Comparison of Monte Carlo EM REML and Bayesian estimation by Gibbs sampling in estimation of genetic parameters for a test day model

K. Matilainen, M. Lidauer, I. Strandén, R. Thompson, E. A. Mäntysaari

MTT Agrifood Research Finland, Biotechnology and Food Research, Biometrical Genetics Biomathematics and Bioinformatics, Rothamsted Research, United Kingdom.

Introduction

- More advanced models - laborious estimation of variance components
  - Multiple trait Random Regression Test Day model
  - Example:
    3 traits, 2 random effects 3nd order, herd curves 2nd order, random HTD, 19700 animals
    - 213 (co)variance components
    - 676 000 equations in MME
- Method of choice: Gibbs sampler
  - \+ Very low requirements for memory
  - \- Very long computing time
  - \- Poor mixing properties w/ complicated models

Alternative to Bayesian MCMC
Monte Carlo REML

- MCMC EM *(Wei and Tanner, 1990; Guo & Thompson, 1992), SAEM (Kuhn and Lavielle, 2005, Jaffrézic et al. 2006), Thompson (1994)*
  - Instead of solving prediction error variances (PEV) by inversion of MME, use Gibbs sampling
- REML by resampling
  - *Garcia-Cortés et al. 1992, Garcia-Cortés 1994*
  - Samples all model effects (including residuals) simultaneously
    1. Generate all random effects, form observations (**y** = **Zu+e**)
    2. solve BLUP and attain PEV
  - Each sample independent

Methods: Monte Carlo REML

- EM REML (for random effect u, and no missing traits)

$$\mathbf{G}_u^{(k+1)} = \frac{1}{N_u}\left(\sum_{i=1}^{N_u}\hat{\mathbf{u}}_i^{(k)}\hat{\mathbf{u}}_i^{(k)T} + \sum_{i=1}^{N_u}\sum_{j}^{N_u}\ \mathbf{C}_{ij}^{(k)uu}\right)$$

$$\mathbf{R}^{(k+1)} = \frac{1}{N}\left(\sum_{i=1}^{N}\tilde{\mathbf{e}}_i^{(k)}\tilde{\mathbf{e}}_i^{(k)T} + \sum_{i=1}^{N}\mathbf{W}_i\mathbf{C}^{(k)}\mathbf{W}^{T}{}_i\right)$$

- MC EM (with N_h samples of data)

$$\mathbf{G}_u^{(k+1)} = \frac{1}{N_u}\left(\sum_{i=1}^{N_u}\hat{\mathbf{u}}_i^{(k)}\hat{\mathbf{u}}_i^{(k)T} + \frac{1}{N_h}\sum_{h}^{N_h}\left[\sum_{i=1}^{N_u}(\tilde{\mathbf{u}}_i^{(k)h} - \bar{\tilde{\mathbf{u}}}_i^{(k)h})(\tilde{\mathbf{u}}_i^{(k)h} - \bar{\tilde{\mathbf{u}}}_i^{(k)h})^T\right]\right)$$

$$\mathbf{R}^{(k+1)} = \frac{1}{N}\left(\sum_{i=1}^{N_u}\hat{\mathbf{e}}_i^{(k)}\hat{\mathbf{e}}_i^{(k)T} + \frac{1}{N_h}\sum_{h}^{N_h}\left[\sum_{i=1}^{N_u}(\tilde{\mathbf{e}}_i^{(k)h} - \bar{\tilde{\mathbf{e}}}_i^{(k)h})(\tilde{\mathbf{e}}_i^{(k)h} - \bar{\tilde{\mathbf{e}}}_i^{(k)h})^T\right]\right)$$

More formulas

http://commons.wikimedia.org/wiki

Methods: MC REML details

- MC EM (with relationships)

$$\mathbf{G}_u^{(k+1)} = \frac{1}{N_u}\left(\sum_{i=1}^{N_u}(\mathbf{L}\hat{\mathbf{u}})_i^{(k)}(\mathbf{L}\hat{\mathbf{u}}^{(k)})_i^T + \frac{1}{N_h}\sum_h^{N_h}\left[\sum_i^{N_u}(\mathbf{L}(\tilde{\mathbf{u}}-\tilde{\tilde{\mathbf{u}}})^{(k)h})_i(\mathbf{L}(\tilde{\mathbf{u}}-\tilde{\tilde{\mathbf{u}}})^{(k)h})_i^T\right]\right.$$

