MiX99

Solving Large Mixed Model Equations

Heterogeneous variance adjustment

Martin Lidauer

Motivation

Wanted:

2000: Heterogeneous variance (HV) adjustment for Finnish TDM

  - Should adjust for the main sources of HV
  - Should be suitable for multi-breed herds

2002: HV adjustment for German-Austrian TDM

  - Very different production systems/regions
  - Different milk recoding schemes

- Multiplicative methods were found most appealing
  - Heterogeneous model: Robert-Granié et al. 1999 (Livest. Prod. Sci. 60:)
  - Multiplicative mixed model: Meuwissen et al. 1996 (J. Dairy Sci. 79:)
- Multiplicative mixed model chosen
  - Easier to implement because updates are only done on RHS of the MME

Multiplicative mixed model approach

- 2 models are solved simultaneously
  1. A model to obtain EBVs:

     $\mathbf{y}_{ti}\lambda_{ti} = \mathbf{X}_{ti}\mathbf{b}_t + \mathbf{T}_{ti}\mathbf{h}_t + \mathbf{Z}_{ti}\mathbf{a} + \mathbf{U}_{ti}\mathbf{p} + \mathbf{V}_{ti}\mathbf{w} + \mathbf{e}_{ti}$, where

     $\mathbf{y}_{ti}$ is a vector with observations for trait *t* in stratum *i*

     λ_{ti} is a multiplicative adjustment factor for stratum *i*
  2. A log-linear model that models heterogeneity of variances:

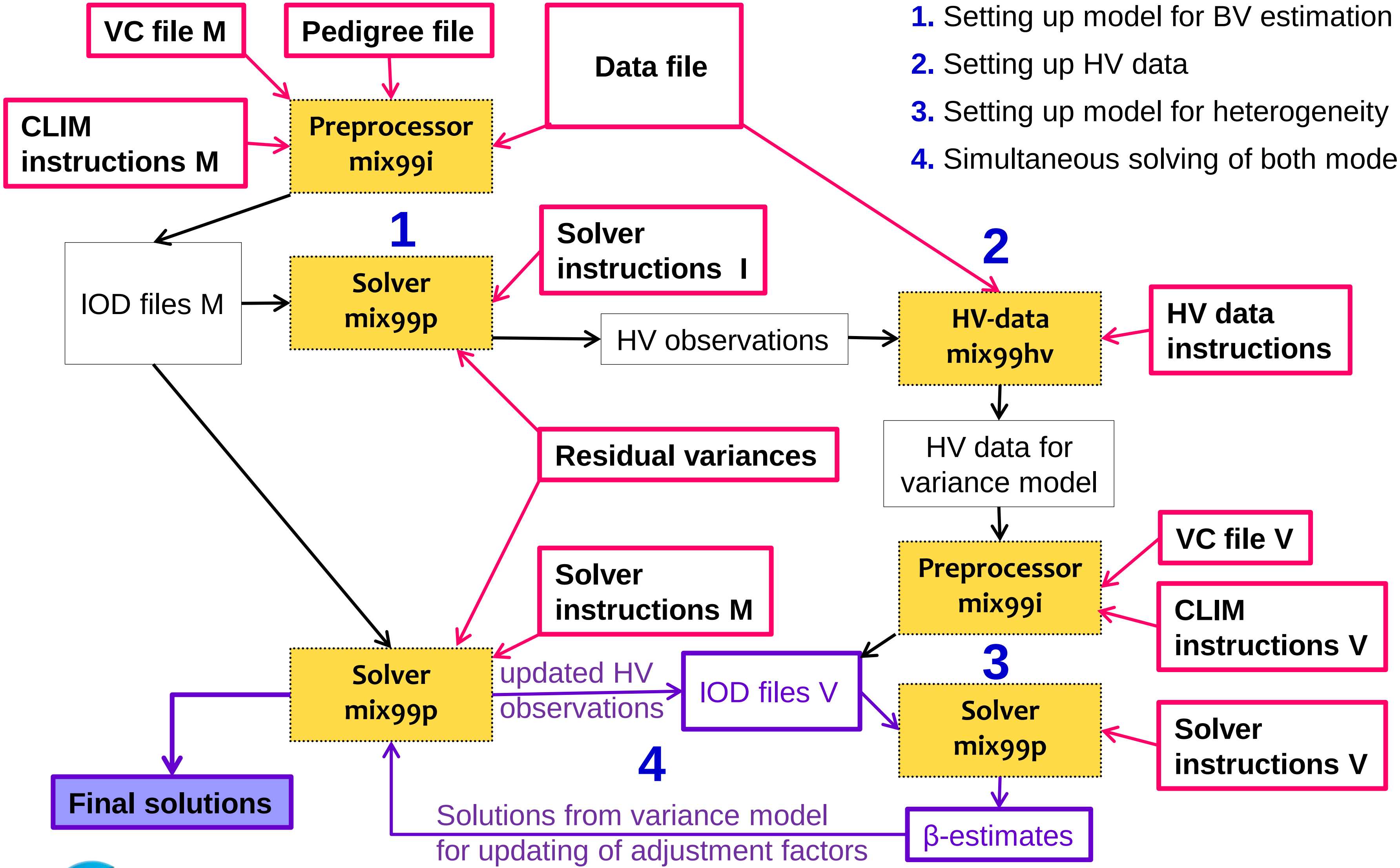
     $s_{thym:j} = \beta_{1_{tym}} + \beta_{2_{thy}} + \varepsilon_{thym:j}$, where

