DEPTH: A Novel Algorithm for Feature Ranking with Application to Genome-Wide Association Studies

Enes Makalic, Daniel F. Schmidt, and John L. Hopper

The University of Melbourne
Centre for MEGA Epidemiology
Carlton VIC 3053, Australia
{emakalic,dschmidt,johnlh}@unimelb.edu.au

Abstract. Variable selection is a common problem in regression modelling with a myriad of applications. This paper proposes a new feature ranking algorithm (DEPTH) for variable selection in parametric regression based on permutation statistics and stability selection. DEPTH is: (i) applicable to any parametric regression task, (ii) designed to be run in a parallel environment, and (iii) adapts naturally to the correlation structure of the predictors. DEPTH was applied to a genome-wide association study of breast cancer and found evidence that there are variants in a pathway of candidate genes that are associated with a common subtype of breast cancer, a finding which would not have been discovered by conventional analyses.

1 Introduction

The problem of selecting predictor variables from a possibly large set of candidate variables occurs in many areas of science. An important recent example is the parametric regression model of genome-wide association studies (GWAS). GWA studies [1] measure thousands of genetic markers, typically single nucleotide polymorphisms (SNPs), for people affected by the disease of interest (cases) and people that are not affected by the disease (controls). The aim of a GWAS is to identify which SNPs, if any, are truly associated with risk of disease.

In the context of a GWAS, selecting potentially interesting variables is challenging due to: (i) the large number of SNPs measured, (ii) the correlation between SNPs, and (iii) the fact that the disease causing variants may not have been measured. The conventional strategy for finding disease associated SNPs is to test each SNP independently of all other measured SNPs using standard hypothesis testing methods. This approach yields a frequentist p-value for each measured SNP which is an indication of the strength of evidence for the association. The p-values are then adjusted for multiple testing [2] using, for example, the Bonferonni procedure and all SNPs whose p-values are less than some pre-specified threshold are in effect considered to be true associations; all the remaining SNPs are effectively discarded.

This paper proposes a new algorithm (see Section 2) for discovering predictors using a regression model based on permutation statistics and stability selection. The basic idea is to rank each variable in terms of evidence for association and then measure the stability of the corresponding ranking by re-sampling the data. Intuitively, one expects the ranking of variables with little or no associations to be highly unstable under minor perturbations of the data. This is because their associations are essentially random and practically indistinguishable. In comparison, the ranking of stronger predictors should remain relatively stable under data permutation, even when there are groups of predictor variables that are highly correlated. The algorithm also adds random noise predictors and ranks these variables alongside the measured variables. Statistics computed for the noise variables correspond to an empirical null distribution and this is used to determine the relative importance of all variables as predictors.

In the context of a GWAS, the algorithm can be used for all SNPs in the genome or for any subset of SNPs. We have found that the algorithm shows good performance when compared to several established procedures using simulated data. When applied to a breast cancer GWAS data set, the proposed algorithm found evidence that there are variants in a pathway of candidate genes that are associated with a common subtype of breast cancer, a finding which would not have been discovered by conventional analyses (see Section 3).

2 Stability Selection Algorithm

Consider a data set $D = \{(\mathbf{x}_1, \mathbf{z}_1, y_1), (\mathbf{x}_2, \mathbf{z}_2, y_2), \ldots, (\mathbf{x}_n, \mathbf{z}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{z}_i \in \mathbb{R}^q$ and $y_i \in \mathbb{R}$ $(i = 1, 2, \ldots, n)$, assumed to be generated by the regression model

$$y_i = f_i(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}) + \varepsilon_i \tag{1}$$

parameterised by $\boldsymbol{\theta} \in \mathbb{R}^k$ with disturbances $\varepsilon_i \sim \pi(\cdot)$. This setup includes common linear as well as non-linear classification and regression models. The task is to rank the p regressor variables $\mathbf{x}$ in terms of the evidence for their strength of association with the target variable $\mathbf{y}$ and thus effectively select which of the p variables constitute signal and which variables are noise. Note that the q variables $\mathbf{z}$ are pre-selected and included in each candidate model.

This paper introduces a novel feature ranking algorithm for parametric regression called `DEPTH` (DEPendency of Association on the number of Top Hits) which is based on re-sampling techniques and stability selection. Briefly, the idea is to first rank all p variables based on their marginal contribution, adjusting for the fixed regressors $\mathbf{z}$, and then evaluate the stability of the corresponding ranking by re-sampling the data. The re-sampling is without replacement and repeated over many iterations. Statistics recorded during each iteration of sampling are then used to automatically select the predictors that are associated with the target. A detailed description of `DEPTH` is given in Algorithm 1.

