Analysis of Genetically Structured Variance Heterogeneity and the Box-Cox Transformation

Ye Yang

Department of Genetics and Biotechnology,

Faculty of Agricultural Sciences,

Aarhus University, Denmark.

Outline

- Introduction
- Box-Cox model with genetically structured variance heterogeneity
- Choice of priors
- Data analysis: simulated data, rabbit and pig litter size data
- Conclusions

Introduction

- Classical model of quantitative genetics

$$y_{ij} = f_{ij} + a_i + p_i + \varepsilon_{ij},\ \varepsilon_{ij} \backsim N\left(0, \sigma^2\right) \quad (f:\ \text{fixed effects})$$

- Structured environmental variance model

$$\varepsilon_{ij} |\ a_i^*, p_i^* \backsim N\left(0, \sigma_{ij}^2\right),\ \log\left(\sigma_{ij}^2\right) = f_{ij}^* + a_i^* + p_i^*,$$

$$(a, a^*)^T | G \sim N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), G \otimes A\right), G = \left(\begin{array}{cc} \sigma_a^2 & \rho\sigma_a\sigma_{a^*} \\ \rho\sigma_a\sigma_{a^*} & \sigma_{a^*}^2 \end{array}\right),$$

$$p \mid \sigma_p^2 \sim N(0, \sigma_p^2 I),\ p^* \mid \sigma_{p^*}^2 \backsim N(0, \sigma_{p^*}^2 I).$$

Objective

Problem with the structured environmental variance model

- Skewness is $\frac{E\left[\left(y_{ij}-f_{ij}\right)^3 \mid f_{ij}, f_{ij}^*\right]}{Var\left(y_{ij} \mid f_{ij}, f_{ij}^*\right)^{\frac{3}{2}}} = \rho \frac{3\sigma_a \sigma_{a^*} \exp\left(f_{ij}^* + \frac{\sigma_{a^*}^2}{2} + \frac{\sigma_{p^*}^2}{2}\right)}{\sigma_a^2 + \sigma_p^2 + \exp\left(f_{ij}^* + \frac{\sigma_{a^*}^2}{2} + \frac{\sigma_{p^*}^2}{2}\right)}$

  showing that both skewed and symmetric distributions can be accommodated.

  Can skewed sampling distributions for data lead to spurious ρ ?

Objective: are results from structured environmental variance model an artifact of the scale of measurement?

Box-Cox model with genetically structured variance heterogeneity

- Box-Cox transformation: choose the scale that provides best fit to data
- Box-Cox model is $y_{ij}^{(\lambda)} \mid \lambda, \mu_{ij}, \sigma_{ij}^2 \backsim N\left(\mu_{ij}, \sigma_{ij}^2\right)$, $\log \sigma_{ij}^2 = \mu_{ij}^*$,

  with
$$\begin{aligned} \mu_{ij} &= f_{ij} + a_i + p_i, \\ \mu_{ij}^* &= f_{ij}^* + a_i^* + p_i^*, \end{aligned}$$

  and
$$y_{ij}^{(\lambda)} = \begin{cases} \frac{y_{ij}^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log y_{ij} & (\lambda = 0) \end{cases}, \text{holds for } y_{ij} > 0$$

Choice of scale: Box-Cox transformation

Sampling distribution of untransformed data

$$P\left(y_{ij}|\ \mu_{ij}, \mu_{ij}^{*}, \lambda\right) = P\left(y_{ij}^{(\lambda)}|\ \mu_{ij}, \mu_{ij}^{*}, \lambda\right) J\left(y_{ij}, \lambda\right),$$

with

$$J\left(y_{ij}, \lambda\right) = \left|y_{ij}^{\lambda-1}\right|.$$

Log-posterior, excluding additive constant

$$\begin{aligned}
\log P\left(\theta \mid y, \lambda\right) &= \sum_{i,j}^{n} \log P\left(y_{ij}^{(\lambda)}|\ \mu_{ij}, \mu_{ij}^{*}, \lambda\right) \\
&\quad + (\lambda - 1) \sum_{i,j}^{n} \log y_{ij} + \sum_{i,j}^{n} \log p\left(\mu_{ij}, \mu_{ij}^{*}, \lambda\right), \\
p\left(\mu_{ij}, \mu_{ij}^{*}, \lambda\right) &= p\left(\mu_{ij}, \mu_{ij}^{*}|\ \lambda\right) P\left(\lambda\right).
\end{aligned}$$

