Bayesian evolutionary optimisation for parental selection in animal breeding

Jia Liu

Agriculture and Food, CSIRO
Biological Data Science Institute, ANU
3 Sep 2024 at Perth, **AASC2024**
Jia.liu@csiro.au
Jia.liu@anu.edu.au

Motivation

Plant and animal breeders increasingly face significant challenges such as variability in the rate of productivity growth due to various factors including climate change and global competition.

We want maintain the stability to achieve of long-term genetic gains.

Solution: Optimising parental selections by finding best mates combinations in each animal breeding generation.

Genomic Parental Selection and Prediction

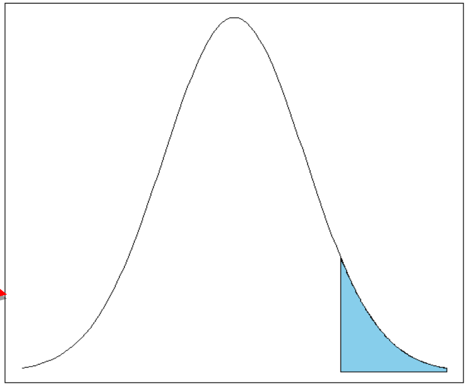
Input data

Genomic data for individuals

Linear mixed model (LLM)

$$Y = XB + ZU + E$$

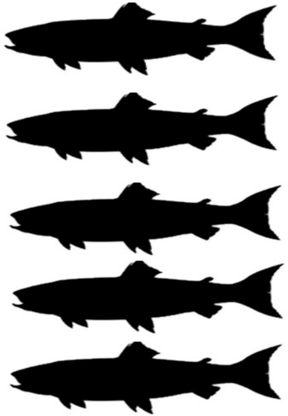
Prediction and Selection



Genomic estimated breeding values

GEBV

Phenotypic traits



Mate plan optimisation

$$P(X) = \prod_{t=1}^{n} p(X_{t-1} | X_{t-2}, \cdots, X_0)$$

Linear mixed model

- **Single trait model**

$$Y = 1_1\mu + g_1 + \sum_k g_k + e, g_1 \sim N(0, K_1\sigma_1^2), g_k \sim N(0, K_k\sigma_k^2), e_1 \sim N(0, I\sigma_e^2),$$

$$K_1 = Z_g G Z_g', \quad G = WW'/c, c = 2\sum_{k=1}^{p} p_k(1 - p_k); \quad K_2 = (Z_k \phi(X) Z_k') \circ (Z_t Z_t'),$$

- **Multi traits**

$$Y_j = 1_1\mu_j + g_{1j} + \sum_k g_{kj} + e, g_{1j} \sim N(0, K_{1j}\sigma_{1j}^2), g_{kj} \sim N(0, K_{kj}\sigma_{kj}^2), e_{1j} \sim N(0, I\sigma_e^2),$$

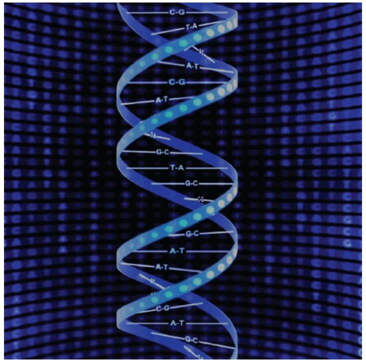
where j is jth trait.

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{g_1} + \boldsymbol{e}, \boldsymbol{g_1} \sim MN(0_p, G, \boldsymbol{\Sigma_T}), \boldsymbol{e_i} \sim MN(0_p, I_{np}, \boldsymbol{\Sigma_e}) \quad \otimes$$

