Summary statistics estimation based on LD structure

2017/06/06

Yosuke Tanigawa (ytanigaw@stanford.edu)
Biomedical Informatics Ph.D. Program

Motivation: summary statistics imputation from LD

- We have population-level genomic data
  - Linkage-disequilibrium (LD) structure

- GWAS summary statistics
  - Multi-institutional data

- Can we solve the following inference problem?
  - Input: GWAS summary stats for limited loci, LD
  - Output: GWAS summary stats for loci of interest

Model

Notation

- $X \in \mathbb{R}^{N \times M}$: genotype. We assume X is normalized such that is has zero-mean and unit variance.
- $Y \in \mathbb{R}^{N \times 1}$: traits. We assume Y is normalized such that is has zero-mean and unit variance.
- $\beta \in \mathbb{R}^{M}$: effect size.
- N: number of individuals.
- M: number of SNP markers.
- $V \in \mathbb{R}^{M \times M}$: LD matrix. This can be found by $V = X^T X$.

Model

Linear regression model

Our regression model is:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, I) \tag{1}$$

Least square and marginal effect model

Least square

$$\nabla_\beta (Y - X\beta)^T (Y - X\beta) = -X^T (Y - X\beta) = 0 \tag{2}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = V^{-1} X^T Y; \quad \text{Var}\left[\hat{\beta}\right] = \sigma_j^2 V^{-1} \tag{3}$$

where, σ_j^2 is residual variance.

Marginal effect

$$\hat{\beta}_M = D^{-1} X^T Y; \quad \text{Var}\left[\hat{\beta}_M\right] = \sigma_M^2 D^{-1} \tag{4}$$

where, D is the diagonal matrix of V.

The relationship between two models

Since we have $V\hat{\beta} = X^T Y = D\hat{\beta}_M$, we have

$$\hat{\beta} = V^{-1} D \hat{\beta}_M \tag{5}$$

Z-score

We define z-score:

$$Z := \frac{\hat{\beta}_M}{\sqrt{\text{Var}\left[\hat{\beta}_M\right]}} = \frac{X^T Y}{\sqrt{N}} \tag{6}$$

We assume

$$Z \sim N(0, V) \tag{7}$$

Imputation of Z-scores

Let's consider to divide Z into two blocks:

1. Z_t: Z-score for typed SNPs
2. Z_i: Z-score for untyped SNPs

i.e.

$$Z^T = (Z_t^T \quad Z_i^T); \quad V = \begin{pmatrix} V_{tt} & V_{ti} \\ V_{it} & V_{ii} \end{pmatrix} \tag{8}$$

Since we modeled the Z-scores as multi-variate normal, the conditional distribution $p(Z_i \mid Z_t)$ is also normal:

$$Z_i \mid Z_t \sim N(V_{it} V_{tt}^{-1} Z_t, V_{ii} - V_{it} V_{tt}^{-1} V_{ti}) \tag{9}$$

Z-score to the summary statistics

Z-scores to summary statistics

We can convert our estimate of z-score into estimate of marginal effect size $\hat{\beta}_M$

$$\hat{\beta}_M = Z_i \sigma_Z + \mu_Z \tag{10}$$

We can estimate the variance of marginal effects by allele frequency (f):

$$\text{Var}\left[\hat{\beta}_M\right] = \sqrt{2Nf(1-f)} \tag{11}$$

We can use these estimates in Eq. (5).

