**iMODA: Integrative Multi-Omics Data Analysis**

**Overview of the iMODA package**

iMODA is a command-line program for integrative analysis of incomplete multi-omics data. Both continuous and binary phenotypes may be analyzed with iMODA. The current implementation fits the following pair of models for the continuous phenotype:

S=α’X + N(0,τ2)

Y=β’Z + γS + N(0,σ2)

For the binary phenotype, iMODA fits the following pair of models:

S=α’X + N(0,τ2)

logit{(P(Y=1)}=β’Z + γS

In the above models, S is the quantitative omics measurement, Y is the phenotype, and X and Z are sets of covariates (plus the constant 1). Typically, X includes the genotype G; Z may or may not include G. The regression parameters in α and β that are associated with G are denoted by αG and βG, respectively. The measurements on S are allowed to be missing and subject to detection limits (lower, upper, or both). iMODA performs maximum likelihood estimation of all the parameters via an EM algorithm.

**Compilation of the iMODA package**

A Linux executable named iMODA can be found in the iMODA directory. The program may also be compiled from the files in the source directory using the make utility. The default make target in makefile will result in the iMODA executable in the base directory and a number of object files in the src subdirectory. The executable and object files may be removed with the “make clean” command. The Eigen libraries and the Boost Math Toolkit are required to compile the source code.

**Program Usage**

iMODA is run by giving the name of a configuration file containing parameters after the executable name. There are example configuration files in the examples subdirectory. The examples/continuous subdirectory contains examples for the continuous phenotype while the examples/binary subdirectory contains examples for the binary phenotype. In the examples/continuous subdirectory, iMODA may be run as follows:

../../iMODA rand\_pheno\_1.txt.no\_snp.config.txt

…/../iMODA rand\_pheno\_1.txt.snp.config.txt

The geno\_filename line in these files gives the name of a file containing genotypes while pheno\_filename is used to give the name of a file containing values for the phenotype, the quantitative omics measurement S, and covariates. In these configuration files, geno\_filename is set to geno\_random\_new.txt and pheno\_filename is set to rand\_pheno\_1.txt. The phenotype and omics measurements are required while covariates are optional. If covariates are given they will be included in both X and Z. For the examples, there are three covariates which are given in the PC, Gender, and Age columns of rand\_pheno\_1.txt. The configuration files also contain a setting pheno\_covar\_snp, which is used to include or exclude G from Z. That is, if this parameter is set to true then G will be included, and if this is set to false then G will be excluded. The format of the configuration files and the other input file types is described in more detail in the “Input Files” section of this document.

The parameter estimates from the steps given above are found in the files rand\_pheno\_1.txt.no\_snp.out.txt and rand\_pheno\_1.txt.snp.out.txt in the examples/continuous subdirectory. There is one line in the output files for each SNP in the genotype file geno\_random\_new.txt. These files contain standard errors and p-values as well as estimates for the effect of genotype on the biomarker and for the effect of the biomarker on the phenotype. If pheno\_covar\_snp is set to true in the configuration file, then the effect of the genotype on the biomarker will also be included in the output file. The format of the output files is described in the “Output Parameter Files” section of this document. Additional parameters may be written to a file with the write\_covariance and write\_full\_estimate settings described below.

The program will produce messages to standard output which may be redirected to a file. The example files rand\_pheno\_1.txt.no\_snp.config.txt.log and rand\_pheno\_1.txt.snp.config.txt.log contain the standard output messages produced by the program runs listed above.

In addition to the two parameter estimate files produced by the calculations described above, there are other example parameter estimate files given in the examples/continuous subdirectory. These files all have names ending in “out.txt”. For example, the files rand\_pheno\_5.txt.no\_snp.out.txt and rand\_pheno\_5.txt.snp.out.txt were generated using rand\_pheno\_5.txt as the phenotype file and geno\_random\_new.txt as the genotype file, and similarly for the other estimate files given in the examples/continuous subdirectory.

The examples/binary subdirectory contains examples for the binary phenotype. These examples correspond closely to the examples given in the examples/continuous subdirectory. For example, in the examples/binary subdirectory iMODA may be run as follows:

../../iMODA rand\_pheno\_binary\_1.txt.no\_snp.config.txt

…/../iMODA rand\_pheno\_binary\_1.txt.snp.config.txt

In the first example, pheno\_covar\_snp is false while it is true in the second example. These examples will produce the parameter estimate files rand\_pheno\_binary\_1.txt.no\_snp.out.txt and

rand\_pheno\_binary\_1.txt.snp.out.txt, respectively. These output files have the same format as the files produced in the continuous phenotype examples. There are also other output files in the examples/binary subdirectory, such as rand\_pheno\_binary\_9.txt.snp.out.txt, which was produced using rand\_pheno\_binary\_9.txt as the input phenotype and omics data and geno\_random\_new.txt for genotypes.

If iMODA is run with no command line arguments, a short description of the configuration file format is printed and then the program terminates.

