Application Tutorial: OmicKriging

Keston Aquino-Michaels, Heather E. Wheeler, Vassily V. Trubetskoy and Hae Kyung Im

January 22, 2014

Method citation: Wheeler HE, et al. (2013) Poly-Omic Prediction of Complex Traits: OmicKriging. arXiv:1303.1788 `http://arxiv.org/abs/1303.1788`

NOTE: Currently, OmicKriging is not built for Windows due to dependency on the doMC library.

---

1 Running OmicKriging with Example Data

To install from CRAN:

```
> install.packages("OmicKriging")
```

Start by loading OmicKriging functions into R:

```
> library(OmicKriging)
```

Define paths to the genotype (plink binary pedigree format), gene expression, and phenotype data files (paths may differ based on where the files are located). The path.package() function returns the package installation directory. These files will later be passed to upcoming functions:

```
> library(OmicKriging)
> "%&%" <- function(a, b) paste(a, b, sep="")
> gdsFile <-"gdsTemp.gds"
> ok.dir <-  path.package('OmicKriging') %&% "/doc/vignette_data/"
> bFile <- ok.dir %&% "ig_genotypes"
> expFile <- ok.dir %&% "ig_gene_subset.txt.gz"
> phenoFile <- ok.dir %&% "ig_pheno.txt"
```

Load the phenotype data into R:

```
> pheno <- read.table(phenoFile, header = T)
```

Load a pre-computed GCTA GRM into R (recommended):

```
> grmMat <- read_GRMBin(bFile)
```

Alternatively, to compute the GRM in R start by converting the genotype data from plink binary format into GDS format:

```
> convert_genotype_data(bFile = bFile, gdsFile = gdsFile)
```

```
Start snpgdsBED2GDS ...
        open /tmp/RtmpdD17qT/Rinst6b8f73805df7/OmicKriging/doc/vignette_data/ig_genotypes.bed in
        open /tmp/RtmpdD17qT/Rinst6b8f73805df7/OmicKriging/doc/vignette_data/ig_genotypes.fam DON
        open /tmp/RtmpdD17qT/Rinst6b8f73805df7/OmicKriging/doc/vignette_data/ig_genotypes.bim DON
Wed Jan 22 10:19:03 2014        store sample id, snp id, position, and chromosome.
        start writing: 99 samples, 43555 SNPs ...
         Wed Jan 22 10:19:03 2014        0%
         Wed Jan 22 10:19:03 2014        100%
Wed Jan 22 10:19:03 2014        Done.
```

Subsequently, compute a genetic relatedness matrix (GRM) from the GDS file:

```
> grmMat <- make_GRM(gdsFile = gdsFile)
```

By default, grmFilePrefix is set to NULL, however if specified, this function will save the computed GRM to disk in GCTA binary format. Additionally by default both snpList and sampleList are set to NULL. However you may restrict the GRM calculation by specifying a vector of sample IDs or a vector of SNP IDs here.

Load and calculate a gene expression relatedness matrix (GXM) with the following function:

```
> gxmMat <- make_GXM(expFile = expFile)
```

Similarly, by default, gxmFilePrefix is set to NULL, however if specified, this function will save the computed GXM to disk in GCTA binary format.

Additional convenience functions are included to perform principal components analysis (PCA):

```
> pcMatXM <- make_PCs_irlba(gxmMat, n.top = 10)
> pcMatGM <- make_PCs_irlba(grmMat, n.top = 10)
> pcMat <- cbind(pcMatGM, pcMatXM[match(rownames(pcMatGM), rownames(pcMatXM)),])
```

The following convenience function allows the user to perform n-fold cross-validation. Specify the number of cores you wish to use (default = "all"), the number of cross-validation folds desired (default = 10), covariates (by default covar.mat = NULL), the phenotype object, pheno.id (by default = 1 (the first phenotype in the file)), the h2 vector and a list of the correlation matrices to be included.

Note: The sum of the h2 vector must be between 0 and 1. In this example, we will give each matrix equal weight.

```
> result <- krigr_cross_validation(pheno.df = pheno,
+         cor.list = list(grmMat, gxmMat),
+         h2.vec = c(0.5, 0.5),
+         covar.mat = pcMat,
+         ncore = 2,
+         nfold = "LOOCV")

Detected 99 samples...
Set leave-one-out cross-validation...
With 2 logical core(s)...
Running OmicKriging...
```

```
Call:
lm(formula = Ytest ~ Ypred, data = res)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26112 -0.62180  0.05058  0.59039  1.86660

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0008919  0.0867842    0.01    0.992
Ypred       0.6544171  0.1166565    5.61 1.91e-07 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.8635 on 97 degrees of freedom
Multiple R-squared:  0.245,        Adjusted R-squared:  0.2372
F-statistic: 31.47 on 1 and 97 DF,  p-value: 1.914e-07

Finished OmicKriging in 0.225 seconds
```

This function will return a data.frame with column Ypred corresponding to the predicted values and column Ytest corresponding to the measured phenotypes.

Congratulations!
You have just completed the OmicKriging tutorial!
