Single-step

Ole F. Christensen

Aarhus University, Center for Quantitative Genetics and Genomics

GenSAP meeting 2014 in Korsør, day3

Single-step: theory

- A joint model for genomic value and marker genotypes (coded 0, 1, 2):

$$g = \sum_j (M_j - 2\rho_j)\beta_j$$

$$\mathrm{E}[M_j] = 2\rho_j 1, \quad \mathrm{Var}[M_j] = v_j A$$

(based on idea by Gengler et al. 2007 to infer missing genotypes).

- Assume M is multivariate normal

- Individuals with missing and observed genotypes

$$M = \begin{bmatrix} M^{miss} \\ m^{obs} \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

Single-step - extension of G to non-genotyped individuals

- Marginalisation (integration) of M^{miss} gives

$$\mathrm{E}[g \mid m^{obs}] = 0,$$

$$\mathrm{Var}[g \mid m^{obs}] = \sigma_g^2 \begin{bmatrix} A_{12}A_{22}^{-1}G & A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} + A_{11} \\ G & GA_{22}^{-1}A_{21} \end{bmatrix}$$

$$= \sigma_g^2 H.$$

where $G = \sum_j (m_j^{obs} - 2\rho_j 1)(m_j^{obs} - 2\rho_j 1)^T / \sum_j v_j$ and $\sigma_g^2 = \sum_j v_j \sigma_\beta^2$.

Sparse inverse

$$H^{-1} = \begin{bmatrix} 0 & 0 \\ G^{-1} - A_{22}^{-1} & 0 \end{bmatrix} + A^{-1}.$$

Single-step: a practical problem

- $$G = \sum_j (m_j^{obs} - 2\rho_j 1)(m_j^{obs} - 2\rho_j 1)^T / \sum_j v_j$$

  where ρ_j is allele frequency in base population, $s = \sum_j v_j$ is a scaling.

- These are unknown!

- Allele frequencies: in principle both phenotypes and marker genotypes provide information about these. Inferring them is computationally challenging, and in addition there will be uncertainty in such estimation.

- Practical solution: use observed allele freq $\hat{\rho}_j = 1^T m_j^{obs}/n$ and adjust G towards A_{22},

$$G_{adjust} = G\beta + 11^T\alpha$$

Single-step: SNP-model

- $$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} A_{12}A_{22}^{-1}\sum_j (m_j^{obs} - 2\rho_j)\beta_j \\ \sum_j (m_j^{obs} - 2\rho_j)\beta_j \end{bmatrix} + \begin{bmatrix} \epsilon \\ 0 \end{bmatrix}$$

  where $\text{Var}(\epsilon) = A_{11} - A_{12}A_{22}^{-1}A_{21}$.

- Substituting $\rho = \hat{\rho} + (\rho - \hat{\rho})$, then

  $$g = \begin{bmatrix} A_{12}A_{22}^{-1} \\ 1 \end{bmatrix} \mu_g + \begin{bmatrix} A_{12}A_{22}^{-1}\sum_j (m_j^{obs} - 2\hat{\rho}_j 1)\beta_j \\ \sum_j (m_j^{obs} - 2\hat{\rho}_j 1)\beta_j \end{bmatrix} + \begin{bmatrix} \epsilon \\ 0 \end{bmatrix}$$

  where $\mu_g = 2\sum_j (\hat{\rho}_j - \rho_j)\beta_j$.

- μ_g related to genetic drift

Single-step: genetic drift

- $\mu_g = 2\sum_j(\hat{\rho}_j - \rho_j)\beta_j$ (genetic drift)
- Fixed effect μ_g, and inferred from phenotypes (Fernando et al. 2014, Vitezica et al. 2011)
- Proper prior on μ_g,

$$\text{Var}(\mu_g 1 + (m^{obs} - 2\hat{\rho}1^T)\beta) = \sigma^2_{\mu_g} 11^T + \sigma^2_g G$$

  and infer $\sigma^2_{\mu_g}$ based on matching to A_{22}, i.e. infer from marker genotypes (Vitezica et al. 2011).

- Important point: adjustments like $G_{adjust} = G\beta + 11^T\alpha$ are related to genetic drift.

Adjusting G to A - some theory

(Powell et al. 2010, Vitezica et al. 2011, Meuwissen et al. 2011)

- $G_{adjust} = G(1 - \alpha/2) + 11^T\alpha$
- G reflects relationships relative to genotyped individuals.
- Idea: translate relationships such they are relative to base population
- Note: assumption of random assignment of alleles in each population, i.e. theory does not really support a population across generations.

Adjusting A to G - some theory

(Christensen, 2012)

- Alternatively: adjust A to G.
- $A^\gamma = A(1-\gamma/2) + \gamma 11^T$ is relationship matrix with base individuals being related and inbreed. Computing formulas are as usual.
- Estimate γ from $G_{1/2} = (m^{obs} - 1)(m^{obs} - 1)^T/s$ where s is unknown scaling parameter.
- A Bayesian derivation (prior on ρ_j and v_j), very sketchy:

$$\mathrm{E}[M_j] = 2E[\rho_j]1 = 1, \;\; \mathrm{Var}[M_j] = \mathrm{E}[v_j]A + 4\,\mathrm{Var}[\rho_j]11^T \propto A(\gamma)$$

Conclusion

- Genomic and pedigree relationships are incompatible. So adjustments are needed.
- Main problem is the allele frequencies.
- Three solutions:
  - Estimate genetic drift in model for phenotypes.
  - Adjust G: $G_{adjust} = G\beta + 11^T\alpha$
  - Adjust A: $A^\gamma = A(1-\gamma/2) + \gamma 11^T$

Topics that I did not mention:

- Additional polygenic effect.
- Multiple breeds, genetic groups
- LDLA approach (Meuwissen et al. 2011).
- Computing !
- Computing !
- Computing !