Application of Belief Propagation to Genome-Wide Association Studies

Vittorio Perduca [1], **Grégory Nuel** [2]

1. *Lab. of Applied Maths (MAP5-CNRS 8145)*
   UPD, Sorbonne Paris Cité
   Paris, France
   vittorio.perduca@parisdescartes.fr
2. *Lab. of Probability (LPMA-CNRS 7599)*
   UPMC, Sorbonne Universités
   Paris, France
   gregory.nuel@math.cnrs.fr

ABSTRACT. *Case-control association studies test for the association between a binary trait and a number of genetic markers, varying from a single candidate to hundred of thousands of markers in modern Genome-Wide Association Studies (GWAS). The development of statistical tests for such associations is a classical domain in biostatistics. If the distribution of a single marginal test is known under the hypothesis of no association (eg chi-square distribution), the joint distribution of marginal tests is not known in general. This is why association statistics are usually controlled empirically through simulation where case-control status are randomly affected independently from the genotypes by permutation. In this communication, we suggest a new simple association statistic for which we derive the asymptotic joint distribution in the permutation framework. We also extend this results to H1 simulations using weighted permutations and constrained Markov chains in a Bayesian network framework. The new association statistic proves itself to be highly versatile (can work with very general covariates), as powerful as state-of-art association statistics, and with the noticeable advantage to be theoretically controlled under H0 and H1 using close or recursive formulas.*

KEYWORDS: *GWAS, statistical power, p-value, Bayesian network, forward-backward recursions*

–

1. Introduction

Case-control association studies test for the association between a binary trait and a number of genetic markers, varying from a single candidate to hundred of thousands of markers in modern GWAS. The development of statistical tests for such associations is a classical domain in biostatistics. Apart from a few simple situations, such as the chi-square test for a candidate gene, the distribution of the test statistics under the null H0 is not known, neither exactly nor asymptotically. p-values for either individual or joint multiple tests are then usually computed by permuting the case/control status, thus breaking all ties between status and markers. Similarly, statistical power is assessed by simulating data under the alternative H1 through more or less computationally intensive methods. In this communication, we report on 1) our recent contribution in the field of GWAS power assessment (work already published), and 2) on our recent developments about a new association test with normal (asymptotic) distributions under H0 and H1.

Power assessment via weighted permutations. Standard approches for Monte-Carlo estimation of GWAS power are based on either the simulation of very large datasets of genotypes and phenotypes (so to reach a sufficient number of cases) or the conditional sampling of genotypes given fixed case-control phenotypes (Su *et al.*, 2011). Both approaches are computationally expensive as they require the simulation of very large genotypic matrices.

In a recent paper (Perduca *et al.*, 2012), we introduced an exact linear algorithm, called `waffect`, for the weighted permutation of case/controls status. `waffect` makes is possible to simulate new phenotypic datasets such that a) the phenotypes are in accordance with the corresponding observed genotypes under the chosen model H1; b) the total number of cases is the same as in the observed dataset. Indeed, we speak of *weighted permutations* for the very reason that we shuffle the case/control status as in plain permutations (so to keep constant the number of cases) but in doing so we are also able to take into account individual probabilities to be a case. Our method is based on belief propagation recursions in a simple Bayesian network and showed to be dramatically faster than other methods.

New test of association. Following on this previous work, the present communication introduces a new versatile test statistic which has already proved its value in the context of familial association analysis (Lange *et al.*, 2004). In the simplest case of an individual marker (or covariate) we suggest to use the score

$$z = \sum_i y_i X_i \tag{1}$$

where y_i is the disease status of individual i (0 for controls and 1 for cases), and X_i is the marker information for individual i (eg number of rare alleles, genotypic encoding, etc). Association is then tested using the Wald statistic $S = \frac{(z-\mu)^2}{\sigma^2}$, where μ and σ^2 are the expected value and variance of S under H0. We extend this statistic to the situation where several markers (and/or covariates) are tested jointly, in which

case we consider the vector $\boldsymbol{z} = (z^1, \ldots, z^p)$ where for each covariate the score z^p is defined as above (see Section 3 for detailed notations).

Our main results are in three parts: a) we prove the asymptotic normality of $\boldsymbol{z}$ both under H0 and H1; b) we use a rigorous Hidden Markov Model formulation of (weighted) permutations of disease status in GWAS to derive exact computation of the expectation and the covariance matrix of $\boldsymbol{z}$ both under H0 (closed-formulas) and H1 (recursive formulas) ; and c) we show that S has chi-square density under H0 and is the weighted sum of non central chi-square random variables under H1. Moreover, we empirically show that our new statistic is at least as powerful as gold standards such as the chi-square and the likelihood-ratio tests for GWAS, and that it also provides a unique way to control the type-I error rate and compute power in the realistic context of permutation based H0 (uniform permutations) and H1 (weighted permutations).

