A unifying framework for summary statistic imputation

Yue Wu[1], Eleazar Eskin[1,2], and Sriram Sankararaman*[1,2]

[1]Department of Computer Science, UCLA
[2]Department of Human Genetics, UCLA

Abstract

Imputation has been widely utilized to aid and interpret the results of Genome-Wide Association Studies(GWAS). Imputation can increase the power to identify associations when the causal variant was not directly observed or typed in the GWAS. There are two broad classes of methods for imputation. The first class imputes the genotypes at the untyped variants given the genotypes at the typed variants and then performs a statistical test of association at the imputed variants. The second class of methods, summary statistic imputation, directly imputes the association statics at the untyped variants given the association statistics observed at the typed variants. This second class of methods is appealing as it tends to be computationally efficient while only requiring the summary statistics from a study while the former class requires access to individual-level data that can be difficult to obtain. The statistical properties of these two classes of imputation methods have not been fully understood. In this paper, we show that the two classes of imputation methods are equivalent, *i.e.*, have identical asymptotic multivariate normal distributions with zero mean and minor variations in the covariance matrix, under some reasonable assumptions. Using this equivalence, we can understand the effect of imputation methods on power. We show that a commonly employed modification of summary statistic imputation that we term summary statistic imputation with variance re-weighting generally leads to a loss in power. On the other hand, our proposed method, summary statistic imputation without performing variance re-weighting, fully accounts for imputation uncertainty while achieving better power.

1 Introduction

Genome-Wide Association Studies(GWAS) has been successfully used to discover genetic variants, typically single nucleotide polymorphisms (SNPs), that affect the trait of interest [1–7]. GWAS measure or type the genotypes of individuals at a chosen set of SNPs and, then, perform a statistical test of association between a given SNP and the trait of interest. SNPs at which the null hypothesis of no association between the genotype and the trait can be rejected are said to be associated with the trait. The threshold that the absolute value of association statistics pass to reject null hypothesis is also referred as significance level.

In a typical GWAS, due to the cost considerations, only a subset of SNPs are genotyped (typed SNPs). Thus, a direct analyses of typed SNPs is likely to have reduced power to detect associations between untyped SNPs and the trait. Thus, imputation methods, that aim to fill in "data" at the untyped SNPs, have emerged as a powerful strategy to increase the power of GWAS. These methods all rely on the correlation or linkage disequilibrium (LD) [8, 9]. between genotypes at untyped SNPs and those at typed SNPs [10–16] Initial work on imputation focused on the problem of genotype imputation, *i.e.*, inferring the genotypes at untyped SNPs given the genotypes at typed SNPs.

*Corresponding author:sriram@cs.ucla.edu, eeskin@cs.ucla.edu

Genotype imputation methods rely a reference panel in which individuals are typed at all SNPs of interest to learn the LD patterns across SNPs. Given a target dataset in which genotypes are typed at a subset of the SNPs, these methods rely on the LD patterns learned from the reference panel to infer the genotypes at the remaining untyped SNPs.

In the context of GWAS, there are two broad classes of imputation methods to estimate the association statistics at untyped SNPs. The first class relies on genotype imputation to infer the genotypes at the untyped SNPs followed by computing association statistics at the imputed genotypes [10–14, 16]. We refer to this class of imputation methods as **Two-step imputation** methods. In practice, the most successful methods for the first step of genotype imputation are based on discrete Hidden Markov Models (HMM) [10, 16]. The second class of methods directly imputes the association statistics at the untyped SNPs given the association statistics at the typed SNPs. As shown in previous work [17, 18], the joint distribution of marginal statistics at the typed SNPs and untyped SNPs follow a multivariate normal distribution (MVN) [17–21]. This class of methods utilizes the correlation between the association statistics induced by their dependence on the underlying genotypes [22, 23]. This class of methods is termed *summary statistic imputation* (**SSI**). Summary statistic imputation is appealing as it tends to be computationally efficient while only requiring the summary statistics from a study while the first class requires access to individual-level data which can be difficult to obtain in practice. Current summary-statistic based imputation methods calibrate the imputed statistics using a technique we call *variance re-weighting* (**SSI-VR**). Despite recent progress, the statistical properties of summary statistic imputation methods (including the impact of variance re-weighting) and the connection between the two classes of summary statistic imputation methods has not been adequately understood.

In this paper, we show that the two classes of imputation methods, **Two-step imputation** and **SSI** are asymptotically multivariate normal with small differences in the underlying covariance matrix. Using this asymptotic equivalence, we can understand the effect of the imputation method on power. Our new method, SSI, perfoms summary statistic imputation without variance re-weighting. The resulting statistics do not then have unit variance as in traditional summary statistic imputation but instead correctly take into account the ambiguity of the imputation process.

We compared the peroformance of the imputations methods on the Northern Finland Birth Cohort (NFBC) data set [24] to show that SSI increases power over no imputation while SSI-VR can sometimes lead to lower power. Finally, we ran SSI, SSI-VR and Two-step imputation on the NFBC dataset and show that the resulting statistics are close thereby justifying the theory.

2 Results

2.1 Overview of Summary Statistics

Assume we have a total of $M = (U + O)$ SNPs that are partitioned into O observed (or tag) SNPs $\{snp_1, snp_2, snp_3 ... snp_O\}$ and U missing SNPs $\{snp_1, snp_2, snp_3, ... snp_U\}$ for N individuals. For the O tag SNPs, let $\boldsymbol{s}_O$ be a vector of association statistics of length O, $\boldsymbol{\lambda}_O$ be a vector of non-centrality (NCP) parameters of length O, and let $\boldsymbol{\Sigma}_O$ be a $O \times O$ matrix of their pairwise correlation coefficients. For the U missing SNPs, let $\boldsymbol{s}_U$ be a vector of association statistics of length U, $\boldsymbol{\lambda}_U$ be a vector of NCP parameters also of length U, and let $\boldsymbol{\Sigma}_U$ be a $U \times U$ matrix of their pairwise correlation coefficients.

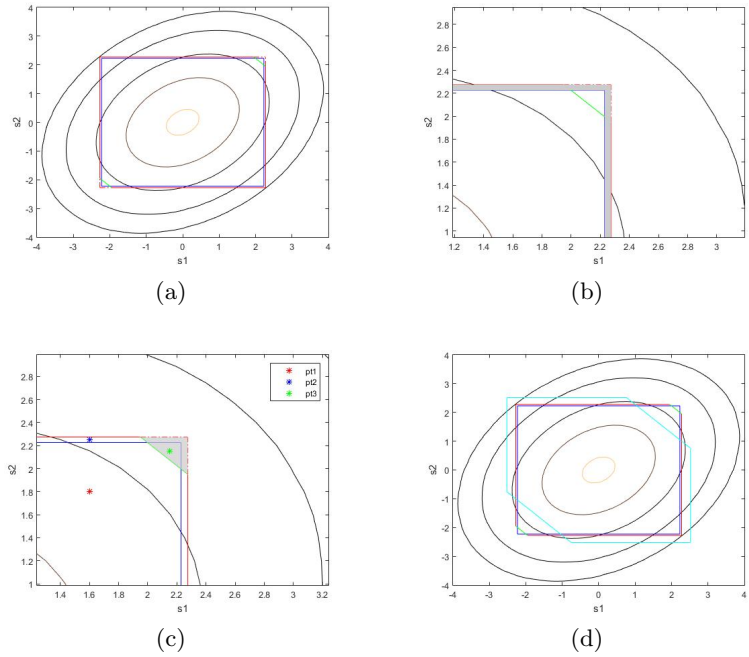
Let $\boldsymbol{\Sigma}_{UO}$ be a $U \times O$ matrix of the pairwise correlation, *i.e.*, linkage disequilibrium (LD), between missing SNPs and observed SNPs. Thus, we have a $M \times M$ LD matrix, $\boldsymbol{\Sigma}_{LD}$. We can partition the LD matrix as: $\boldsymbol{\Sigma}_{LD} = \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UO} \\ \boldsymbol{\Sigma}_{OU} & \boldsymbol{\Sigma}_O \end{bmatrix}$. For large sample sizes, the association

statistics follow a multivariate normal distribution,

$$\begin{bmatrix} \boldsymbol{s}_U \\ \boldsymbol{s}_O \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\lambda}_U \\ \boldsymbol{\lambda}_O \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UO} \\ \boldsymbol{\Sigma}_{OU} & \boldsymbol{\Sigma}_O \end{bmatrix} \right) \tag{1}$$

Under the null where we assume that none of the SNPs is causal, $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_O$ are equal to $\mathbf{0}$.