Dataset description

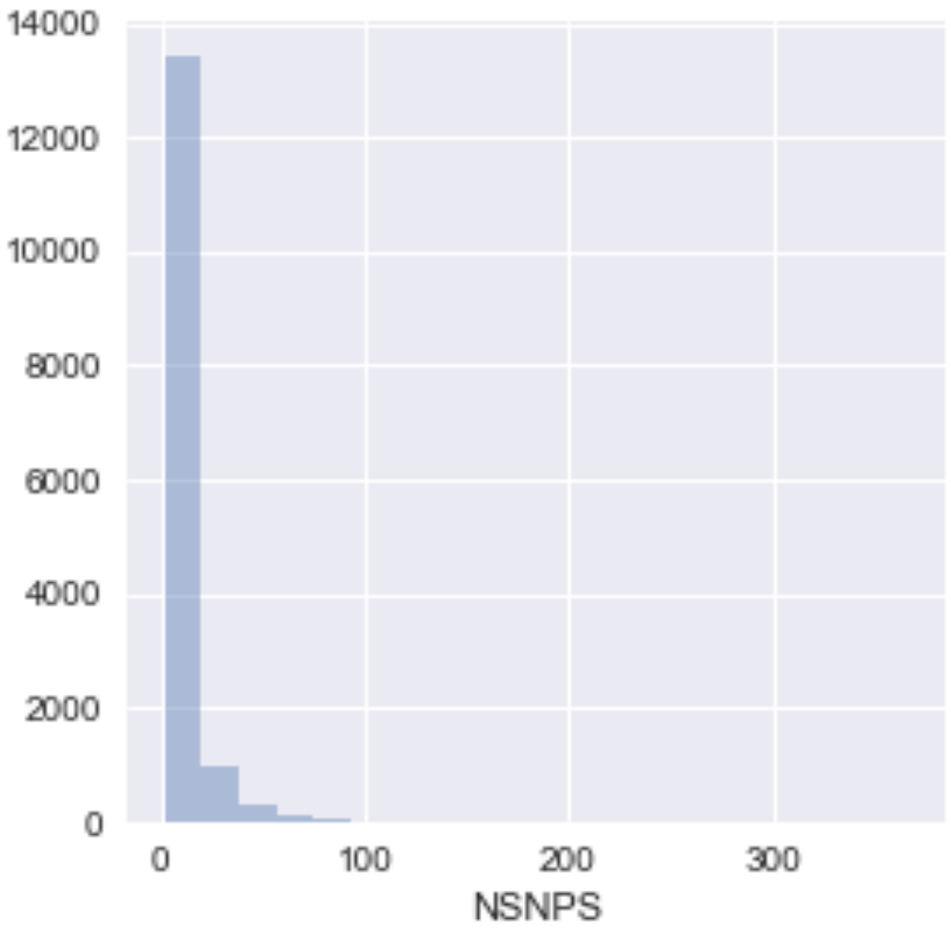
- Genotype info:
  - UKBB (with population stratification): 112,338 individuals
  - Focusing on chromosome 20