If an error results from the calculations for a SNP, then that SNP will be skipped in all data output files, an informational message will be produced, and execution will continue with the next SNP in the genotype file.

**Input Files**

Each line of the configuration file should have the form

keyword value

where “keyword” is one of several strings indicating a program parameter and “value” is a setting for that parameter. The keyword and value may be separated by any combination of tabs and spaces at least one character long.

The files named rand\_pheno\_1.txt.no\_snp.config.txt and rand\_pheno\_1.txt.snp.config.txt in the examples/continuous subdirectory are example configuration files. There are also configuration files in the examples/binary subdirectory named rand\_pheno\_binary\_1.txt.no\_snp.config.txt and rand\_pheno\_binary\_1.txt.snp.config.txt.

The keywords and values that may be given in the configuration file are the following:

**geno\_filename**

The required geno\_filename value should be the name of a text file containing genotype data for all subjects. Each line of the file should contain the name of a SNP followed by a string of whitespace delimited values. The following values are allowed for the SNP data: 0, 1, or 2 for subjects with known genotype information for the SNP and the string “NA” for subjects missing genotype data. If a subject has missing value for a SNP, then the value used in the calculations will be the arithmetic mean of the values with data for that SNP. Note that the treatment of data with missing SNP data can be modified by the min\_call\_rate parameter described below.

The file geno\_random\_new.txt in the examples/continuous subdirectory is a sample genotype file.

**pheno\_filename**

The pheno\_filename parameter is required, and it should be set to the name of a text file containing the phenotype, quantitative omics measurements S, and covariate information. This file should contain at least three columns. The file must include a header line. All other lines correspond to a given subject and the subjects should be listed in the same order as in the genotype data file. The first column is a floating-point value for the phenotype which is required. Any subject with missing phenotype data, which is represented by the string “NA” in the first column, will be excluded from the calculations. For the binary phenotype, this field should contain a value of 0 or 1 for all subjects with phenotype data. The second column, which is also required, is a floating-point number that may be a quantitative omics measurement, a detection limit, or a dummy value (which may be set to any floating-point value without affecting the results). The third column is also required and is an integer indicating the type of the S measurement associated with that subject. The following table gives the relationships between the type of S measurement and the second and third columns of the phenotype file.

|  |  |  |
| --- | --- | --- |
| S Measurement Type | Second Column of File | Third Column of File |
| Known | S Value | 0 |
| Below Detection Limit | Detection Limit | 1 |
| Above Detection Limit | Detection Limit | 2 |
| Missing | Dummy Value | 3 |

The remaining columns of a phenotype file should be any covariates that are to be used and are optional. If present in this file, both X and Z will include these covariates. The phenotype file is assumed to have a first row giving column names. The column names are unused.

The phenotype is automatically determined to be either continuous or binary according to the actual values given in the phenotype file. The program will write a message to standard output indicating which type of phenotype was detected.

The files rand\_pheno\_1.txt, rand\_pheno\_2.txt, etc. in the examples/continuous subdirectory are sample phenotype files. Each of these files contains three covariates which are given in the PC, Gender, and Age columns. There are sample files for the binary phenotype in the examples/binary subdirectory named rand\_pheno\_binary\_1.txt, rand\_pheno\_binary\_2.txt, etc. Each of these files also has three covariates which are given in the PC, Gender, and Age columns.

**pheno\_covar\_snp**

This required string parameter should be set to either “true” or “false”. If this parameter is set to true, then G will be included in Z. If it is set to false, then G will not be included in Z.

**min\_call\_rate**

This is an optional parameter with a default value of 0.9. If given, min\_call\_rate must be set to a positive floating point value less than or equal to 1.0. The default value of 0.9 indicates that for a given SNP to be considered in the calculations, at least 90% of the subjects must have a non-missing value for that SNP.

**out\_freq**

The out\_freq parameter is optional. If present, it must be set to a positive integer. If this parameter is set to 10, then there will be an output message every tenth SNP. This parameter may be useful to monitor the calculations.

**out\_suffix**

The name of the output data file containing the results of several estimates (including p-values and standard errors) is formed by appending the optional out\_suffix parameter to the phenotype input file name. If this parameter is not given in the configuration file, it is automatically set to a default value (“.out.txt”).

The output file that is produced by default does not contain all estimated parameters. The data that is written is described below under “Output Parameter Files”. To save more estimates the write\_covariance and write\_full\_estimate options described below should be used.