     $s_{thym:j}$ is a heterogeneity observation for stratum *ti (i=hym)*

  AND $\lambda_{ti} = exp\left(-0.5\left(\hat{\beta}_{1_{tym}} + \hat{\beta}_{2_{thy}}\right)\right)$
- The system converges to satisfy

  $(\mathrm{n}_{\mathrm{ti}} - \mathrm{r}_{\mathrm{ti}})\hat{\sigma}^2_{\mathrm{e}_t} = \mathbf{y}^{\mathrm{T}}_{ti}\lambda_{\mathrm{ti}}\hat{\mathbf{e}}_{ti}$, where $\hat{\sigma}^2_{\mathrm{e}_t}$ is the residual variance for trait *t*

Computational approach

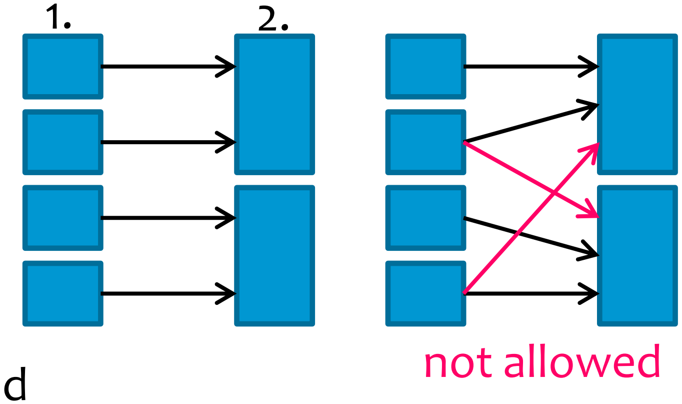


1. Setting up model for BV estimation
2. Setting up HV data
3. Setting up model for heterogeneity
4. Simultaneous solving of both models

Modelling of heterogeneity

- To avoid extensive increase in I/O work and computations, some restrictions on possible variance models are imposed
  - The model must contain 2 effects of which the second effect can be a fixed or random
  - Optionally, an AR(1) process can be applied for the random effect
  - The second effect must contain a BLOCK interaction, where BLOCK is identically with the BLOCK variable specified in the DATASORT command
  - Within a data BLOCK, the levels of the 1. effect must be nested within the levels of the 2. effect

REASON: Within a data BLOCK, for each level of the first effect only one HV observation is built (to reduce I/O-work and computing time)



- In future, these restrictions can be relaxed

Adjusting for heterogeneity

- The multiplicative mixed model scales all model effects in proportion

$$\mathbf{y}_{\mathrm{ti}} = (\mathbf{X}_{\mathrm{ti}}\mathbf{b}_{t} + \mathbf{T}_{\mathrm{ti}}\mathbf{h}_{t} + \mathbf{Z}_{\mathrm{ti}}\mathbf{a} + \mathbf{U}_{\mathrm{ti}}\mathbf{p} + \mathbf{V}_{\mathrm{ti}}\mathbf{w} + \mathbf{e}_{\mathrm{ti}})/\lambda_{\mathrm{ti}}$$