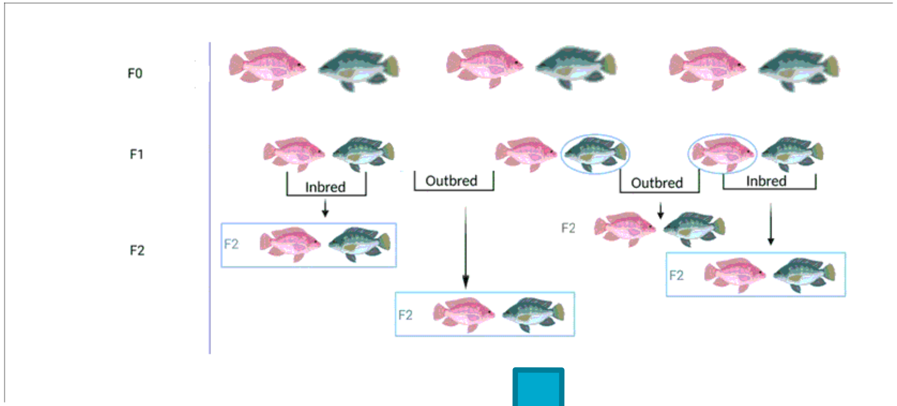
Example in Salmon breeding

Biomass, body length, etc

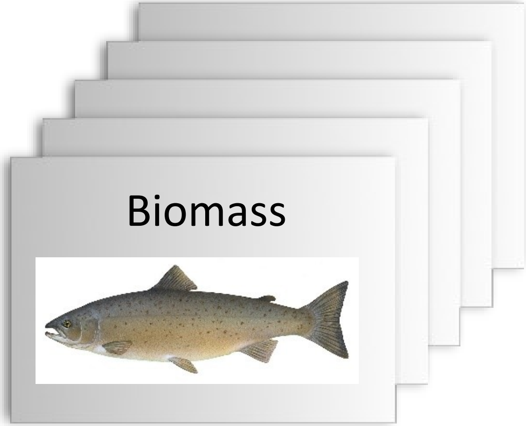
Fish lavea selection



Optimal mate plan

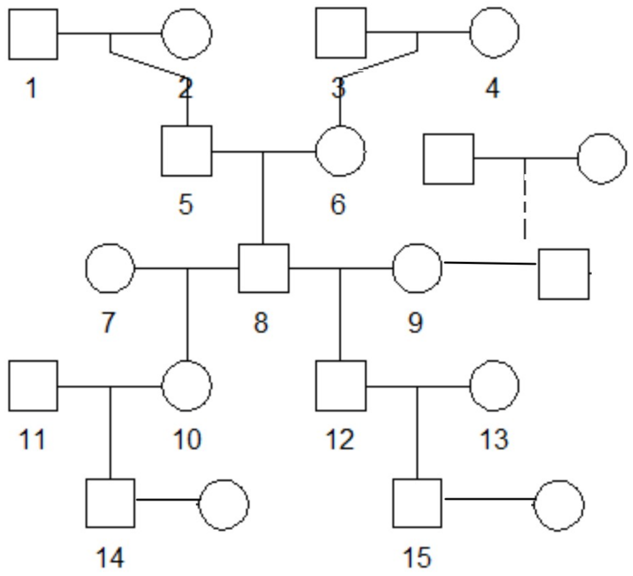


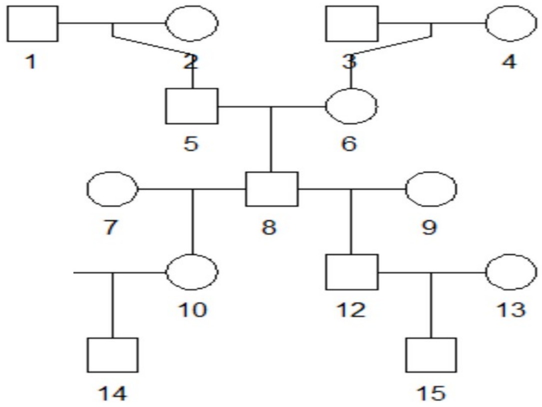
Offspring generation



Optimal mating strategies

$$X = X_0 \cup X_2 \cdots \cup X_n$$





Bayesian inference

$$p(X_k, \boldsymbol{\theta}) = \prod_{l=1} p(\theta_l) \prod_{t=1}^{k-1} \prod_{i=1}^{|X_t|} N(y^k | X_{k-1}, \cdots, X_0; \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \{\mu, \sigma^2, \boldsymbol{\beta}\}$$