Algorithm 1 A description of the DEPTH algorithm

Require: Data $D = \{(\mathbf{x}_1, \mathbf{z}_1, y_1), (\mathbf{x}_2, \mathbf{z}_2, y_2), \ldots, (\mathbf{x}_n, \mathbf{z}_n, y_n)\}$, $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{z}_i \in \mathbb{R}^q$, $y_i \in \mathbb{R}$, number of iterations $T > 0$

1: $\mathbf{y}^p \leftarrow$ random permutation of target variable $\mathbf{y} \in \mathbb{R}^n$
2: $\mathbf{r}_0 \leftarrow$ initial ranking of variables using data D {see Algorithm 2}
3: $\mathbf{r}_0^p \leftarrow$ initial ranking of variables using D with $\mathbf{y}^p$ instead of $\mathbf{y}$ {see Algorithm 2}
4: $\mathbf{c} \leftarrow \mathbf{0}_p$, $\mathbf{c}^p \leftarrow \mathbf{0}_p$ {measure of variable significance}
5: **for** $t = 1$ to T **do**
6: $D_* \leftarrow$ re-sample data D without replacement
7: $D_*^p \leftarrow D_*$ where the target vector is randomly permuted
8: Append an extra p columns of noisy variables to data D_* and D_*^p
9: $\mathbf{r}_t \leftarrow$ new ranking based on re-sampled data D_*
10: $\mathbf{r}_t^p \leftarrow$ new ranking based on permuted data D_*^p
11: $c_j \leftarrow c_j + 1$, $\forall \mathbf{x}$ variables in D_* ranked before the best ranked noise variable
12: $c_j^p \leftarrow c_j^p + 1$, $\forall \mathbf{x}$ variables in D_*^p ranked before the best ranked noise variable
13: $\mathbf{o}_t \leftarrow$ ranking overlap between $\mathbf{r}_0$ and $\mathbf{r}_t$ {see text}
14: $\mathbf{o}_t^p \leftarrow$ ranking overlap between $\mathbf{r}_0^p$ and $\mathbf{r}_t^p$
15: **end for**
16: Ranking stability plot based on $\{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$ and $\{\mathbf{o}_1^p, \ldots, \mathbf{o}_T^p\}$
17: $\mathbf{R} \leftarrow$ index of $\mathbf{x}$ variables obtained by sorting $\mathbf{c}$ in descending order
18: **return** final ranking of the p $\mathbf{x}$ variables $\mathbf{R}$

2.1 DEPTH Algorithm

`DEPTH` first creates a copy of the data set $\mathbf{D}$ where the target vector $\mathbf{y}$ is randomly permuted (Step 1). The new data set is denoted by D^p and corresponds to an empirical null distribution. Since the target vector in D^p is essentially random, all `DEPTH` statistics will be compared to the corresponding statistics obtained using this random data. `DEPTH` then calculates statistics from the data which are compared to the empirical null distribution to minimise false positive findings.

`DEPTH` ranks the p variables $\mathbf{x}$ in D and the $\mathbf{x}$ variables in D^p (Steps 2–3). `DEPTH` employs marginal variable ranking where each variable is examined independently of all other variables to be ranked. The ranking algorithm is described in Algorithm 2. For each of the p variables (Steps 3–7, Algorithm 2), the ranking function fits a regression model using one $\mathbf{x}$ variable at a time (Step 4, Algorithm 2), adjusting for all the $\mathbf{z}$ variables, and computes the corresponding log-likelihood (Step 5, Algorithm 2). The regression model is fitted using maximum likelihood estimation, though in principle another estimation technique can be used. Following maximum likelihood fitting, a ranking statistic is computed for each of the p $\mathbf{x}$ variables (Step 6, Algorithm 2). The ranking statistic is the difference between the log-likelihood of a model with one regressor x_j (and q regressors $\mathbf{z}$) and the log-likelihood of a model with only the q regressors $\mathbf{z}$. All p $\mathbf{x}$ variables are then ranked in ascending order of the ranking statistic (Step 8, Algorithm 2); the regressor x_j that results in the best improvement to the log-likelihood over the model with $\mathbf{z}$ regressors only is ranked first; the second

Algorithm 2 A description of the marginal feature ranking function

Require: Data $D_* = \{(\mathbf{x}_1, \mathbf{z}_1, y_1), (\mathbf{x}_2, \mathbf{z}_2, y_2), \ldots, (\mathbf{x}_{n_*}, \mathbf{z}_{n_*}, y_{n_*})\}$, $\mathbf{x}_i \in \mathbb{R}^{p_*}$, $\mathbf{z}_i \in \mathbb{R}^q$, $y_i \in \mathbb{R}$