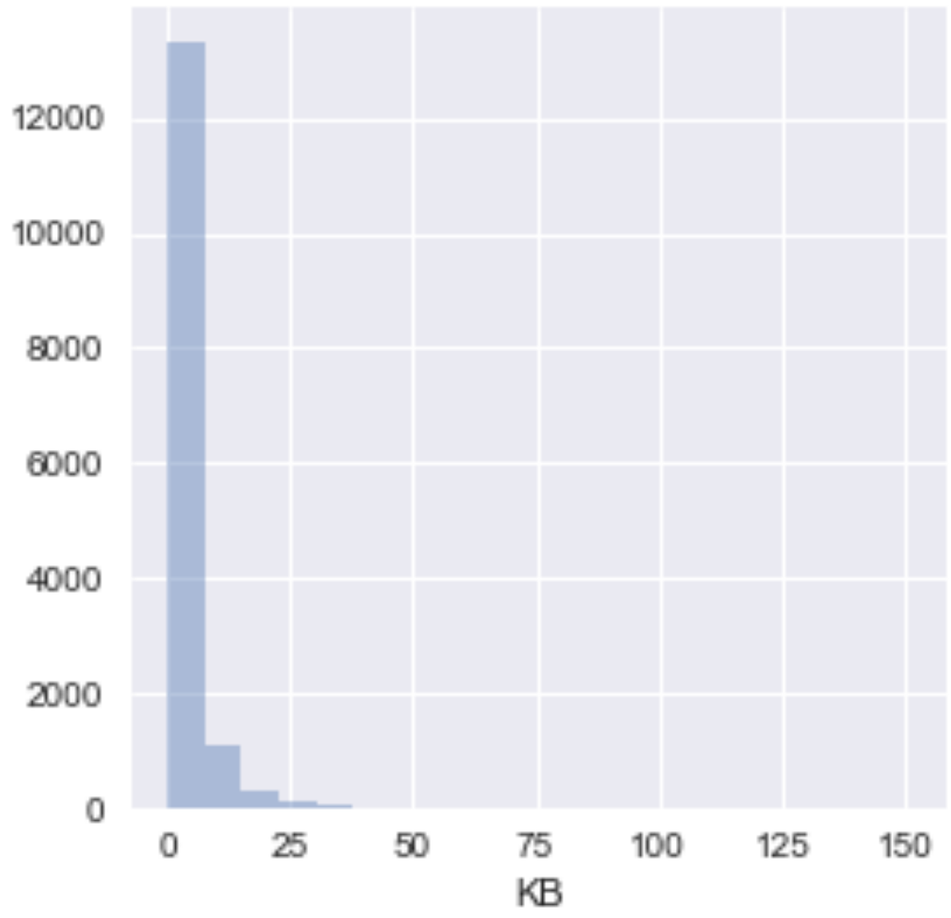
- LD block (plink)
  - --blocks no-pheno-req
  - --blocks-max-kb 1000
  - --blocks-min-maf .05

- GWAS summary statistics
  - ADD, age, sex, C1-C4 (first 4 principle components)
  - Focusing on ADD (additive effects)