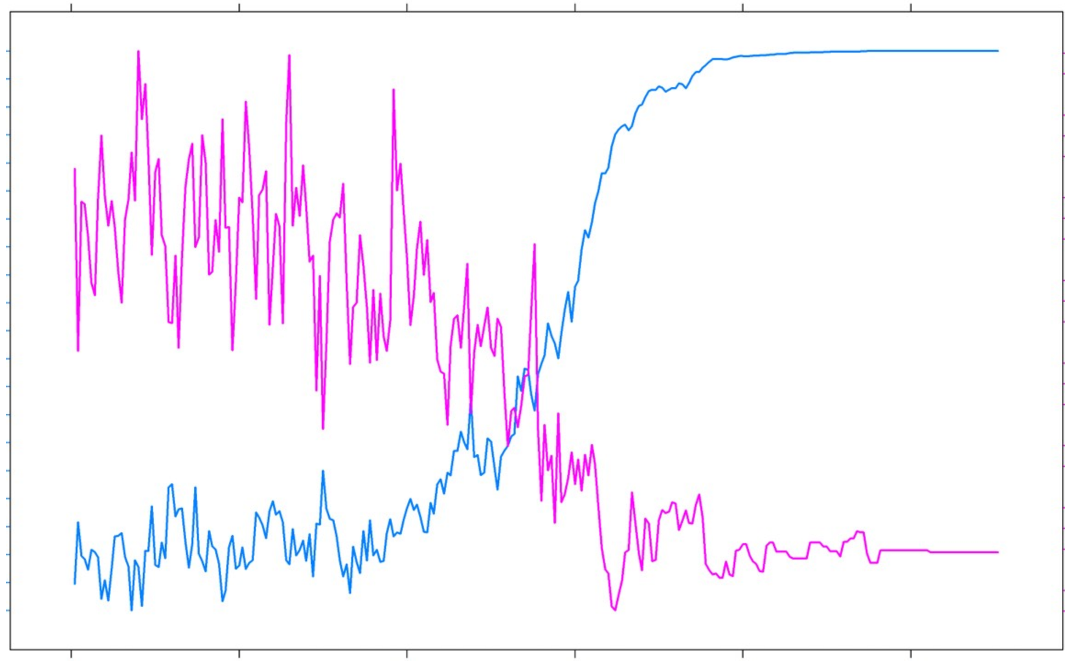
Simulate $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} | \mathbf{Y}^{t-1}, \mathbf{Y}^{t-2}, \cdots Y^0, X_{t-1}, X_{t-2}, \cdots X_0)$

Predicted $\hat{\mathbf{Y}}^t$

$$p(y^t | \mathbf{Y}^{\hat{t}-1}, \mathbf{Y}^{\hat{t}-2}, \cdots \mathbf{Y}^0, X_{t-1}, X_{t-2}, \cdots X_0, \hat{\boldsymbol{\theta}})$$

Design objective?

Maximising GEBV and minimising inbreeding value



Multi Objective Functions

GEBV (Genetic estimated breeding value**)**

Y=XB +ZU+ E

Inbreeding value (Kinship—pedigree)

$$\boldsymbol{F} = \sum_{k} h\left(x_k, p_k\right).$$

Constraints

Expected Utilities

Expected utility

$$U(X_k)=\int_Y \int_{\boldsymbol{\theta}} U(Y^k,\boldsymbol{\theta},|\{X_{k-1},\cdots X_0,Y^{k-1},\cdots Y^0,\})\,p(Y^k|\boldsymbol{\theta},\{\boldsymbol{X}_s,\boldsymbol{Y}_s\})\,p(\boldsymbol{\theta}|\boldsymbol{Y}_s,\boldsymbol{X}_s)\,d\,y^k d\,\boldsymbol{\theta}$$

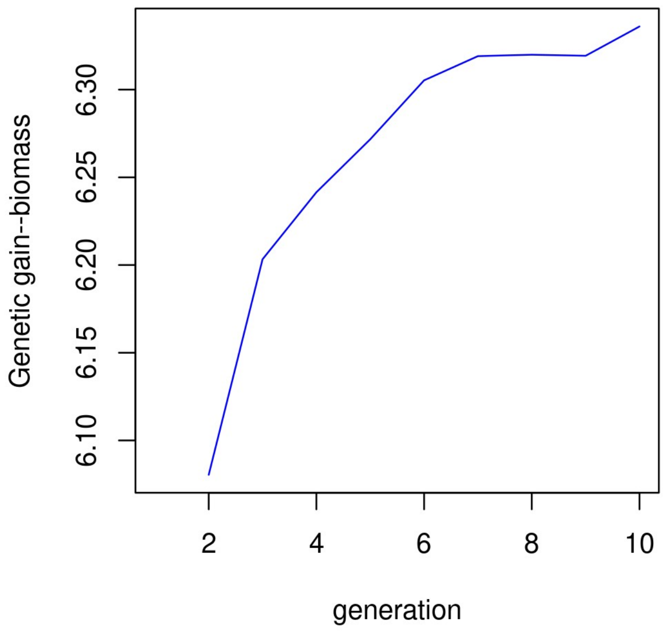
$$U(X_k)=\frac{1}{M}\sum_{m=1}^{M} U(Y^k,\boldsymbol{\theta},|\{X_{k-1},\cdots X_0,Y^{k-1},\cdots Y^0\})$$