**write\_covariance**

This is an optional string parameter that should be set to either “true” or “false”. The default setting is false, which indicates that the covariance matrices for the SNPs will not be written to a file. If set to true, then the full covariance matrices for all SNPs will be saved to a file with a file name ending in “.covarmat.txt”. There are examples of this file type in the examples/continuous directory named rand\_pheno\_1.txt.no\_snp.out.txt.covar.txt and rand\_pheno\_1.txt.snp.out.txt.covarmat.txt . These files were generated using the settings given in rand\_pheno\_1.txt.no\_snp.config.txt and rand\_pheno\_1.txt.snp.config.txt respectively. There are also examples of this file type in the subdirectory examples/binary named rand\_pheno\_binary\_1.txt.no\_snp.out.txt.covarmat.txt and rand\_pheno\_binary\_1.txt.snp.out.txt.covarmat.txt which were produced using the configuration files rand\_pheno\_binary\_1.txt.no\_snp.config.txt and rand\_pheno\_binary\_1.txt.snp.config.txt respectively.

**write\_full\_estimate**

This is an optional string parameter that should be set to either “true” or “false”. The default setting is false. If set to true, then the estimates, standard errors, and p-values for all parameters will be saved to a file with a file name ending in “.fullest.txt”. There are examples of this file type in the examples/continuous directory named rand\_pheno\_1.txt.no\_snp.out.txt.fullest.txt and rand\_pheno\_1.txt.snp.out.txt.fullest.txt. These files were generated using the settings given in rand\_pheno\_1.txt.no\_snp.config.txt and rand\_pheno\_1.txt.snp.out.txt respectively. There are additional example files of this type in the examples/binary subdirectory named rand\_pheno\_binary\_1.txt.no\_snp.out.txt.fullest.txt and rand\_pheno\_binary\_1.txt.snp.out.txt.fullest.txt, which were produced using the configuration files rand\_pheno\_binary\_1.txt.no\_snp.config.txt and rand\_pheno\_binary\_1.txt.snp.config.txt , respectively.

**Output Parameter Files**

iMODA will create an output file containing estimated parameters. The output file will have a name combining the name of the phenotype file and the out\_suffix parameter. For example, for a phenotype file named pheno\_example\_1.txt, the corresponding output file will be named pheno\_example\_1.txt.out.txt, assuming that the default value for out\_suffix is used.

The first line of the output file is a header line. The column labeled SNP gives the name of each SNP used in the calculations. The columns labeled alpha\_G, alpha\_G\_se, and alpha\_G\_pval give the estimate, standard error, and p-value, respectively, for the effect of the genotype on the biomarker, i.e., αG. The gamma, gamma\_se, and gamma\_pval columns are the estimate, standard error, and p-value for the biomarker effect on the phenotype, i.e.,γ. Finally, the beta\_G, beta\_G\_se, and beta\_G\_pval columns are present only when pheno\_covar\_snp is set to true and are the estimate, standard error, and p-value for the genotype effect on the phenotype, i.e., βG .

Example output files for parameter estimates are given in the examples/continuous subdirectory, such as the files rand\_pheno\_1.txt.no\_snp.out.txt and rand\_pheno\_1.txt.snp.out.txt, which were generated using the rand\_pheno\_1.txt.no\_snp.config.txt and rand\_pheno\_1.txt.snp.config.txt configuration files, respectively. The other files in the examples/continuous subdirectory ending in “out.txt” are further examples of parameter estimate output files. There are also example parameter output files for the binary phenotype in the examples/binary subdirectory named rand\_pheno\_binary\_1.txt.no\_snp.out.txt and rand\_pheno\_binary\_1.txt.snp.config.txt, which were generated using the configuration files rand\_pheno\_binary\_1.txt.no\_snp.config.txt and rand\_pheno\_binary\_1.txt.snp.config.txt, respectively.

If the write\_full\_estimate option is set to true in the configuration file, then estimates, standard errors, and p-values for all parameters will be written to a file with a name combining the phenotype file name, the out\_suffix parameter setting, and the extension “.fullest.txt”. For example, if the phenotype input file is named pheno\_example\_1.txt and the default out\_suffix setting is used, then the full estimates will be written to pheno\_example\_1.txt.out.txt.fullest.txt. The examples/continuous directory contains the files rand\_pheno\_1.txt.no\_snp.out.txt.fullest.txt and rand\_pheno\_1.txt.snp.out.txt.fullest.txt, which are examples of this file type. The examples/binary subdirectory contains files of this type for the binary phenotype. The file produced by this option will contain all the columns that are written to the parameter output file described in previous paragraphs, along with estimates, standard errors, and p-values for all other parameters as well.

Estimated covariances for all parameters may be produced by setting write\_covariance to true in the configuration file. The name given to the covariance matrix output file will be formed from the input phenotype file name, the out\_suffix parameter setting, and the extension “.covarmat.txt”. For example, if the phenotype input file is named pheno\_example\_1.txt and the default out\_suffix setting is used, then the full estimates will be written to pheno\_example\_1.txt.out.txt.covarmat.txt. The examples/continuous subdirectory contains the files rand\_pheno\_1.txt.no\_snp.out.txt.covarmat.txt and rand\_pheno\_1.txt.snp.out.txt.covarmat.txt, which are examples of this file type. Examples for the binary phenotype are given in the examples/binary subdirectory. There will be one covariance matrix given in a covariance matrix file for each SNP in the genotype.