Bibliography

J. M. Chambers. *Programming with Data. A guide to the S language*. Springer, 1998. URL http://cm.bell-labs.com/cm/ms/departments/sia/Sbook/.

M. Kohl, P. Ruckdeschel, and T. Stabla. General Purpose Convolution Algorithm for Distributions in S4-Classes by means of FFT. Technical Report. Also available under http://www.uni-bayreuth.de/departments/math/org/mathe7/RUCKDESCHEL/pubs/comp.pdf, Feb. 2005.

J. A. Rice. *Mathematical statistics and data analysis*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California, 1988.

P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen. S4 *Classes for Distributions— a manual for packages* distr, distrSim, distrTEst, *version 1.7*, distrEx, *version 0.4-3*, May 2006. http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR.

Peter Ruckdeschel
Matthias Kohl
Thomas Stabla
Florian Camphausen
Mathematisches Institut
Universität Bayreuth
D-95440 Bayreuth
Germany
peter.ruckdeschel@uni-bayreuth.de

The regress function

An R function that uses the Newton Raphson algorithm for fitting certain doubly linear Gaussian models.

by David Clifford and Peter McCullagh

Introduction

The purpose of this article is to highlight the many uses of the `regress` function contained in the `regress` package. The function can be used to fit linear Gaussian models in which the mean is a linear combination of given covariates, and the covariance is a linear combination of given matrices. A Newton-Raphson algorithm is used to maximize the residual log likelihood with respect to the variance components. The regression coefficients are subsequently obtained by weighted least squares, and a further matrix computation gives the best linear predictor for the response on a further out-of-sample unit.

Many Gaussian models have a covariance structure that can be written as a linear combination of matrices, for example random effects models, polynomial spline models and some multivariate models. However it was our research on the nature of spatial variation in crop yields, McCullagh and Clifford (2006), that motivated the development of this function.

We begin this paper with a review of the kinds of spatial models we wish to fit and explain why a new function is needed to achieve this. Following this we discuss other examples that illustrate the broad range of uses for this function. Uses include basic random effects models, multivariate linear models and examples that include prediction and smoothing. The techniques used in these examples can also be applied to growth curves.

Spatial Models

McCullagh and Clifford (2006) model spatial dependence of crop yields as a planar Gaussian process in which the covariance function at points z, z' in the plane has the form

$$\sigma_0^2 \delta_{z-z'} + \sigma_1^2 K(|z - z'|) \qquad (1)$$

with non-negative coefficients σ_0^2 and σ_1^2. In the first term, Dirac's delta is the covariance function for white noise, and is independent on non-overlapping sets. The second term is a stationary isotropic covariance function from the Matérn class or power class.

McCullagh and Clifford (2006) proposed that the spatial correlation for any crop yield is well described by a convex combination of the Dirac function and the logarithmic function associated with the de Wijs process. The de Wijs process (de Wijs, 1951, 1953) is a special case of both the Matérn and power classes and is invariant to conformal transformation. The linear combination of white noise and the de Wijs process is called the conformal model. We examined many examples from a wide variety of crops worldwide, comparing the conformal model against the more general Matérn class. Apart from a few small-scale anisotropies, little evidence was found of systematic departures from the conformal model. Stein (1999) is a good reference for more details on spatial models.

It is important to point out that the de Wijs process and certain other boundary cases of the Matérn family are instances of generalized covariance functions corresponding to intrinsic processes. The fitting

of such models leads to the use of generalized covariance matrices that are positive definite on certain contrasts. Such matrices usually have one or more negative eigenvalues, and as such the models cannot be fitted using standard functions in R.

Clifford (2005) shows how to compute the generalized covariance matrices when yields are recorded on rectangular plots laid out on a grid. For the de Wijs process this is an analytical calculation and the `spatialCovariance` package in R can be used to evaluate the covariance matrices.