The paper is organized accordingly as follows: in Section 2 we recall our algorithm for weighted permutations; in Section 3 we introduce the new statistics and state results about its distribution under H0 and H1; and in Section 4 we present a simulation study. Concluding remarks follow in Section 5.

2. Weighted permutations

Let n be the number of individuals in the study and n_1 the fixed number of cases. We denote π_i the probability that individual i is a case given his genotype: $\mathbb{P}(y_i = 1|\boldsymbol{X}_i) = \pi_i$. Let N_i, $i = 1, \ldots, n$, the random variable counting the total number of cases among individuals indexed by $\{1, \ldots, i\}$: $N_i = \sum_{j=1}^{i} y_j$, with the convention that $N_0 = 0$. Observe that $N_i = N_{i-1} + y_i$ and therefore $N_i \in \{0, \ldots, i\}$ for each i. The dependencies between $\boldsymbol{y}$ and $\boldsymbol{N}$ variables are depicted by the DAG in figure 1. Indeed, $\{\boldsymbol{y}, \boldsymbol{N}\}$ constitutes a very simple Bayesian Network whose underlying directed graph is similar the one of an Hidden Markov Model (the difference being that the chain is on the $\boldsymbol{N}$ variables and not on the $\boldsymbol{y}$ variables as in HMMs). The conditional dependencies encoded by the directed edges determine the following factorization of the joint distribution:

$$\mathbb{P}(\boldsymbol{y}, \boldsymbol{N}) = \mathbb{P}(y_1)\mathbb{P}(N_1|y_1) \prod_{i=2}^{n} \mathbb{P}(y_i)\mathbb{P}(N_i|N_{i-1}, y_i).$$

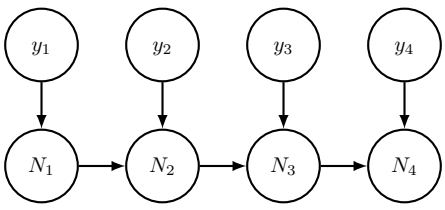

Figure 1. Bayesian network for Theorem 1

When all the weights π_i are given, we are interested in sampling the values of the y_is, given the condition that the total number of cases N_n must be equal to n_1, i.e. in sampling the distribution $\mathbb{P}(y_1, \ldots, y_n|\mathcal{E})$, where $\mathcal{E}$ is the *evidence* $\{N_n = n_1\}$. To achieve this goal we find recursive formulas for the probabilities $\mathbb{P}(y_i = 1|N_{i-1} = m, \mathcal{E})$, $i = 1, \ldots, n$:

THEOREM 1. — *For each individual* $i = 1, \ldots, n$*:*

$$\mathbb{P}(y_i = 1|N_{i-1} = m, \mathcal{E}) = \frac{\pi_i B_i(m+1)}{B_{i-1}(m)}, \tag{2}$$

where the backward quantities B_i *can be computed using the recursive formulas*

$$B_i(m) = \pi_{i+1} B_{i+1}(m+1) + (1 - \pi_{i+1}) B_{i+1}(m), \tag{3}$$

with the following edge conditions: $B_0(0) = \pi_1 B_1(1) + (1 - \pi_1) B_1(0)$ *and* $B_q(m) = \delta(m, r)$*,* δ *being the Kronecker's symbol.*

Note that these recursions can be seen as an application of the belief update algorithm for general Bayesian Networks. The theorem gives the recursive algorithm `waffect`, to sample in the space of all possible configurations of the y_is under the condition that the number of cases must be n_1 and knowing for each individual i his weight π_i.

Here is an outline of the algorithm: starting from $i = n$, simply compute all the backward quantities with Eq. (3). Then starting from the first individual $i = 1$, affect a status for the individual i accordingly to the binomial distribution which depends on the previous affectations

$$y_i|N_{i-1}, \mathcal{E} \sim \mathcal{B}\left(\frac{\omega_i B_i(N_{i-1}+1)}{B_{i-1}(N_{i-1})}\right).$$

Observe that if $\pi_i = \pi_0$ for each i, then `waffect` outputs a permutation of the phenotypes; this is equivalent to simulating under H0.

Our algorithm is implemented in the R package `waffect` available on CRAN. Interested users who want to learn about its main function are invited to read the tutorial which can be obtained with the command

```
>vignette('waffect-tutorial')
```

3. New association statistics

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the covariate matrix, $\boldsymbol{y} \in \{0, 1\}^n$ be the individual status, then we introduce the score $\boldsymbol{z} \in \mathbb{R}^{p \times 1}$ defined by:

$$\boldsymbol{z} = \boldsymbol{X}^T \boldsymbol{y}$$

The first purpose of this section is to obtain the asymptotic distribution of $\boldsymbol{z}$ both under the H0 hypothesis (permutations) and under the H1 hypothesis (weighted permutations).