- Alternatively, only random effects can be scaled

$$\mathbf{y}_{\mathrm{ti}} = \mathbf{X}_{\mathrm{ti}}\mathbf{b}_{t} + \mathbf{T}_{\mathrm{ti}}\mathbf{h}_{t} + (\mathbf{Z}_{\mathrm{ti}}\mathbf{a} + \mathbf{U}_{\mathrm{ti}}\mathbf{p} + \mathbf{V}_{\mathrm{ti}}\mathbf{w} + \mathbf{e}_{\mathrm{ti}})/\lambda_{\mathrm{ti}}$$

  - Scaling random effects only improves convergence (**Pena et al. 2002**. Interbull Bulletin 29:)
  - But it may reduce fit of the model (**Pena et al**. 2005. Interbull Bulletin 33:)
  - Similarly we found significant differences between both adjustment methods in the ability of the variance model to model heterogeneity

Comparison of multiplicative heterogeneous variance adjustment models for genetic evaluations.
Márkus et al., 2013. J. Anim. Breed. Genet. 131:

- MiX99 allows to exclude or include any effect into the scaling

Preparation of the model for BV estimation

Pre-processor mix99i run:

  - Model needs to account for changes in the means due to scaling
  - CLIM instructions as usual

2 parallel solver mix99p runs:

- Initialization run **I**
  - Only 20 iterations
  - HV adjustment START-UP CYCLE
  - Scaling of model effects by HV adjustment
    1= scaling; 0=no scaling
  - No solutions are written
- Multip. solving of both models **M**
  - Stopping criteria as usual
  - HV adjustment: CYCLE (C) between MODELS
  - Scaling of model effects by HV adjustment
  - Stopping criteria for adjustment factors

Solver instructions I

```
# RAM: RAM demand: H=high, M=medium, L=low
  H
# STOP: max. iterations, stopping criteria (CR)
  20                      1.0e-7
# RESID: Calculate residuals? (Y/N)
  N
# VALID:
  N
# VAROPT: START-UP CYCLE for HV adjustment
  S
  # ADJUST: HerdYear YearMonth, Age, PE, Animal
                1        1       1    1    1
# SOLTYP: type of solution files? (N,Y,A)
  N
```

Solver instructions M

```
# RAM: RAM demand: H=high, M=medium, L=low
  H
# STOP: max. iterations, stopping criteria (CD)
  2500                    1.0e-5     d      f
# RESID: Calculate residuals? (Y/N)
  N
# VALID:
  N
# VAROPT: CYCLE between MODELS for HV adjustment
  C
  # ADJUST: HerdYear YearMonth, Age, PE, Animal
                1        1       1    1    1
  # STOPC: max. cycles, convergence of lambda values
  94                      1.0e-8
# SOLTYP: type of solution files? (N,Y,A)
  Y
```

Setting up HV data for the variance model

mix99hv program sets up working files with all required information for the variance model

Required instructions

- Basic information about the data file
- Trait group code (optional)
- BLOCK variable
- Classification information for building strata and the variance model effects

Useful options

- AR(1) information file for REML
- Standardization of adjustment factors
- Approximation of rank
- Combining of second effect across traits
- Working file directory

HV data instructions

```
#  Name of the MiX99 input data file
   /home/ejo31/BVestimation.dat
#  Number of integer columns; type of file (f/u)
   5  f
#  Column with the trait group code?
   0
#  Column with the BLOCK variable (Herd)
   3
#  Classification information for strata building
#  Across block strata; Within block strata
#  (YearMonth)           (Year)
        5                  4
#  Create AR(1) information file for REML? (y/n)
   n
#  Standardization of adjustment factors? (y/n)
   y
   # Name of file with the class codes of the base
     ID.Base.Classes
#  Approximation of rank for residual variance estim.
   y
#  Combining of effect across traits
   1
#  Directory for temporary working files
   /home/ejo31/tmpHV
```

Preparation of the variance model

Pre-processor mix99i run:

- Most of input information is pre-defined
- CLIM instructions as usual
- Only information **in red** has to be specified
- In case of multiple trait models, traits are assumed to be uncorrelated
- Same number of cores as for solving the breeding value model

Parallel solving mix99p run **V**

- Solver options are predefined
- For each multiplicative cycle 50 PCG iterations are done on the variance model
- Solutions from the variance models are accelerated. Recommended option: H (staggered Aitken half-Chebychev)

