The Multivariate Normal Distribution Framework for Analyzing Association Studies

Jose A. Lozano[1,8], Farhad Hormozdiari[2,3], Jong Wha (Joanne) Joo[4], Buhm Han[5], Eleazar Eskin[6,7,*]

1 Intelligent Systems Group, University of the Basque Country UPV/EHU, Donostia Spain

2 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

3 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

4 Department of Computer Science, Computer Science Department Dongguk University-Seoul Campus, Seoul, Republic of Korea

5 Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

6 Department of Computer Science, University of California, Los Angeles, California 90095, USA

7 Department of Human Genetics, University of California, Los Angeles, California 90095, USA

8 Basque Center for Applied Mathematics (BCAM), Bilbao, Spain

* E-mail: eeskin@cs.ucla.edu

Abstract

Genome-wide association studies (GWAS) have discovered thousands of variants involved in common human diseases. In these studies, frequencies of genetic variants are compared between a cohort of individuals with a disease (cases) and a cohort of healthy individuals (controls). Any variant that has a significantly different frequency between the two cohorts is considered an associated variant. A challenge in the analysis of GWAS studies is the fact that human population history causes nearby genetic variants in the genome to be correlated with each other. In this review, we demonstrate how to utilize the multivariate normal (MVN) distribution to explicitly take into account the correlation between genetic variants in a comprehensive framework for analysis of GWAS. We show how the MVN framework can be applied to perform association testing, correct for multiple hypothesis testing, estimate statistical power, and perform fine mapping and imputation.

1 Introduction

In the last decade, genome-wide association studies (GWAS) have discovered thousands of common variants implicated in genetic diseases [31]. Technological developments in microarray and sequencing technologies fueled these discoveries, which paved the way for cost-effective collection of genetic information in large amounts [32, 10]. Specifically, these technologies enabled the collection

of genetic information at the scale of half a million variants spread throughout the genome. Further, the affordable cost of these technologies made feasible the study of thousands of individuals simultaneously. The initial GWAS studies [42] established essential groundwork for subsequent larger studies, which have identified the majority of today's known common variants implicated in diseases [43].

While collecting genetic information on half a million variants is a technical marvel in itself, the actual amount of common variants present in the human genome is an order of magnitude larger. Fortunately, the correlation structure between genetic variants, referred to as "linkage-disequilibrium" (LD) in the genetics literature [36], makes the half million variants sufficient for GWAS [5]. The first large scale maps of human genetic variation [39, 13] partly aimed to identify this correlation structure. In fact, foundational literature enabling GWAS focused on identifying approaches to select the subset of variants that should be collected in a GWAS [5]. Even if a disease-causing variant is not collected in a GWAS, the locus (region of the genome) would be identified as associated given that a correlated variant is collected, which is referred to as a tag.

However, there are two sides of the coin with respect to LD. While the correlation structure of the human genome enabled important GWAS discoveries, the same correlation structure complicates GWAS analyses. Hypothesis tests of association at each variant are not independent due to this structure. Implications of LD include complicating multiple testing, complicating estimating statistical power, and introducing ambiguities to interpretation of association study results. This complication will only be exacerbated with the advance of next generation sequencing, which will enable future studies to collect virtually all of the genetic variants in the genome that are tightly correlated [38].

In this review, we describe a comprehensive approach to analyzing GWAS that uses the multivariate normal (MVN) distribution to model correlations in the genome. This approach offers an advantageous ability to model the effect of LD on all of the statistics simultaneously. The approach presented here provides a framework encompassing many different types of analyses related to GWAS, including multiple testing correction [3, 11, 19], estimation of statistical power [11], statistical fine mapping [16, 15, 24], and imputation [26, 34, 44] while taking into account the LD structure of the human genome.

2 GWAS at One SNP and Hypothesis Testing

2.1 Association Testing for Case/Control Studies

We first consider GWAS with case/control study design where information on genetic variants is collected from a dataset containing individuals with the disease (cases) and healthy individuals (controls). In this case, a hypothesis test is performed for each collected variant. This hypothesis test compares the frequency of a variant between the cases and controls in order to identify associated variants.

Here, we consider a GWAS study with a total of n individuals that are genotyped at m SNPs. In order to simplify notation, we assume a balanced case-control study where we have $\frac{n}{2}$ cases and $\frac{n}{2}$ controls. Since each individual has two chromosomes, we have a total of n case chromosomes and n control chromosomes.

For each group and each SNP, we count the number of times that the minor allele appears and calculate the corresponding frequencies. Let $\hat{p}_i^+$ and $\hat{p}_i^-$ be the observed case and control frequencies, respectively, of SNP i. The true frequencies will be denoted as p_i^+ and p_i^-. Assuming that n is large enough, the observed frequencies follow a Gaussian distribution $\hat{p}_i^+ \sim \mathcal{N}(p_i^+,\ p_i^+(1-p_i^+)/n)$ and $\hat{p}_i^- \sim \mathcal{N}(p_i^-,\ p_i^-(1-p_i^-)/n)$ where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance

σ^2. We can then convert the observed frequencies into the following statistic

$$s_i = \frac{(\hat{p}_i^+ - \hat{p}_i^-)}{\sqrt{2/n}\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{1}$$

where $\hat{p}_i = (\hat{p}_i^+ + \hat{p}_i^-)/2$. The statistic will follow the normal distribution

$$s_i \sim \mathcal{N}\left(\frac{(p_i^+ - p_i^-)}{\sqrt{2/n}\sqrt{p_i(1-p_i)}}, 1\right) \tag{2}$$

We denote the mean of this distribution as the non-centrality parameter $\lambda\sqrt{n}$ where

$$\frac{(p_i^+ - p_i^-)}{\sqrt{2p_i(1-p_i)}}\sqrt{n} = \lambda\sqrt{n}$$

The probability density function of the normal distribution at point x for mean zero and variance σ^2 is

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{3}$$

The statistic s_i takes into account the difference between the observed frequencies. When this difference is significantly high, we will assume an association between the SNP i and the disease. In GWAS, this is done under the framework of hypothesis testing. This framework allows us to control for Type I errors or quantify the error we can commit while implicating SNPs.

2.2 Association Testing for Continuous Phenotypes

The same framework can also be applied to continuous phenotypes such as cholesterol levels. We assume that our genetic study collects n individuals and the continuous phenotype of individual j is denoted as y_j. We assume that the study collects m variants. We denote the frequency of variant i in the population as p_i. We denote the genotype of the ith variant in the jth individual as $g_{ij} \in \{0,1,2\}$, which encodes the number of minor alleles for that variant present in the individual. In order to simplify the formulas and without loss of generality, we standardize the genotype values such that $x_{ij} \equiv \frac{g_{ij}-2p_i}{\sqrt{2p_i(1-p_i)}} \in \{\frac{-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{1-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{2-2p_i}{\sqrt{2p_i(1-p_i)}}\}$ since the mean and variance of the column vector of genotype (g_i) is $2p_i$ and $2p_i(1-p_i)$, respectively. Due to the standardization, the sample mean and sample variance of the vector of genotypes at a specific variant i denoted as X_i are 0 and 1, respectively.

For the association at SNP i, the following model for the effect of SNP i on the phenotype is utilized

$$y_j = \mu + \beta_i x_{ij} + \epsilon_j \tag{4}$$

where μ is the population mean of the phenotype, β_i is the effect size of the SNP and $\epsilon_j \sim \mathcal{N}(0,\sigma_e^2)$ is the contribution of the environment to the phenotype for individual j. σ_e^2 is referred to as the environmental variance. In vector notation, this model is

$$Y = \mu\mathbf{1} + \beta_i X_i + \mathbf{e} \tag{5}$$

where X_i is a column vector of standardized genotypes for variant i and $\mathbf{e} \sim \mathcal{N}(0,\sigma_e^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of dimension n.

Using equation (5), we can obtain an estimate of β_i with the observed data. This reduces the equation to a simple regression problem where the resulting estimates are $\hat{\mu} = \frac{1}{n}\mathbf{1}^T Y$, $\hat{\beta}_i =$

$(X_i^T X_i)^{-1} X_i^T Y = \frac{X_i^T Y}{n}$ since X_i is standardized so $X_i^T X_i = n$. The estimated residuals $\hat{\mathbf{e}} = Y - \hat{\mu}\mathbf{1} - \hat{\beta}_i X_i$ can be used to estimate the standard error $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T \hat{\mathbf{e}}}{n-2}}$. Since these studies are often quite large, the association statistic will approximately follow a normal distribution such that

$$s_i = \frac{\hat{\beta}_i}{\hat{\sigma}} \sqrt{n} \sim \mathcal{N}\left(\frac{\beta_i}{\sigma_e}\sqrt{n}, 1\right) \tag{6}$$

where the non-centrality parameter $\lambda\sqrt{n} = \frac{\beta_i}{\sigma_e}\sqrt{n}$.

2.3 Hypothesis Testing

If we assume that SNP i is not involved in the disease, referred to as the null hypothesis, then $p_i^+ = p_i^-$, and therefore s_i follows a standard normal distribution $\mathcal{N}(0, 1)$. Typically, a false positive rate α (called type I error) is determined in advance (common values are 0.05 or 0.001 for a single hypothesis and 5×10^{-8} for a GWAS). From α, a threshold is calculated using the inverse of the standard normal cumulative distribution function Φ^{-1}, i.e. $\theta_\alpha = \Phi^{-1}(1 - \alpha/2)$. A SNP i is declared as associated if $|s_i| > \theta_\alpha$. In this framework, α is the probability that s_i is in the tails of the standard Gaussian distribution under the assumption that the null hypothesis is true and the mean value of s_i is 0 (see Figure 1).

Often, a p-value is reported as the result of a statistical test. In this case, the p-value of a SNP i is the probability of observing s_i or a more extreme value assuming that SNP i is not associated. This value can be calculated using $p = 2(1 - \Phi^{-1}(s_i))$ in case of $s_i > 0$ or $p = 2\Phi^{-1}(s_i)$ in case $s_i < 0$. Comparing s_i with the threshold is equivalent to comparing the p-value with α. If $p < \alpha$ then we declare the SNP as associated.

2.4 Statistical Power

When performing association testing, the null hypothesis assumes that the SNP is not associated with the disease. However, our objective is to discover the SNPs that are involved in the disease. In order to discover a SNP involved in the disease, when we perform the statistical test on collected data, we must reject the null hypothesis and declare the association as significant. We are interested in computing the probability of rejecting the null hypothesis for the SNPs which actually affect the disease. Intuitively, this is a measure of how likely an association study will succeed in finding the true disease causing variants. This quantity is referred to as the statistical power and can be calculated using the hypothesis testing framework. The power is a function of the effect size, which measures the effect of the SNP on the disease and the significance threshold. If SNP i is associated then $p_i^+ - p_i^- \neq 0$ or $\beta_i \neq 0$ and s_i is normally distributed with mean $\lambda\sqrt{n}$ (the non-centrality parameter) which is non-zero, and unit variance.

In order to declare a SNP i as associated, it has to happen that $|s_i| > \theta_\alpha$, so to know the probability of detecting it we have to calculate $Pr(|s_i| > \theta_\alpha)$. However, in contrast with the case where SNP i was not associated and s_i followed a standard normal distribution, now s_i follows the Gaussian distribution $\mathcal{N}(\lambda\sqrt{n}, 1)$. The power is visualized as the green area shown in the graphics presented in Figure 1. The power is a function of α and the non-centrality parameter $\lambda\sqrt{n}$ and can be computed using the following formula

$$P(\alpha, \lambda\sqrt{n}) = \Phi\left(\Phi^{-1}(\alpha/2) - \lambda\sqrt{n}\right) + 1 - \Phi\left(-\Phi^{-1}(\alpha/2) - \lambda\sqrt{n}\right) \tag{7}$$

Implicitly, the power depends on factors such as the significance threshold, effect size, the minor allele frequency, and the number of individuals. Furthermore, a higher non-centrality parameter produces a higher power.

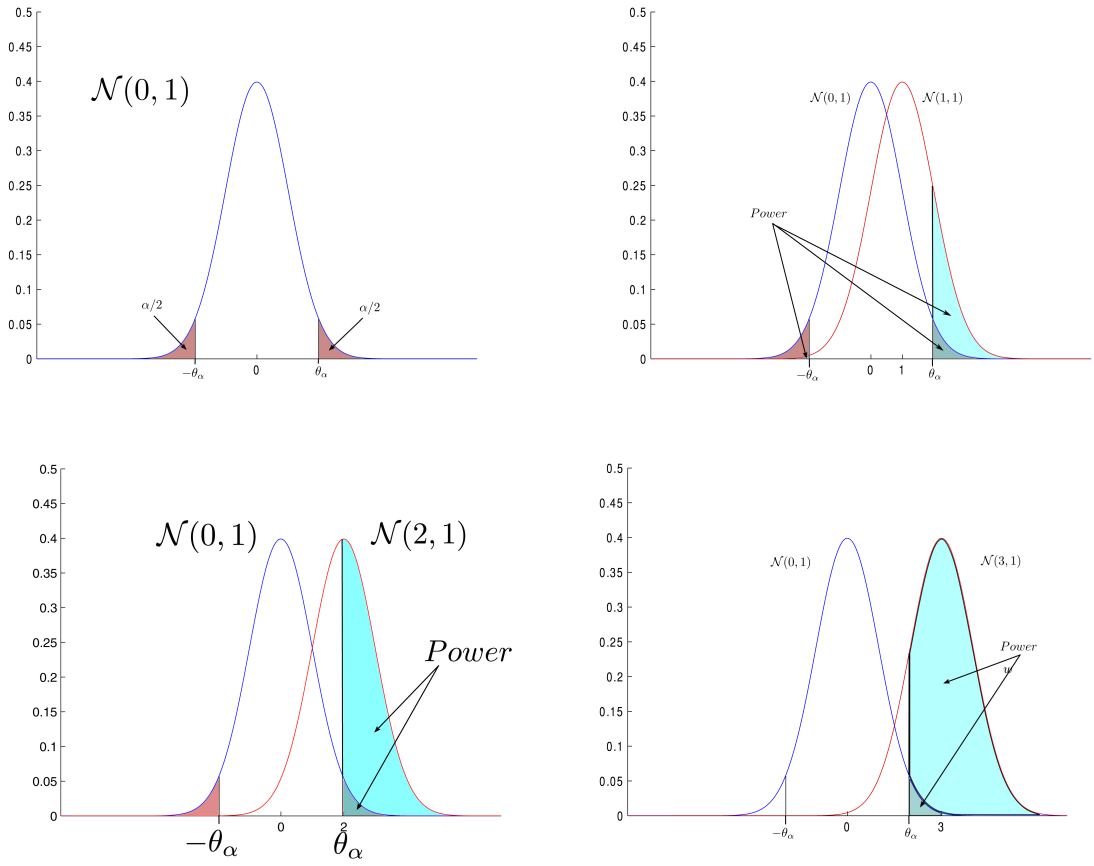


Figure 1: (a) A standard normal distribution with the reject region for an error α. (b) A $\mathcal{N}(1,1)$ for an associated SNP with $\lambda = 0.1$ and $n = 100$. (c) A $\mathcal{N}(2,1)$ for an associated SNP where $\lambda = 0.2$ and $n = 100$. (d) A $\mathcal{N}(3,1)$ for an associated SNP where $\lambda = 0.2$ and $n = 325$

3 GWAS for Multiple SNPs

3.1 Correlated SNPs– Multivariate Normal Distribution Model

In a region of the genome that is involved in a disease, some of the variants will have a direct effect on the disease. We refer to these variants as causal variants. However, due to the correlation between the variants, many more of the variants will be associated and have non-zero non-centrality parameters. Association studies perform an association test at each SNP. Resulting statistics are dependent due to the underlying correlation structure, referred to as Linkage Disequilibrium (LD), of the SNPs themselves. A natural measure of the correlation between two SNPs (i and j) is simply their correlation coefficient, which can be calculated as follows

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}$$

where p_{ij}, p_i, and p_j are the joint minor allele frequency of SNPs i and j, and the minor allele frequency of SNPs i and j, respectively.

Obviously, if two SNPs i and j are correlated then the probability distributions of the SNPs statistics, s_i and s_j, should also be related. In fact, the correlation coefficient plays a central role

in the joint distribution of the statistics of the m SNPs, $S^T = (s_1, s_2, \ldots, s_m)$, which follows a multivariate normal distribution (MVN) [11, 25, 34, 15, 14, 17, 19, 44, 18]

$$S \sim \mathcal{N}(\Lambda, \Sigma) \tag{8}$$

where $\Lambda = (\lambda_1\sqrt{n}, \lambda_2\sqrt{n}, \ldots, \lambda_m\sqrt{n})$ is the vector of non-centrality parameters and Σ is the variance-covariance matrix with $\sigma_{ii}^2 = 1$ and $\sigma_{ij} = r_{ij}$ for all $i \neq j$. Σ is referred to as the LD matrix. The probability density function of the MVN at point X for mean vector μ and variance-covariance matrix Σ is

$$f(X; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right) \tag{9}$$

A derivation of the covariance is provided in Appendix A.

The non-centrality parameters for the SNPs depend on which SNPs are causal, the SNP effect sizes, and the correlation between the SNPs. If we consider two SNPs (i and j) where SNP j is a causal SNP with non-centrality parameter $\lambda^c\sqrt{n}$, the non-centrality parameter of another SNP statistic i, $\lambda_i\sqrt{n}$ is

$$\lambda_i\sqrt{n} = r_{ij}\lambda^c\sqrt{n} \tag{10}$$

The previous equality has important implications for GWAS. Given that SNP j is a causal SNP, then $\lambda^c\sqrt{n} \neq 0$ and, therefore, for each SNP i correlated with SNP j its non-centrality parameter $\lambda_i\sqrt{n} = r_{ij}\lambda_j\sqrt{n}$ is also non-zero. This means that a GWAS study has a high probability of discovering both causal SNPs and highly correlated SNPs among the identified associated SNPs. Thus, not all SNPs must be collected in a GWAS for the purpose of identifying an associated region; only a subset of the SNPs (referred to as tag SNPs) need to be collected that are correlated with the remaining uncollected SNPs.

When we consider more than one SNP at a time in order to identify responsibility for the association, a distinction arises between the effect of the variants and the variants themselves. Observed effects of these variants can simply be due to the correlation between the effects and the causal variants. This distinction parallels the distinction between direct and indirect effects in the causal inference literature [35] and has been discussed at length in the genetics literature [36]. We use the notation λ_j^c and the term effect size to denote the actual causal effect of SNP j. We note that the correlated SNP i has a non-centrality parameter $r_{ij}\lambda_j^c$ due to the correlation. We also note that if SNP j is causal, the actual non-centrality parameter at SNP j may differ from $\lambda_j^c\sqrt{n}$ due to the effect of other variants in the region.