  - GWAS from array data
  - GWAS from inputed genotype data

LD block structure on chromosome 20

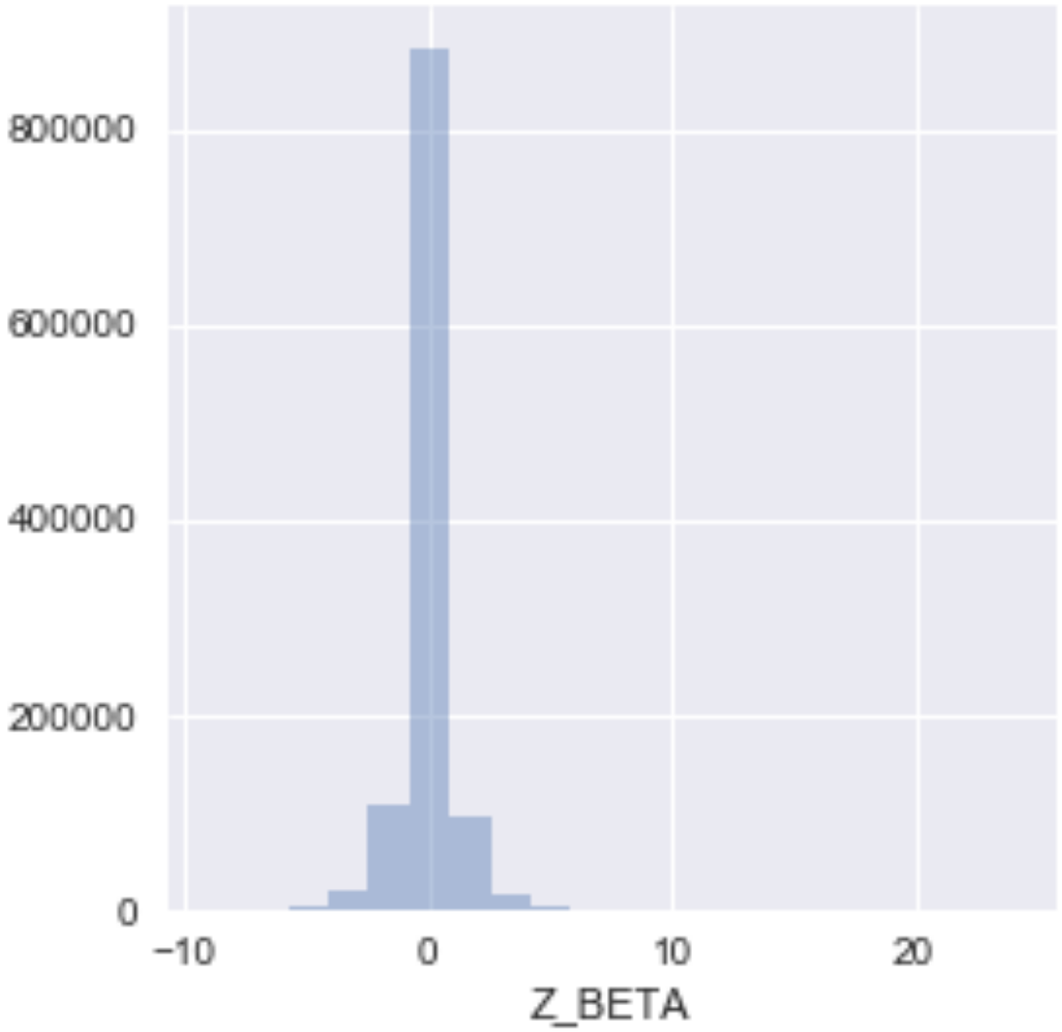
- (left) Number of SNPs in a LD block (median 4.0)
  - Note: MAF 5%
- (right) Size of LD block (median 1.1865)