2.2 Example



(a) (b) (c) (d)

Figure 1: **The effect of imputation on the rejection boundary**: This figure shows rejection boundary with no imputation, with imputation (SSI), and variance re-weighted imputation (SSI-VR) for an example containing two observed SNPs snp_1, snp_2 and an unobserved SNP snp_3. The contours represent the probability density of the statistics for the observed SNPs: s_1 and s_2 projected in the plane. In Figure 1a, the blue box is the rejection boundary with FWER 0.05 for snp_1 and snp_2 before imputation. The polygon with red and green colored boundaries is the rejection boundary after imputation. Figure 1b and Figure 1c are a zoomed in version of Figure 1a to show the rejection boundaries changes. Figure 1b shows the power change on two observed SNPs. Figure 1c shows the power change on the imputed SNP and has 3 points corresponding to different scenarios. Figure 1d shows the rejection boundary of imputation with SSI-VRin cyan color in addition to the rejection boundaries seen in Figure 1a.

We consider a simple example to illustrate how imputation affects the rejection threshold at a given set of SNPs. We consider three SNPs: snp_1, snp_2, and snp_3. In this example, snp_1, snp_2 are observed, and $snp3$ is imputed. We assume the statistics of the tag SNPs (snp_1, snp_2), $\begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$ follows $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$ where $|\rho| \leq 1$ and we use $\pi(s_1, s_2)$ to denote this distribution. We also assume that the statistics of the tag SNPs snp_1, snp_2 and the unobserved SNP snp_3 jointly follow the distribution $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix} \right)$ where $|\rho| \leq 1$, $|\alpha| \leq 1$.

Thus having the joint distribution of the statistics s_1, s_2, and s_3, we can compute the conditional distribution of the untyped SNP conditioned on the marginal statistics of the typed SNPs s_1 and s_2:

$$P\left(s_3|s_1,s_2\right) \sim \mathcal{N}\left(\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}^T \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}^{-1} \begin{bmatrix}s_1\\ s_2\end{bmatrix}, 1 - \begin{bmatrix}\alpha\\ \alpha\end{bmatrix}^T \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}^{-1} \begin{bmatrix}\alpha\\ \alpha\end{bmatrix}\right)$$

Typically, summary statistic imputation uses the posterior mean of the statistic s_3 given the observed values of $\hat{s}_1$ and $\hat{s}_2$ to estimate s_3. In our example, this leads to the statistic s_3 for snp_3 being imputed as a function of $\hat{s}_1, \hat{s}_2$:

$$\hat{s}_3(\hat{s}_1,\hat{s}_2) = \frac{\alpha}{1+\rho}\left(\hat{s}_1+\hat{s}_2\right)$$

.

We choose thresholds t for rejecting each of the statistics $(\hat{s}_1, \hat{s}_2, \hat{s}_3)$ such that the family-wise error rate, *i.e.*, the probability of at least one false positive, is controlled at a level 0.05 . For each tested SNP, we choose the threshold to be the same.

In the case where no imputation is performed, we only test two SNPs. We use the same threshold t for SNPs snp_1 and snp_2. Figure 1a shows the rejection boundary (the blue box) for two SNPs with correlation $\rho = 0.36$ where the region outside this box corresponding to the rejection region. Given the joint density $\pi(s_1, s_2)$ of the association statistics (s_1, s_2), we determined the rejection boundary by computing the length of the side of the blue box such that the cumulative density in the rejection area, *i.e.*, the area under the density $\pi(s_1, s_2)$ outside the box is equal to 0.05. Mathematically, we need to find t such that $FWER(t) = 0.05$ where:

$$FWER(t) \quad \equiv \quad 1 - \int \pi(s_1,s_2)\mathbf{1}\left\{s_1 \in -[t,t]\right\}\mathbf{1}\left\{s_2 \in [-t,t]\right\} ds_1 ds_2$$

Here $\mathbf{1}\left\{s_1 \in -[t,t]\right\}\mathbf{1}\left\{s_2 \in [-t,t]\right\}$ defines the acceptance region, *i.e.*, the set of points $(s_1, s_2) \in \mathbb{R}^2$ where the null hypothesis at both SNPs are accepted.

We now consider the effect of testing imputed SNPs in addition to the tag SNPs. The rejection region for snp_1, snp_2, snp_3 are the regions outside the intervals $R_1 = [-t,t], R_2 = [-t,t], R_3 = [-t,t]$ respectively. We can compute the FWER for a given t by determining the probability mass outside the rejection region. To do this, we note that the joint sampling distribution of $(s_1, s_2, \hat{s}_3)$ is determined only by the distribution of (s_1, s_2) since $\hat{s}_3$ is a deterministic function of s_1 and s_2.

$$\begin{aligned} FWER(t) &\equiv 1 - \int \pi(s_1,s_2)\mathbf{1}\left\{s_1 \in -[-t,t]\right\}\mathbf{1}\left\{s_2 \in [-t,t]\right\}\mathbf{1}\left\{s_3 \in [-t,t]\right\} ds_1 ds_2 ds_3 \\ &= 1 - \int \pi(s_1,s_2)\mathbf{1}\left\{s_1 \in [-t,t]\right\}\mathbf{1}\left\{s_2 \in [-t,t]\right\}\mathbf{1}\left\{\frac{\alpha}{1+\rho}(s_1+s_2) \in [-t,t]\right\} ds_1 ds_2 \end{aligned}$$

Notice that, in the setting with imputation, the acceptance region $\mathbf{1}\left\{s_1 \in [-t,t]\right\}\mathbf{1}\left\{s_2 \in [-t,t]\right\}$ $\mathbf{1}\left\{\frac{\alpha}{1+\rho}(s_1+s_2) \in [-t,t]\right\}$ can never increase relative to the setting where only the tag SNPs are tested. Now consider the case where the null hypothesis at both the observed SNPs is accepted. This happens when $|\hat{s}_1| \leq t$ and $|\hat{s}_2| \leq t$. Then the statistic at the imputed SNP:

$$\begin{aligned} |\hat{s}_3(\hat{s}_1,\hat{s}_2)| &= |\frac{\alpha}{1+\rho}(\hat{s}_1+\hat{s}_2)| \\ &\leq |\frac{\alpha}{1+\rho}|\left(|\hat{s}_1|+|\hat{s}_2|\right) \quad \text{(triangle inequality)} \\ &\leq 2|\frac{\alpha}{1+\rho}|t \end{aligned}$$

Thus, if $2|\frac{\alpha}{1+\rho}| \leq 1$, then we have $|\hat{s}_3(\hat{s}_1 + \hat{s}_2)| \leq t$. Thus, the imputed SNP will never be rejected when neither of the observed SNPs is rejected. Thus, the acceptance region remains the same as the setting when only the tag SNPs are tested. In other words, imputation does not change the rejection boundary.