In general, where we can assume that the SNPs $j_1, j_2, \ldots, j_k$ are causal with individual effect size $\lambda_{j_r}^c$ for $r = 1, \ldots, k$, we can consider a vector $C^T = (c_1, c_2, \ldots, c_m)$ such that

$$c_i = \begin{cases} 0 & \text{if SNP } i \text{ is not a causal} \\ \lambda_i^c & \text{if SNP } i \text{ is causal} \end{cases}$$

. Accounting for all this information, the multivariate normal can be written as

$$(s_1, s_2, \ldots, s_n)^T \sim \mathcal{N}(\Sigma C\sqrt{n}, \Sigma) \tag{11}$$

We note that $\sqrt{n}$ is scalar in the above equation that multiplies each entry in the mean vector, and we use the notation above for clarity. For example, suppose we have four SNPs where SNPs 2 and 4 are causal with effect sizes λ_2^c and λ_4^c respectively. In this case vector C is as follows: $C^T = (0, \lambda_2^c, 0, \lambda_4^c)$ and the distribution of the statistics is

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} r_{21}\lambda_2^c + r_{41}\lambda_4^c \\ \lambda_2^c + r_{42}\lambda_4^c \\ r_{23}\lambda_2^c + r_{43}\lambda_4^c \\ r_{24}\lambda_2^c + \lambda_4^c \end{pmatrix} \sqrt{n}, \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{pmatrix} \right)$$

where we can see how the non-centrality parameter of each SNP is affected by the causal SNPs.

We note that many association studies today utilize linear mixed models for computing the assication statistics to take into account population structure or relatedness in the sample [21, 28, 29, 46, 45, 30]. When linear mixed models are utilized, the correlation between statistics is affected [20] and a derivation of the correlation when applying mixed models is shown in Appendix B.

3.2 Multiple Hypotheses Testing under the MVN Model

In a GWAS study, we carry out a hypothesis test for each SNP i. Although each of these hypotheses tests has associated a particular false positive rate α_s, we would like to account for the overall false positive rate of the whole process of testing m non-associated SNPs. In the case of two SNPs i and j, the false positive rate can be viewed as the probability under the distribution

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{bmatrix} \right)$$

of the external part of a rectangle defined by the points $(-\theta_{\alpha_s}, -\theta_{\alpha_s})$ and $(\theta_{\alpha_s}, \theta_{\alpha_s})$ where

$\theta_{\alpha_s} = -\Phi^{-1}(\alpha_s/2)$ is the per-maker threshold associated to the SNPs (Figure 2). In the case of multiple SNPs, this region, which is the external part of a hypercube, is denoted as R_{α_s}. This region is the rejection region, where we reject at least one of the null hypotheses if our vector of statistics falls inside it. In the case of m SNPs, and a given per SNP threshold α_s, the overall false positive rate of the study can be written as follows

$$\alpha = \int_{R_{\alpha_s}} f(X; 0, \Sigma) dx \tag{12}$$

and the per SNP threshold α_s can be set so that the overall false positive rate of the study is at the desired level. From the equation it is clear that the false positive rate depends on the per SNP threshold, the number of SNPs and the variance-covariance matrix. When applied to a GWAS, equation (12) requires an integration in a space that may have over a million dimensions. An efficient method for computing this integration is described in Han et al., (2009)[11].

3.3 Power under the MVN Model

In order to calculate power under the MVN model, we assume that we know the vector of true effect sizes C. Therefore, the statistics $(s_1, s_2, \ldots, s_m)^T$ follow a multivariate normal distribution $\mathcal{N}(\Sigma C\sqrt{n}, \Sigma)$ with C and Σ as in (11). Therefore, the power is the probability of the rejection region, R_{α_s}, under the previously defined distribution. Specifically, the power then generalizes equation (7) to multiple SNPs

$$P(\alpha_s, C\sqrt{n}, \Sigma) = \int_{R_{\alpha_s}} f(X; \Sigma C\sqrt{n}, \Sigma) dx \tag{13}$$

For example, in the case of two SNPs where the first SNP is the causal SNP with non-centrality parameter λ_1^c, the correlation between the two SNPs is r_{12}, and the per SNP significance threshold

is α_s, then the power of the association study testing both SNPs is

$$\int_{R_{\alpha_s}} f\left(X; \begin{bmatrix} \lambda_1^c \\ r_{12}\lambda_1^c \end{bmatrix} \sqrt{n}, \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\right) dx \tag{14}$$

where R_{α_s} is the region defined outside of the square defined by the points $(-\theta_{\alpha_s}, -\theta_{\alpha_s})$ and $(\theta_{\alpha_s}, \theta_{\alpha_s})$. See Figure 2 for an example.

The traditional notion of statistical power assumes a specific alternative hypothesis, which defines which variants are causal and makes an assumption on their effect sizes. However, in practice, we are interested in the concept of "average power" which is the average of the statistical power computed for each variant given a specific effect size.

In order to estimate the power of GWAS, we typically assume that each SNP has equal chance of being causal. We then compute the power for each SNP and report the average value over all the SNPs. This approach assumes a probability model over the causal vectors C where each SNP has equal chance of being the causal variant. A simple probability model is to assume that at most 1 SNP is causal and each SNP i has a probability of c_i (with $\sum_{i=1}^m c_i < 1$) of being causal, all with the same effect size of λ^c. In this scenario there are $m+1$ possible models. We can define this set of possible models as $\mathcal{C}_1$. This set contains $m+1$ vectors that can be denoted as $C^{(i)}$ with $i = 0, \ldots, m$, where $C^{(0)} = (0, \ldots, 0)$ represents the situation in which no SNP is causal (and has probability $1 - \sum_{i=1}^m c_i < 1$) and $C^{(i)}$, for $i = 1, \ldots, m$ is such that all the elements are 0 except the ith component that is λ^c. The power of an association study in the case of using this probability model is

$$P(\alpha, \mathcal{C}_1, \sqrt{n}, \Sigma) = \frac{1}{\sum_{i=1}^m c_i} \sum_{i=1}^m P(\alpha, C^{(i)}\sqrt{n}, \Sigma) c_i \tag{15}$$

We note that each variant can be assigned a different prior probability based on additional data such as functional genomic data. We can then modify our association testing approach to maximize the statistical power in equation (15) as described in Eskin (2008) [7] and related publications [4, 6].

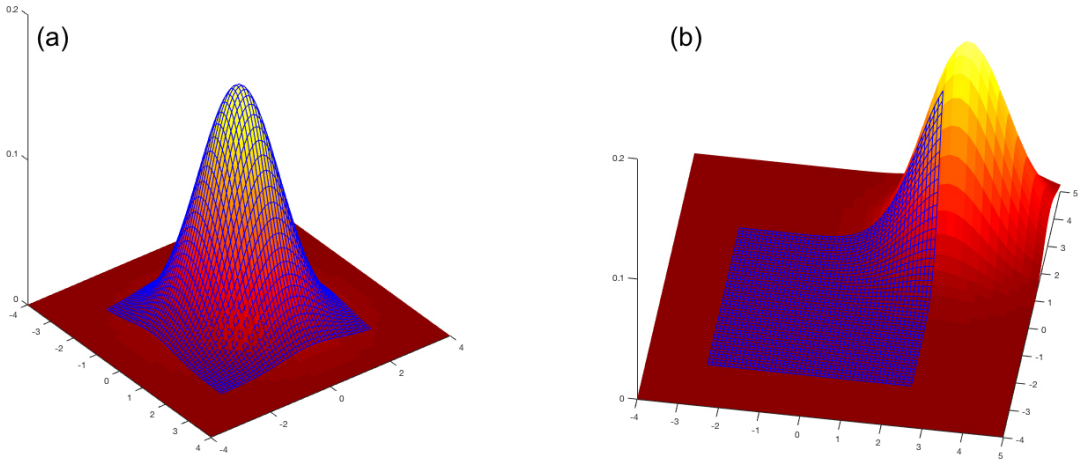


Figure 2: An illustration of the multivariate normal model (a) Type I Error (b) Power

4 Statistical Fine Mapping

A central problem in GWAS is identifying the actual causal variants responsible for the association. This approach is referred to as statistical fine mapping. In this problem, we observe the results of an association study represented as a vector of observed statistics S, and we aim to obtain some information about the actual causal vector C. This requires assuming a distribution over possible values of C. We then define the prior over possible values of C as $P(C)$. Given the value of C, we can then define the probably over the statistics as decribed above and we denote this as $P(S|C)$.

The simplest formulation of fine mapping is to consider the set of probability models in $\mathcal{C}_1$. In this case, the posterior for each model $C^{(i)}$, given the value of the statistics $S = \hat{S}$, is

$$P(C^{(i)}|\hat{S}) = \frac{P(\hat{S}|C^{(i)})c_i}{\sum_{C^{(j)} \in \mathcal{C}_1} P(\hat{S}|C^{(j)})c_j}$$

It can be noted that the ranking of $P(C^{(i)}|\hat{S})$ is the same as the one provided by the p-values of association statistics.

We can use these posteriors to define a "confidence set". A confidence set is the set of possible causal variants that capture a sufficient fraction of the posterior distribution. Intuitively, this is the set of variants that has a high probability of containing the causal variant. We define the posterior for a set of k SNPs $j_1, j_2, \ldots, j_k$ as the sum of the posteriors $\sum_{i=1}^{k} P(C^{(j_i)}|\hat{S})$. A typical "confidence set" is defined as a set that accounts for a posterior probability higher than .95. A confidence set computed for a GWAS locus intuitively is the set of SNPs that, with high probability, contains the causal variants responsible for the association at the locus.

Since the posterior computation requires assumptions, the interpretation of the "confidence sets" assumes that the priors over the causal vector C are consistent with what is actually occurring at the locus. In this sense, the model for statistical fine mapping $\mathcal{C}_1$ above is unrealistic for two reasons.

First, there are often multiple causal variants in the same locus, and our previous model only considers the scenario where one SNP affects the trait at a given locus. In general, this model is a good approximation if the causal variants are not in LD with each other. Secondly, it assumes that all the causal SNPs have the same effect size λ^c.

Hormozdiari et al.,(2014)[16] introduced a hierarchical model based on the multivariate normal model which allows for multiple variants with effect sizes drawn from a normal distribution. Most recently developed fine mapping methods build upon this model[23, 2, 22, 1, 27]. In this case, we assume that each SNP has a probability c_i of being causal independently of the rest of SNPs. Here, the set of models, denoted as $\mathcal{C}_m$ contains 2^m elements. If we define a binary variable γ_i that takes a value of 1 when SNP i is causal, and otherwise takes a value of 0, then the prior probability for a casual status is as follows

$$\prod_{i=1}^{m} c_i^{\gamma_i}(1 - c_i)^{1-\gamma_i} \tag{16}$$

Once we know which SNPs are causal, we use a Gaussian model inspired by the classic Fisher's polygenic model to get the effect sizes of the causal SNPs. The vector C is drawn from the distribution

$$C \sim \mathcal{N}(0, \Sigma_C) \tag{17}$$

where Σ_C has elements

$$sigma_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ \sigma_g & \text{if SNP } i \text{ is causal} \\ \epsilon & \text{otherwise} \end{cases}$$

Under this model, the fine-mapping is carried out as follows. Given a set of SNPs $\mathcal{K}$, we denote the set of causal SNP configurations rendered by $\mathcal{K}$ with $\mathcal{C}_\mathcal{K}$, which accounts for all possible models with causal SNPs in $\mathcal{K}$. These are $2^{|\mathcal{K}|}$ models, and the posterior probability of $\mathcal{C}_\mathcal{K}$ can be calculated as follows

$$P(\mathcal{C}_\mathcal{K}|\hat{S}) = \sum_{C \in \mathcal{C}_\mathcal{K}} P(C|\hat{S})$$

where to compute $P(C|\hat{S})$, we utilize the Bayse's rule and compute $P(\hat{S}|C)$ which is the likelihood of observed statistics given the vector of causal status. The details for this calculation is provided in Appendix C.

The fine-mapping consists in given a threshold ρ find the smallest subset of SNPs $\mathcal{K}^*$ such that $P(\mathcal{C}_{\mathcal{K}^*}|\hat{S}) \geq \rho$. An algorithm to compute such a set is presented in Hormozdiari et al.,(2014) [16]. We can also incorporate functional genomics data to set the prior probabilities (c_i) in the fine mapping model [24].

5 Inference about Uncollected SNPs

In the context of GWAS, one of the advantages of this multivariate framework is that it allows us to carry out inference about uncollected SNPs (i.e., given a non-collected SNP u we can use the statistics from correlated SNPs to obtain some information about the statistic of SNP u).

Given an non-collected or unobserved SNP u, we consider its O most strongly correlated collected SNPs (this information can be obtained from a database of SNPs such as the HapMap). Let R_{uO} denote the $O \times 1$ vector of the correlation coefficients between u and the O tag SNPs. Similarly, let S_O and C, respectively, be the $O \times 1$ vectors of the association statistics; let effect sizes of the tag SNPs and Σ_O be the $O \times O$ matrix of their pairwise correlation coefficients. The joint distribution of the association statistics of the unobserved SNP u and the O tag SNPs follows a multivariate normal distribution, which can be obtained by developing equation 11 and expressed as follows

$$\begin{pmatrix} s_u \\ S_O \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 1 & R_{uO}^T \\ R_{uO} & \Sigma_O \end{pmatrix} \begin{pmatrix} C_u \\ C \end{pmatrix} \sqrt{n} \ , \ \begin{pmatrix} 1 & R_{uO}^T \\ R_{uO} & \Sigma_O \end{pmatrix} \right)$$

The previous joint distribution allows us to make inference about the statistic of the unobserved SNP u. Departing from that distribution, we can calculate the distribution of the statistic of the uncollected SNP u given the statistic of the collected SNPs $\hat{S}_O$

$$s_u|S_O = \hat{s}_O \sim \mathcal{N}\left((c_u + R_{uO}^T C)\sqrt{n} + R_{uO}^T \Sigma_O^{-1} (\hat{s}_O - \Sigma_O C \sqrt{n}) \ , \ 1 - R_{uO}^T \Sigma_O^{-1} R_{uO} \right)$$

This distribution can be used in different ways. This approach can be thought of as a method for imputation [26, 34]. For example, we could fill the value of s_u with the mean value of the previous distribution. A second application is to calculate the probability of that SNP being causal given the value of the collected SNPs [25, 44].

6 Discussion

We have presented how the multivariate normal distribution can be utilized to explicitly model the correlation between variants in the analysis of GWAS studies. We have demonstrated how the MVN framework can be applied to correct for multiple testing, estimate the statistical power of an association study, and perform fine mapping and imputation.

References

[1] Christian Benner, Chris C.A. Spencer, Aki S. Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, jan 2016.

[2] Wenan Chen, Beth R. Larrabee, Inna G. Ovsyannikova, Richard B. Kennedy, Iana H. Haralambieva, Gregory A. Poland, and Daniel J. Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736, may 2015.

[3] Karen N. Conneely and Michael Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet*, 81(6):1158–68, 12 2007.

[4] Gregory Darnell, Dat Duong, Buhm Han, and Eleazar Eskin. Incorporating prior information into association studies. *Bioinformatics*, 28(12):i147–i153, 6 2012.

[5] Paul I. W. de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B. Gabriel, Mark J. Daly, and David Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–23, 11 2005.

[6] Dat Duong, Jennifer Zou, Farhad Hormozdiari, Jae Hoon Sul, Jason Ernst, Buhm Han, and Eleazar Eskin. Using genomic annotations increases statistical power to detect egenes. *Bioinformatics*, 32(12):i156–i163, 6 2016.

[7] Eleazar Eskin. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res*, 18(4):653–60, 4 2008.

[8] Jonathan Flint and Eleazar Eskin. Genome-wide association studies in mice. *Nat Rev Genet*, 13(11):807–17, 11 2012.

[9] Matthew L. Freedman, David Reich, Kathryn L. Penney, Gavin J. McDonald, Andre A. Mignault, Nick Patterson, Stacey B. Gabriel, Eric J. Topol, Jordan W. Smoller, Carlos N. Pato, Michele T. Pato, Tracey L. Petryshen, Laurence N. Kolonel, Eric S. Lander, Pamela Sklar, Brian Henderson, Joel N. Hirschhorn, and David Altshuler. Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36(4):388–93, 4 2004.

[10] Kevin L. Gunderson, Frank J. Steemers, Grace Lee, Leo G. Mendoza, et al. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, 37(5):549–54, 5 2005.

[11] Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, 5(4):e1000456, 4 2009.

[12] Agnar Helgason, Brynds Yngvadttir, Birgir Hrafnkelsson, Jeffrey Gulcher, and Kri Stefnsson. An icelandic example of the impact of population structure on association studies. *Nat Genet*, 37(1):90–5, 1 2005.

[13] David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox. Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–9, 2 2005.

[14] Farhad Hormozdiari, Eun Yong Kang, Michael Bilow, Eyal Ben-David, Chris Vulpe, Stela McLachlan, Aldons J. Lusis, Buhm Han, and Eleazar Eskin. Imputing phenotypes for genome-wide association studies. *The American Journal of Human Genetics*, 99(1):89–103, jul 2016.

[15] Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc, and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 31(12):i206–i213, jun 2015.

[16] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 10 2014.

[17] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V. Segrè, Xiao Li, Jong Wha J. Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, dec 2016.

[18] Farhad Hormozdiari, Anthony Zhu, Gleb Kichaev, Chelsea J.-T. Ju, Ayellet V. Segrè, Jong Wha J. Joo, Hyejung Won, Sriram Sankararaman, Bogdan Pasaniuc, Sagiv Shifman, and Eleazar Eskin. Widespread allelic heterogeneity in complex traits. *The American Journal of Human Genetics*, 100(5):789–802, may 2017.

[19] Jong Wha J. Joo, Farhad Hormozdiari, Buhm Han, and Eleazar Eskin. Multiple testing correction in linear mixed models. *Genome Biol*, 17:62, 4 2016.

[20] Jong Wha J. Joo, Eun Yong Kang, Elin Org, Nick Furlotte, Brian Parks, Farhad Hormozdiari, Aldons J. Lusis, and Eleazar Eskin. Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure. *Genetics*, 204(4):1379–1390, 12 2016.

[21] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Y. Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–54, 4 2010.

[22] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2):260–271, aug 2015.

[23] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10(10):e1004722, oct 2014.

[24] Gleb Kichaev, Wen-Yun Y. Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*, 10(10):e1004722, 10 2014.

[25] Emrah Kostem and Eleazar Eskin. Efficiently identifying significant associations in genome-wide association studies. *J Comput Biol*, 20(10):817–30, 9 2013.

[26] Donghyung Lee, T. Bernard Bigdeli, Brien Riley, Ayman Fanous, and Silviu-Alin . A. Bacanu. Dist: Direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, page btt500, 8 2013.

[27] Yue Li and Manolis Kellis. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research*, 44(18):e144–e144, jul 2016.

[28] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nat Methods*, 8(10):833–5, 2011.

[29] Jennifer Listgarten, Christoph Lippert, Carl M. Kadie, Robert I. Davidson, Eleazar Eskin, and David Heckerman. Nat methods, 5 2012.

[30] Po-Ru R. Loh, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, Paul M. Ridker, Benjamin M. Neale, Bonnie Berger, Nick Patterson, and Alkes L. Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*, 47(3):284–90, 3 2015.

[31] Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins. A hapmap harvest of insights into the genetics of common disease. *J Clin Invest*, 118(5):1590–605, 5 2008.

[32] Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods*, 1(2):109–11, 11 2004.

[33] D. L. Newman, M. Abney, M. S. McPeek, C. Ober, and N. J. Cox. Am j hum genet, 11 2001.

[34] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P. Strachan, Nick Patterson, and Alkes L. Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–14, 10 2014.

[35] Judea Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K.; New York, 2000.

[36] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 7 2001.

[37] S. R. Seaman and B. Müller-Myhsok. Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *Am J Hum Genet*, 76(3):399–408, 3 2005.

[38] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nat Biotechnol*, 26(10):1135–45, 10 2008.

[39] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299, 10 2005.

[40] Benjamin F. Voight and Jonathan K. Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, 1(3):e32, 9 2005.

[41] Bruce S. Weir, Amy D. Anderson, and Amanda B. Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[42] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 6 2007.

[43] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, et al. The NHGRI GWAS catalog, a curated resource of snp-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–6, 1 2014.

[44] Yue Wu, Farhad Hormozdiari, Jong Wha J. Joo, and Eleazar Eskin. Improving imputation accuracy by inferring causal variants in genetic studies. In *Lecture Notes in Computer Science*, pages 303–317. Springer International Publishing, 2017.

[45] Jian Yang, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46(2):100–6, 2 2014.

[46] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 44(7):821–4, 7 2012.

A Derivation of covariance of association statistics

In this section, we show the derivation of the covariance between association statistics. Let m be the number of SNPs, s_i be a statistic for the ith SNP, and $\Sigma = \{\texttt{Cov}(s_i, s_j)\}$ be the $m \times m$ covariance matrix between the statistics.

For the association at SNP i, the following model for the effect of SNP i on the kth individual is utilized

$$y_k = \mu + \beta_i x_{ik} + \epsilon_k$$

and in vector notation

$$Y = \mu\mathbf{1} + \beta_i X_i + \mathbf{e}$$

Here, X_i is a column vector of normalized genotypes for variant i and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of dimension n and $\mathbf{1}$ is a column vector of 1's. Then, the phenotype follows a MVN with a mean and variance as follows:

$$Y \sim \mathcal{N}(\mu\mathbf{1} + X_i\beta_i, \sigma_e^2\mathbf{I})$$

The ordinary least-squares solutions of β for SNP i and SNP j are as follows:

$$\hat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y \sim \mathcal{N}\left(\beta_i, \frac{\sigma_e^2}{X_i^T X_i}\right)$$

$$\hat{\beta}_j = (X_j^T X_j)^{-1} X_j^T Y \sim \mathcal{N}\left(\beta_j, \frac{\sigma_e^2}{X_j^T X_j}\right)$$

The association statistics of the two SNPs are computed as follows:

$$s_i = \frac{\hat{\beta}_i}{\hat{\sigma}_e}\sqrt{X_i^T X_i} \sim \mathcal{N}\left(\beta_i \frac{\sqrt{X_i^T X_i}}{\sigma_e}, 1\right)$$
$$s_j = \frac{\hat{\beta}_j}{\hat{\sigma}_e}\sqrt{X_j^T X_j} \sim \mathcal{N}\left(\beta_j \frac{\sqrt{X_j^T X_j}}{\sigma_e}, 1\right)$$

Here, the estimated values for μ, $\mathbf{e}$, and σ for the SNP i are as follows: $\hat{\mu} = \frac{\mathbf{1}^T X_i}{X_i^T X_i}, \hat{\mathbf{e}} = Y - \hat{\mu}\mathbf{1} - X\hat{\beta}$ and $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{n-2}}$. Then, we can prove that the covariance of the two statistics, $\texttt{Cov}(s_i, s_j)$, is equal to the correlation between the genotypes, r_{ij}, as follows:

$$\begin{aligned}
\texttt{Cov}(s_i, s_j) =& \texttt{Cov}\left(\frac{\hat{\beta}_i}{\sigma_e}\sqrt{X_i^T X_i}, \frac{\hat{\beta}_j}{\sigma_e}\sqrt{X_j^T X_j}\right) \\
=& \frac{1}{\sigma_e^2}\texttt{Cov}\left(\frac{X_i^T Y}{\sqrt{X_i^T X_i}}, \frac{X_j^T Y}{\sqrt{X_j^T X_j}}\right) \\
=& \frac{X_i^T X_j}{\sqrt{X_i^T X_i}\sqrt{X_j^T X_j}} \\
=& \texttt{Cor}(X_i, X_j) \equiv r_{ij}
\end{aligned} \tag{S1}$$

This relationship between genotype correlation and MVN covariance holds for case/control studies as well [37, 11].

B Covariance of association statistics taking into account for population structure

Because of each population's own genetic and social history, allele frequencies are known to vary widely from population to population. This creates genetic similarity between individuals in the study population, referred to as "population structure". Individuals within a population have more similar phenotype values than individuals in distant populations. Population structure, along with this correlation of a phenotype with its populations, may cause spurious correlations between genotypes and a phenotype and induce an inflation of the values of association statistics leading to false positives [33, 9, 40, 12, 41, 8]. Linear mixed model (LMM) has emerged as a general approach to address this problem by explicitly modeling population structure in its association statistic [21, 28, 29, 46, 45, 30, 20].

For LMM, equation (S1) is no longer valid. That is, we cannot use the genotype correlation matrix as the covariance matrix of association statistics for mixed model. To derive the covariance matrix of association statistics under structured population, we assume a mixed model instead of

the linear model equation shown in the previous section. For the association at SNP i, the following LMM for the effect of SNP i on the kth individual is utilized

$$y_k = \mu + \beta_i^M x_{ik} + u_k + \epsilon_k$$

and in vector notation

$$Y = \mu\mathbf{1} + \beta_i^M X_i + \mathbf{u} + \mathbf{e}$$

Here, X_i is a column vector of normalized genotypes for variant i, $\mathbf{u} \sim \mathcal{N}(0, \sigma_g^2\mathbf{K})$ is a column vector modeling population structure effects, where $\mathbf{K}$ is the genetic relative matrix, referred to as "kinship matrix", that explains the correlation between the individuals induced by population structure. Since the genotypes are normalized, the kinship matrix can be expressed as $\mathbf{K} = XX^T/m$, where m is the number of genotypes and X is the $n \times m$ matrix of the normalized genotypes. $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$ is a column vector modeling residual errors, where $\mathbf{I}$ is the identity matrix of dimension n.

Under this model, the phenotype follows a MVN with a mean and variance as follows:

$$Y \sim \mathcal{N}(\mu\mathbf{1} + X_i\beta_i^M, \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I})$$

Given the observed data, it is straightforward to fit a LMM and estimate the parameters σ_g^2 and σ_e^2 using standard strategies, which define the covariance matrix of phenotypes, $\texttt{Cov}(Y) = \hat{V} = \hat{\sigma}_g^2\mathbf{K}, +\hat{\sigma}_e^2\mathbf{I}$. Now we utilize the fact that after obtaining $\hat{V}$, the remaining regression procedure is equivalent to performing ordinary least-squares in the transformed space,

$$\hat{V}^{-1/2}Y \sim \mathcal{N}(\hat{V}^{-1/2}\mu\mathbf{1} + \hat{V}^{-1/2}X_i\beta_i^M, \mathbf{I})$$

where both genotypes and phenotypes are transformed by a factor $\hat{V}^{-1/2}$. Assuming that $\hat{V}^{-1/2}X_i$ and $\hat{V}^{-1/2}Y$ are normalized as mean 0 and variance 1 (without loss of generality), the ordinary least-squares solution of β_i^M for ith SNP and jth SNP are as follows:

$$\begin{aligned}
\hat{\beta}_i^M &= (X_i^T\hat{V}^{-1}X_i)^{-1}X_i^T\hat{V}^{-1}Y \sim \mathcal{N}\left(\beta_i^M, (X_i^T\hat{V}^{-1}X_i)^{-1}\right) \\
\hat{\beta}_j^M &= (X_j^T\hat{V}^{-1}X_j)^{-1}X_j^T\hat{V}^{-1}Y \sim \mathcal{N}\left(\beta_j^M, (X_j^T\hat{V}^{-1}X_j)^{-1}\right)
\end{aligned}$$

The statistics are computed as follows:

$$\begin{aligned}
s_i^M &= \hat{\beta}_i^M\sqrt{X_i^T\hat{V}^{-1}X_i} \sim \mathcal{N}\left(\beta_i^M\sqrt{X_i^T\hat{V}^{-1}X_i}, 1\right) \\
s_j^M &= \hat{\beta}_j^M\sqrt{X_j^T\hat{V}^{-1}X_j} \sim \mathcal{N}\left(\beta_i^M\sqrt{X_j^T\hat{V}^{-1}X_j}, 1\right)
\end{aligned}$$

Accordingly, the correlation between the statistics changes from Equation (S1) to the following and the correlation between the statistics are equal to the correlation between the SNP transformed by the inverse square root of $\hat{V}$,

$$\begin{aligned}
\texttt{Cov}(s_i^M, s_j^M) &= \texttt{Cov}\left(\frac{X_i^T\hat{V}^{-1}Y}{\sqrt{X_i^T\hat{V}^{-1}X_i}}, \frac{X_j^T\hat{V}^{-1}Y}{\sqrt{X_j^T\hat{V}^{-1}X_j}}\right) \\
&= \frac{X_i^T\hat{V}^{-1/2}(\hat{V}^{-1/2})^TX_j}{\sqrt{X_i^T(\hat{V}^{-1/2})^T\hat{V}^{-1/2}X_i}\sqrt{X_j^T(\hat{V}^{-1/2})^T\hat{V}^{-1/2}X_j}} \\
&= \texttt{Cor}(\hat{V}^{-1/2}X_i, \hat{V}^{-1/2}X_j) = r_{ij}^M
\end{aligned}$$

Thus, we can account for population structure in analyses related to GWAS, including multiple testing correction [3, 11, 19], estimation of statistical power [11], statistical fine mapping [16, 15, 24], and imputation [26, 34].