Continued

Choice of prior under $y^{(\lambda)}$ must be consistent for different values of λ. Box and Cox suggested

$$y_{ij}^{(\lambda)} \approx k + l_\lambda y_{ij},$$

as basis for choice of priors under $y^{(\lambda)}$, where

$$l_\lambda = \left(J\left(y,\lambda\right)\right)^{\frac{1}{n}} = \left(\prod_{i,j}^{n} \left|y_{ij}^{\lambda-1}\right|\right)^{\frac{1}{n}}$$

Continued

This leads to the following prior specifications:

$$
\begin{aligned}
P\left(f_\lambda \mid \lambda\right) &\propto \left(J\left(y, \lambda\right)\right)^{\frac{p}{n}}, \\
P\left(\exp\left(f_\lambda^*\right) \mid \lambda\right) &\propto \left(J\left(y, \lambda\right)\right)^{\frac{2}{n}}, \\
P\left(\sigma_{a,\lambda}^2 | \lambda\right) &\propto P\left(\sigma_a^2\right)\left(J\left(y, \lambda\right)\right)^{\frac{2}{n}}, \\
P\left(\sigma_{a*,\lambda}^2 | \lambda\right) &\propto P\left(\sigma_{a*}^2\right), \\
P\left(\rho_\lambda | \lambda\right) &\propto P\left(\rho\right), \\
P\left(\sigma_{p,\lambda}^2 | \lambda\right) &\propto P\left(\sigma_p^2\right)\left(J\left(y, \lambda\right)\right)^{\frac{2}{n}}, \\
P\left(\sigma_{p*,\lambda}^2 | \lambda\right) &\propto P\left(\sigma_{p*}^2\right), \\
p\left(a_\lambda, a_\lambda^* | \sigma_{a,\lambda}^2, \sigma_{a*,\lambda}^2, \rho_\lambda\right) &\propto p\left(a, a^* | \sigma_a^2, \sigma_{a*}^2, \rho\right), \\
p\left(p_\lambda, p_\lambda^* | \sigma_{p,\lambda}^2, \sigma_{p*,\lambda}^2\right) &\propto p\left(p, p^* | \sigma_p^2, \sigma_{p*}^2\right), \\
\lambda &\backsim Un\left(-3, 3\right)
\end{aligned}
$$

Simulation study

Identifiability of λ and ρ. True $\lambda = 1$

number of records	mean(λ)	HPD interval(λ)	corr(λ,ρ)
1	0.89	(0.01,1.84)	0.69
2	0.54	(-0.05,1.2)	0.48
3	0.99	(0.45,1.4)	0.35
5	0.78	(0.42,1.12)	0.42
10	0.93	(0.71,1.18)	0.23
rabbit data	0.82	(0.48,1.51)	0.34

Litter size data in rabbits and pigs

- Rabbit litter size data: 2996 litters, average of 3.2 litters per female, pedigree file: 1281 individuals. Average litter size: 7.22
- Pig litter size data: 9778 litters, average of 2.4 litters per female, pedigree file: 6437 individuals. Average litter size: 10.28

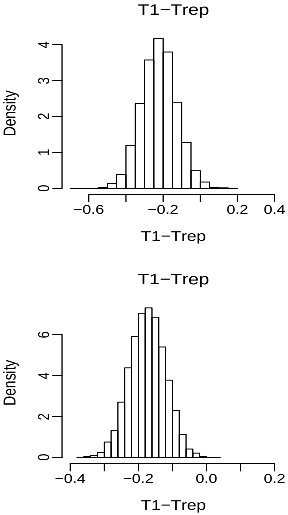
Results

Posterior means and 95% posterior intervals for variance components

Models	σ^2_a	σ^2_{a*}	ρ	σ^2_p	σ^2_{p*}
Rabbits $\lambda = 1$	0.805	0.133	-0.73	0.38	0.052
	(0.475,1.216)	(0.056,0.23)	(-0.89,-0.5)	(0.15,0.66)	(0.025,0.099)
Rabbits $\lambda = 1.4134$	2.59	0.056	0.285	2.858	0.042
	(1.47,4.2)	(0.027,0.11)	(-0.236,0.789)	(1.53,4.22)	(0.02,0.084)
Pigs $\lambda = 1$	1.63	0.071	-0.642	0.52	0.021
	(1.24,2.05)	(0.038,0.11)	(-0.82,-0.45)	(0.25,0.83)	(0.01,0.038)
Pigs $\lambda = 1.393$	8.17	0.037	0.7	4.15	0.017
	(5.9,10.63)	(0.02,0.06)	(0.44,0.98)	(2.17,6.03)	(0.0078,0.026)