On the other hand, when $\frac{\alpha}{1+\rho} > \frac{1}{2}$, then imputation will change the rejection region. Figure 1 shows the effect of imputation with $\alpha = 0.80$ and $\rho = 0.36$ so that $\hat{s}_3(\hat{s}_1, \hat{s}_2) = 0.5882(\hat{s}_1 + \hat{s}_2)$. The rejection boundary of the observed SNPs snp_1 and snp_2 after imputation are shown by the red lines. The rejection region for snp_3 corresponds to the region where $|0.5882(s_1 + s_2)| > t$ which corresponds to the green line. Thus the cumulative density outside the polygon of red and green lines is the same as the rejection area outside the blue box. In Figure 1b, the shaded area indicates the power loss on the observed SNPs, and in Figure 1c the shaded area is the power gained from imputation.

Thus assume we have three points, $p1$, $p2$ and $p3$ in Figure 1c, which are three different pairs of association statistics of observed SNPs $snp1$ and $snp2$. The first point is in both the blue rectangle and the polygon, which means we will accept null with or without imputation. The second point $p2$ is the case that, without imputation we will reject null, and after imputation we will accept null because of the change of boundary on observed SNPs. The third point $p3$ is the special case. In this case, the observed SNPs don't have significant association because it lies inside the blue box, but after imputation, the imputed SNP has a significant association since it lies outside the polygon and thus we reject the null.

2.3 Simulation Results

As shown in previous work on summary statistics [22], the marginal statistics at typed SNPs and untyped SNPs follow a multivariate normal distribution. With the assumption that none of the SNP is significantly associated with train, the mean of the multivariate normal distribution is 0.

As in the previous simple case having 3 SNPs, snp_1, snp_2 and snp_3, under the null hypothesis of no association, the summary statistics follow the distribution $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix}\right)$.

Thus having the joint distribution of the statistics s_1, s_2, and s_3, we can compute the conditional distribution of the untyped SNP conditioned on the marginal statistics of the typed SNPs s_1 and s_2:

$$P\left(s_3|s_1, s_2\right) \sim \mathcal{N}\left(\begin{bmatrix} \alpha \\ \alpha \end{bmatrix}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, 1 - \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}\right) \tag{2}$$

Summary statistic imputation estimates s_3 using the mean of the above distribution $\hat{s}_3$. The variance of the imputed statistic: $var(\hat{s}_3) = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}$ is smaller than 1 (since Equation 2 shows that the variance of $s_3|s_1, s_2$ is $1 - \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}$ and the variance is non-negative). Thus, in most summary statistics imputation [22, 23], snp_3 is imputed as $\hat{z}_3 = \frac{\hat{s}_3}{\sqrt{var(\hat{s}_3)}}$ so that all the association statistics have variance 1. Since the variance of $\hat{s}_3$ is ≤ 1, the new statistic $|\hat{z}_3| \geq |\hat{s}_3|$. As a result, for a given threshold, the acceptance region in SSI-VR is never greater than with SSI. In other words, to achieve a given FWER, the threshold t needs to be larger for SSI-VR than without as shown in Figure 1d.

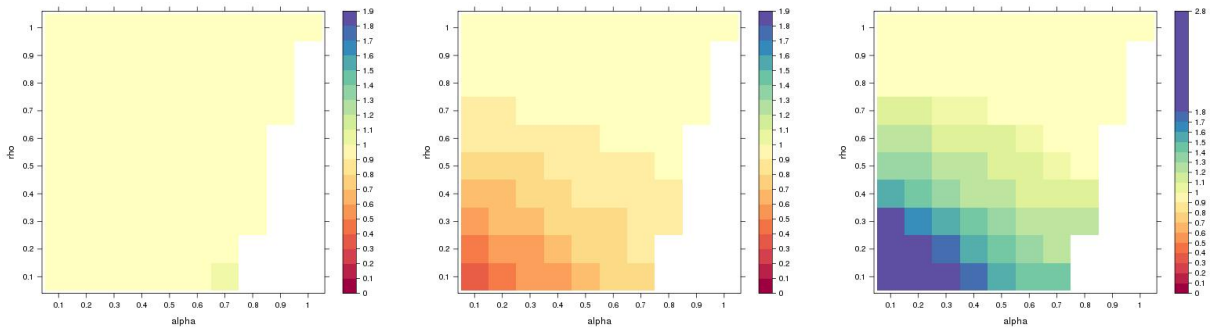
Now having snp_3 imputed using summary statistics, we want to find out how power is affected by SSI and SSI-VR. In section 3 of the Supplementary Information, we analytically compute the

average marginal power function for both methods. In order to assess power, we assume that 3 SNPs, snp_1, snp_2 and snp_3 are drawn from a region associated with a trait. We assume that the untagged variant, snp_3, is causal with non-centrality parameter (NCP) so that (s_1, s_2, s_3) follow a non-zero mean multivariate normal distribution: $\mathcal{N}\left(\begin{matrix}2.31\alpha \\ 2.31\alpha, \\ 2.31\end{matrix}\begin{bmatrix}1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & 1\end{bmatrix}\right)$. We choose the NCP to be 2.31 so that the maximum power of no imputation will be around 0.5, which will happen when both α and ρ are 1. We let the correlation between untagged and tag SNPs α and the correlation between tag SNPs ρ vary across: $[0.1, 0.2, \ldots, 0.9, 1]$.

For each combination of $[\alpha, \rho]$, we determined a set of 3 thresholds i) for no imputation, ii) for imputation, and iii) imputation with variance correction. We drew 10^8 samples from each distribution, and the power is defined as the the probability that we reject the null hypothesis based on thresholds for each method.

In all the combinations except the cases that the LD matrix is no longer positive definite, we find the power of no imputation, SSI and SSI-VR (Figure 2). In Figure 2a, we compared SSI versus no imputation, and we show that SSI always increases power when $\frac{\alpha}{1+\rho} > \frac{1}{2}$ as the ratio is always larger in 1. Since the power of no imputation depend more on the correlation between tagged and untagged SNP, we see the power being sensitive to α. For instance, if $\alpha = 0.7$ and $\rho = 0.3$, the average power of no imputation is 0.4918 while the average of the power of imputation with no correction is 0.6614. In figure 2(b), we compared SSI-VR versus no imputation. We see comparing to 2(a), the power increasing much less significant. In fact, in some cases, we observe SSI-VR has less power than no imputation. For example, when $\alpha = 0.7$ and $\rho = 0.1$, the average power of imputation with variance correction is 0.4639, and null has an average power of 0.5154.

Then, we compare imputation and imputation with variance re-weighting in Figure 2c and we notice that SSI-VR will always cause power loss and in figure the value of ratio are all larger than 1. For instance, when $\alpha = 0.7$ and $\rho = 0.3$, the average power of imputation is 0.6614, and the average power of imputation with variance correction is 0.5403.



(a) (b) (c)

Figure 2: **A comparison of the power of imputation (SSI) v.s. no imputation(a), SSI-VR v.s. no imputation(b), and SSI v.s. SSI-VR in a simple example consisting of three SNPs of which only two are observed**. In each panel, we plot the ratio of the power of the two methods under all configurations of α and ρ. In each figure, the configuration of α and ρ that results in a covariance matrix that is not positive definite, e.g. $\alpha = 1$, $\rho = 0.1$, is left empty. Figure 2a shows that for values of $\alpha \leq \frac{1+\rho}{2}$, the ratio is near one since the rejection boundary is unchanged (as predicted by our theory) while for values of $\alpha > \frac{1+\rho}{2}$, the power of SSI is greater than that of no imputation. Figure 2b and 2c show that SSI-VR can lose power relative to both no imputation as well as SSI for a range of configurations of LD.