- Where $\mathbf{L}\mathbf{L}^T = \mathbf{A}^{-1}$

- If missing traits, then residuals weighted

$$\sum_{j=1}^{N} tr(\mathbf{R}_j^{-1}\mathbf{E}_{mn}\mathbf{R}_j^{-1}\hat{\mathbf{R}}_0^{k+1}) = \sum_{j=1}^{N} tr(\mathbf{R}_j^{-1}\mathbf{E}_{mn}\mathbf{R}_j^{-1}\overline{\mathbf{R}}_j^k)$$

where $\overline{\mathbf{R}}_j^k = \tilde{\mathbf{e}}_j^k\tilde{\mathbf{e}}_j^{kT} + \frac{1}{N_h}\sum_h^{N_h}(\tilde{\mathbf{e}}_j^h - \tilde{\tilde{\mathbf{e}}}^h{}_j)(\tilde{\mathbf{e}}_j^h - \tilde{\tilde{\mathbf{e}}}^h{}_j)^T$

$$\text{where, } \mathbf{E}_{mn} = \frac{\partial \mathbf{R}_j}{\partial r_{mn}}$$

MC EM implementation details

- Requires frequent BLUP solves: We used PCG
- Alternatives to use $\tilde{\tilde{\mathbf{u}}}$: PEV1, PEV2, PEV3 (Garcia-Cortéz et al. 1995)
- Assesment of convergence
  - For algorithm fine tuning: $\left\|\tilde{\Theta}^k - \Theta^{pseudo\ true}\right\| / \left\|\Theta^{pseudo\ true}\right\|$
  - For practical use: $c_d{}^{(k)} = \frac{\left\|\tilde{\Theta}^{(k)}(\rho_i) - \tilde{\Theta}^{(k)}(\rho_{i-1})\right\|}{\left\|\tilde{\Theta}^{(k)}(\rho_i)\right\|}$

where $\tilde{\Theta}^{(k)}(\rho_i)$ is a linear predictor of solutions in round k using solutions from rounds k-p to k

MC EM implementation details

- Different number of samples were used per round
- Different convergence requirements for PCG
  - No results shown: High convergence level important

Test data sets

- Data set and model 1:
  - Small data that was possible analyze with analytical EM (EM REML)
  - 5,339 first lactation cows
  - 51,004 TDs modeled, with 155,300 equations
  - Were run by MC REML, EM REML (DMUv.6) and Bayesian MCMC
- Data set and model 2:
  - 85,007 TD records of milk, protein and fat
  - Pedigree with 31,255 animals
  - 19,709 first lactation cows, 675,700 equations
  - Were run by MC REML and Bayesian MCMC

Data set 2 (large)

Heritabilities, genetic (above) and phenotypic (below) correlations

MC EM	Milk	Prot	Fat
Milk	.38	.85	.67
Prot	.93	.32	.79
Fat	.80	.84	.34

GS	Milk	Prot	Fat
Milk	.34	.86	.69
Prot	.93	.31	.78
Fat	.80	.84	.33

Data set 2 (large)

Heritabilities, genetic (above) and phenotypic (below) correlations

MC EM	Milk	Prot	Fat
Milk	.38	.85	.67
Prot	.93	.32	.79
Fat	.80	.84	.34

GS	Milk	Prot	Fat
Milk	.34	.86	.69
Prot	.93	.31	.78
Fat	.80	.84	.33

Data set 2 (small)

Heritabilities, genetic (above) and phenotypic (below) correlations

MC EM	Milk	Prot	Fat
Milk	.34	.88	.68
Prot	.94	.26	.78
Fat	.82	.85	.26

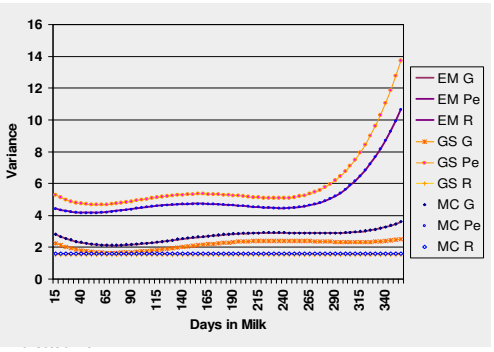
GS	Milk	Prot	Fat
Milk	.27	.89	.66
Prot	.95	.23	.77
Fat	.82	.85	.24

Data set 2 (small)
Heritabilities, genetic (above) and phenotypic correlations with EM REML

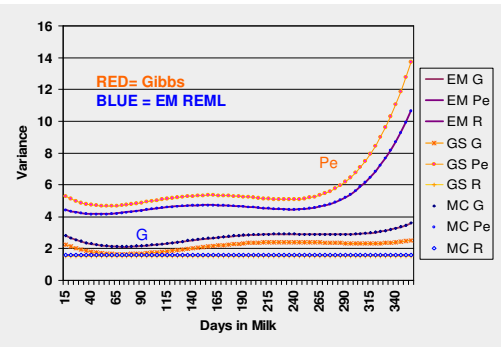
MC EM	Milk	Prot	Fat
Milk	.34	.88	.68
Prot	.94	.26	.78
Fat	.82	.85	.26