Under H0

THEOREM 2. — *If we randomly affect n_1 cases among n patients, with n large enough, then*

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\mu} = \frac{n_1}{n}\boldsymbol{X}^T\boldsymbol{u} \quad \textit{and} \quad \boldsymbol{\Sigma} = \boldsymbol{X}^T\boldsymbol{N}\boldsymbol{X} - \boldsymbol{\mu}^T\boldsymbol{\mu}$$

where $\boldsymbol{u} = (1, \ldots, 1)^T$ is a size n vector and $\boldsymbol{N}$ is a order n matrix where off-diagonal terms are equal to $\frac{n_1(n_1-1)}{n(n-1)}$ and diagonal terms are equal to $\frac{n_1}{n}$.

Under H1

For all $i \in \{1, \ldots, n\}$ and for all $j \in \{1, \ldots, p\}$, we introduce the partial sum:

$$z_i^j = y_1 X_{1,j} + \ldots + y_i X_{i,j}$$

such as $\boldsymbol{z} = (z_n^1, \ldots, z_n^p)^T$. We then introduce the following forward quantities:

$$\begin{cases} F_i^0(m) = \sum_{y_1,\ldots,y_i} \mathbb{P}(y_1, \ldots, y_i, N_i = m) & \\ F_i^j(m) = \sum_{y_1,\ldots,y_i} z_i^j \mathbb{P}(y_1, \ldots, y_i, N_i = m) & \text{for any } j \in \{1, \ldots, p\} \\ F_i^{j,k}(m) = \sum_{Y_1,\ldots,Y_i} z_i^j z_i^k \mathbb{P}(y_1, \ldots, y_i, N_i = m) & \text{for any } j, k \in \{1, \ldots, p\} \end{cases}.$$

Having modeled the joint distribution of $\{\boldsymbol{y}, \boldsymbol{N}\}$ with the Bayesian Network in Figure 1, the forward quantities can be computed by belief propagation:

THEOREM 3. — *The forward quantities can be computed recursively with initialization $F_0^0(m) = \mathbf{1}_{m=0}$ and $F_0^j(m) = F_0^{j,k}(m) = 0$ and the following recursive formulas:*

$$\begin{cases} F_i^0(m) = F_{i-1}^0(m-1)\pi_i + F_{i-1}^0(m)(1-\pi_i) \\ F_i^j(m) = F_{i-1}^0(m-1)X_{i,j}\pi_i + F_{i-1}^j(m-1)\pi_i + F_{i-1}^j(m)(1-\pi_i) \\ F_i^j(m) = F_{i-1}^0(m-1)X_{i,j}X_{i,k}\pi_i + F_{i-1}^j(m-1)X_{i,k}\pi_i \\ \qquad\quad + F_{i-1}^k(m-1)X_{i,j}\pi_i + F_{i-1}^{j,k}(m-1)\pi_i + F_{i-1}^{j,k}(m)(1-\pi_i) \end{cases}.$$

COROLLARY 4. — *For any $j, k \in \{1, \ldots, p\}$ we have:*

$$\mathbb{E}\left[z_n^j\right] = \frac{F_n^j(n_1)}{F_n^0(n_1)} \quad \textit{and} \quad \mathbb{E}\left[z_n^j z_n^k\right] = \frac{F_n^{j,k}(n_1)}{F_n^0(n_1)}.$$

If n is large enough, then $\boldsymbol{z}$ has a multivariate gaussian distribution with expectation and covariance matrix which are computed using the corollary above.

The statistic

Testing the association between $\boldsymbol{y}$ and $\boldsymbol{X}$ turns out to be equivalent to testing the hypothesis H0: $E[\boldsymbol{z}] = \boldsymbol{\mu}$ against H1: $E[\boldsymbol{z}] \neq \boldsymbol{\mu}$. In order to do so we use the statistic

$$S = (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}).$$

As a direct consequence of Theorem 2, if n is large enough and H0 holds, then S has a χ^2 distribution with p degrees of freedom. This makes it possible to compute (asymptotic) p-values.

Under H1, the quadratic form S turns out to be a weighted sum of non central chi-square densities (Duchesne, De Micheaux, 2010) whose quantiles can for example be computed using the R package `CompQuadForm` . This remarkable property of S, makes it possible to do theoretical (rather than empirical) power studies.

4. Simulation study

Methods

We simulated a simple genotypic dataset consisting of $p = 5$ SNPs for $n = 1000$ individuals, with possible values in $\{0, 1, 2\}$. SNP 1 and SNP 4 were simulated according to Hardy-Weinberg equilibrium with MAF $= 0.2$ and 0.1 respectively. SNP 2 and SNP 5 were obtained from SNP 1 and SNP 4 respectively, by setting 150 randomly chosen positions to 1. Similarly, SNP 3 was obtained from SNP 2 by setting 100 randomly chosen positions to 0. As a result, we have two blocks of highly correlated SNPs as depicted in Figure 2.