In order to illustrate exactly what the spatial model looks like let us consider the Fairfield Smith wheat yield data (Fairfield Smith, 1938). We wish to fit the conformal model to this dataset and to include row and column effects in our model of the mean. Assume at this point that the following information is loaded into R: the yield `y` and factors `row` and `col` that indicate the position of each plot.

The plots are 6 inches by 12 inches in a regular grid of 30 rows and 36 columns with no separations between plots. Computing the covariance matrix involves two steps. The first stage called `precompute` is carried out once for each dataset. The second stage called `computeV` is carried out for each member of the Matérn family one wishes to fit to the model.

```
require("spatialCovariance")
foot <- 0.3048 ## convert from ft to m
info <- precompute(nrows=30,ncols=36,
                   rowwidth=0.5*foot,
                   colwidth=1*foot,
                   rowsep=0,colsep=0)
V <- computeV(info)
```

The model we wish to now fit to the response variable `y` observed on plots of common area $|A| = 0.0465m^2$ is

$$y \sim \text{row} + \text{col} + \text{spatial} \;\; \text{error}$$

where the spatial error has generalized covariance structure

$$\Sigma = \sigma_0^2|A|I + \sigma_1^2|A|^2V. \qquad (2)$$

The fact that yield is an extensive variable means one goes from Equation 1 to Equation 2 by aggregating yield over plots, i.e. by integrating the covariance function over plots. We fit this model using the `regress` function. One can easily check that V has a negative eigenvalue and hence the model cannot be fitted using other packages such as `lme` for example.

```
require("regress")
model1 <- regress(y~row+col,~V,identity=TRUE)
```

In order to determine the residual log likelihood relative to a white noise baseline, we need to fit the model in which $\sigma_1^2 = 0$, i.e. $\Sigma = \sigma_0^2|A|I$. The residual log likelihood is the difference in log likelihood between these two models.

```
> BL <- regress(y~row+col,~,identity=TRUE)
> model1$llik - BL$llik
[1] 12.02406
> summary(model1,fixed.effects=FALSE)

Maximised Resid. Log Likelihood is -4373.721

Linear Coefficients: not shown

Variance Coefficients:
             Estimate Std. Error
   identity  39243.69   2139.824
          V  60491.21  19664.338
```

The summary of the conformal model shows that the estimated variance components are $\hat{\sigma}_0^2|A| = 39.2 \times 10^4$ and $\hat{\sigma}_1^2|A|^2 = 6.05 \times 10^4$. The standard errors associated with each parameter are based on the Fisher Information at the point of convergence.

Random Effects Models

Exchangeable Gaussian random effects models, also known as linear mixed effects models Pinheiro and Bates (2000), can also be fitted using `regress`. However the code is not optimised to use the `groupedData` structure like `lmer` is and `regress` cannot handle the large datasets that one can fit using `lmer`, Bates (2005).

The syntax for basic models using `regress` is straightforward. Consider the example in Venables and Ripley (2002)[p. 286] which uses data from Yates (1935). In addition to the treatment factors N and V, corresponding to nitrogen levels and varieties, the model has a block factor B and a random B:V interaction.

The linear model is written explicitly as

$$y_{bnv} = \mu + \alpha_n + \beta_v + \eta_b + \zeta_{bv} + \epsilon_{bnv} \qquad (3)$$

where the distribution of the random effects are $\eta_b \sim N(0, \sigma_B^2)$, $\zeta_{bv} \sim N(0, \sigma_{B:V}^2)$ and $\epsilon_{bnv} \sim N(0, \sigma^2)$, all independent with independent components, Venables and Ripley (2002). The `regress` syntax mirrors how the model is defined in (3).

```
oats.reg <- regress(Y~1+N+V,~B+I(B:V),
                    identity=TRUE,data=Oats)
```

Similar syntax is used by the `lmer` function but the residual log likelihoods reported by the two functions are not the same. The difference is attributable in part to the term $+\frac{1}{2}\log|X^\top X|$, which is constant for comparison of two variance-component models, but is affected by the basis vectors used to define the subspace for the means. The value reported by `regress` is unaffected by the choice of basis vectors.