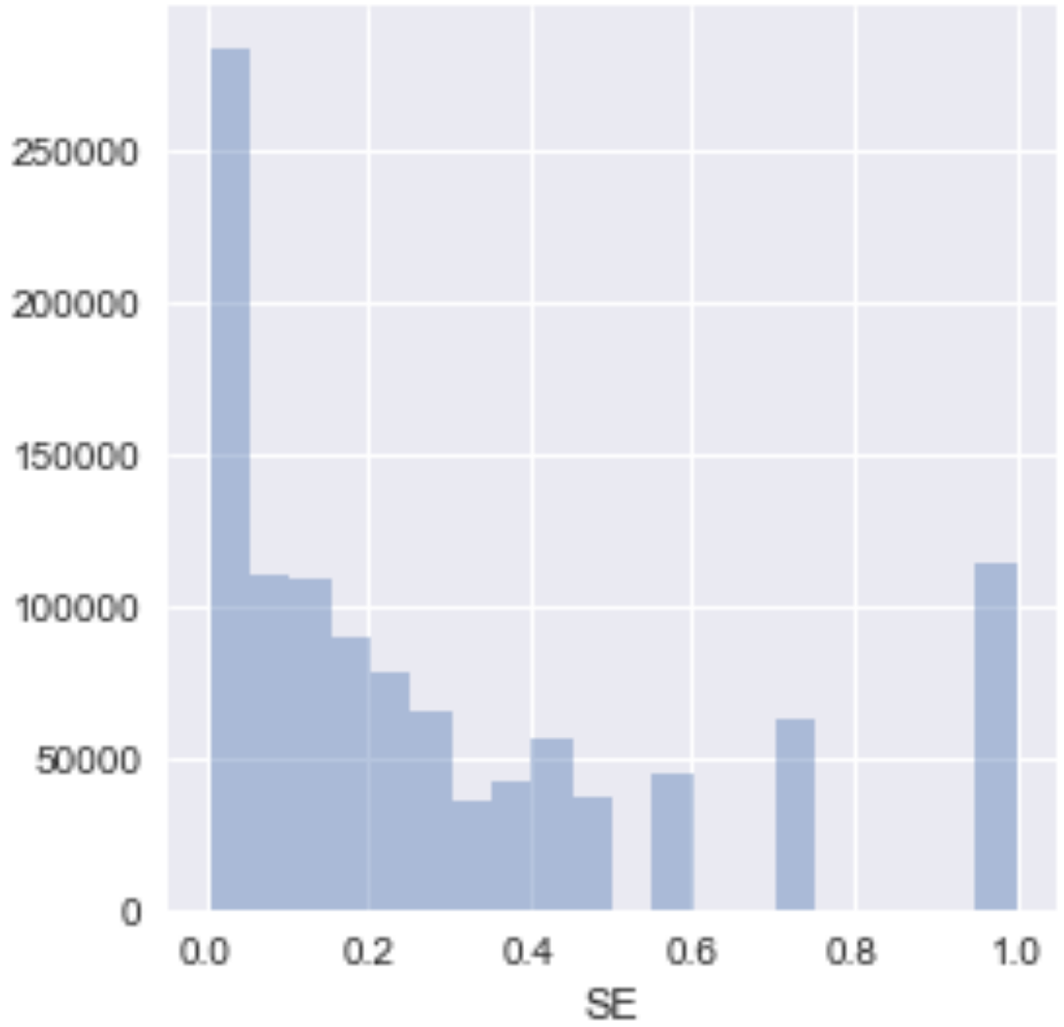




Z-score distribution (for effect size for imputed data)

- Zero-mean and unit-variance normalization for Z-score





Examples of LD blocks:

```
In [5]: df_block.loc[[9, 40]]
```

Out[5]:

	CHR	BP1	BP2	KB	NSNPS	SNPS
9	20	98930	102181	3.252	5	rs6116135|rs6116236|rs6139361|rs71870630|rs605...
40	20	170642	171189	0.548	2	rs1469781|rs6078096

Example 1: LD block chr20:rs6116135-rs6052493

LD #9

```
In [6]: df_beta_select_range(df_beta_i, df_block.BP1[9], df_block.BP2[9])
```

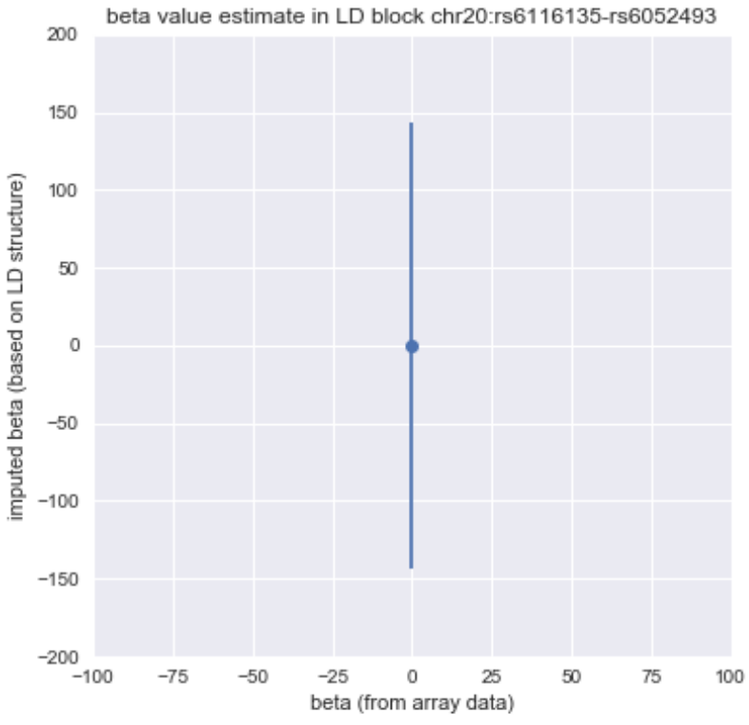
Out[6]:

	#CHROM	POS	ID	REF	ALT1	TEST	OBS_CT	BETA	SE	T_STAT	P	ALT_FREQS
7	20	101362	Affx-16857388	2	1	ADD	104655	0.00086	0.003512	0.244958	0.80649	0.762089

```
In [7]: df_beta_select_range(df_beta_t_zscore, df_block.BP1[9], df_block.BP2[9])
```

Out[7]:

	Unnamed: 0	#CHROM	POS	ID	Z_BETA	ALT_FREQS
54	927	20	98930	rs6116135	-0.020121	0.925180
55	972	20	100699	rs6116236	-0.000057	0.851703
56	986	20	101362	rs6139361	-0.026851	0.762089
57	1005	20	102080	rs71870630	-0.000697	0.701485
58	1007	20	102181	rs6052493	-0.003476	0.824852



- beta_hat = 0.004614
- beta_si = 143.3
- beta (array) = 0.0008603
- condition_num(V_tt) = 26.83

Example 2: LD block chr20:rs1469781-rs6078096

LD #40

```
In [10]: df_beta_select_range(df_beta_i, df_block.BP1[40], df_block.BP2[40])
```

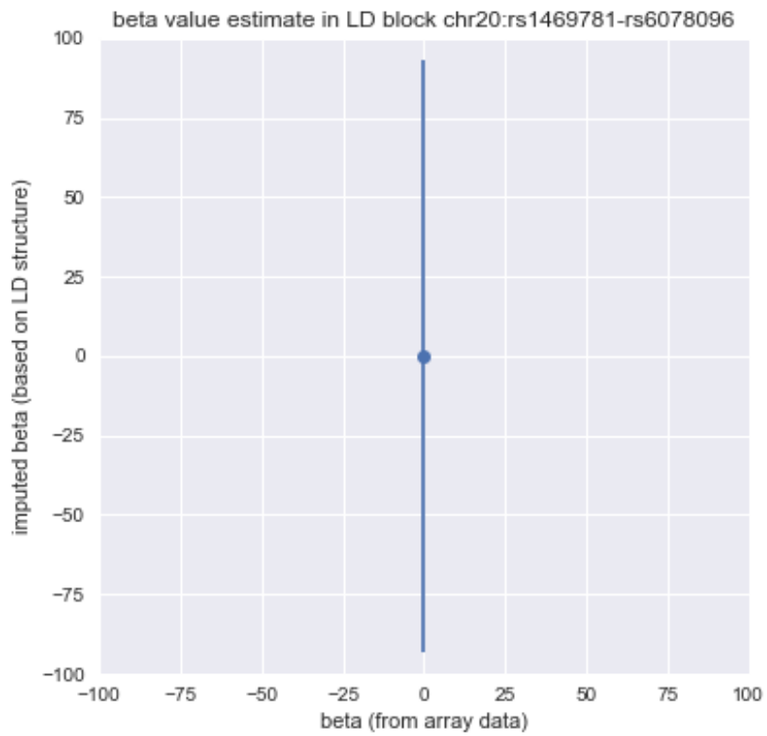
Out[10]:

	#CHROM	POS	ID	REF	ALT1	TEST	OBS_CT	BETA	SE	T_STAT	P	ALT_FREQS
20	20	170642	Affx-16356907	1	2	ADD	109894	-0.001834	0.00296	-0.619535	0.535565	0.441836

```
In [11]: df_beta_select_range(df_beta_t_zscore, df_block.BP1[40], df_block.BP2[40])
```

Out[11]:

	Unnamed: 0	#CHROM	POS	ID	Z_BETA	ALT_FREQS
296	2904	20	170642	rs1469781	-0.019409	0.441836
297	2923	20	171189	rs6078096	-0.012905	0.554101



- beta_hat = -0.002247
- beta_si = 92.96
- beta (array) = -0.001834
- condition_num(V_tt) = 1.0

Discussion: What should be refined?

- The variance of estimate is huge
- Numerical instability
  - Inverse operation, condition number, ..
- Missing values in genetic information
  - currently: mean imputation
- MAF >= 0.05 threshold