Parameter estimation in the ancestral regression with missing parental or grandparental genotypes

Rodolfo J. C. Cantet, Natalia S. Forneris

Departamento de Producción Animal,
Facultad de Agronomía
Universidad de Buenos Aires;
INPA-CONICET

CONICET

Introduction

- Animal model of Henderson-Quaas-Bulmer is a Data Generating Process with a Markovian and p.d. covariance matrix.

- Cantet et al (2017) proposed an individual based model ("ancestral regression", AR) that is autoregressive (causal), Markovian and easy to fit, but no parameter estimation.

- Model parameters (the individual´s β_S and β_D) are *identifiable only* with information from a dense set of SNPs to estimate the sufficient statistics.

- Bayesian estimators require the distribution of β_S & β_D.

- Goal: To present a Gibbs sampler for β_S and β_D.

Ancestral regression

$$a_X = 0.5\, a_S + 0.5\, a_D + \beta_S \left(a_{SS} - a_{DS}\right) + \beta_D \left(a_{SD} - a_{DD}\right) + \phi_X$$

$$\beta_S = \frac{\Sigma_{X,PGP} - \Sigma_{PGP,R}\, \Sigma_R^{-1}\, \Sigma_{R,PGP}}{\Sigma_{PGP} - \Sigma_{PGP,R}\, \Sigma_R^{-1}\, \Sigma_{R,PGP}} \qquad \beta_D = \frac{\Sigma_{X,MGP} - \Sigma_{MGP,R}\, \Sigma_R^{-1}\, \Sigma_{R,MGP}}{\Sigma_{MGP} - \Sigma_{MGP,R}\, \Sigma_R^{-1}\, \Sigma_{R,MGP}}$$

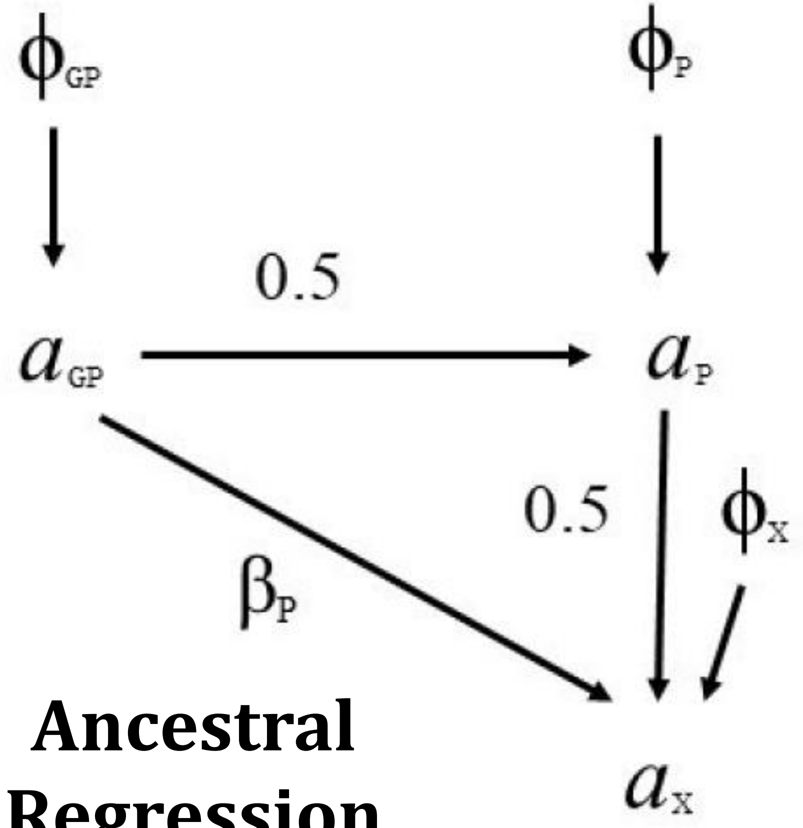
where the sufficient statistics are

$$\Sigma_{X,PGP} = 0.5\left(\Sigma_{X,SS} - \Sigma_{X,DS}\right) \qquad \Sigma_{X,MGP} = 0.5\left(\Sigma_{X,SD} - \Sigma_{X,DD}\right)$$