C Efficient likelihood computation

We show in Equation (11) that the joint distribution of marginal statistics given the causal status is as follows:

$$(S|C) \sim \mathcal{N}(\Sigma C\sqrt{n}, \Sigma)$$

In addition, we use Equation (17) that is inspired by the classic Fisher's polygenic model to get the effect sizes of the causal SNPs. This distribution is as follows:

$$C \sim \mathcal{N}(0, \Sigma_C).$$

Utilizing the MVN conjugate prior and applying it to Equations (11) and (17), we can obtain the joint distribution of marginal statistics as follows:

$$S \sim \mathcal{N}(0, \Sigma + n\Sigma\Sigma_C\Sigma) \tag{S2}$$

To compute the likelihood of the casual status, we utilize the probability density function of the MVN as shown in Equation (9). Unfortunately, a naive method to compute the likelihood is computationally intensive. In the naive method, we need to compute $S^T(\Sigma + n\Sigma\Sigma_C\Sigma)^{-1}S$ and $|\Sigma + n\Sigma\Sigma_C\Sigma|$ that both require $O(m^3)$ operations. We use Woodbury matrix identity formula to speedup the computation of $S^T(\Sigma + n\Sigma\Sigma_C\Sigma)^{-1}S$ and use Sylvester's determinant identity to speedup the computation of $|\Sigma + n\Sigma\Sigma_C\Sigma|$.

We reduce the time complexity by only computing the values that change with matrix C. We can factor out the Σ matrix as follows:

$$S^T(\Sigma + n\Sigma\Sigma_C\Sigma)^{-1}S = S^T\Sigma^{-1}(\boldsymbol{I} + n\Sigma_C\Sigma)^{-1}S$$
$$|\Sigma + n\Sigma\Sigma_C\Sigma| = |\Sigma||\boldsymbol{I} + n\Sigma_C\Sigma|$$

where $|\Sigma|$ and $S^T\Sigma^{-1}$ can be computed once and can be used many times. Thus, we need to compute $(\boldsymbol{I} + n\Sigma_C\Sigma)^{-1}$ and $|\boldsymbol{I} + n\Sigma_C\Sigma|$ for every causal status. It is worth mentioning that we require Σ to be full rank. Unfortunately, in some loci, Σ, can be low rank. In this section, we assume that the LD matrix is full rank and in Appendix D we deal with low rank LD matrices. To ease the notation, we introduce two matrices U and V where U has $(m \times k)$ elements and V has $(k \times m)$ elements. We set elements of U and V such that $n\Sigma_C\Sigma = UV$. Let α_i indicate the index of ith causal variant. We set elements of V as follows: $V(i,j) = r_{\alpha_i,j}$. We set $U(\alpha_i, i)$ to $n\sigma$ while the rest of elements in U are set to zero.

We use the Woodbury matrix identity formula to compute $(\boldsymbol{I} + n\Sigma_C\Sigma)^{-1}$. The Woodbury matrix identity formula is as follows:

$$(A + UEV)^{-1} = A^{-1} - A^{-1}U(E^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

where we set A to $\boldsymbol{I}_{m\times m}$ and E to $\boldsymbol{I}_{k\times k}$. As a result, we have:

$$\begin{aligned}(\boldsymbol{I}_{m\times m} + n\Sigma_C\Sigma)^{-1} &= (\boldsymbol{I}_{m\times m} + UV)^{-1} \\ &= \boldsymbol{I}^{-1}_{m\times m} - \boldsymbol{I}^{-1}_{m\times m}U(I^{-1}_{k\times k} + V\boldsymbol{I}^{-1}_{k\times k}U)^{-1}V\boldsymbol{I}_{m\times m} \\ &= \boldsymbol{I}_{m\times m} - U(\boldsymbol{I}_{k\times k} + VU)^{-1}V\end{aligned}$$

Interestingly, to compute $(\boldsymbol{I}_{k\times k} + VU)^{-1}$ we need to inverse a $k \times k$ matrix that is much smaller than inverting a $m \times m$ matrix. Thus, we reduce the computation of $S^T(\Sigma + n\Sigma\Sigma_C\Sigma)^{-1}S$ from $O(m^3)$ to $O(m^2k)$ where k is the number of causal variants for a given causal status ($k << m$).

We use the Sylvester's determinant identity to speedup $|\boldsymbol{I}+n\Sigma_C\Sigma|$ computation. The Sylvester's determinant identity formula is as follows:

$$|\boldsymbol{I}_{m\times m} + UV| = |\boldsymbol{I}_{k\times k} + VU|$$

Thus, instead of computing the determinate of a $n \times n$ matrix, we can compute the determinate of a $k \times k$ matrix. We set matrices U and V such that $UV = n\Sigma_C\Sigma$. Thus, we can compute $|\boldsymbol{I} + n\Sigma_C\Sigma|$ in $O(k^3)$ operations [18].

D Handling Low Rank LD Matrices

As mentioned in the above section, we assume that the LD matrix, Σ, is full rank. However, in some loci the LD matrix can be low rank due to linear dependency between different variants (e.g., two variants that are in perfect LD). We recall that we use Equation (S2) to compute the likelihood of each causal status. The LD matrix is computed from genotype data ($\Sigma = X^TX$), thus the LD matrix is semi-positive definite. Using the fact that the LD matrix is semi-positive definite, we can use the eigenvalue decomposition of the LD matrix which is as follows:

$$\Sigma = Q\Omega Q^T$$

where Q is the matrix of eigenvectors and the ith column of Q is the i-th eigenvector of matrix Σ. Matrix Q is an orthogonal matrix ($Q^TQ = QQ^T = \boldsymbol{I}$). Let Ω be a diagonal matrix that consists of eigenvalues of Σ where the ith diagonal element of Ω is the ith eigenvalue of matrix Σ. We introduce a new set of marginal statistics $S' = \Omega^{-1/2}Q^TS$ such that the joint distribution is computed as follows:

$$S' = \Omega^{-1/2}Q^TS \sim \mathcal{N}(0, \Omega^{-1/2}Q^T\Sigma Q\Omega^{-1/2} + n\Omega^{-1/2}Q^T\Sigma\Sigma_C\Sigma Q\Omega^{-1/2})$$

where we can simplify $\Omega^{-1/2}Q^T\Sigma Q\Omega^{-1/2}$ to $\boldsymbol{I}$ and $n\Omega^{-1/2}Q^T\Sigma\Sigma_C\Sigma Q\Omega^{-1/2}$ to $n\Omega^{1/2}Q^T\Sigma_C Q\Omega^{1/2}$ which can be shown as follows:

$$\Omega^{-1/2}Q^T\Sigma Q\Omega^{-1/2} = \Omega^{-1/2}Q^TQ\Omega Q^TQ\Omega^{-1/2} = \Omega^{-1/2}\Omega\Omega^{-1/2} = \boldsymbol{I}$$

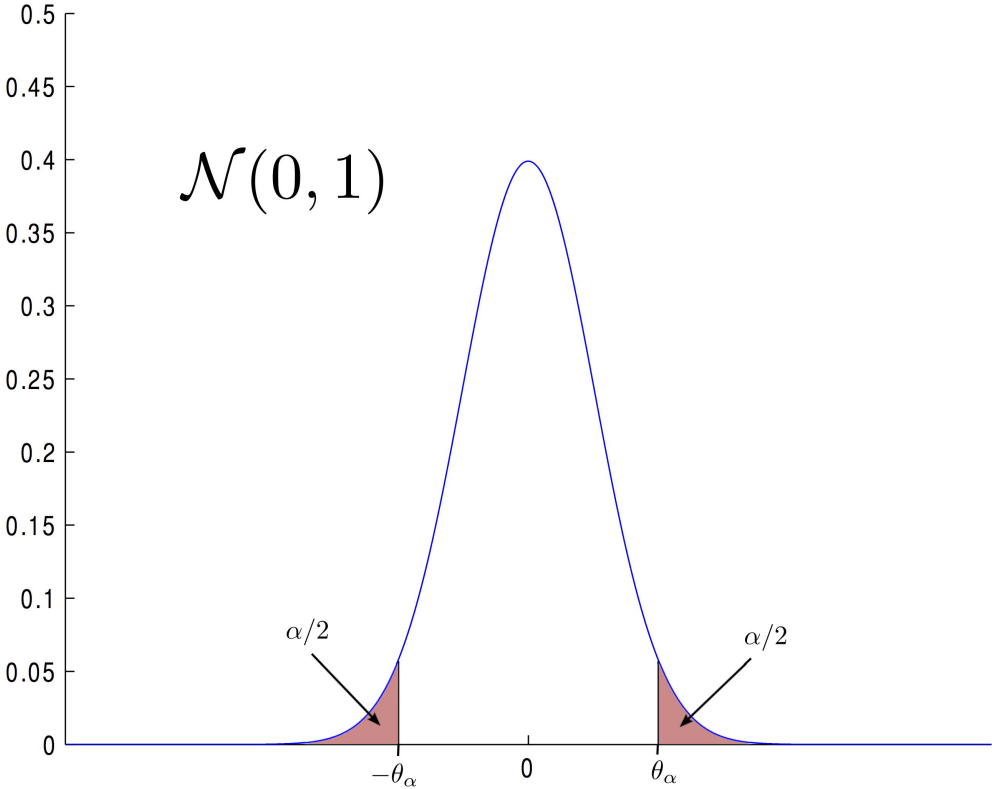
Similarly, we have:

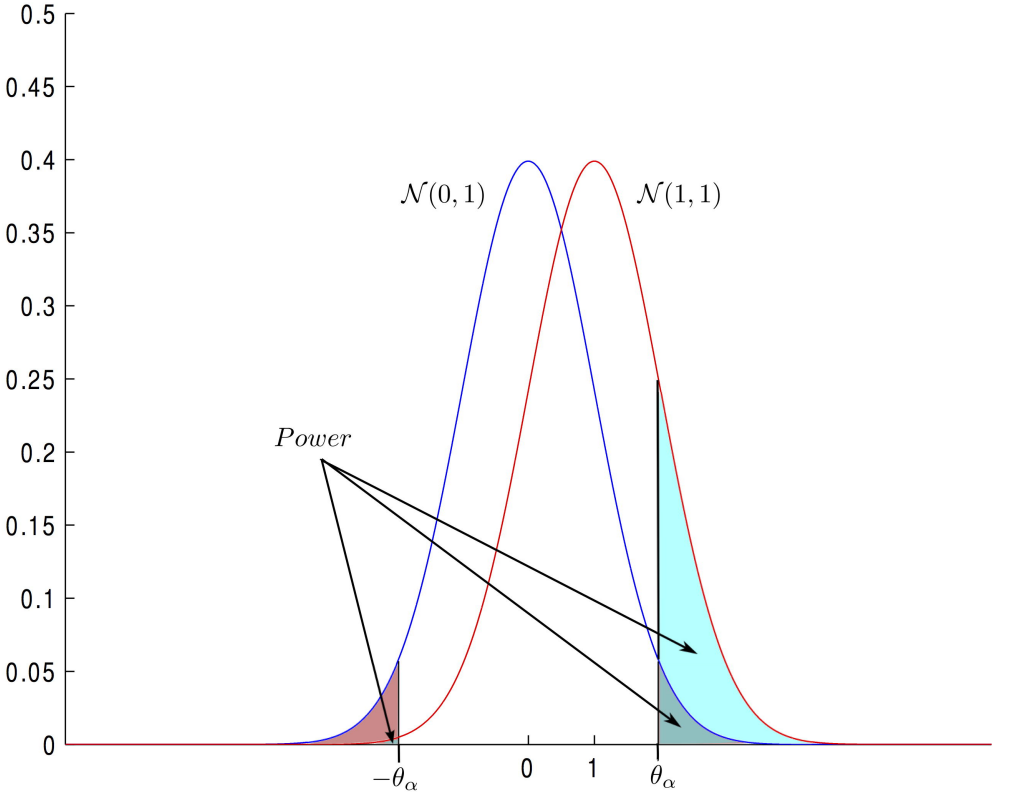
$$\begin{aligned} n\Omega^{-1/2}Q^T\Sigma\Sigma_C\Sigma Q\Omega^{-1/2} &= n\Omega^{-1/2}Q^TQ\Omega Q^T\Sigma_C Q\Omega Q^TQ\Omega^{-1/2} \\ &= n\Omega^{-1/2}\Omega Q^T\Sigma_C Q\Omega\Omega^{-1/2} \\ &= n\Omega^{1/2}Q^T\Sigma_C Q\Omega^{1/2} \end{aligned}$$

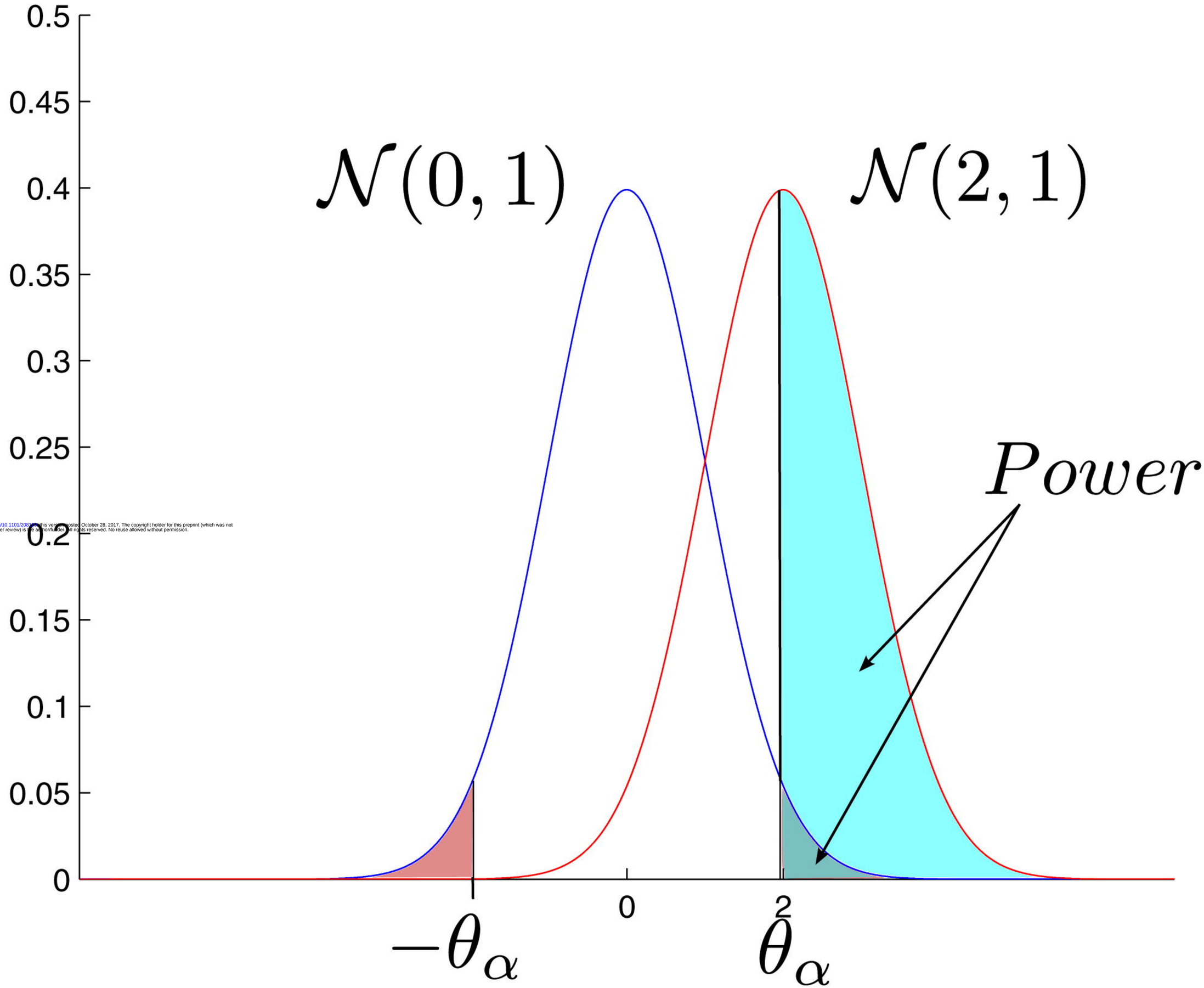
Thus, the joint distribution of S' is as follows:

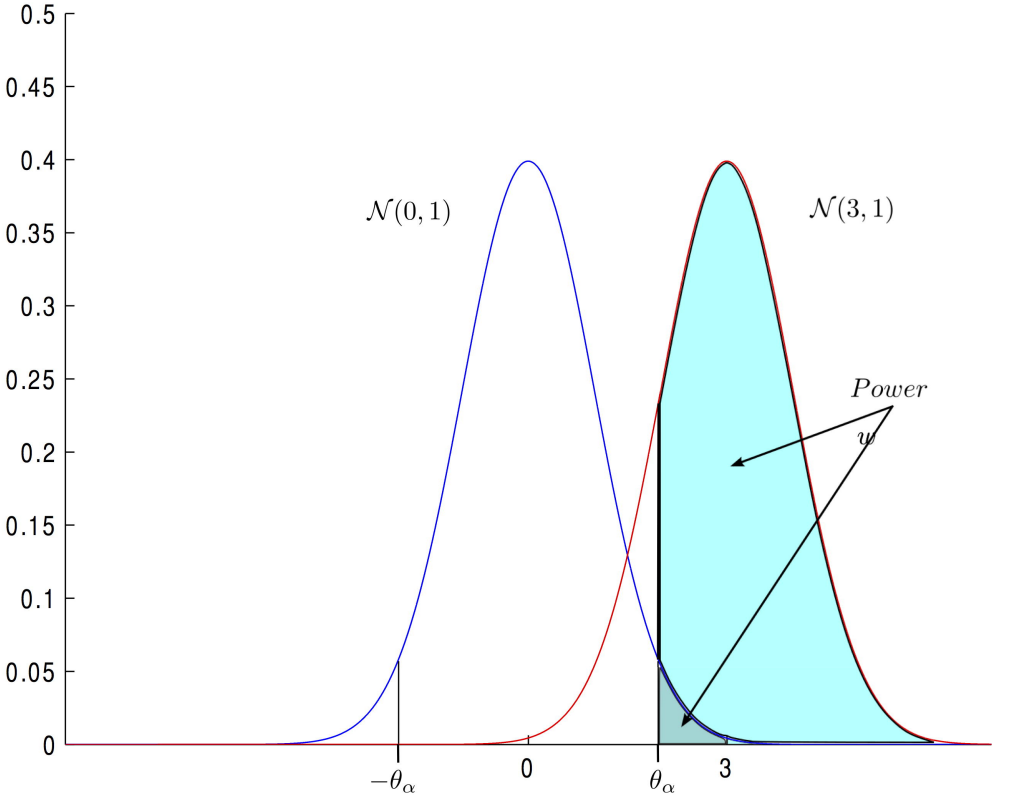
$$S' = \Omega^{-1/2}Q^TS \sim \mathcal{N}(0, \boldsymbol{I} + nB\Sigma_C B^T) \tag{S3}$$