While the `regress` function can be used for basic random effects models, the `lmer` function can be used to fit models where the covariance structure cannot be written as a linear combination of known matrices. The `regress` function does not apply to such models. An example of such a model is one that includes a random intercept and slope for a covariate for each level of a group factor. A change of scale for the covariate would result in a different model unless the intercept and slope are correlated. In `lmer` the `groupedData` structure automatically includes a correlation parameter in such a model. See Bates (2005) for an example of such a model.

Multivariate Models Using regress

In this section we show how to fit a multivariate model. The example we choose to work with is a very basic multivariate example but the method illustrates how the complex models can be fitted. Suppose that the observations are i.i.d bivariate normal, i.e. $Y_i \sim N(\mu, \Sigma)$ where $\mu = \binom{\mu_1}{\mu_2}$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. The goal is to estimate the parameter $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \gamma = \rho\sigma_1\sigma_2)$.

Let Y denote a vector of the length $2n$ created by concatenating the observations for each unit. The model for the data can be written as $Y \sim N(X\mu, I_n \otimes \Sigma)$ where $X = 1 \otimes I_2$ and 1 is a vector of ones of length n. Next we use the fact that one can write Σ as

$$\Sigma = \sigma_1^2 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \gamma \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \sigma_2^2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

to express the model as $Y \sim N(X\mu, D)$ where

$$D = \sigma_1^2 I_n \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \gamma I_n \otimes \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \sigma_2^2 I_n \otimes \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

The code to generate such data and to fit such a model is given below, note that $A \otimes B$ = `kronecker(A,B)` in R.

```
library("regress")
library("MASS")  ## needed for mvrnorm
n <- 100
mu <- c(1,2)
Sigma <- matrix(c(10,5,5,10),2,2)
Y <- mvrnorm(n,mu,Sigma)
## this simulates multivariate normal rvs

y <- as.vector(t(Y))
X <- kronecker(rep(1,n),diag(1,2))

V1 <- matrix(c(1,0,0,0),2,2)
V2 <- matrix(c(0,0,0,1),2,2)
V3 <- matrix(c(0,1,1,0),2,2)

sig1 <- kronecker(diag(1,n),V1)
sig2 <- kronecker(diag(1,n),V2)
gam <- kronecker(diag(1,n),V3)

reg.obj <- regress(y~X-1,~sig1+sig2+gam,
            identity=FALSE,start=c(1,1,0.5))
```

The summary of this model shows that the estimated mean parameters are $\hat{\mu}_1 = 1.46$ and $\hat{\mu}_2 = 1.78$. The estimates for the variance components are $\hat{\sigma}_1^2 = 9.25$, $\hat{\sigma}_2^2 = 10.27$ and $\hat{\gamma} = 3.96$. The true parameter is $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \gamma = \rho\sigma_1\sigma_2) = (1, 2, 10, 10, 5)$.

```
Maximised Residual Log Likelihood is -315.48

Linear Coefficients:
      Estimate Std. Error
   X1    1.461      0.304
   X2    1.784      0.320

Variance Coefficients:
      Estimate Std. Error
 sig1    9.252      1.315
 sig2   10.271      1.460
 gam     3.963      1.058
```

A closed form solution exists for the estimate of Σ in this case, $\hat{\Sigma} = \frac{1}{n-1}(Y - \bar{Y})^\top (Y - \bar{Y})$ where Y denotes the $n \times 2$ matrix of observations. Our computed result can be checked using:

```
> Sig <- var(Y)
> round(Sig, 3)

      [,1]   [,2]
[1,] 9.252  3.963
[2,] 3.963 10.271
```