2.4 SSI achieves better power compared to existing methods in Northern Finland Birth Cohort (NFBC)

In order to assess the power of imputation and the effect of SSI-VR on imputation in a real dataset, we simulated marginal statistics utilizing the Northern Finland Birth Cohort (NFBC) dataset.

We assume that every other SNP on chromosome 22 is missing. Thus, we observe half of SNPs on chromosome 22 and perform imputation on the rest. We find the per-SNP threshold for only observed SNPs (*i.e.* no imputation), for SSI and for SSI-VR with the constraint that FWER is controlled at 0.05. We sampled association statics from the multivariate distribution on the observed SNPs from the genome. Then we used the sampled statistics to find the per-SNP significance threshold on the observed SNPs. We found the threshold to be **4.59705**. Having this threshold, we then assume that there are causal SNPs in the genome, *i.e.* the mean of statistics on these SNPs are not 0, and assess the power with no imputation. For no imputation, we found an average power of **0.4946**.

For the imputation methods, SSI and SSI-VR we impute the association statistics using the samples statistics. We impute in two ways, one utilizing the MVN of equation(4), and the other one use variance re-weighting technique as equation(5). Under the null, we found per-SNP thresholds for SSI and SSI-VR to be **4.5977** and **4.6891**. We then assume that there are causal SNPs, and used the thresholds to compute the power of each of the imputation methods. We found the average power to be **0.50124** for SSI and **0.4346** for SSI-VR. Notice that the threshold we found for no imputation, SSI, and SSI-VR are more accurate than Bonferroni correction and thus less conservative.

In the Table 1, we also impute the most significantly associated SNPs reported in previous studies using SSI, SSI-VR and a Two-step imputationusing IMPUTE2 to perform genotype imputation. We find the association statistics are similar across the three methods validating our theoretical results.

Table 1: We show that the two classes of imputation method, SSI and Two-step imputation have similar imputation statistics on the NFBC data set. We consider SNPs that were reported significant in a previous study [24]. Then, we treat these SNPs as untyped and impute the marginal statistics using SSI, SSI-VR, and Two-step imputation using IMPUTE2 to impute genotype of untyped SNPs.

Phenotype	chr	rsID	True Statistics	SSI	| True - SSI|	SSI-VR	| True -SSI-VR|	IMPUTE2	| True - IMPUTE2|
TG	2	rs673548	-5.444	-5.37	0.074	-5.37	0.074	-4.46	0.984
	8	rs10096633	-5.679	-5.63	0.049	-5.76	0.082	-5.17	0.509
	15	rs2624265	4.22	3.55	0.67	-3.85	0.37	3.60	0.62
HDL	15	rs1532085	7.13	5.59	1.54	6.33	0.8	6.47	0.66
	16	rs3764261	12.01	8.23	3.78	10.19	1.82	6.47	5.54
	16	rs255049	6.06	5.11	0.95	5.5	0.56	5.70	0.36
	17	rs9891572	4.25	3.99	0.26	4.02	0.23	4.40	0.15
LDL	1	rs646776	-7.70	-7.7	0	-7.81	0.11	-6.96	0.74
	2	rs693	6.81	6.27	0.54	6.34	0.47	5.91	0.9
	11	rs102275	-4.51	-4.43	0.08	-4.45	0.06	-4.54	0.03
	11	rs174546	-4.52	-4.43	0.09	-4.45	0.07	-4.58	0.06
	11	rs174556	-4.69	-4.73	0.04	-4.85	0.16	-4.62	0.07
	11	rs1535	-4.43	-4.46	0.03	-4.66	0.23	-4.45	0.02
	19	rs11668477	-5.96	-3.78	2.18	-4.4	1.56	-5.33	0.63
	19	rs157580	-5.161	-2.6	2.561	-3.11	2.051	-4.20	0.961
CRP	12	rs2650000	-7.08	-5.25	1.83	-6.54	0.54	-6.05	1.03
GLU	2	rs560887	-6.97	-6.21	0.76	-6.3	0.67	-5.69	1.28
	7	rs10244051	5.31	4.34	0.97	4.45	0.86	4.97	0.34
	7	rs2191348	5.30	4.33	0.97	4.47	0.83	4.97	0.33
	11	rs1447352	-6.35	-5.08	1.27	-5.21	1.14	-4.75	1.6
	11	rs7121092	-5.50	-4.93	0.57	-5.31	0.19	-4.60	0.9

3 Methods

3.1 Summary Statistics

Under the null hypothesis, the joint distribution of the association statistics of the U untagged SNP $\boldsymbol{s}_U$ and the O tag SNPs $\boldsymbol{s}_O$ follows a multivariate normal distribution:

$$\begin{bmatrix} \boldsymbol{s}_U \\ \boldsymbol{s}_O \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\lambda}_U \\ \boldsymbol{\lambda}_O \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UO} \\ \boldsymbol{\Sigma}_{UO}^T & \boldsymbol{\Sigma}_O \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UO} \\ \boldsymbol{\Sigma}_{UO}^T & \boldsymbol{\Sigma}_O \end{bmatrix}\right) \tag{3}$$

Since none of the $M = (U + O)$ SNPs are associated, the non-centrality parameters of both $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_O$ are $\mathbf{0}$. Further, the statistics are standardized so that the diagonal elements of the covariance matrix are 1, *i.e.*, $\Sigma_{U_{i,i}} = \Sigma_{O_{j,j}} = 1$.

3.1.1 Summary statistic imputation

Under the null assumption where $\boldsymbol{s}_O$ and $\boldsymbol{s}_U$ are not associated, $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_O$ are each $\mathbf{0}$. Using the joint distribution, we can compute the distribution of the true statistics at the untagged SNPs, $\boldsymbol{s}_U$ conditioned on the statistics observed at the tag SNPs, $\boldsymbol{s}_O$. The conditional distribution follows a multivariate normal distribution, which is computed as follows,

$$P\left(\boldsymbol{s}_U | \boldsymbol{s}_O\right) \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{UO}\boldsymbol{\Sigma}_O^{-1}\boldsymbol{s}_O, \boldsymbol{\Sigma}_U - \boldsymbol{\Sigma}_{UO}{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{OU}\right) \tag{4}$$

The observed statistics are denoted $\hat{\boldsymbol{s}}_O$.Thus $\boldsymbol{s}_U$ is imputed using a function of observed statistics:

$$\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{\Sigma}_{UO}\boldsymbol{\Sigma}_O^{-1}\hat{\boldsymbol{s}}_O \tag{5}$$

Let $\boldsymbol{A} = \boldsymbol{\Sigma}_{UO}{\boldsymbol{\Sigma}_O}^{-1}$ and thus $\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{A}\hat{\boldsymbol{s}}_O$.

3.1.2 Summary statistic imputation with variance re-weighting (SSI)

From the previous result, we have $\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{A}\hat{\boldsymbol{s}}_O$. Notice that the underlying joint distribution over the test statistics assumes that each of the statistics at the observed as well as unobserved SNPs has variance one. On the other hand, Equation 5 shows that the variance of the imputed statistic is less than 1. Variance re-weighting proposes standardizing the statistics at the untagged SNPs.