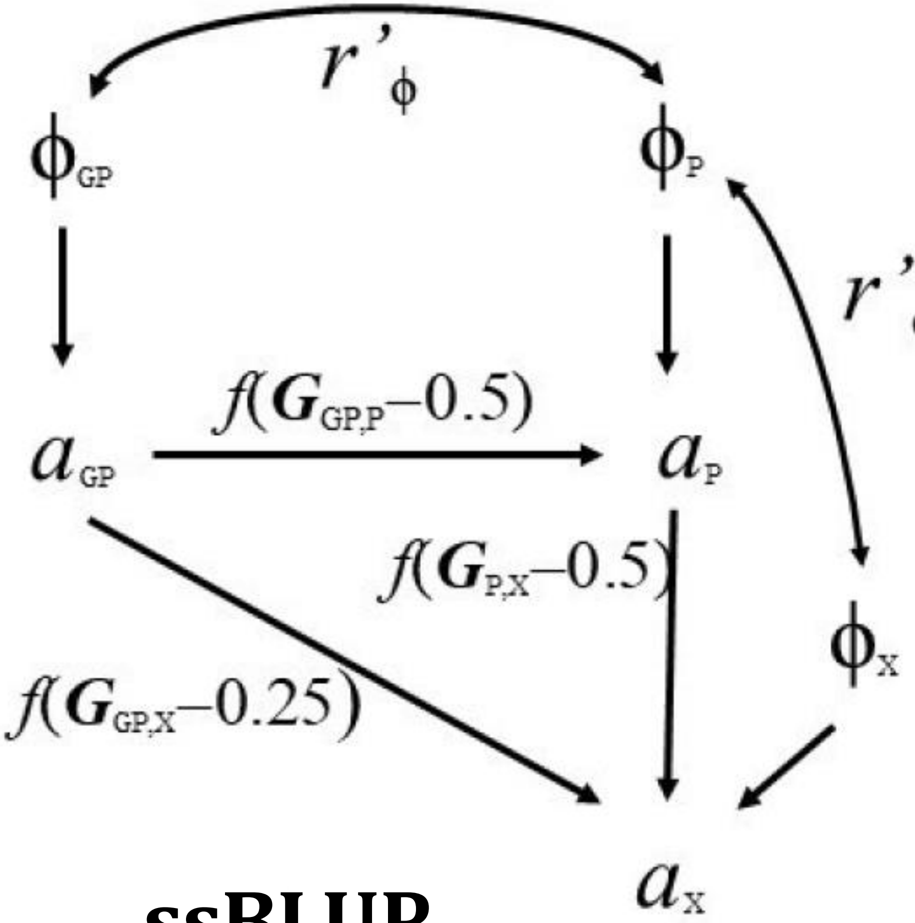
with corresponding variances

$$\Sigma_{PGP} = 0.25\left(\Sigma_{SS} + \Sigma_{DS} - \Sigma_{SS,DS}\right) \qquad \Sigma_{MGP} = 0.25\left(\Sigma_{SD} + \Sigma_{DD} - \Sigma_{SD,DD}\right)$$

Path coefficient view of the Ancestral Regression and ssBLUP



Ancestral Regression


r'_ϕ
ϕ_{GP}
ϕ_P
r'_ϕ
$f(\mathbf{G}_{GP,P}-0.5)$
a_{GP}
a_P
$f(\mathbf{G}_{P,X}-0.5)$
ϕ_X
$f(\mathbf{G}_{GP,X}-0.25)$
a_X

ssBLUP

Covariance matrix of BV under AR

$$\boldsymbol{B}_{\mathrm{X}(i)} = \begin{bmatrix} 0 \ldots \beta_{\mathrm{S}} & -\beta_{\mathrm{S}} & \ldots 0 \ldots \beta_{\mathrm{D}} & -\beta_{\mathrm{D}} & 0.5 & 0.5 \end{bmatrix} \Rightarrow \boldsymbol{B}$$

Autoregressive causal model

$$\boldsymbol{a} = \boldsymbol{B}\boldsymbol{a} + \phi \quad \Rightarrow \quad \Sigma = \left(\boldsymbol{I} - \boldsymbol{B}\right)^{-1} \boldsymbol{D} \left(\boldsymbol{I} - \boldsymbol{B}'\right)^{-1}$$

- The distribution of ***a*** is MVN (proved elsewhere), such that Σ^{-1} from an autoregressive structure, follows an inverted Wishart, and the betas (in ***B***) are standard normal (Roverato, 2000).

Covariances between BV under AR

A = Ancestor, S = Sire , SS = Sire of Sire , DS = Dam of Sire , D = Dam, SD = Sire of Dam, DD = Dam of Dam.