1: Initialise score vector $\mathbf{s} = (s_1, s_2, \ldots, s_{p_*})' = \mathbf{0}_{p_*}$
2: $s_0 \leftarrow$ log-likelihood of model with regressors $\mathbf{z}_i$ $(i = 1, 2, \ldots, n_*)$
3: **for** $j = 1$ to p_* **do**
4: Fit regression model using data $(x_{ij}, \mathbf{z}_i, y_i)$ $(i = 1, 2, \ldots, n_*)$
5: $l_j \leftarrow$ log-likelihood of the fitted model
6: $s_j \leftarrow l_j - s_0$ {difference in log-likelihood}
7: **end for**
8: $\mathbf{r} \leftarrow$ sort all variables in descending order of $\mathbf{s}$
9: **return r** {ranked list of all p_* variables}

best regressor $x_{j'}$ $(j \neq j')$ is the one that results in the second best improvement in log-likelihood, etc.

`DEPTH` performs $T > 0$ steps of data re-sampling and re-ranking in order to assess variable selection stability (Steps 5–15). `DEPTH` keeps two count vectors $\{\mathbf{c} \in \mathbb{R}^p, \mathbf{c}^p \in \mathbb{R}^p\}$ for each of the p variables in D and D^p which are used to measure variable importance. During each sampling iteration, `DEPTH` creates a new data sample D_* by sampling the original data without replacement (Step 6); the new data contains 66% of the original data points. In a similar fashion to Step 1, a copy of D_* with the target vector randomly permuted is stored in D_*^p (Step 7). Additional p columns of (noisy) data are then appended to D_* and D_*^p (Step 8). These extra columns will be used to determine the total number of significant $\mathbf{x}$ variables (a similar idea was mentioned in passing by Miller [3]). The data sets D_* and D_*^p therefore have $n_* = \lfloor 2n/3 \rfloor$ samples and $p_* = 2p$ variables that will be used for ranking. Algorithm 2 is used to compute a new ranking list r_t for the p_* variables in D_* (Step 9); the same procedure is also applied to the permuted data D_*^p resulting in $\mathbf{r}_t^p$ (Step 10). The count vector $\mathbf{c}$ is updated in Step 11 as follows: for each $\mathbf{x}$ variable j $(1 \leq j \leq p)$, if the variable ranks ahead of the best ranked noisy variable, add one to the count c_j, otherwise proceed to step 12. The counts $\mathbf{c}^p$ are updated in the same fashion using ranking list $\mathbf{r}_t^p$ instead of $\mathbf{r}_t$ (Step 13).

`DEPTH` uses the percentage of overlap between ranking vectors r_0 (based on data D) and r_t (based on re-sampled data D_*) as a metric of variable selection stability. Further, the percentage of overlap between $\mathbf{r}_0^p$ and $\mathbf{r}_t^p$ is used to estimate the empirical null distribution of overlap. The procedure to compute percentage overlap between two lists of ranks, say $\mathbf{r}_0$ and $\mathbf{r}_t$, is as follows (Steps 13–14). First, the p extra (noisy) variables that were added in Step 8 are removed. We then compute the intersection between the first j components of r_0 and r_t for all $(j = 1, 2, \ldots, p)$. The number of variables that remain in the intersection set for each j is stored in $\mathbf{o}_t$ and is a metric of ranking repeatability. As an example, consider two lists of rankings $\{3, 2, 5, 4, 1\}$ and {2,3,4,1,5}. Following the calculation of overlap percentage, the vector $\mathbf{o}_t$ for these two lists would be $\{0, 1, 0.67, 0.75, 1\}$. The first entry of $\mathbf{o}_t$ is 0 as the top ranked variable is

different in the two lists. Similarly, the second entry of $\mathbf{o}_t$ is 1 since both lists rank variables $\{2, 3\}$ as most significant.

Following T re-sampling iterations, `DEPTH` produces: (i) an estimate of the number of significant predictors in the data γ $(0 \leq \gamma \leq p)$, (ii) a plot of variable selection stability (Step 16), and (3) a ranking of all $\mathbf{x}$ variables in terms of their strength of association with the target $\mathbf{y}$. The total number of significant predictors is estimated as

$$\gamma = \frac{1}{T} \sum_{j=1}^{p} \max(c_j - c_j^p, 0), \quad (0 \leq \gamma \leq p) \tag{2}$$

where γ is the mean number of $\mathbf{x}$ variables ranked below the noise variables over T sampling iterations. Note that the number of significant variables in the permuted data D_p^*, which is an estimate of the empirical null distribution under the assumption that all variables are noise, is subtracted from the estimate obtained using D_*. This ensures that γ controls the type I error rate by recording the number of significant variables above what would be expected by chance.