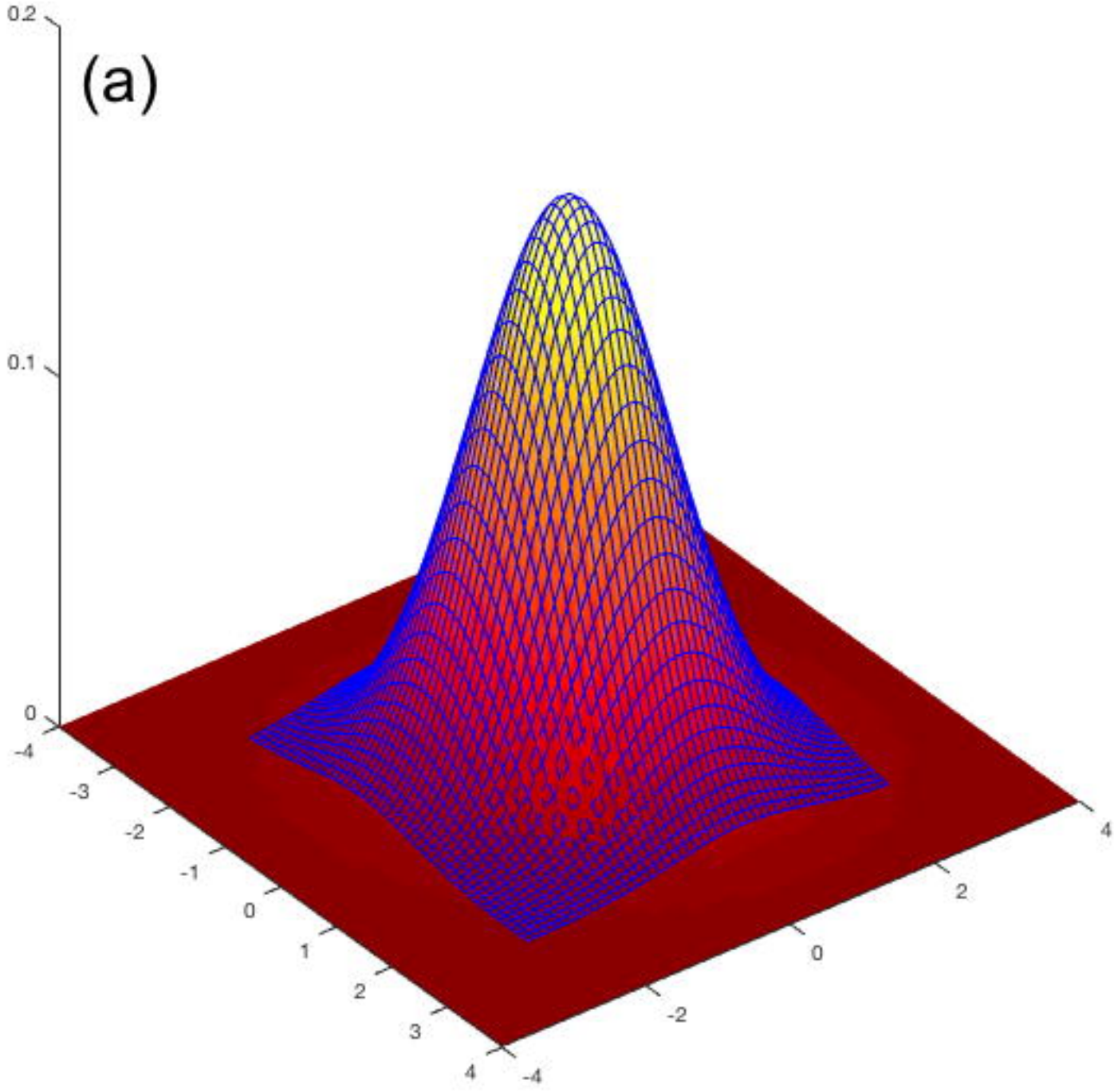
where $B = \Omega^{1/2}Q^T$. It is worth mentioning that $\boldsymbol{I} + nB\Sigma_C B^T$ is full rank. Thus, we can compute the likelihood of causal status for a locus where the LD matrix is not full rank.

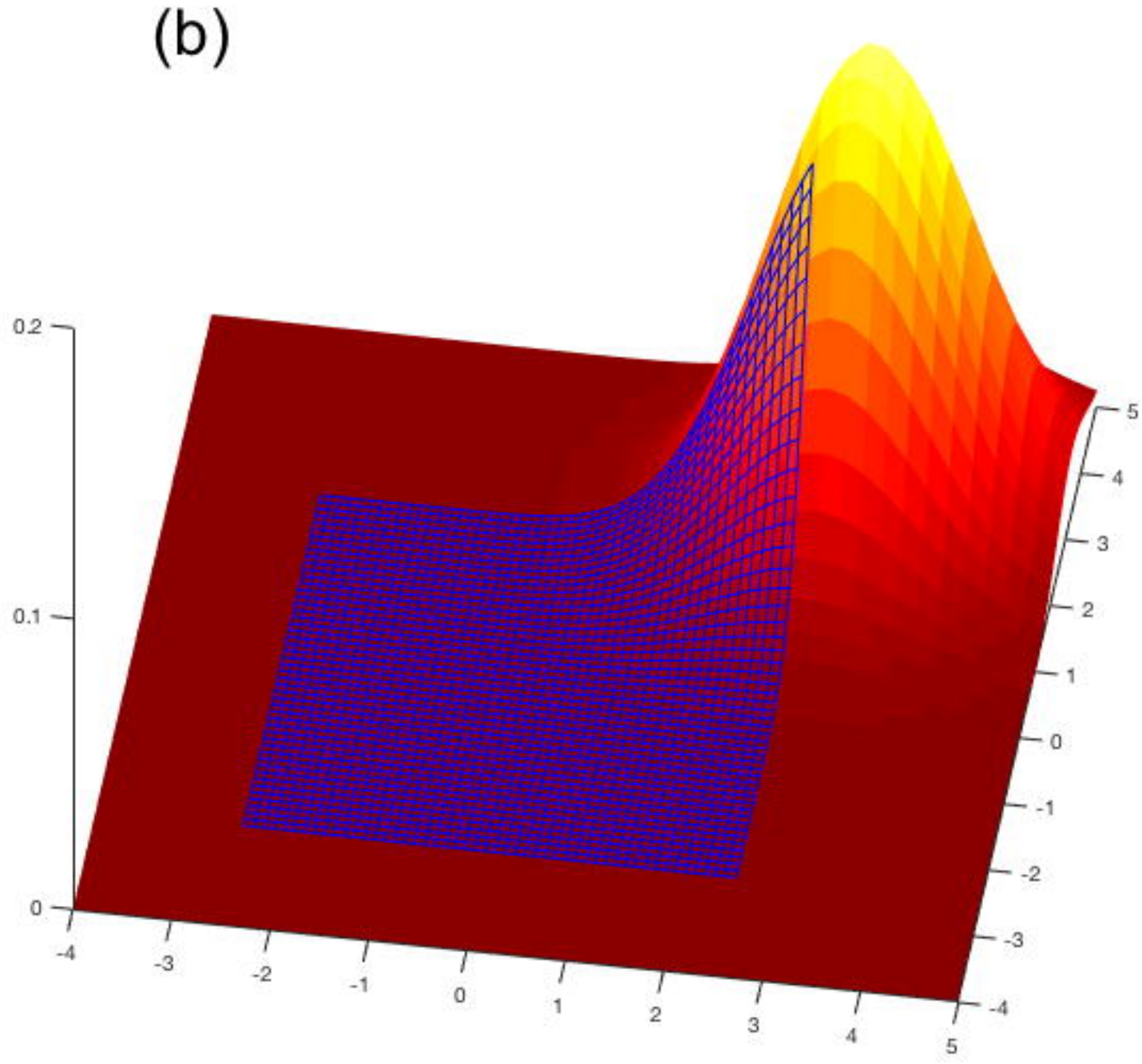

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
$\mathcal{N}(0,1)$
$\alpha/2$
$\alpha/2$
$-\theta_\alpha$
0
θ_α


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
$\mathcal{N}(0,1)$
$\mathcal{N}(1,1)$
Power
$-\theta_\alpha$
0
1
θ_α




0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
$\mathcal{N}(0,1)$
$\mathcal{N}(3,1)$
$Power$
ψ
$-\theta_\alpha$
0
θ_α
3


(a)
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
-4
-2
0
2
4


(b)
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
5
5
4
3
2
1
0
-1
-2
-3
-4