Design Document for Relevance-Weighting: R Package glmm

Sydney Benson

March 9, 2019

Abstract

This design document will give an overview of the changes made to the R package `glmm` with respect to the relevance-weighted likelihood method mentioned in Wang (2001). We use relevance-weighting to better reflect the real-world occurrence of more or less informative observations.

1 Introduction

This project is meant to enable the user of the `glmm` function in the `glmm` R package to include an optional relevance-weighting scheme. A common assumption of linear models is that each observation in a data set is equally informative and trustworthy; however, in real-world data sets, this is rarely the case. Thus, the optional relevance-weighting scheme will allow the user to place a heavier weight on the more informative and/or trustworthy observations in their data set so that those data points that are less informative affect the model to a lesser degree.

2 The Process

The weighting scheme for the observations can be implemented in the `objfun.R` function. First, the function will need to establish whether the user has supplied a proper weighting scheme. Next, the weighting scheme will need to be applied to the respective elements of the log-likelihood. After defining the weighting vector, the remainder of this section will illustrate how this weighting scheme will be applied to the log-likelihood values.

2.1 The Weighting Vector

This vector, called Λ, must be a vector with the same length as the response vector and must contain all positive values.

2.2 Weighted Log-Likelihood

From Hu and Zidek's (1997) method, mentioned in Wang (2001), we know the relevance-weighted likelihood (REWL) is

$$\text{REWL}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)^{\lambda_i} \tag{1}$$

where $f(x_i;\theta)$ is the probability distribution for x_i and λ_i is the weight for x_i. Then, the relevance-weighted log-likelihood (RWLL) is

$$\begin{aligned}
\text{RWLL}(\theta) &= \log\left(\prod_{i=1}^{n} f(x_i;\theta)^{\lambda_i}\right) && (2)\\
&= \sum_{i=1}^{n} \log\left(f(x_i;\theta)^{\lambda_i}\right) && (3)\\
&= \sum_{i=1}^{n} \lambda_i \log\left(f(x_i;\theta)\right) && (4)
\end{aligned}$$

Then, knowing that $l_m(\theta|y) = \log\left(\frac{1}{m}\sum_{i=1}^{r} j_i e^{v_i}\right)$ where m is the number of rows of the u matrix, j_i is the number of rows of the chunk of the u matrix being processed by the ith core and v_i is the value obtained by each core, as described in the design document for parallel computing, we have

$$\begin{aligned}
\text{RWLL}(\theta) &= \sum_{i=1}^{n} \lambda_i \log\left(\frac{1}{m}\sum_{i=1}^{r} j_i e^{v_i}\right) && (5)\\
&= a + \sum_{i=1}^{n} \lambda_i \log\left(\frac{1}{m}\sum_{i=1}^{r} j_i e^{v_i - a}\right) && (6)
\end{aligned}$$

where $a = \max(v_i)$.