It may be of interest to fit the sub-model with equal variances, $\sigma^2 = \sigma_1^2 = \sigma_2^2$, a case for which no closed form solution exists. This model can be fitted using the code shown below. The estimate for σ^2 is $\hat{\sigma}^2 = 9.76$ and we can see that the difference in residual log likelihood between the sub-model and the full model is only 0.16 units.

```
> sig <- sig1+sig2
> reg.obj.sub <- regress(y~X-1,~sig+gam,
 +              identity=FALSE,start=c(1,.5))
> summary(reg.obj.sub)

Maximised Residual Log Likelihood is -315.65

Linear Coefficients:
     Estimate Std. Error
  X1    1.461      0.312
  X2    1.784      0.312

Variance Coefficients:
     Estimate Std. Error
 sig    9.761      1.059
 gam    3.963      1.059
```

The argument `start` allows one to specify the starting point for the Newton-Raphson algorithm. Another argument `pos` allows one to force certain parameters to be positive or negative, in this example we may only be interested in cases with negative correlation. By default the support for the variance components is the real line.

Prediction and Smoothing

`regress` can also be used to implement smoothing and best linear prediction, McCullagh (2005)[example 5]. Stein (1999) notes that in the geostatistical literature best linear prediction is also known as kriging and in effect

$$\mathrm{BLP}(x) = \mathrm{E}(Y(x)|\mathrm{data})$$

for an out-of-sample unit whose x-value is x.

Smoothing usually requires the choice of a bandwidth in each example, Efron (2001), but bandwidth selection occurs automatically through REML estimation of polynomial spline models. This is best illustrated by the cubic spline generalized covariance function (Wahba, 1990; Green and Silverman, 1994). The cubic spline model has a two-dimensional kernel spanned by the constant and linear functions of x. Consequently the model matrix X must include $\sim 1 + x$, but could also include additional functions.

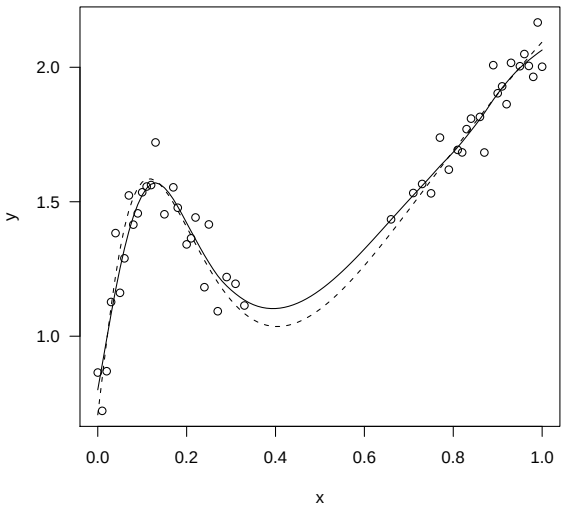

Figure 1: Smoothing spline and best linear predictor fitted by REML.

In the code below we simulate data as in Wahba (1990)[p. 45], but after generating data for one hundred points in the unit interval we take a sample of these points to be used for fitting purposes. The goal of fitting a smooth curve to the data is achieved by the best linear predictor, which is a cubic spline. The spline is computed at all x-values and has knots at the observed values. This experiment can be carried out in the following four steps.

```
## 1: Data simulation
n <- 101
x <- (0:100)/100
prb <- (1 + cos(2*pi*x))/2
## indicator which x-values are observed
obs  <- (runif(n) < prb)
mu <-  1 + x +  sin(3*log(x+0.1))/(2+x)
## full data
y <- mu + rnorm(n,0,0.1)

## 2. Fit the cubic spline to observed data
one <- rep(1,n)
d <- abs(x %*% t(one) - one %*% t(x))
V <- d^3
X <- model.matrix(y~1+x)
fit <- regress(y[obs]~X[obs,],~V[obs,obs],
                identity=TRUE)

## 3. BLP at all x given the observed data
wlsfit <- X%*%fit$beta
blp <- wlsfit + fit$sigma[1]*V[,obs]
               %*% fit$W%*%(y-wlsfit)[obs]

## 4. Plot of results
plot(x[obs],y[obs], xlab="x",ylab="y")
lines(x,mu,lty=2)
lines(x,blp)
```

The results of this experiment can be seen in Figure 1. The points are the observed data, the solid line is the best linear predictor and the dashed line shows the function used to generate the data.

