “entranhas de sistemas operacionais”

2.15 Alguns pontos fortes do projeto

R é largamente usado por grupos em países onde um sistema comercial é proibitivo e roda bem em hardware “quase obsoleto”

Listas (R-help, R-packages, R-announce e R-devel)

Quase todo contato por internet

Fácil adição de novos aspectos pelo usuário

- Possibilidades didáticas
- Encontrou um bug: arrume a prossiga!

Mais aspects de orientação a objetos nas novas versões

Sinergia com DBMS’s & mais uso/integração via XML

Ênfase em compatibilidade com várias plataformas

Disponibilidade de documentação e materiais

Desenvolve senso de apreciação pelo desenho de software e suporte

2.16 Estrutura Atual

Pacotes

- **base** (parte do ”source code”): base*, datasets*, grDevices*, graphics*, methods*, stats*, utils*, grid, splines, stats4, tcltk, tools
- **recommended**:
  boot, class, cluster, codetools, foreign, lattice, mgcv, nlme, nnet, rcompgen, rpart, survival, VR *bundle* (nnet, MASS, spatial)
- **contributed packages**:
  fontes: CRAN, OMEGAHAT, BIOCONDUCTOR
- *unofficial packages*

Pacotes disponívies

- para listar: `library()`
- para carregar: `require(pacote)`

3 Usando o R

Dados, objetos e gráficos

3.1 R, (X)emacs e ESS

emacs/xemacs : editor genérico com facilidades para diversas linguagens

ess: **e**macs **s**peaks **s**tatistics

módulo para integrar e facilitar o uso de programas estatísticos com (x)emacs

suporte para: R, S-plus, SAS, Stata, BUGS

para carregar coloque em `.xemacs/init.el`:

```
(require 'ess-site)
```

3.2 Entrada de dados

como vetores

vetores ”automáticos”

estilo planilha

importação de outros formatos

conexões, compartilhamento de espaços de memória

3.3 Estruturas de dados e tipo de objetos

vetores

matrizes

data-frames

listas

funções

- ver página e script

3.4 Interface - Função em C

Arquivo test.c:

```
#include <math.h>
#include <R.h>
#include <Rmath.h>

void cormatern(int *n, double *uphi, double *kappa, double *ans)
 {
  int register i;
  double cte;
  for (i=0; i<*n; i++){
    if (uphi[i]==0) ans[i] = 1;
    else{
      if (*kappa==0.5)
      ans[i] = exp(-uphi[i]);
      else {
        cte = R_pow(2, (-(*kappa-1)))/gammafn(*kappa);
        ans[i] = cte * R_pow(uphi[i], *kappa) * bessel_k(uphi[i],*kappa,1);
    }}}}
```

3.5 Interface - *wrapper* em R

Arquivo teste.R:

```
"matern" <- function(u, kappa){
  out <- .C("cormatern",
            as.integer(length(u)),
            as.double(u),
            as.double(kappa),
            res = as.double(rep(0,length(u))))$res
  return(out)
}
```

3.6 Interface com C - uso no R

Comandos:

```
$ R CMD SHLIB teste.c
$ R
> source('teste.R')
> dyn.load('teste.so')
> matern(0.1, 1)
> matern(seq(0,1,l=11), 1)
```

4 Uma visão pessoal e institucional

O sistema R como parte do sistema computacional de um Departamento acadêmico

4.1 Como comecei e porque uso o R

1998 – S e início da **geoS** – *S-PLUS* Ambiente Unix

Dificuldades

- Rotinas númericas e Bayesianas
- Evitar *loops*
- uso de memória
- velocidade

Soluções

- Programação eficiente (em S)
- Transcrição de partes do código para *C*
- R

Outras Motivações: Sistemas LINUX, código aberto, custos, e perspectivas na volta ao Brasil.

mudança inicialmente “subversiva” depois largamente adotada

4.2 Uso do R na UFPR

Como parte do *Projeto de Recursos Computacionais no Apoio ao Ensino e Pesquisa*.

Concepção: projeto de baixíssimo custo com aproveitamente de *hardware* obsoleto, modelo cliente–servidor, com uso exclusivo de programas gratuitos (e de preferência com código aberto), administração facilitada

Básico: Linux + R + LaTeX+ Openoffice

Vantagens: distribuição livre, integração, multi-plataforma, arquivos de comandos

4.3 Projetos

LEG : Laboratório de Estatística e Geoinformação

- **geoR** e **geoRglm**
- **aRT**: API R-Terralib
- myR
- **Rcitrus**
- parte do projeto URR (Ultimate Research Resources)
- pacote com funções de apoio ao ensino

Parcerias

- Projeto SAUDAVEL (**S**istema de **A**poio **U**nificado para **D**etecção e **A**companhamento em **V**igilância **E**pidemio**L**ógica)
- FUNDECITRUS
- CESO - DUKE

5 Sweave e Pacotes

Material de apoio didático/científico:

5.1 Sweave

Integração de R com LaTeX

Conceito de Ciência Reproduzível

artigos, livros, apostilas, etc

Documentos dinâmicos

5.2 Construindo pacotes

Modelos para pacote

Aprenda com os outros!

Estrutura organizada

Testes e documentação

pacotes “oficiais” e não oficiais

Ideal para divulgação de trabalhos de pesquisa

Ideal para instrumentos de apoio didático, produção de cursos e materiais como livros apostilas, etc

pequena demo do pad: montando, instalando e usando