Genetically structured variance heterogeneity - modelling and computation

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

November 24, 2007

Breeding for homogeneity

Animal breeding so far focused on *increasing* output:

- litter size (number of piglets)
- body weight
- milk yield
- ...

However *homogeneous* production important too.

Is it possible to breed for small variance - i.e. is the variance of a trait controlled by genes ?

Recent empirical evidence of genetic variance heterogeneity found for pig litter size, snail body weight, bristle number of Drosophila, body weights of poultry

(San Cristobal et al. 2001, Sorensen and Waagepetersen 2003, Ros et al. 2004, Mackay and Lyman 2005, Rowe et al. 2006)

Outline

1. Linear models with variance heterogeneity in animal breeding
2. Posterior predictive model assessment
3. MCMC computation
4. Genetic variance heterogeneity and heavy tailed distributions

1. Linear models with variance heterogeneity in animal breeding

2. Posterior predictive model assessment

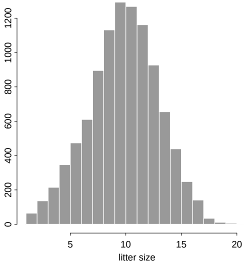
3. MCMC computation

4. Genetic variance heterogeneity and heavy tailed distributions

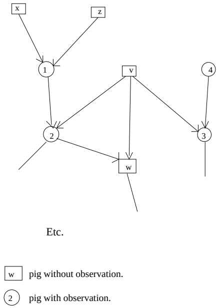
Data example (litter sizes for 4149 sows)

y_{ij} : jth litter size (# piglets) for ith sow $j = 1, \ldots, n_i$ (most $n_i \leq 3$)

Histogram



Pedigree



(+ covariate information herd/year/season)

10060 observations, 6437 pigs in pedigree

Linear models in quantitative genetics

Standard model:

$\mathbf{a}$, $\mathbf{p}$ vectors of genetic and environmental random effects affecting litter size

$$
\begin{aligned}
y_{ij} =& \mu_{ij} + \epsilon_{ij} \\
& \mu_{ij} = f_{ij} + a_i + p_i \\
& \qquad \epsilon_{ij} \sim N(0, \sigma^2) \qquad (f: \text{ fixed effects})
\end{aligned}
$$

Heterogeneous residual variance model (San-Cristobal *et al.*, 98)

$\mathbf{a}^*, \mathbf{p}^*$ genetic/environmental random effects affecting residual variation

$$
\begin{aligned}
\epsilon_{ij} | \mathbf{a}^*, \mathbf{p}^* \sim N(0, \sigma^2_{ij}) \\
\log(\sigma^2_{ij}) = f^*_{ij} + a^*_i + p^*_i
\end{aligned}
$$

Models for random effects

Environmental:

$$\mathbf{p} \sim N(0, \sigma_p^2 I) \quad \mathbf{p}^* \sim N(0, \sigma_{p^*}^2 I) \quad \text{(independent)}$$

Genetic:

$$(\mathbf{a}, \mathbf{a}^*) \sim N\big((0,0), G \otimes A\big)$$

where

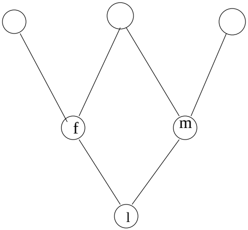
$$G = \begin{bmatrix} \sigma_a^2 & \rho\sigma_a\sigma_{a^*} \\ \rho\sigma_a\sigma_{a^*} & \sigma_{a^*}^2 \end{bmatrix}$$

and A additive genetic correlation matrix (pedigree).

Central parameters: $\sigma_a^2 = \mathbb{V}\mathrm{ar}\, a_l$, $\rho = \mathbb{C}\mathrm{orr}(a_l, a_l^*)$, $\sigma_{a^*}^2 = \mathbb{V}\mathrm{ar}\, a_l^*$ (genetic covariance).

NB: A 6437×6437 in example.