Bibliography

D. Bates. Fitting linear mixed models in R. *RNews*, 5 (1), May 2005.

D. Clifford. Computation of spatial covariance matrices. *Journal of Computational and Graphical Statistics*, 14(1):155–167, 2005.

H. de Wijs. Statistics of ore distribution. Part I: frequency distribution of assay values. *Journal of the Royal Netherlands Geological and Mining Society*, 13: 365–375, 1951. New Series.

H. de Wijs. Statistics of ore distribution. Part II: Theory of binomial distribution applied to sampling and engineering problems. *Journal of the Royal Netherlands Geological and Mining Society*, 15:125–24, 1953. New Series.

B. Efron. Selection criteria for scatterplot smoothers. *Annals of Statistics*, 2001.

H. Fairfield Smith. An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28:1–23, January 1938.

P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.

P. McCullagh. *Celebrating Statistics: Papers in honour of Sir David Cox on his 80th birthday*, chapter Exchangeability and regression models, pages 89–113. Oxford, 2005.

P. McCullagh and D. Clifford. Evidence for conformal invariance of crop yields. Accepted at *Proceedings of the Royal Society A*, 2006.

J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. New York : Springer-Verlag, 2000.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer-Verlag New York, Inc., 4th edition, 2002.

G. Wahba. *Spline Models for Observational Data*. SIAM, Philadephia, 1990.

F. Yates. Complex experiments. *Journal of the Royal Statistical Society (Supplement)*, 2:181–247, 1935.

David Clifford, CSIRO
Peter McCullagh, University of Chicago
David.Clifford@csiro.au
pmcc@galton.uchicago.edu

Processing data for outliers

by Lukasz Komsta

The results obtained by repeating the same measurement several times can be treated as a sample coming from an infinite, most often normally distributed population.

In many cases, for example quantitative chemical analysis, there is no possibility to repeat measurement many times due to very high costs of such validation. Therefore, all estimates and parameters of an experiment must be obtained from a small sample.

Some repeats of an experiment can be biased by crude error, resulting in values which do not match the other data. Such values, called outliers, are very easy to be identified in large samples. But in small samples, often less than 10 values, identifying outliers is more difficult, but even more important. A small sample contaminated with outlying values will result in estimates significantly different from real parameters of whole population (Barnett, 1994).

The problem of identifying outliers in small samples properly (and making a proper decision about removing or leaving suspicious data) is very old and the first papers discussing this problem were published in the 1920s. The problem remained unresolved until 1950s, due to lack of computing technology to perform valid Monte-Carlo simulations. Although current trends in outlier detection rely on robust estimates, the tests described below are still in use in many cases (especially chemical analysis) due to their simplicity.

Dixon test

All concepts of outlier analysis were collected by Dixon (1950) :

1. chi-squared scores of data
2. score of extreme value
3. ratio of range to standard deviation (two opposite outliers)
4. ratio of variances without suspicious value(s) and with them
5. ratio of ranges and subranges

The last concept seemed to the author to have the best performance and Dixon introduced his famous test in the next part of the paper. He defined several coefficients which must be calculated on an **ordered** sample $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. The formulas below show these coefficients in two variants for each of them - when the suspicious value is lowest and highest.

$$r_{10} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \quad (1)$$

$$r_{11} = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \quad (2)$$

$$r_{12} = \frac{x_{(2)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(3)}} \quad (3)$$

$$r_{20} = \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} \quad (4)$$

$$r_{21} = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \quad (5)$$

$$r_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}, \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}} \quad (6)$$

The critical values of the above statistics were not given in the original paper, but only discussed. A year later (Dixon, 1951) the next paper with critical