Let s_i be the statistic at the i^{th} untagged SNP. Thus, instead of imputing s_i using $\hat{s}_i$, we impute using $\hat{z}_i = \frac{\hat{s}_i}{\sqrt{var(\hat{s}_i)}}$, so that all the imputed $\hat{z}_i$ have variance equal to 1.We have: $var(\hat{s}_i) = \mathbb{E}[\boldsymbol{\Sigma}_{U_i,O}{\boldsymbol{\Sigma}_{O,O}}^{-1}\hat{\boldsymbol{s}}_O\hat{\boldsymbol{s}}_O{}^T{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{OU_i}] = \boldsymbol{\Sigma}_{U_i,O}{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{O,U_i}$. Thus we have:

$$\hat{\boldsymbol{z}}_i(\hat{\boldsymbol{s}}_{\boldsymbol{O}}) = \frac{\boldsymbol{\Sigma}_{UO}{\boldsymbol{\Sigma}_O}^{-1}\hat{\boldsymbol{s}}_{\boldsymbol{O}}}{\sqrt{\boldsymbol{\Sigma}_{U_i,O}{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{O,U_i}}} \tag{6}$$

3.2 The impact of imputation on the rejection boundary

SSIuses the following function to impute statistics at the unobserved statistics: $\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{A}\hat{\boldsymbol{s}}_O$. Let $\boldsymbol{A}_i$ be the i^{th} row of matrix $\boldsymbol{A}$, $\boldsymbol{A}_i = {\boldsymbol{\Sigma}_{U_iO}}^T{\boldsymbol{\Sigma}_O}^{-1}$, where $\boldsymbol{\Sigma}_{U_iO}^T$ is the correlation vector between untagged variant snp_i and all the observed SNPs. We choose thresholds t for rejecting statistics at each of the observed and imputed SNP, *i.e.*, we reject the null hypothesis at observed SNP O_j if $|\hat{s}_{O_j}| > t$ while we reject the null hypothesis at unobserved SNP U_i if $|\hat{s}_{U_i}| > t$ where t is chosen

to control the FWER. We would like to understand the conditions the threshold t for SSIrelative to the threshold t when no imputation was performed, *i.e.*, we want to provide conditions when imputation changes the rejection boundary.

Theorem 1. *The imputed statistic at snp_i computed using SSI will change the rejection boundary iff the sum of the absolute values of all the entries of $\boldsymbol{A}_i$, $\sum_j |A_{ij}| > 1$.*

Proof. See Section 2 in Supplementary Information. □

In SSI-VR, instead of using $\hat{s}_i$ as the imputed statistic for variant i, we use

$$\hat{z}_i = \frac{\hat{s}_i}{\sqrt{var(\hat{s}_i)}} = \frac{\sum_j A_{ij}\hat{s}_{O_j}}{\sqrt{\sum_j A_{ij}^2 + 2\sum_{j\neq k} A_{ij}A_{ik}\Sigma_{O_j,O_k}}} \tag{7}$$

In SSI-VR, untagged variant i will effect the rejection boundary iff $\frac{\sum_j |A_{ij}|}{\sqrt{\sum_j A_{ij}^2 + 2\sum_{j\neq k} A_{ij}A_{jk}\Sigma_{O_j,O_k}}} > 1$.

3.3 Two-step imputation

The two-step approach to summary statistic imputation first performs genotype imputation followed by testing for association using the imputed genotypes. Genotype imputation fills in the genotypes at the unobserved SNPs, $\boldsymbol{G}_U$ given the genotypes at observed SNPs $\boldsymbol{G}_O$ [15]. Typically, this involves defining a probability distribution for the missing genotypes given the observed genotypes $P(\boldsymbol{G}_u|\boldsymbol{G}_O)$. Let $p_i(\boldsymbol{g}) = P(\boldsymbol{G}_{U_i} = \boldsymbol{g}|\boldsymbol{G}_O)$ denote the posterior probability at unobserved SNP i. Given a vector $\boldsymbol{g}$ of N genotypes at a SNP, let the association statistic $s(\boldsymbol{g})$ be a function of the genotypes $\boldsymbol{g}$. We can then compute the association statistic at unobserved SNP i as the posterior mean of the association statistic: $\mathbb{E}\left[s(\boldsymbol{G}_{U_i})|\boldsymbol{G}_O\right] = \sum_{\boldsymbol{g}} s(\boldsymbol{g})p_i(\boldsymbol{g})$. In practice, instead of the posterior mean, association statistics are restricted to imputed SNPs at which the imputation is confident (*e.g.* using the INFO score reported by software such as IMPUTE2 [16]) followed by using the maximum *a posteriori* estimate of the genotype at each SNP. We focus on the posterior mean as it accounts for the uncertainty in imputation and is easier to analyze. We first consider a simple genotype imputation strategy that uses the pairwise correlation among SNPs in a multivariate normal distribution [25] (Section 3.3.1). In Section 3.3.2, we consider the use of hidden Markov Models (HMMs) for genotype imputation.

3.3.1 Genotype imputation using multivariate normal distribution

First, we consider a multivariate normal distribution with mean zero and covariance matrix given by the LD matrix to model the distribution of the genotype vector at the observed and unobserved SNPs for each individual [25]. We can then impute the genotypes for missing SNPs $\hat{\boldsymbol{G}}_U$ as a function of observed genotypes $\boldsymbol{G}_O$ using the conditional mean for the multivariate normal distribution (Equation 4). Denoting the $N \times O$ matrix of standardized genotypes as $\boldsymbol{X}_O$ and the imputed genotype vector across N individuals at unobserved SNP i as $\widehat{\boldsymbol{x}}_{U_i}$, we have:

$$\widehat{\boldsymbol{x}}_{U_i}(\boldsymbol{X}_O) = \left(\boldsymbol{\Sigma}_{U_iO}{\boldsymbol{\Sigma}_O}^{-1}{\boldsymbol{X}_O}^{\mathrm{T}}\right)^{\mathrm{T}} = \boldsymbol{X}_O{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{OU_i} \tag{8}$$

where $\boldsymbol{\Sigma}_{U_iO}$ is the i^{th} row of matrix $\boldsymbol{\Sigma}_{UO}$.

Given a vector of continuous phenotypes $\boldsymbol{y} \in \mathbb{R}^N$ measured across N individuals, the effect size $\hat{\beta}_j$ for observed SNP j can be estimated by a linear regression of $\boldsymbol{y}$ on the genotypes at SNP

j: $\hat{\beta}_j = \frac{{\boldsymbol{x}_{O_j}}^{\mathrm{T}}\boldsymbol{y}}{N}$ so that the association statistic s_j at this SNP j: $\hat{s}_j = \frac{\hat{\beta}_j}{\sqrt{var(\hat{\beta}_j)}} = \frac{{\boldsymbol{x}_{O_j}}^{\mathrm{T}}\boldsymbol{y}}{\sigma\sqrt{N}}$. Here σ denotes the standard deviation of the phenotype. Analogously, the association statistic $\hat{s}_i$ at unobserved SNP i is $\hat{s}_i = \frac{\widehat{\boldsymbol{x}}_{U_i}^{\mathrm{T}}\boldsymbol{y}}{\sqrt{var(\widehat{\boldsymbol{x}}_{U_i}^{\mathrm{T}}\boldsymbol{y})}}$. From Equation 8, we have:

$$\hat{s}_i = \frac{\boldsymbol{\Sigma}_{U_iO}{\boldsymbol{\Sigma}_O}^{-1}{\boldsymbol{X}_O}^{\mathrm{T}}\boldsymbol{y}}{\sigma\sqrt{\boldsymbol{\Sigma}_{U_iO}{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{X}_O^{\mathrm{T}}\boldsymbol{X}_O{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{OU_i}}} = \frac{\boldsymbol{\Sigma}_{U_iO}{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{s}_O}{\sqrt{\boldsymbol{\Sigma}_{U_iO}{\boldsymbol{\Sigma}_O}^{-1}\boldsymbol{\Sigma}_{OU_i}}} \tag{9}$$

Here we used $\frac{\boldsymbol{X}_O^{\mathrm{T}}\boldsymbol{X}_O}{N} = \boldsymbol{\Sigma}_O$.

This function is identical to SSI-VR as seen in Equation 7. Thus, applying the imputation function in Equation 8 to directly impute genotypes is equivalent to SSI-VR.

3.3.2 Genotype imputation using hidden Markov models

We consider the use of a hidden Markov model (HMM) for genotype imputation. These models assume that a reference panel $\boldsymbol{M}$ is available that contains genotype data across $M = (U + O)$ SNPs [26, 16, 10, 14]. The HMM models the conditional distribution of each of the pair of haplotypes $(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)})$ in each of the N individuals in the study at the O observed and U unobserved SNPs by the conditional distribution $P(\boldsymbol{h}|\boldsymbol{M})$. Specifically, $\boldsymbol{h}_n^{(a)} \overset{iid}{\sim} P(\boldsymbol{h}|\boldsymbol{M})$ for $n \in \{1, \ldots, N\}$, $\boldsymbol{h}_n^{(a)} \in \{0, 1\}^M$ $a \in \{1, 2\}$.

The effect size estimate for SNP j: $\hat{\beta}_j = \frac{cov(\boldsymbol{h}_j, \boldsymbol{y})}{var(\boldsymbol{h}_j)}$ and the association statistic $s_j = \frac{cov(\boldsymbol{h}_j, \boldsymbol{y})}{\sigma\sqrt{var(\boldsymbol{h}_j)}}$.

We show (in Supplementary Information Section 1) that the vector of association statistics asymptotically follows a multivariate normal distribution:

$$\boldsymbol{s} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_S) \tag{10}$$

The asymptotic covariance matrix of the association statistics $\boldsymbol{\Sigma}_S$ depends on the specific HMM used. Under the commonly used Li-Stephens model [27], this covariance matrix is:

$$\Sigma_{S,ij} = \begin{cases} (1-\theta)^2 + \frac{\theta}{2}\left(1 - \frac{\theta}{2}\right)\frac{1}{\sigma_i^2}, i = j \\ \exp(-\frac{\rho_{ij}}{2N})\Sigma_{ij} \end{cases}, i \neq j \tag{11}$$

Here Σ_{ij} is the LD or the correlation between SNPs i and j, θ is a parameter related to the mutation rate, and ρ_{ij} is an estimate of the population-scaled recombination rate between SNPs i and j . Thus, the association statistics computed using genotypes imputed using a HMM follows a multivariate normal distribution with mean zero and covariance matrix equal to a LD matrix with shrinkage applied according to the recombination rate between SNPs.

4 Discussion

In this paper, we showed the connection between the two broad classes of imputation, Two-step imputation and SSI. We also showed that a commonly employed modification of SSI, variance re-weighting, will cause power loss using simulation and real data. Thus, this leads us to conclude that SSI (with no variance re-weighting) is more powerful.

Summary statistic imputation assumes that statistics follow multivariate normal distribution: this assumption breaks down for small sample sizes and for rare SNPs. Compared to summary statistics, current HMM methods are likely to be more accurate for rare variation. A possible future direction is to improve accuracy on rare variants and small sample sizes.

References

1. Eleftheria Zeggini, Michael N Weedon, Cecilia M Lindgren, et al. Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341, 2007.

2. Robert Sladek, Ghislain Rocheleau, Johan Rung, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

3. Hakon Hakonarson, Struan FA Grant, Jonathan P Bradfield, et al. A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene. *Nature*, 448(7153):591–594, 2007.

4. Jian Yang, Teri A Manolio, Louis R Pasquale, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature genetics*, 43(6):519–525, 2011.

5. Anna Köttgen, Eva Albrecht, Alexander Teumer, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature genetics*, 45(2):ng–2500, 2012.

6. Yi Lu, Veronique Vitart, Kathryn P Burdon, et al. Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nature genetics*, 45(2):155–163, 2013.

7. Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10):1150–1159, 2013.

8. David E Reich, Michele Cargill, Stacey Bolk, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

9. Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

10. S.R. Browning and B.L. Browning. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097, 2007.

11. Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–959, 2012.

12. Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

13. Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387–406, 2009.

14. Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

15. Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

16. J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39:906–913, 2007.

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

18. Emrah Kostem, Jose A Lozano, and Eleazar Eskin. Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. *Genetics*, 188(2):449–460, 2011.

19. 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, 2014.

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

21. Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segre, et al. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

22. Donghyung Lee, T Bernard Bigdeli, Brien P Riley, Ayman H Fanous, and Silviu-Alin Bacanu. Dist: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927, 2013.

23. Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

24. Chiara Sabatti, Anna-Liisa Hartikainen, Anneli Pouta, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46, 2009.

25. Xiaoquan Wen and Matthew Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158, 2010.

26. P Scheet and M Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, 78(4), 04 2006.

27. Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003.

Supplementary Information: A unifying framework for summary statistic imputation

Yue Wu[1], Eleazar Eskin[1,2], and Sriram Sankararaman*[1,2]

[1]Department of Computer Science, UCLA
[2]Department of Human Genetics, UCLA

1 Genotype Imputation using a hidden Markov model

We assume a hidden Markov model for genotype imputation. These models assume that a reference panel $\boldsymbol{M}$ is available that contains genotype data across $(U+O)$ SNPs [1–4]. The HMM models the conditional distribution of each of the pair of haplotypes $(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)})$ in each of the N individuals in the study at the O observed and U unobserved SNPs by the conditional distribution $P(\boldsymbol{h}|\boldsymbol{M})$. Specifically, $\boldsymbol{h}_n^{(a)} \overset{iid}{\sim} P(\boldsymbol{h}|\boldsymbol{M})$ for $n \in \{1,\ldots,N\},$, $\boldsymbol{h}_n^{(a)} \in \{0,1\}^{U+O}$ $a \in \{1,2\}$.

In GWAS, given a vector of continuous phenotypes $\boldsymbol{y} \in \mathbb{R}^N$ measured across N individuals, the effect size $\hat{\beta}_j$ for an observed SNP j can be estimated by a linear regression of $\boldsymbol{y}$ on the genotypes at SNP j to obtain an estimate of the effect size. The effect size for SNP j is estimated as $\hat{\beta}_j = \frac{cov(\boldsymbol{h}_j,\boldsymbol{y})}{var(\boldsymbol{h}_j)}$ and the association statistic $s_j = \sqrt{2N}\frac{cov(\boldsymbol{h}_j,\boldsymbol{y})}{\sigma\sqrt{var(\boldsymbol{h}_j)}}$.

Let $\boldsymbol{p} = \frac{\sum_{n=1}^{N}(\boldsymbol{h}_n^{(1)}+\boldsymbol{h}_n^{(2)})}{2N}$. For large sample sizes N, $\boldsymbol{p}$ is asymptotically distributed as a multivariate normal distribution with mean $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{h}|\boldsymbol{M}]$ and covariance matrix $\boldsymbol{\Sigma} = \frac{1}{2N}\mathrm{Var}[\boldsymbol{h}|\boldsymbol{M}]$. The specific form for the mean and the covariance matrix depends on the form of the HMM used. For example, in the Li-Stephens HMM [5], [6] showed that

$$\boldsymbol{\mu}_{LS} = \boldsymbol{f}(1-\theta) + \frac{\theta}{2}\mathbf{1} \quad (1)$$

$$\boldsymbol{\Sigma}_{LS} = \boldsymbol{S}(1-\theta)^2 + \frac{\theta}{2}(1-\frac{\theta}{2})\boldsymbol{I} \quad (2)$$

Here $\boldsymbol{f}$ is the mean allele frequency in a panel, θ is a parameter related to the mutation rate, and $\boldsymbol{S}$ is an estimator of the covariance

$$S_{ij} = \begin{cases} D_{ij}, i = j \\ \exp(-\frac{\rho_{ij}}{2N})D_{ij}, i \neq j \end{cases} \quad (3)$$

Here ρ_{ij} is an estimate of the population-scaled recombination rate between SNPs i and j while D_{ij} is an estimate of the empirical covariance between SNPs i and j in a reference panel.

*Corresponding author:sriram@cs.ucla.edu, eeskin@cs.ucla.edu

Under the null, we have $y_n \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y^2)$ where y_n is independent of the haplotype $(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)})$. Let

$$\begin{aligned} \boldsymbol{a}_N &= \sqrt{N}\left(\frac{\sum_{n=1}^N(\boldsymbol{h}_n^{(1)}+\boldsymbol{h}_n^{(2)})y_n}{N} - \frac{\sum_{n=1}^N(\boldsymbol{h}_n^{(1)}+\boldsymbol{h}_n^{(2)})}{N}\frac{\sum_{n=1}^N y_n}{N}\right) \\ &= \sqrt{N}\left(\frac{\sum_{n=1}^N(\tilde{\boldsymbol{h}}_n^{(1)}+\tilde{\boldsymbol{h}}_n^{(2)})\tilde{y}_n}{N} - \frac{\sum_{n=1}^N(\tilde{\boldsymbol{h}}_n^{(1)}+\tilde{\boldsymbol{h}}_n^{(2)})}{N}\frac{\sum_{n=1}^N \tilde{y}_n}{N}\right) \end{aligned} \tag{4}$$

Here $\tilde{\boldsymbol{h}}_n^{(a)} = \boldsymbol{h}_n^{(a)} - \boldsymbol{\mu}$ and $\tilde{y}_n = y_n - \mu_y$, $n \in \{1, \ldots, N\}, a \in \{1, 2\}$.

Now using the Central Limit Theorem and the fact that under the null, the phenotype y_n and the haplotypes $(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)})$ are independent:

$$\sqrt{N}\frac{\sum_{n=1}^N(\tilde{\boldsymbol{h}}_n^{(1)}+\tilde{\boldsymbol{h}}_n^{(2)})\tilde{y}_n}{N} \xrightarrow{d} \mathcal{N}(\mathbf{0}, 2\sigma_y^2\boldsymbol{\Sigma}) \tag{5}$$

$$\sqrt{N}\frac{\sum_{n=1}^N \tilde{y}_n}{N} \xrightarrow{d} \mathcal{N}(0, \sigma_y^2) \tag{6}$$

$$\frac{\sum_{n=1}^N(\tilde{\boldsymbol{h}}_n^{(1)}+\tilde{\boldsymbol{h}}_n^{(2)})}{N} \xrightarrow{p} \mathbf{0} \tag{7}$$

$$\tag{8}$$

$$\sqrt{N}\frac{\sum_{n=1}^N(\tilde{\boldsymbol{h}}_n^{(1)}+\tilde{\boldsymbol{h}}_n^{(2)})}{N}\frac{\sum_{n=1}^N \tilde{y}_n}{N} \xrightarrow{p} \mathbf{0} \tag{9}$$

This follows from Equations 6 and 7 by application of Slutsky's lemma and the continuous mapping theorem [7].

Thus, again applying Slutsky's lemma to Equations 5 and 9, we can write $\boldsymbol{a}_N$ as (Equation 4):

$$\boldsymbol{a}_N \xrightarrow{d} \mathcal{N}(\mathbf{0}, 2\sigma_y^2\boldsymbol{\Sigma}) \tag{10}$$

$$\sqrt{2N}\frac{\sum_{n=1}^N(\boldsymbol{h}_n^{(1)}+\boldsymbol{h}_n^{(2)})y_n}{2N} \xrightarrow{d} \mathcal{N}(\mu_y\boldsymbol{\mu}, \sigma_y^2\boldsymbol{\Sigma}) \tag{11}$$

$$\tag{12}$$

Let $\hat{\sigma_y^2}$ be a consistent estimator of σ_y^2. Similarly $\hat{\sigma}_j^2 = \frac{\sum_{n=1}^N\left(h_{n,j}^{(1)}+h_{n,j}^{(2)}-2\overline{h}_j\right)^2}{2N}$ is a consistent estimator of Σ_{jj}.

Let $\boldsymbol{\Lambda} = diag\left(\frac{1}{\hat{\sigma}_1}, \ldots, \frac{1}{\hat{\sigma}_{(U+O)}}\right)$ be a diagonal matrix with diagonal entries corresponding to the inverse of the standard deviation of the genotype at each SNP. We can then derive the asymptotic distribution of the asymptotic statistics as :

$$\boldsymbol{s} = \frac{1}{\sqrt{2\hat{\sigma}_y}}\Lambda\boldsymbol{a}_N \tag{13}$$

The entry corresponding to SNP j is the association statistic for SNP j:

$$s_j = \sqrt{2N}\frac{\frac{\sum_{n=1}^{N}(h_{n,j}^{(1)}+h_{n,j}^{(2)})y_n}{2N} - \frac{\sum_{n=1}^{N}(h_{n,j}^{(1)}+h_{n,j}^{(2)})}{2N}\frac{\sum_{n=1}^{N}y_n}{N}}{\frac{\sum_{n=1}^{N}({h_{n,j}^{(1)}}^2+{h_{n,j}^{(2)}}^2)}{2N} - \left(\frac{\sum_{n=1}^{N}(h_{n,j}^{(1)}+h_{n,j}^{(2)})}{2N}\right)^2} \tag{14}$$

We then have

$$\begin{aligned} \boldsymbol{s} &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_S) \\ &= \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}) \end{aligned} \tag{15}$$

The asymptotic covariance matrix of the association statistics $\boldsymbol{s}$ under the Li-Stephens model is given by

$$\Sigma_{S,ij} = \begin{cases} (1-\theta)^2 + \frac{\theta}{2}\left(1-\frac{\theta}{2}\right)\frac{1}{\sigma_i^2}, i = j \\ \exp(-\frac{\rho_{ij}}{2N})\Sigma_{ij} \end{cases}, i \neq j \tag{16}$$

Here Σ_{ij} is the LD or the correlation between SNPs i and j. Thus, the association statistics computed using genotypes imputed using a HMM follows a multivariate normal distribution with mean zero and covariance matrix equal to a LD matrix with shrinkage applied according to the recombination rate between SNPs.

2 Impact of summary statistic imputation on the rejection boundary

In general, summary statistic imputation uses a linear function of the observed statistics $\boldsymbol{s}_O$ to impute the statistic at the unobserved SNP U_i: $\hat{s}_i(\hat{\boldsymbol{s}}_O) = \boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{s}}_O$ for some weight vector $\boldsymbol{w}$.

We choose thresholds t for rejecting statistics at each of the observed and imputed SNP, *i.e.*, we reject the null hypothesis at observed SNP O_j if $|s_{O_j}| > t$ while we reject the null hypothesis at unobserved SNP U_i if $|s_{U_i}| > t$ where t is chosen to control the FWER. The acceptance region (the region where all statistics are accepted so that there are no false positives) is the vector of O values of the observed statistics that satisfies the following constraints:

$$|s_{O_j}| \leq t \quad j \in \{1, \ldots, O\} \tag{17}$$

$$|\boldsymbol{w}\boldsymbol{s}_O| - t \leq 0 \tag{18}$$

Equations 17 together define a O-dimensional hyper-cube with vertices defined by $(\pm t, \ldots, \pm t)$ while equation 18 defines two hyperplanes $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{s}_O - t = 0$ and $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{s}_O + t = 0$.

Theorem 1. *Unobserved SNP U_i that is imputed using the linear function $\hat{s}_i(\hat{\boldsymbol{s}}_O) = \boldsymbol{w}^T\hat{\boldsymbol{s}}_O$ will alter the rejection boundary iff $\sum_j |w_j| > 1$.*

Proof. For an untagged variant snp_i to have an effect on the rejection boundary, the two hyperplanes that define the imputed statistic at SNP i : $|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{s}_O| - t \leq 0$ must intersect the O-dimensional hypercube with vertices defined by $(\pm t, \pm t, \ldots, \pm t)$. This occurs iff there exists a vertex of the hypercube and the origin lie on different sides of the hyperplane. Given a hyperplane defined by the equation $h(\boldsymbol{x})) = 0$ and a point $\boldsymbol{x}_0$, $h(\boldsymbol{x}_0)$ is proportional to the signed distance of $\boldsymbol{x}_0$ from the hyperplane. Denoting a vertex of the hypercube as $t\boldsymbol{x}$ where $\boldsymbol{x} \in \{-1, +1\}^O$, the above condition is equivalent to the product of signed distances of the origin and one of the vertices being negative, *i.e.*, there exists a $\boldsymbol{x} \in \{-1, +1\}^O$ such that

$$\begin{aligned} &\left(\boldsymbol{w}^{\mathrm{T}}(t\boldsymbol{x}) - t\right)\left(\boldsymbol{w}^{\mathrm{T}}(\mathbf{0}) - t\right) < 0 \\ &\Rightarrow -\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \mathbf{1}\right)t^2 < 0 \\ &\Rightarrow \left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \mathbf{1}\right) > 0 \end{aligned} \tag{19}$$

Equation 19 holds iff $\sum_j |w_j| > 1$. □

3 A comparison of the power of summary statistic imputation methods

Consider three SNPs, snp_1, snp_2, snp_3 where SNPS snp_1 and snp_2 are observed while snp_3 is unobserved. Assuming that SNP snp_3 is the causal SNP with non-centrality parameter λ, the summary statistics at the three SNPs $(s_1, s_2, s_3))$ follow the distribution:

$$\mathcal{N}\left(\begin{bmatrix} \lambda\alpha \\ \lambda\alpha \\ \lambda \end{bmatrix}, \begin{bmatrix} 1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix}\right) \tag{20}$$

For this distribution to be well-defined, the covariance matrix must be positive-definite. A necessary condition for this is that the determinant of the covariance matrix (which is equal to the product of the eigenvalues) must be positive. Thus, we require the determinant $(1-\rho)(1+\rho-2\alpha^2) > 0$. $|\rho| < 1$ for the marginal distribution over SNPs 1 and 2 to represent a valid distribution. Further, we require $\alpha^2 < \frac{1+\rho}{2}$.

Consider the case of SSI, *i.e.*, summary statistic imputation (with no variance re-weighting). In this case,

$$\hat{s}_3 = \frac{\alpha}{1+\rho}(s_1 + s_2) \tag{21}$$

The mean, the variance and the coefficient of variation of $\hat{s}_3$ can be computed under the distribution (Equation 20):

$$\begin{aligned} \mu_1 &= \lambda\frac{2\alpha^2}{1+\rho} \\ \sigma_1^2 &= \frac{2\alpha^2}{1+\rho} \\ c_{v,1} &= \frac{\mu_1}{\sigma_1} \\ &= \lambda\alpha\sqrt{\frac{2}{(1+\rho)}} \end{aligned} \tag{22}$$

Now consider the case of SSI-VR, *i.e.*, summary statistic imputation (with variance re-weighting). In this case,

$$\hat{z}_3 = \frac{1}{\sqrt{2(1+\rho)}}(s_1 + s_2) \tag{23}$$

The mean, the variance and the coefficient of variation of $\hat{z}_3$ can be computed under the distribution (Equation 20):

$$\begin{aligned} \mu_2 &= \lambda\frac{\sqrt{2}\alpha}{\sqrt{(1+\rho)}} \\ \sigma_2^2 &= 1 \\ c_{v,2} &= \frac{\mu_2}{\sigma_2} \\ &= \lambda\alpha\sqrt{\frac{2}{(1+\rho)}} \end{aligned} \tag{24}$$

The power to reject the null hypothesis at a threshold t is

$$\beta_1(t) = P(|s_1| > t \cup |s_2| > t \cup |\hat{s}_3| > t) \tag{25}$$

The power function depends on the joint distribution of s_1, s_2, and $\hat{s}_3$. To simplify this expression, we will analyze the average power, a notion that is easier to analyze than the power as defined in Equation 25:

$$\begin{aligned} \gamma_1(t) &= \frac{P(|s_1| > t) + P(|s_2| > t) + P(|\hat{s}_3| > t)}{3} \\ &= \frac{2f(\lambda\alpha, t) + f(\lambda\alpha\sqrt{\frac{2}{(1+\rho)}}, t\sqrt{\frac{1+\rho}{2\alpha^2}})}{3} \end{aligned} \tag{26}$$

Here f is the power function defined in Section A.

Analogously, we can compute the average power γ_2 for SSI-VR.

$$\begin{aligned} \gamma_2(t) &= \frac{P(|s_1| > t) + P(|s_2| > t) + P(|\hat{z}_3| > t)}{3} \\ &= \frac{2f(\lambda\alpha, t) + f(\lambda\alpha\sqrt{\frac{2}{(1+\rho)}}, t)}{3} \end{aligned} \tag{27}$$

Comparing $\gamma_1(t)$ and $\gamma_2(t)$ allows us to understand the power of the two imputation methods. To analyze the power of each method, we need to do so at a threshold t such that each method attains the same FWER.

If t_1 and t_2 are the thresholds for SSI and SSI-VR, we have $t_1 \leq t_2$. Thus, comparing Equations 26 and 27, we see that the first term for $\gamma(t_1)$ is greater than the first term for $\gamma(t_2)$. The second term for $\gamma(t_1)$ will be greater than or equal to the second term for $\gamma(t_2)$ if $t_1\sqrt{\frac{1+\rho}{2\alpha^2}} < t_2$. Thus, when the correlation between the unobserved and observed SNPs is close to its maximum possible value, *i.e.*, $\frac{2\alpha^2}{1+\rho} \approx 1$, then SSI-VR has lower power than SSI.

A Power function

Given a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, the probability that the absolute value of X exceeds a given threshold t:

$$\begin{aligned} P(|X| > t) &= \Phi(-c_v - \frac{t}{\sigma}) + \Phi(c_v - \frac{t}{\sigma}) \\ &\equiv f(c_v, \frac{t}{\sigma}) \end{aligned} \tag{28}$$

where $c_v = \frac{\mu}{\sigma}$ is the coefficient of variation and Φ is the normal CDF.

Note that f increases with c_v and σ and decreases with t.

References

1. P Scheet and M Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, 78(4), 04 2006.

2. J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39:906–913, 2007.

3. S.R. Browning and B.L. Browning. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097, 2007.

4. Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

5. Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003.

6. Xiaoquan Wen and Matthew Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158, 2010.

7. Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.