Structure of genetic correlation matrix A



$$a_l = (a_f + a_m)/2 + \eta_l$$
$$\mathbf{a} = T\boldsymbol{\eta}$$

$\boldsymbol{\eta} \sim N(0, \sigma_a^2 D)$ Mendelian sample noise (D diagonal)

Factorization: $A = TDT^{\mathsf{T}}$

$\boldsymbol{\eta} = T^{-1}\mathbf{a}$ where T^{-1} sparse ($\eta_l = a_l - a_f/2 - a_m/2$)

Factorization using sparse matrices:

$$A^{-1} = (T^{-1})^{\mathsf{T}} D^{-1} T^{-1}$$

$\mathbf{a}$: Markov random field (sparse A^{-1})

Mean-variance relation

Ignoring fixed and environmental effects:

$$\mathbb{E}(y_{ij}|\mathbf{a}, \mathbf{a}^*) = a_i \qquad \log \mathbb{V}\mathrm{ar}(y_{ij}|\mathbf{a}, \mathbf{a}^*) = a_i^* = \beta a_i + u_i$$

where

$$\beta a_i = \mathbb{E}(a_i^*|\mathbf{a}) = \frac{\rho\sigma_{a^*}^2}{\sigma_a^2} a_i$$

and

$$\mathbf{u} = \mathbf{a}^* - \mathbb{E}(\mathbf{a}^*|\mathbf{a}) \sim N(0, \sigma_{a^*}^2(1-\rho^2)A)$$

is independent of $\mathbf{a}$.

$|\rho| = 1$: deterministic mean-variance relationship.

$\sigma_{\mathbf{a}^*}^2 = 0$: no relation.

Moderate $\sigma_{a^*}^2$: $\mathbb{V}\mathrm{ar}(y_{ij}|\mathbf{a}, \mathbf{a}^*) \approx 1 + \beta a_i + u_i$

Model assessment: genetic variance heterogeneity

Model assessment from raw data useless (too noisy - many sources of variation).

Standardized residuals under standard linear mixed model

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sigma}$$

Conditional on a_i and $u_i = a_i^* - \beta a_i$:

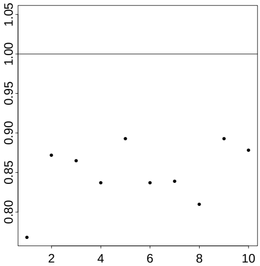
$$\mathbb{E}(r_{ij}^2 | a_i, u_i) = \exp(\beta a_i + u_i) \approx 1 + \beta a_i + u_i$$

Averaged squared residuals for observations grouped according to a_i:

$$T_k(\mathbf{y}, \boldsymbol{\mu}, \sigma, \mathbf{a}) = \frac{1}{n_k} \sum_{ij: v_k \le a_i \le v_{k+1}} r_{ij}^2, \quad -\infty = v_0 < v_1 < \cdots < v_K = \infty$$

$\mathbb{E}\, T_k = 1$ if $\mathbf{a}^* = 0$.

Plots of $T_k(\mathbf{y}, \hat{\boldsymbol{\mu}}, \hat{\mathbf{a}}, \hat{\sigma})$ obtained with point estimates of $\boldsymbol{\mu}$, $\mathbf{a}$, $\log \sigma$ (posterior means)



Overfitting: conditional expectation follow data too closely

$\mathbb{E}[\mathbf{a}|\mathbf{y}]$ best MSE predictor but *far from typical* realisation of $\mathbf{a}$.

Posterior predictive model assessment

$T(\mathbf{y}, \psi)$ summary statistic for *observed data* $\mathbf{y}$ where $\psi = (\boldsymbol{\mu}, \mathbf{a}, \sigma, \ldots)$.

Idea:

- ψ known: compare $T(\mathbf{y}, \psi)$ with sampling/predictive distribution of $T(\mathbf{Y}, \psi)$.
- ψ unknown: consider posterior predictive distribution of

$$T(\mathbf{y}, \psi) - T(\mathbf{Y}, \psi)$$

i.e. $(\mathbf{Y}, \psi)$ generated from posterior predictive distribution given $\mathbf{y}$.

In practice: consider distribution of

$$T(\mathbf{y}, \psi^{(l)}) - T(\mathbf{Y}^{(l)}, \psi^{(l)})$$

where $(\mathbf{Y}^{(l)}, \psi^{(l)})$ posterior predictive simulations (MCMC).