CLIM instructions V

```
TMPDIR          ../tmpMiX/B1
PEDFILE         ../tempHV/HV.pedi
DATAFILE        ../tempHV/HV.data
INTEGER         BLOCK BLOCKxYear TRGRP YearMonth
REAL            s_Milk wgt_Milk
MISSING         -8192.0
DATASORT BLOCK=BLOCK PEDIGREECODE=BLOCKxYear
TRAITGROUP      TRGRP
PEDIGREE        BLOCKxYear ar
AR              0.75
RANDOM          BLOCKxYear
PARFILE         variance.model.var
MODEL
                s_Milk(1) = YearMonth BLOCKxYear
PARALLEL        2 0
```

Solver instructions V

```
# RAM: RAM demand: H=high, M=medium, L=low
  H
# STOP: max. iterations, stopping criteria (CD)
  50                     1.0e-7      d
# RESID: Calculate residuals? (Y/N)
  N
# VALID:
  N
# VAROPT:
  N
# SOLTYP: type of solution files? (N,Y,A,H)
  H
```

Solving of the multiplicative model

- Both models are solved simultaneously using MiX99 parallel solver
- The whole solving process controlled by a solving script
    - Only a few key parameters need to be specified (indicated in red)
- A certain, optimized number of iterations on the model for breeding values follow 50 iterations on the variance model
- Once adjustment factors have converged, iterations continue on the breeding value model only until convergence

Solving script

```
#---------------------------------------------------------------------------#
echo ' STEP 0: PARAMETERS '
#---------------------------------------------------------------------------#
  # number of processes used.
  Nproc=2
  # path for binaries.
  BIN=/home/ejo31/MiX99/MiX99work/bin
  # path for mpi call
  CALLMPIrun=/opt/mpich/intel/bin/mpirun
  echo ' Number of processes used: ' $Nproc
#---------------------------------------------------------------------------#
echo ' STEP 1: INITIALIZE MEAN MODEL '
#---------------------------------------------------------------------------#
  date
# run the preprocessing for the mean model
  $BIN/mix99i <MEAN_MODEL.DIR >Mi.log
# make the index file for the mean model
  $BIN/imake99 >imake99.log
# run the first iterations on the mean model
  $CALLMPIrun -np $Nproc $BIN/mix99p <stpA>Ma.log
#---------------------------------------------------------------------------#
echo ' STEP 2: CREATE DATA FOR VARIANCE MODEL '
#---------------------------------------------------------------------------#
  $BIN/mix99hv<HVDATA.DIR>HVd.log
#---------------------------------------------------------------------------#
echo ' STEP 3: INITIALIZE VARIANCE MODEL '
#---------------------------------------------------------------------------#
date
# go into the variance model directory
  cd B1
    # run the preprocessing for the variance model
    $BIN/mix99i <VARIANCE_MODEL.DIR>Vi.log
    # make the index file
    $BIN/imake99>imake99.log
    # run the first iterations on the variance model
    rm Vs.log
    cat > solve_variance_model <<!!
       $CALLMPIrun -np $Nproc $BIN/mix99p<stpV>>Vs.log
!!
    chmod u+x solve_variance_model
    ./solve_variance_model
    # go back to the mean model directory
  cd ../
#---------------------------------------------------------------------------#
echo ' STEP 4: CYCLE BETWEEN MEAN MODEL AND VARIANCE MODEL '
#---------------------------------------------------------------------------#
  date
  $CALLMPIrun -np $Nproc $BIN/mix99p<stpC
#---------------------------------------------------------------------------#
echo ' MULTIPLICATIVE MIXED MODEL SOLVED ‘
  date
#---------------------------------------------------------------------------#
```

Example A:

HV adjustment for German-Austrian test-day model evaluation

- Random regression test-day model for milk yield in 1. to 3. lactation
- Large heterogeneity between herds and regions

Region		Number of herds	Contemporary group size		1st parity milk yield	
			All parities	1st parity	Mean	SD
G3	Münchner Raum	2945	23.8	7.8	18.8	4.26
G8	Franken	4258	23.4	8.1	19.1	4.59
G12	Baden-Württemberg	1879	21.6	6.9	17.8	4.01
A2	Hochalpen	1629	8.0	2.4	17.9	3.71
A3	Wald- u. Mühlviertel	4220	13.2	3.8	18.2	4.18