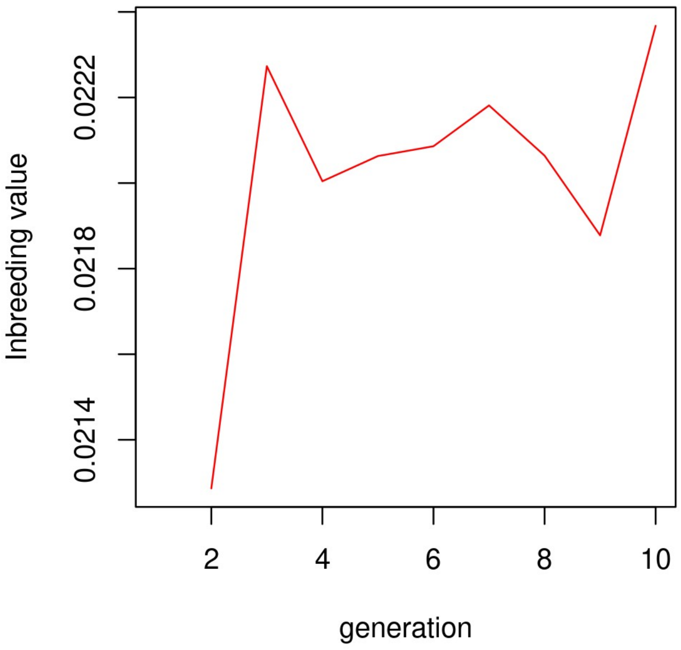
Selection Criteria → **EI**

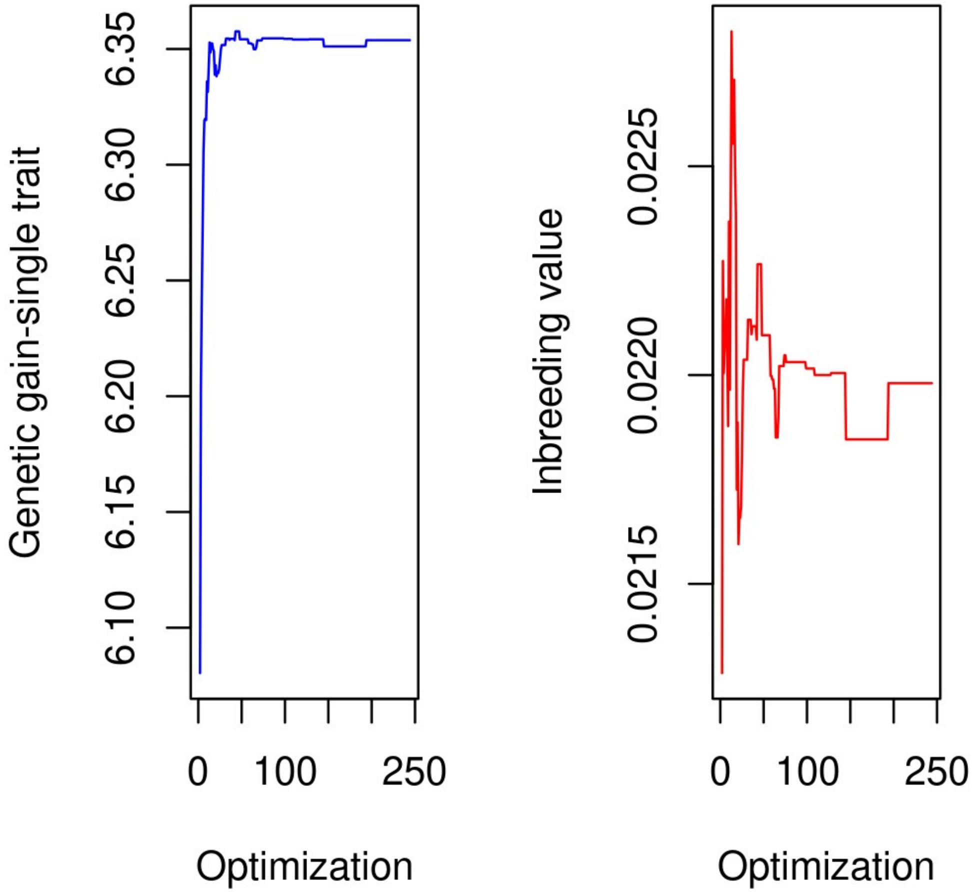
Optimization strategy → Stochastic optimiser **MCMC**

Simulation results

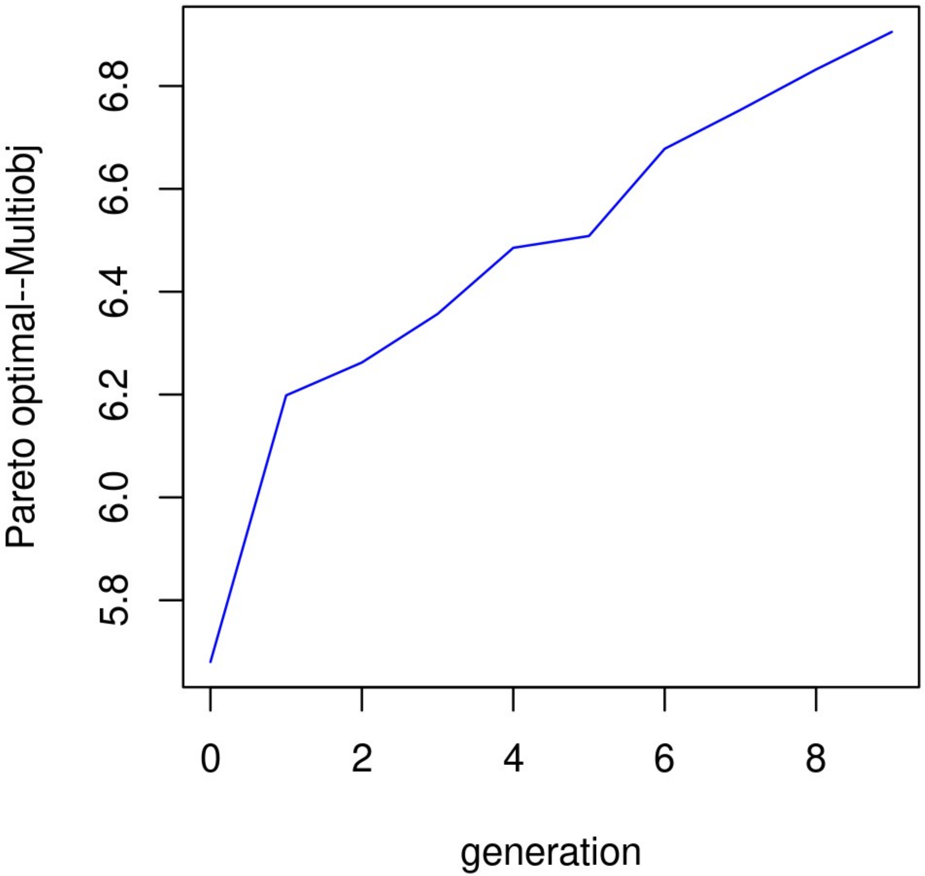
Simulation data were generated by mimic real breading scenarios







Genetic gain-single trait
6.10
6.15
6.20
6.25
6.30
6.35
0
100
250
Optimization
Inbreeding value
0.0215
0.0220
0.0225
0
100
250
Optimization



Pareto frontier

Summary

- The complexity of likelihood model can be increased by taking care of more variables, such environment, management.
- Design objective and the choices of utility functions.
- There may be no optimal can be achieved. (High dimensional, multi-objectives).
- Optimisation problems: choices of optimiser; multi modals; mixing rate; computational efficiency.
- **We need good designs to achieve good results!**

Reference

Liu, J., & Vanhatalo, J. (2020). Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process. Spatial Statistics, 35, 100392. https://doi.org/10.1016/j.spasta.2020.100392

Kuck, H., de Freitas, N., and Doucet, A. (2006). “Smc samplers for bayesian optimal nonlinear design”, in Nonlinear Statistical Signal Processing Workshop, 2006 IEEE, 99–102 (IEEE).

Kullback, S. (1987). “Letter to the editor: The Kullback-Leibler distance”, American Statistician 41, 340–341.

Müller, P. (1999). “Simulation based optimal design”, Handbook of Statistics 6, 459–474.

Müller, P., Sansó, B., and De Iorio, M. (2004). “Optimal bayesian design by inhomogeneous markov chain simulation”, Journal of the American Statistical Association 99, 788–798.

Australian National University

CSIRO

Thank you very much!

Wagdy Mekkawy,
Hobart Tasmania, CSIRO