The ranking stability plot is obtained from the overlap vectors $\{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$ and $\{\mathbf{o}_1^p, \ldots, \mathbf{o}_T^p\}$. `DEPTH` overlays the median percentage overlap computed from data sets D_* and D_*^p in one figure. This gives the experimenter the ability to compare median replicability between the real data rankings and the random data rankings over T iterations of re-sampling. The area between the two median curves is a surrogate statistic for the amount of signal present in the data. If the area is small (that is, the curves are virtually overlapping), `DEPTH` may not able to distinguish between signal and noise variables. In contrast, a large area between two median curves may indicate presence of strong signal variables.

Finally, `DEPTH` produces a ranking of all $\mathbf{x}$ variables which is used in conjunction with the estimate of the number of significant predictors (2) for model selection. The `DEPTH` ranking is computed from the count vectors $\mathbf{c}$ and $\mathbf{c}^p$. The variable j with the largest count $(c_j - c_j^p)$ is ranked first; the variable j' with the second largest count $(c_{j'} - c_{j'}^p)$ is ranked second, etc. As the `DEPTH` ranking is an average over T re-sampling iterations, it is expected to be more stable than a single ranking.

3 Application to Breast Cancer GWAS Data

This section examines the application of `DEPTH` to a real GWAS of 204 women with breast cancer obtained from the Australian Breast Cancer Family Study [4] and 287 controls from the Australian Mammographic Density Twins and Sisters Study [5]. All women were genotyped using a Human610-Quad beadchip array resulting in over $600,000$ SNPs per woman. Recommended GWAS data cleaning and quality control procedures (e.g., checks for SNP missingness, duplicate relatedness, population outliers [6]) were performed prior to analysis. We selected SNPs for `DEPTH` analysis from genes encoding a candidate susceptibility pathway. All SNPs were validated in the Caucasian population and were downloaded from

the HapMap Consortium [7]. This particular pathway was chosen due to biological considerations and because previous GWAS research in the pathway has detected potentially interesting SNPs. The final data set consisted of 366 SNPs selected from a genomic region of approximately two million base pairs. The correlation structure was approximately block-diagonal where blocks of highly correlated SNPs are interspersed with regions of low correlation.

`DEPTH` ranking of the data was done using $T = 1,000$ re-sampling iterations and logistic regression for marginal variable ranking. The difference in area between two overlap curves showed the possible presence of multiple risk-associated SNPs. `DEPTH` selected five SNPs ($\gamma = 5.11$) as important. To examine whether there is any difference in SNP rankings across different types of breast cancer, we stratified the GWAS data into two groups. Breast cancer type was determined from the collected cancer pathology data and `DEPTH` ranking was then performed for all SNPs in the two subgroups. `DEPTH` showed that SNPs in the pathway are only associated with one common type of breast cancer and not the other. This is an important discovery which we are currently attempting to replicate using a large, independent breast cancer GWAS data set.

Due to time constraints, initial `DEPTH` tests have concentrated on subsets of the human genome, chosen by biological consideration. The `DEPTH` algorithm is now being implemented on the IBM BlueGene/Q supercomputer, a Victorian government initiative in partnership with the University of Melbourne and the IBM Research Collaboratory for Life Sciences. The BlueGene/Q comprises 4,096 compute nodes with 65,536 user processor cores in four racks. The authors have been granted a significant amount of compute time on the supercomputer and have been funded by the National Health and Medical Research Council to perform `DEPTH` analyses of GWAS data. A manuscript detailing `DEPTH` analyses of a large international breast cancer GWAS data set, obtained through the Breast Cancer Association Consortium, is currently in preparation.

References

1. Manolio, T.A.: Genomewide association studies and assessment of the risk of disease. The New England Journal of Medicine **363**(2) (2010) 166–176
2. Dudoit, S., Shaffer, J.P., Boldrick, J.C.: Multiple hypothesis testing in microarray experiments. Statistical Science **18**(1) (2003) 71–103
3. Miller, A.J.: Selection of subsets of regression variables. Journal of the Royal Statistical Society (Series A) **147**(3) (1984) 389–425
4. Dite, G., Jenkins, M., Southey, M., Hocking, J., Giles, G., McCredie, M., Venter, D., Hopper, J.: Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations. J Natl Cancer Inst **95** (2003) 448–457
5. Odefrey, F., Gurrin, L., Byrnes, G., Apicella, C., Dite, G.: Common genetic variants associated with breast cancer and mammographic density measures that predict disease. Cancer Research **70** (2010) 1449–1458
6. Weale, M.: Quality control for genome-wide association studies. Methods Mol Biol. **628** (2010) 341–372
7. Consortium, I.H.: A second generation human haplotype map of over 3.1 million snps. Nature **449** (2007) 851–861