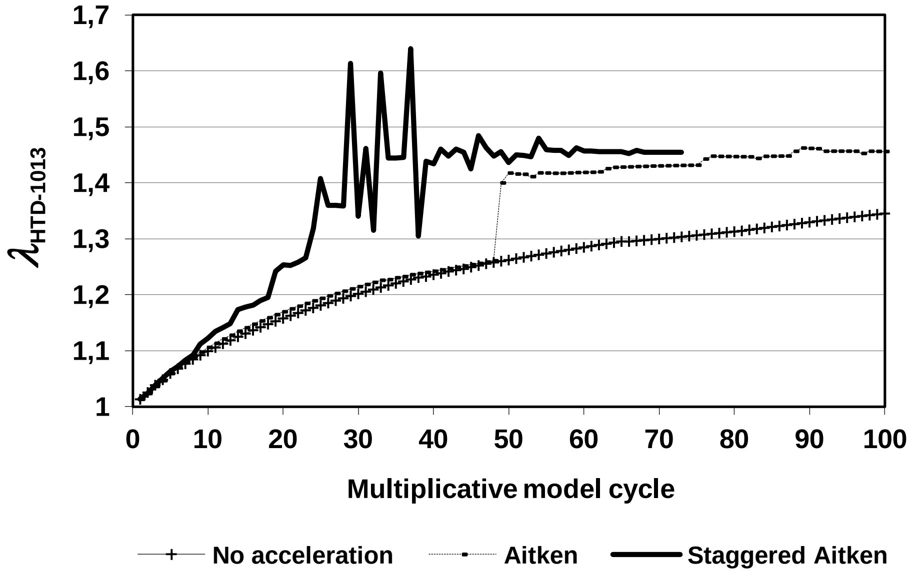
- Applied variance model:
  - region × year × month (fixed) + herd × test-month (random) with AR(1) process within herds and $\rho=0.99$
  - each lactation as own trait
- 10 million variance model equ.

Multiplicative random regression model for heterogeneous variance adjustment in genetic evaluation for milk yield in Simmental .
Lidauer et al., 2008. J.Anim.Breed.Genet. 125:

Example A:

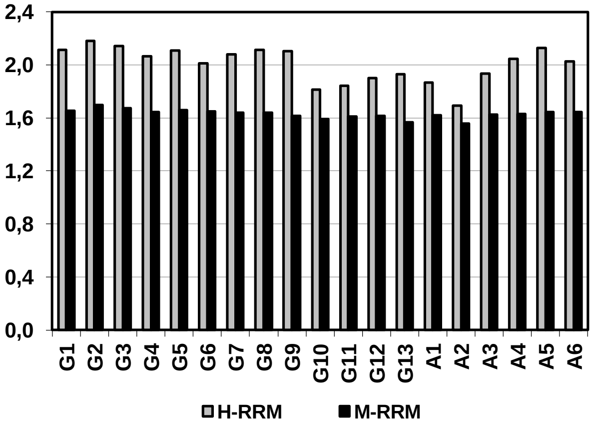
Convergence of adjustment factors

- Acceleration of variance model solutions reduces computation time significantly
- Staggered Aitken has fixed acceleration points and fits better for routine runs

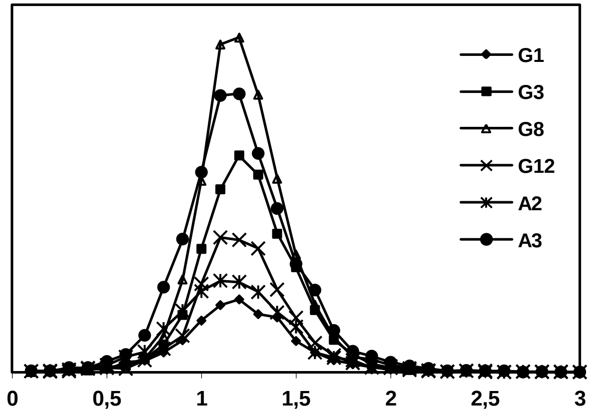


Residual SD after adjustment

- Same across regions



- Vary across herds because of random HTM effect



Example B:

HV adjustment for Nordic across-country test-day model evaluation

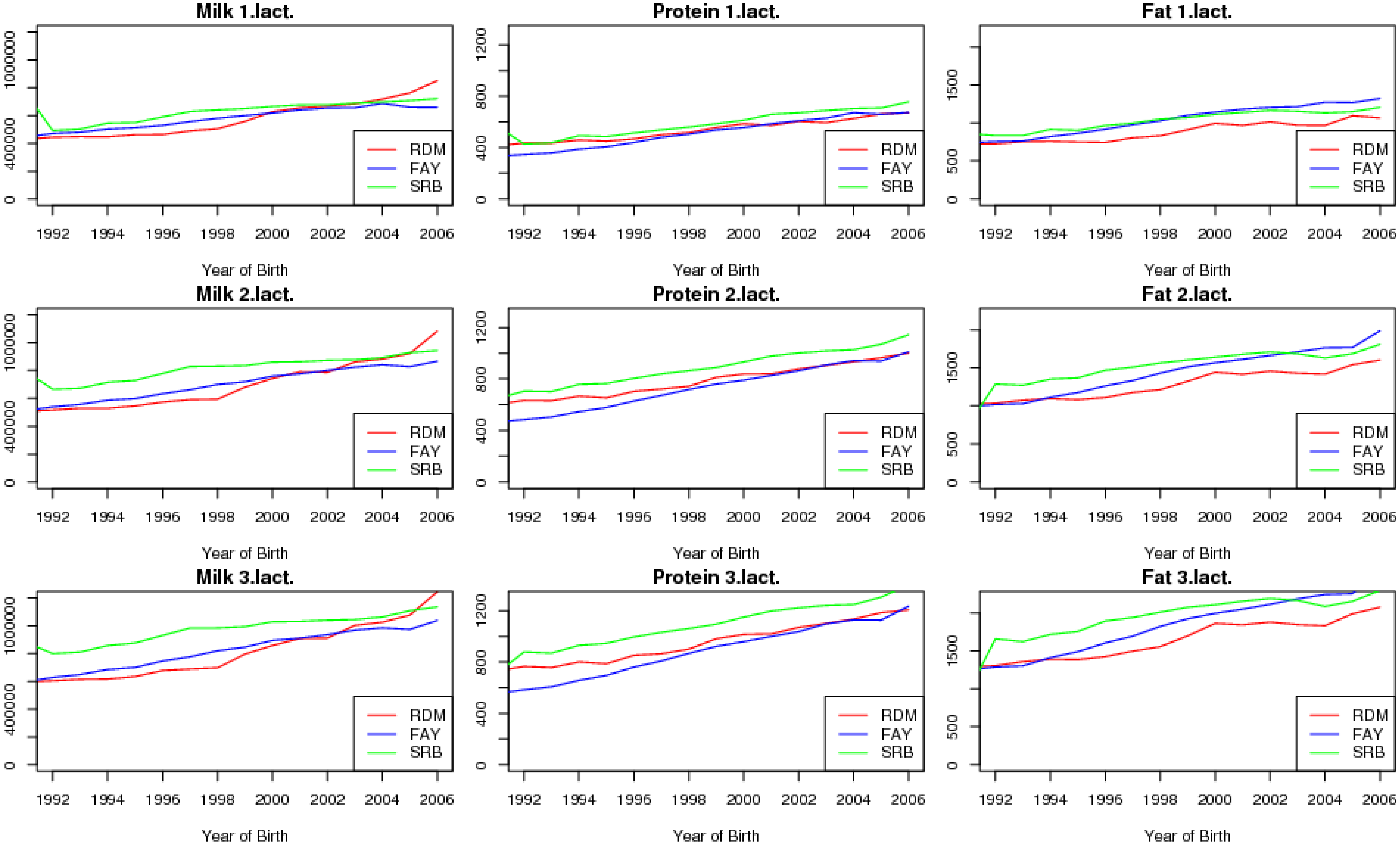
- Multiple-trait RRM with 9 traits
- 27 traits for HV adjustment (3× 9)

Across-country test-day model evaluations for Holstein, Nordic Red Cattle, and Jersey
Lidauer et al., 2015. J.Dairy Sci. 98:

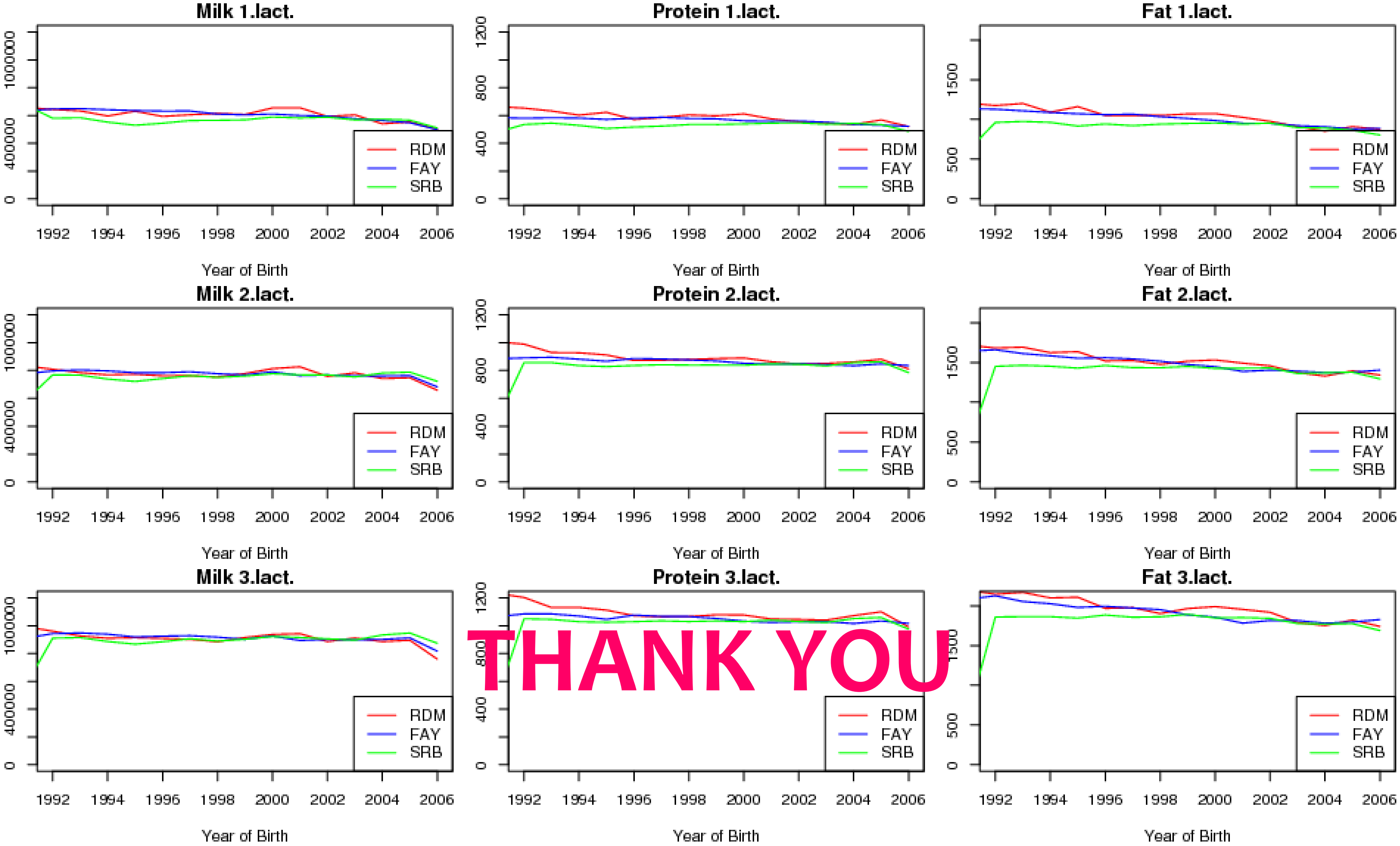
- Applied variance model:
  - year × month (fixed) + herd × test-year (random) with AR(1) process within herds and ρ 0.70 and 0.80, for first and later lactations
- For each country own set of standardization (residual) variances
  - Allows to get homogeneous genetic variances across countries
- **Under development:**
  - Own sets of standardization (residual) variances for observations from automatic milking systems

Simulation study on heterogeneous variance adjustment for observations with different measurement error variance.
Pitkänen et al., 2013. Interbull Bulletin. 47:

Example B: Genetic variances without HV adjustment



Example B: Genetic variances with HV adjustment



THANK YOU