DMU EM	Milk	Prot	Fat
Milk	.34	.88	.68
Prot	.94	.26	.77
Fat	.81	.85	.26

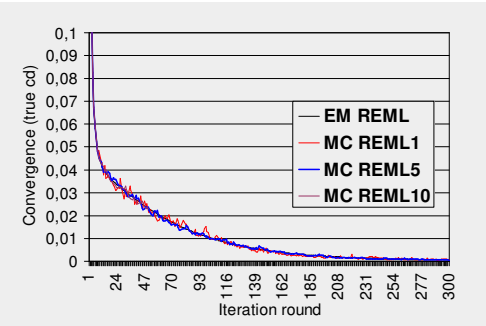
Daily variance estimates for milk (small data)
Genetic (G), Permanent env. (Pe), Residual (R)



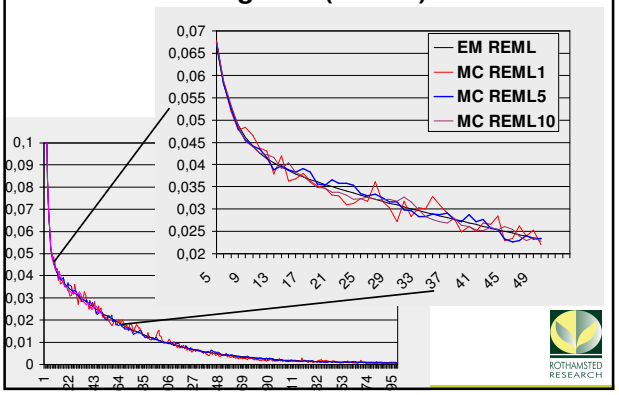
Daily variance estimates for milk (small data)
Genetic (G), Permanent env. (Pe), Residual (R)



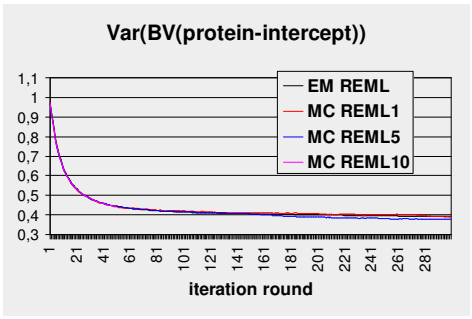
General convergence



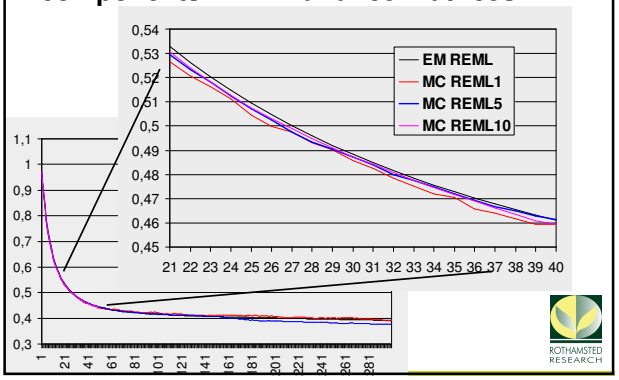
General convergence (..zoom)



Convergence of individual components in RR variance matrices



Convergence of individual components in RR variance matrices



Convergence and stability

- Surprisingly large round-to-round variation remains even after 10+ samples
  - Does not affect convergence
- EM REML
  - Needed 335 rounds to reach Cd = 10^{-4}
  - Needed 935 rounds to reach Cd = 10^{-5}

  (MC REML took 1219 REML rounds to reach of $10^{-4.5}$)
- GS was done 200,000 rounds
  - with the first 50,000 as burn-in period

Results: Computing requirements

EM REML§	110 days 5 hours	300 rounds
MC REML5	16 hours	300 rounds
Gibbs sampler	13 days 18 hours	200 000 iterates

§ EM REML had MME 6.4 milj non-zeros

Conclusions

- Different solutions for parameters from Bayesian and likelihood methods
- Solutions from small analysis confirmed differences between REML and Gibbs sampler
  - Reasons caused by either by implementation or the model
- The MC EM analysis was found reliable
  - with small memory need
  - relatively short computing time
  - Typical EM convergence (slow…)
- Computing strategy for MC EM and convergency criterion remained still to be optimized
- Both the EM approaches were found slow to converge
  - RR model is very pathological

http://commons.wikimedia.org/wiki/

Acknowledgements

Development and implementation of Monte Carlo EM REML is part of a Ph.D. study financed by Rothamsted Research and MTT

Test data described here was received from NAV, the Nordic Cattle Genetic Evaluation

Nordisk Avlsvärdivurdering