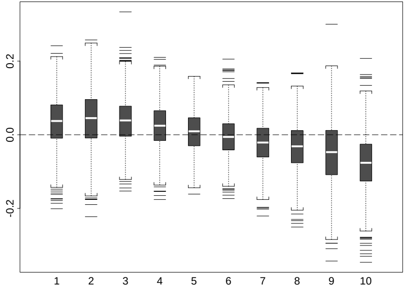
Genetic variance heterogeneity:

$$T_k(\mathbf{y}, \boldsymbol{\mu}, \mathbf{a}, \sigma) = \frac{1}{n_k} \sum_{ij: v_k \le a_i \le v_{k+1}} r_{ij}^2$$

Simulate posterior (predictive) realizations $\boldsymbol{\mu}^{(l)}$, $\mathbf{a}^{(l)}$, $\sigma^{(l)}$, and $\mathbf{Y}^{(l)}$ under standard linear mixed model and compute

$$D_k^{(l)} = T_k(\mathbf{y}, \boldsymbol{\mu}^{(l)}, \mathbf{a}^{(l)}, \sigma^{(l)}) - T_k(\mathbf{Y}^{(l)}, \boldsymbol{\mu}^{(l)}, \mathbf{a}^{(l)}, \sigma^{(l)})$$

Posterior predictive distributions of D_k



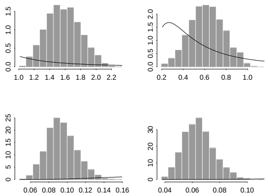
Posterior results

Posterior means and 95 % credibility intervals:

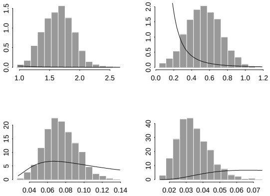
σ_a^2	σ_{a*}^2	ρ
1.62	.09	-.62
1.20;2.05	.06;.13	-.80;-.43

Posterior distributions (two choices of priors):

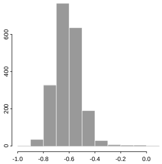
σ_a^2, σ_p^2, σ_{a*}^2, σ_{p*}^2



σ_a^2, σ_p^2, σ_{a*}^2, σ_{p*}^2



ρ



Bayesian inference using MCMC

Explore posterior ($\theta = (\sigma^2_a, \sigma^2_{a*}, \rho, \dots)$)

$$p(\mathbf{a}, \mathbf{a}^*, \theta | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{a}, \mathbf{a}^*, \theta) p(\mathbf{a}, \mathbf{a}^*; \theta) p(\theta)$$

using MCMC sample: $(\mathbf{a}^1, \mathbf{a*}^1, \theta^1)$, $(\mathbf{a}^2, \mathbf{a*}^2, \theta^2), \dots$

Current value: $(\mathbf{a}, \mathbf{a}*)^k$

Proposal: $(\mathbf{a}, \mathbf{a}*)^{\text{prop}} \sim q((\mathbf{a}, \mathbf{a}*)^{\text{prop}} \,|\, (\mathbf{a}, \mathbf{a}*)^k)$

Bayesian inference using MCMC

Explore posterior ($\theta = (\sigma_a^2, \sigma_{a*}^2, \rho, \dots)$)

$$p(\mathbf{a}, \mathbf{a}^*, \theta | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{a}, \mathbf{a}^*, \theta) p(\mathbf{a}, \mathbf{a}^*; \theta) p(\theta)$$

using MCMC sample: $(\mathbf{a}^1, \mathbf{a}^{*1}, \theta^1)$, $(\mathbf{a}^2, \mathbf{a}^{*2}, \theta^2), \dots$

Current value: $(\mathbf{a}, \mathbf{a}^*)^k$

Proposal: $(\mathbf{a}, \mathbf{a}^*)^{\text{prop}} \sim q((\mathbf{a}, \mathbf{a}^*)^{\text{prop}} \,|\, (\mathbf{a}, \mathbf{a}^*)^k)$

With probability

$$\min\left\{1, \frac{p((\mathbf{a}, \mathbf{a}^*)^{\text{prop}}, \theta | \mathbf{y}) \; q((\mathbf{a}, \mathbf{a}^*)^k \,|\, (\mathbf{a}, \mathbf{a}^*)^{\text{prop}})}{p((\mathbf{a}, \mathbf{a}^*)^k, \theta | \mathbf{y}) \; q((\mathbf{a}, \mathbf{a}^*)^{\text{prop}} \,|\, (\mathbf{a}, \mathbf{a}^*)^k)}\right\}$$

new state $(\mathbf{a}, \mathbf{a}^*)^{k+1} = (\mathbf{a}, \mathbf{a}^*)^{\text{prop}}$; otherwise $(\mathbf{a}, \mathbf{a}^*)^{k+1} = (\mathbf{a}, \mathbf{a}^*)^k$.

Problem: efficient update of highdimensional $(\mathbf{a}, \mathbf{a}^*)$.

Choice of proposal density q

Gibbs sampler: full conditional distribution only tractable for **a**.

Random walk:

$$(\mathbf{a}, \mathbf{a}^*)^{\text{prop}} \sim N((\mathbf{a}, \mathbf{a}^*)^k, hI)$$

- small acceptance rates due to high dimension.

Langevin-Hastings (use gradient information):

$$(\mathbf{a}, \mathbf{a}^*)^{\text{prop}} \sim N((\mathbf{a}, \mathbf{a}^*)^k + h\nabla \log p(\mathbf{a}, \mathbf{a}^*|\mathbf{y}, \theta)/2, hI)$$

- better acceptance rates than random walk in high dimensions.

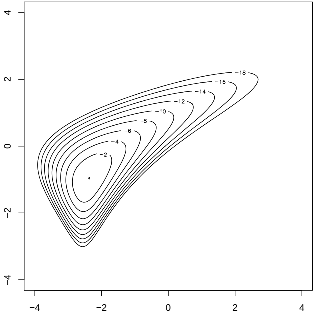
Reparametrization: apply Langevin-Hastings to transformed random effects

$$(\gamma, \gamma^*) = (G \otimes A)^{-1/2}(\mathbf{a}, \mathbf{a}^*)^{\mathsf{T}} \sim N(0, I)$$

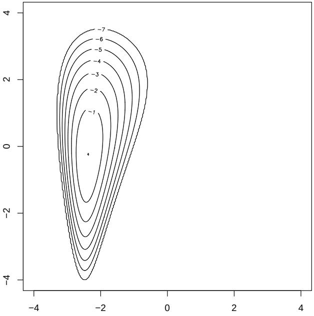
Illustration of MCMC strategies for toy example

a and **a*** each one-dimensional (only one animal in pedigree), simulated data $\mathbf{y} = (-2.62, -2.42)$.

Posterior of $(\mathbf{a}, \mathbf{a}^*)$

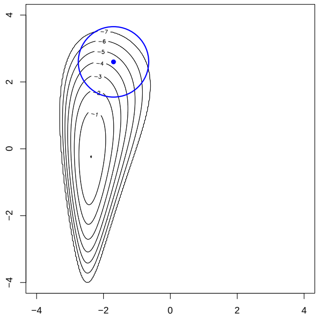


Posterior of (γ, γ^*)

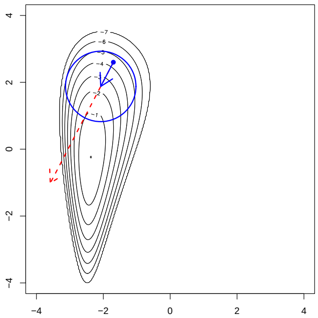


Random walk and Langevin-Hastings updates for (γ, γ^*) (blue dot is current value)

Random walk



Langevin-Hastings



Normal approximation

Idea: approximate posterior of $\mathbf{a}$ (or $\mathbf{a}, \mathbf{a}^*$) using second order Taylor expansion:

$$\log p(\mathbf{a}|\mathbf{y}) \approx \log p(\hat{\mathbf{a}}|\mathbf{y}) + (\mathbf{a} - \hat{\mathbf{a}}) \nabla \log p(\hat{\mathbf{a}}|\mathbf{y})^\mathsf{T} - \frac{1}{2}(\mathbf{a} - \hat{\mathbf{a}}) H(\hat{\mathbf{a}})(\mathbf{a} - \hat{\mathbf{a}})^\mathsf{T}$$

Hence

$$\mathbf{a}^{\text{prop}} \sim N(\hat{\mathbf{a}} + \nabla \log p(\hat{\mathbf{a}}|\mathbf{y}) H(\hat{\mathbf{a}})^{-1}, H(\hat{\mathbf{a}})^{-1})$$

Normal approximation

Idea: approximate posterior of $\mathbf{a}$ (or $\mathbf{a}, \mathbf{a}^*$) using second order Taylor expansion:

$$\log p(\mathbf{a}|\mathbf{y}) \approx \log p(\hat{\mathbf{a}}|\mathbf{y}) + (\mathbf{a}-\hat{\mathbf{a}})\nabla \log p(\hat{\mathbf{a}}|\mathbf{y})^\mathsf{T} - \frac{1}{2}(\mathbf{a}-\hat{\mathbf{a}})H(\hat{\mathbf{a}})(\mathbf{a}-\hat{\mathbf{a}})^\mathsf{T}$$

Hence

$$\mathbf{a}^{\text{prop}} \sim N(\hat{\mathbf{a}} + \nabla \log p(\hat{\mathbf{a}}|\mathbf{y})H(\hat{\mathbf{a}})^{-1}, H(\hat{\mathbf{a}})^{-1})$$

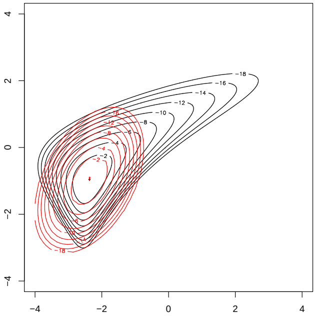
Possibilities for $\hat{\mathbf{a}}$:

- current value $\hat{\mathbf{a}} = \mathbf{a}^k$
- $\hat{\mathbf{a}}$: one-step Newton-Raphson from current value
- $\hat{\mathbf{a}}$ mode of $p(\mathbf{a}|\mathbf{y})$:

$$\mathbf{a}^{\text{prop}} \sim N(\hat{\mathbf{a}}, H(\hat{\mathbf{a}})^{-1})$$

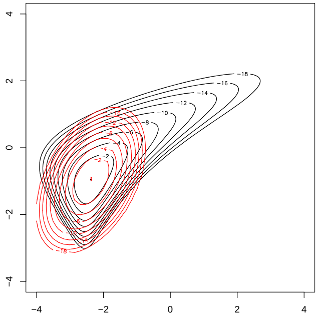
Normal approximation for toy example

Posterior and Normal Approximation for $(\mathbf{a}, \mathbf{a}^*)$

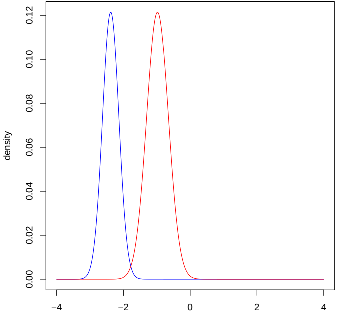


Normal approximation for toy example

Posterior and Normal Approximation for $(\mathbf{a}, \mathbf{a}^*)$



Conditional densities of $\mathbf{a}|\mathbf{a}^*, \mathbf{y}$ and $\mathbf{a}^*|\mathbf{a}, \mathbf{y}$ at mode



Use normal approximation for $\mathbf{a}$ and $\mathbf{a}^*$ separately (conditional distribution of $\mathbf{a}$ given $(\mathbf{a}^*, \mathbf{y})$ exactly normal).

Sampling from normal approximation I

Suppose $\mathbf{y}$ depend on $\mathbf{a}$ through $Z\mathbf{a}$. Then $H(\hat{\mathbf{a}})$ of the form

$$H(\hat{\mathbf{a}}) = A^{-1}/\sigma_a^2 + Z^\mathsf{T}\Sigma^{-1}Z$$

Normal approximation $N(\hat{\mathbf{a}}, H(\hat{\mathbf{a}})^{-1})$ formally equivalent to conditional distribution of $\mathbf{a}$ given $\tilde{\mathbf{y}} = Z\mathbf{a} + \tilde{\epsilon}$ for 'virtual' data $\tilde{\mathbf{y}}$.

Use García-Cortés & Sorensen algorithm based on

$$\mathbf{a} = \left(\mathbf{a} - \mathbb{E}[\mathbf{a}|\tilde{\mathbf{y}}]\right) + \mathbb{E}[\mathbf{a}|\tilde{\mathbf{y}}] = R + \hat{\mathbf{a}}$$

where 'prediction error' $R = (\mathbf{a} - \mathbb{E}[\mathbf{a}|\tilde{\mathbf{y}}])$ and $\hat{\mathbf{a}} = \mathbb{E}[\mathbf{a}|\tilde{\mathbf{y}}], \tilde{\mathbf{y}}$ independent.

Hence if R_{sim} is a simulation of R then

$$\mathbf{a}_{\text{sim}} = R_{\text{sim}} + \hat{\mathbf{a}}$$

is a conditional simulation of $\mathbf{a}$ given $\tilde{\mathbf{y}}$.

Sampling from normal approximation II

'Conditional simulation' of $\mathbf{a}$ given $\tilde{\mathbf{y}}$:

$$\mathbf{a}_{\text{sim}} = R_{\text{sim}} + \hat{\mathbf{a}}$$

Generation of R_{sim}:

1. simulate $(\mathbf{a}_{\text{sim}}, \tilde{\mathbf{y}}_{\text{sim}})$ from joint distribution of $(\mathbf{a}, \tilde{\mathbf{y}})$ (use factorization $A = TDT^{\mathsf{T}}$)
2. compute $\hat{a}_{\text{sim}} = \mathbb{E}[\mathbf{a}|\tilde{\mathbf{y}}_{\text{sim}}]$ (mixed model equations)
3. return $R_{\text{sim}} = \mathbf{a}_{\text{sim}} - \hat{a}_{\text{sim}}$.

Sparse matrix methods

Use general sparse matrix Cholesky decomposition for hessian

$$H(\hat{\mathbf{a}}) = A^{-1}/\sigma_a^2 + Z^{\mathsf{T}}\Sigma^{-1}Z$$

in normal approximation $N(\hat{\mathbf{a}}, H(\mathbf{a})^{-1})$.

GMRFLib (H. Rue): general software in c for MCMC computation in models with sparse precision matrix for random effects. E.g. routines for computing updates using normal approximation.

Lots of useful tricks and advice in book Rue & Knorr-Held (2005).

Comparison of Langevin-Hastings and normal approximation

Estimation of posterior means for various parameters and three data sets.

Ratios (LH/NX) of numbers of iterations needed to obtain given precision of Monte Carlo estimates:

Data	$\mathbf{a}A^{-1}\mathbf{a}^\mathsf{T}$	a_1	$\tilde{a}_1$	σ_a^2	$\sigma_{a^*}^2$	ρ	cost NX/LH
Rabbits	54	105	106	102	81	117	20
Pigs	315	873	673	129	190	278	100
Snails	317	465	253	328	158	401	35

Last column: ratios (NX/LH) of computing times for given number of iterations.

NB: joint update of random effects and covariance parameters.

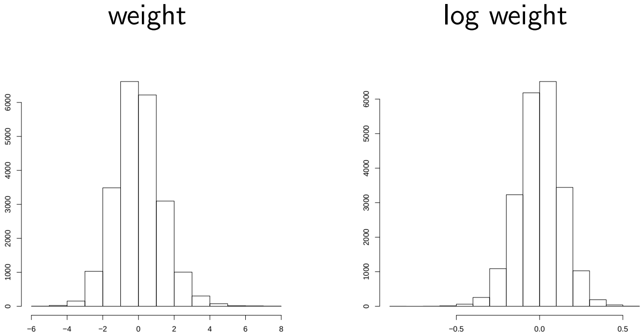
1. Linear models with variance heterogeneity in animal breeding

2. Posterior predictive model assessment

3. MCMC computation

4. Genetic variance heterogeneity and heavy tailed distributions

Body weight and log body weight of snails



$\rho = 0.81$ $\rho = -0.59$

'Paradox': weight positively correlated with variance but log weight negatively correlated with variance.

Skewness

Skewness for model with genetic variance heterogeneity:

$$\frac{\mathbb{E}[(y_i - f_{ij})^3]}{\mathbb{V}\mathrm{ar}[y_i]^{3/2}} = \frac{3\rho\sigma_a\sigma_{a^*}\exp(f^*_{ij} + \sigma^2_{a^*}/2 + \sigma^2_{p^*}/2)}{\mathbb{V}\mathrm{ar}[y_i]^{3/2}}$$

Hence, model can accomodate both heavy tailed and symmetric distributions depending on ρ.

Concern: can heavy-tailed sampling distribution lead to spurious positive ρ ?

Box-Cox transformation

Solution of ‘paradox’: model does not fit weight and log body weight equally well, interpret model on the scale for which it is the best fit.

Ongoing research: consider Box-Cox transformations

$$\tilde{\mathbf{y}} = \begin{cases} (\mathbf{y}^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log \mathbf{y} & \lambda = 0 \end{cases}$$

where λ is an additional parameter to be inferred from data.

Issues:

- identifiability ρ and λ
- averaging over posterior of λ ?

That's it - thanks for your attention !