Covariance between an ancestor and X

$$\Sigma_{A,X} = 0.5\left(\Sigma_{A,S} + \Sigma_{A,D}\right) + \beta_S\left(\Sigma_{A,SS} - \Sigma_{A,DS}\right) + \beta_D\left(\Sigma_{A,SD} - \Sigma_{A,DD}\right)$$

Covariances between two animals, neither of whom is an ancestor of the other

$$\begin{bmatrix} \Sigma_A & \Sigma_A \boldsymbol{B}_Y' \\ \boldsymbol{B}_X \Sigma_A & \boldsymbol{B}_X \Sigma_A \boldsymbol{B}_Y' \end{bmatrix}$$

Inbreeding under AR

S = Sire ,   SS = Sire of Sire , DS = Dam of Sire ,
D = Dam,   SD = Sire of Dam, DD = Dam of Dam.

$$F_{X(AR)} =$$

$$0.5\left[\Sigma_{S,D} + \beta_{S}\left(\Sigma_{SS,D} - \Sigma_{DS,D}\right) + \beta_{D}\left(\Sigma_{SD,S} - \Sigma_{DD,S}\right) + \beta_{S}\,\beta_{D}\left(\Sigma_{SS,SD} - \Sigma_{SS,DD} - \Sigma_{DS,SD} + \Sigma_{DS,DD}\right)\right]$$

Distribution of sufficient statistics and parameters

❑ Jimenez' Thesis: Simulation to obtain the empirical distribution. Pig and beef cattle genotypes to validate the distributions obtained.

❑ Cov(Grandparent, Individual) => Beta.

❑ **Sufficient statistics:** $\Sigma_{X,PGP}$ = Cov(Grandsire, X) – Cov(Grand-dam,X) => Normal

❑ β_S, β_D => Normal. F_{AR} => Exponential.

Estimating equations

$$\boldsymbol{L}\,\boldsymbol{B}_{\mathrm{X}}\,' = \boldsymbol{\beta} = \begin{bmatrix} \beta_{\mathrm{S}} \\ \beta_{\mathrm{D}} \end{bmatrix}$$

$$L' = \begin{bmatrix} 0.5 & 0 \\ -0.5 & 0 \\ 0 & 0.5 \\ 0 & -0.5 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

If all grandparents and parents are genotyped, betas are the solution of the system with 2 equations:

$$L\,\Sigma_{\mathrm{A}}\,L'\,\beta \;=\; L\,\Sigma_{\mathrm{AX}}$$

$$\mathrm{Var}\left(\phi_{\mathrm{X}}\right) = \left[1 + F_{\mathrm{X(AR)}} - L\,B_{\mathrm{X}}\,\Sigma_{\mathrm{A}}\,B_{\mathrm{X}}\,'L'\right]\sigma^2_{\mathrm{A}}$$

Gibbs sampling algorithm for β_S & β_D

1) Estimation of IBD relationships (data) with an algorithm that accounts for a) Pedigree, b) Inbreeding, c) LD: Han & Abney (2011, Genetic Epidemiology 35: 557-567).

2) Build up the reduced system:

$$\boldsymbol{L}\,\boldsymbol{\Sigma}_{\mathrm{A}}\,\boldsymbol{L}'\,\boldsymbol{\beta} = \boldsymbol{L}\,\boldsymbol{\Sigma}_{\mathrm{AX}}$$

3) Sampling β_S and β_D from

$$\beta_t \sim T\,N\left(\beta_{t-1}, (0.25)^2\right), \quad \beta_t \in [-0.25, 0.25]$$

Algorithm of Foulley (2000, GSE 32:631–635).

IBD genomic relationships under AR for Sultán when dams are not genotyped

Sultana Sumaj 566 Comodoro

$$\beta_S = 0.075 \pm 0.036$$

$$\beta_D = 0.086 \pm 0.036$$

$$F_{AR} = 0.037$$

rel_{AR}	SS	DS	SD	DD
Sultán	0.347	0.188	0.23	0.27
F_{ped} = 0.031	Sire	0.57	Dam	0.51

Missing data: work in progress

❑ Patterns of missing data are variable. Most problematic: **missing grand-dams**, because cov(Grand-dam, X) is part of the sufficient statistics (SuSt). Fortunately, the SuSt are Normal.

❑ Estimate the missing covariance using available genotypes from relatives **no more than 2 meioses apart** (***U*** = Uncles, grand-uncles). ***U*** are many in pigs, but not enough in cattle.

❑ Collaterals (FS): $\boldsymbol{\Sigma}_{\mathrm{FS}} = 0.5 + 2\left(\beta_{\mathrm{SX}}\,\beta_{\mathrm{SY}} + \beta_{\mathrm{DX}}\,\beta_{\mathrm{DY}}\right)$

❑ HMM to estimate missing covariances through **IBD segment sharing** with different SNP panels, or even a small number of microsatellites used for paternity.