Continued

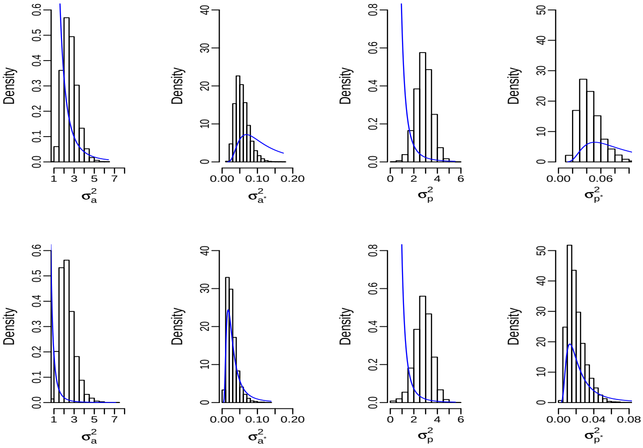
Statistical support for the model -residual skewness. Upper: Rabbits; Lower: Pigs



Prior sensitive analysis (Rabbits)

Top $\nu = 5,\ S_{\sigma^2_a} = 0.492,\ S_{\sigma^2_{a*}} = 0.096,\ S_{\sigma^2_p} = 0.264,\ S_{\sigma^2_{p*}} = 0.072$

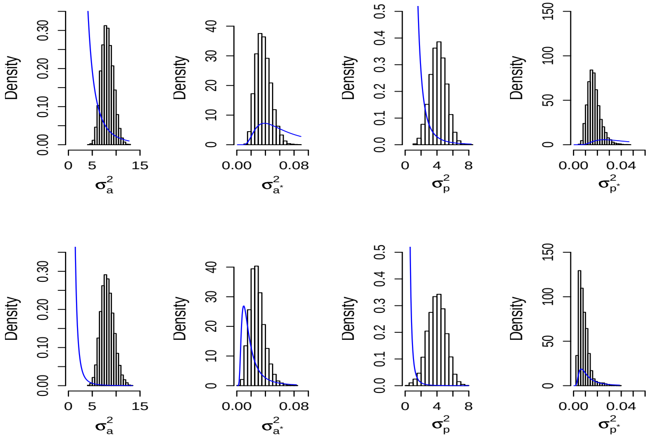
Bottom $\nu = 5,\ S_{\sigma^2_a} = 0.124,\ S_{\sigma^2_{a*}} = 0.024,\ S_{\sigma^2_p} = 0.066,\ S_{\sigma^2_{p*}} = 0.018$



Prior sensitive analysis (Pigs)

Top $\nu = 5$, $S_{\sigma_a^2} = 0.972$, $S_{\sigma_{a*}^2} = 0.054$, $S_{\sigma_p^2} = 0.36$, $S_{\sigma_{p*}^2} = 0.036$

Bottom $\nu = 5$, $S_{\sigma_a^2} = 0.243$, $S_{\sigma_{a*}^2} = 0.0135$, $S_{\sigma_p^2} = 0.09$, $S_{\sigma_{p*}^2} = 0.009$



Conditional predictive ordinate (CPO)

One way of assessing global predictive ability of a set of Models

$$\begin{aligned}\widehat{CPO_{ij}} &= \widehat{p}\left(y_{ij} \mid y_{-ij}, M_r\right) \\ &= \left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p\left(y_{ij} \mid \theta^{(t)}, M_r\right)}\right]^{-1},\end{aligned}$$

The logarithm of the CPO for Model r (M_r) is

$$\log\left[\widehat{CPO}_{M_r}\right] = \sum_{i,j}^{n} \log\left[p\left(y_{ij} \vert\ y_{-ij}, M_r\right)\right]$$

Note: the larger value of $\log\left[\widehat{CPO}_{M_r}\right]$ indicates a better fit of a model.

Continued

Model	Rabbits	Pigs
$\lambda = 1$	-3,930.7	-23,998.0
Mode λ	-3,919.5	-15,269.1
Mode λ, $\sigma^2_{a^*,\lambda} = 0$,$\sigma^2_{p^*,\lambda} = 0$	-3,927.3	-15,297.1

Conclusions

- Statements about variance sensitive to presence of heavy tails.
- The conditional distribution of phenotypic data given all model parameters is normally distributed under the posterior mode of λ, instead of $\lambda = 1$ in both rabbit and pig litter size data.
- The support of additive genetic variance affecting variance is much weaker under the "correct" scale than under the original scale.

Acknowledgement

- Daniel Sorensen
- Ole Christensen
- Guosheng Su

This work was conducted as part of the SABRETRAIN Project, funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission.