# Genome-wide association and sequencing studies

**Pubh 8446**

**3/2/2020**

# Logit factor analysis (LFA) and inverse regression for association test (GCAT)

- Refs
  1. Hao W, Song M, Storey JD. Probabilistic models of genetic variation in structured populations applied to global human studies. Bioinformatics. 2016 Mar 1;32(5):713–21.
  2. Gopalan P, Hao W, Blei DM, Storey JD. Scaling probabilistic models of genetic variation to millions of humans. Nat Genet. 2016 Dec;48(12):1587–90.
  3. Song M, Hao W, Storey JD. Testing for genetic associations in arbitrarily structured populations. Nat Genet. 2015 May;47(5):550–4.
- Given genotype matrix $G = (g_{ij})$ for sample $i = 1, \ldots, n$; and marker $j = 1, \ldots, m$
  - model $g_{ij} \sim Binom(2, \theta_{ij})$, where $\theta_{ij}$ is the marker MAF
  - approx $[logit(\theta_{ij})] = \Gamma F$
    - $\Gamma$: latent factors/variables (nxd)
    - $F$: coefficient/loading matrix (dxm)
    - latent dimension $d \ll n, m$
  - In downstream analyses, say, adjuting for population stratification
    - treat fitted $(\hat{\theta}_{ij}) = [\exp(\hat{\Gamma}\hat{F})]$ as the population stratification component (uni-covariate)
      - smaller number of params, potentially leading to more power
    - Treat the latent factors $\Gamma$ as $d$ ancestry covariates to be adjusted in any statistical models
      - more flexible
  - compared to PCA approx: $G \approx L = \Gamma F$
  - model identifiability: need some constraint on $\Gamma$ and $F$, say, orthogonality
- Estimations
  - PCA estimation: $\min_{\Gamma, F} \|G - \Gamma F\|^2$
    - bilinear regressions: row and column wise LS
  - (composite) likelihood maximization
    - $\max_{\Gamma, F} \sum_{i,j} \Pr(g_{ij}|\theta_{ij})$
    - computationally intensive ($n \sim 10^5, m \sim 10^5$)
  - 2-step PCA approx
    1. PCA/SVD of $G$ to approx $\hat{G}$, used to form $\hat{L} = logit(\hat{G})$ (thresholded to [0,1])
    2. PCA/SVD of $\hat{L} \approx \Gamma F$
    3. Fix $\Gamma$, fit logit model to estimate/update $F$.
       - each column of $F$ can be estimated by fitting a logit model for each marker (column of $G$).
       - Note, $L_{.j} = \sum_{k=1}^{K} \Gamma_{.k}\gamma_{kj}$

# Genotype-conditional association test (GCAT)

- Consider a SNP with genotypes $G = (g_1, \ldots, g_n)$, marginally following $Binom(2, \theta)$
- Given a quantitative or binary trait $Y$,
  - outcome model $Y \sim N(\alpha + G\beta, \sigma^2)$ or $logit[\Pr(Y = 1|G)] = \alpha + G\beta$
    - interested in testing $H_0 : \beta = 0$
- GCAT
  - consider the following GCAT model, $(G|Y) \sim Binom(2, \tau)$
    - $logit(\tau) = \nu + Y\gamma + logit(\theta)$
  - can show that $\beta = 0$ implies $\gamma = 0$
    - so an equivalent test $H_0 : \gamma = 0$
    - check LM and Logit model (TBD)
  - modeling $logit(\theta) = \Gamma\Delta$
    - leverage information across all SNPs to estimate $\Gamma$
    - treat $\Delta$ as params to be estimated jointly with other parameters.

# Association test of imputed SNPs

- Previous model implicitly assumes we directly observe $G$, taking values in $\{0, 1, 2\}$.
- For those not directly genotyped SNPs (not in the chip)
  - we can predict/impute their genotypes pretty accurately by leveraging the local LD and the powerful HMM (TBD)
  - we obtain imputation prob instead, $(p_0, p_1, p_2)$
  - here $\sum_j p_j = 1$, telling us the relative prob of observing each genotype for the imputed SNP
- How to test imputed SNPs?
  - use the best-guess genotypes, $\hat{G} = \arg\max_j p_j$
    - some obvious loss of information
  - how to best account for the imputation scores? (TBD)
    - score functions/imputation scores/GEE
    - quasi-likelihood

# Statistical test in binomial regression model

- Consider e.g. a logistic regression model $logit[\Pr(Y = 1|G)] = \alpha + G\beta$
  - interested in testing $H_0 : \beta = 0$
- Three asymptotically equivalent tests
  - Wald test
    - $\hat{\beta}^2 / \widehat{Var}(\hat{\beta})$, asymptotically a $\chi^2_1$ rv
    - asymptotic covariance from the Fisher information matrix (computed under the MLE)
  - Score test
    - score function $U = G^T(Y - \hat{\theta}_0)$
      - $\hat{\theta}_0$ estimated event prob under the null
    - asymptotic covariance from the Fisher information matrix (computed under the null model)
    - $U^2 / \widehat{Var}(U) \sim \chi^2_1$ under the null
  - LRT

        - fit two models, under $\beta = 0$ and without constraint
        - compute the LRT (likelihood ratio test), asymptotically a $\chi^2_1$ rv
- Wald test
    - computed under the MLE (potential problems due to quasi-separation etc)
    - degenerate power as $n \to \infty$
    - need large sample size $n$ and work well for small-dim test problems ($p \ll n$)
- Score test
    - computed under the null
    - well-behaved, though slightly conservative
    - computationally convenient
    - work well even when dimension of $G$ $p \gg n$!
- LRT
    - Fit two models under the null and alternative
    - computationally intensive
    - well-calibrated and generally the most powerful
    - need large sample size $n$ and work well for small-dim test problems ($p \ll n$)