We simulated $20,000$ phenotypic replicates under both H0 and H1 by using our package `waffect` which implements the algorithm described in section 2. Each replicate consisted of $n_1 = 500$ cases and 500 controls. For H1 simulations, we used SNP 3 as a causal disease marker with dominant effect. We assumed a disease prevalence of $f_0 = 0.01$ for non-carriers and a prevalence $f_1 = \text{RR} f_0$ with RR $= 1.2$ for carriers. H0 replicates were obtained considering no relative risk at all, that is by simply permuting case and control status.

We compared the test introduced in section 3, with two classic methods used in case-control studies: the Fisher exact test of association, and the significance test of logistic regression coefficients. For each method we computed the marginal association statistics for 5 SNPs. For our new statistic, we computed only the global association statistic S.

Control of type I errors and power study

We estimated the false positive rate by computing the proportion of H0 replicates with at least one significant p-value. Significance threshold was set to $\alpha/p =$

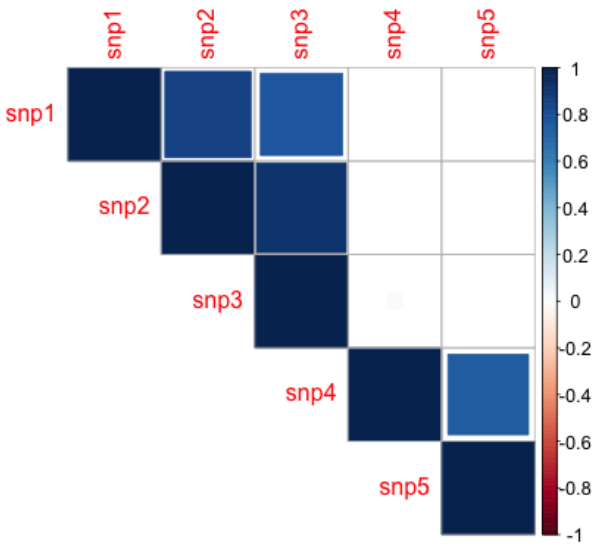

Figure 2. Correlation among simulated SNPs

$0.20/5 = 0.04$ (Bonferroni correction) for Fisher tests and logistic regressions, whereas it remained $\alpha = 0.20$ for our method, as it translates into an individual test. Results are depicted in Table 1.

As expected the Bonferroni correction resulted in a too conservative control for Fisher test and the logistic regression, a clear consequence of the correlation among SNPs. On the other hand, our new testing procedure clearly did not suffer from this problem.

We estimated the power by computing the proportion of H1 replicates with at least one significant p-value. We can see in Table 1 the result of the empirical power study. Thanks to its more accurate type-I error control, our new statistic achieves the highest power. One should however note that the overall AUC remains similar for all three methods which proves that the difference observed in power is due to inaccurate type I error control.

Table 1. Empirical type I error rate, power and AUC for a global test (5 SNPs) with $\alpha = 20\%$ and 90% confidence interval ($20,000$ replications)

Test used	Type I (%)	Power (%)	AUC (%)
Fisher	13.18 [12.80-13.58]	30.67 [30.14-31.20]	64.84 [64.39-65.29]
Logistic	12.54 [12.16-12.93]	34.48 [33.92-35.03]	66.96 [66.52-67.40]
New	20.05 [19.59-20.52]	38.61 [38.05-39.17]	64.07 [63.62-64.52]

5. Conclusion

In this communication, we suggest a new association statistics for GWAS and establish its theoretical distribution under unweighted or weighted permutations of the individual disease status. Under H0, this new statistic allows to account efficiently for multi-testing without risk of over-conservative adjustment in the context of correlated tests. Under H1, the distribution of our new statistic is more delicate to establish since it requires both modified forward recursions and a rather sophisticated algorithm to obtain the tail distribution of a quadratic form in normal variable. However, to the best of our knowledge, this method has the unique ability to provide theoretical power study for multiple SNPs in GWAS. Let us finally point out that our new statistic can easily cope with additional covariates (including continuous variables) and possible interaction effects.

Bibliographie

Duchesne P., De Micheaux P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, vol. 54, n° 4, p. 858–862.

Lange C., DeMeo D., Silverman E. K., Weiss S. T., Laird N. M. (2004). Pbat: tools for family-based association studies. *The American Journal of Human Genetics*, vol. 74, n° 2, p. 367–369.

Perduca V., Sinoquet C., Mourad R., Nuel G. (2012). Alternative methods for H1 simulations in genome-wide association studies. *Human Heredity*, vol. 73, n° 2, p. 95–104.

Su Z., Marchini J., Donnelly P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, vol. 27, n° 16, p. 2304–2305.