Reliable computing in estimation of variances in swine

Ignacy Misztal

University of Georgia

USA

Desirables

- Multi-trait of correlated and important traits
- Reflect nature of traits:
  - Normally distributed?
  - Continuous, categorical
  - censored, hazard (survival),…
  - Discrete or longitudinal?
  - Maternal effects? Correlated?
- Account for "important" as opposed to significant effects
- Different purposes
  - Reflect biology of traits
  - Genetic evaluation

Practical estimation

- Edit data
- Select model according to limitations
- Run program (ASREML?)
- Do results make sense?
- If yes, ☺

- If no:
  - time available: refine model and continue
  - Time ran out: justify and submit!

Typical methodologies

- General REML
  - DF
  - EM
  - AI
- REML by canonical transformation
- MCMC
  - Simple
  - Optimized

General REML (EM & AI)

- Cost t^3, $\text{animal}^{2\ldots3}$
- EM
  - Stable (except RRM)
  - Slow (50-200… rounds)
  - Simple to program
- AI
  - Fast (4-200 … rounds)
  - Heuristics needed
  - Complex to program

REML - cont

- Good with small data sets
- Breaks down with many traits

- Canonical transformation
  - Low cost with large number of traits
  - Model limitations

- Hard to determine formulas / program for complex models (e.g., threshold, censored,…),especially MT

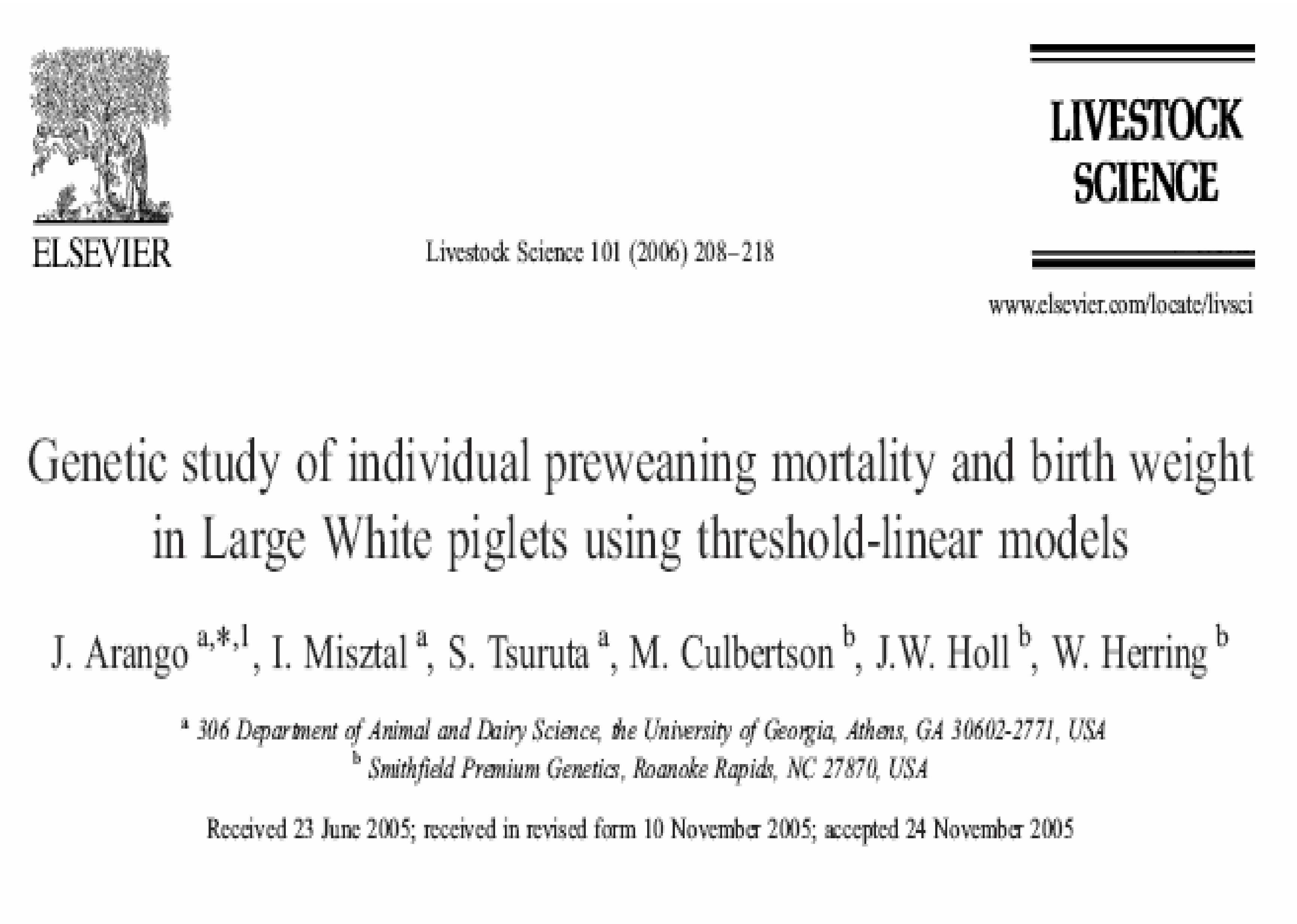
MCMC

- Simple to program incl. complex models
- Small memory requirements
- Speed dependent on optimization
- Details determine quality
- Convergence sometimes hard to determine
- Priors
  - Can make any model converge
  - Flat priors good for large but not small data sets
- No problem with many traits if optimized

Optimization with Gibbs samplers

- Iteration on data
  - Natural choice for RRM
  - Hard with maternal effects or irregular models
- **Storage of only single-trait matrices (a la canonical transformation)**
  - No problem with many random effects
  - Missing traits predicted
  - Different designs through pseudo-random effects
- …..

Selected Projects at UGA



ELSEVIER

Livestock Science 101 (2006) 208–218

LIVESTOCK SCIENCE

www.elsevier.com/locate/livsci

Genetic study of individual preweaning mortality and birth weight in Large White piglets using threshold-linear models

J. Arango [a,*,1], I. Misztal [a], S. Tsuruta [a], M. Culbertson [b], J.W. Holl [b], W. Herring [b]

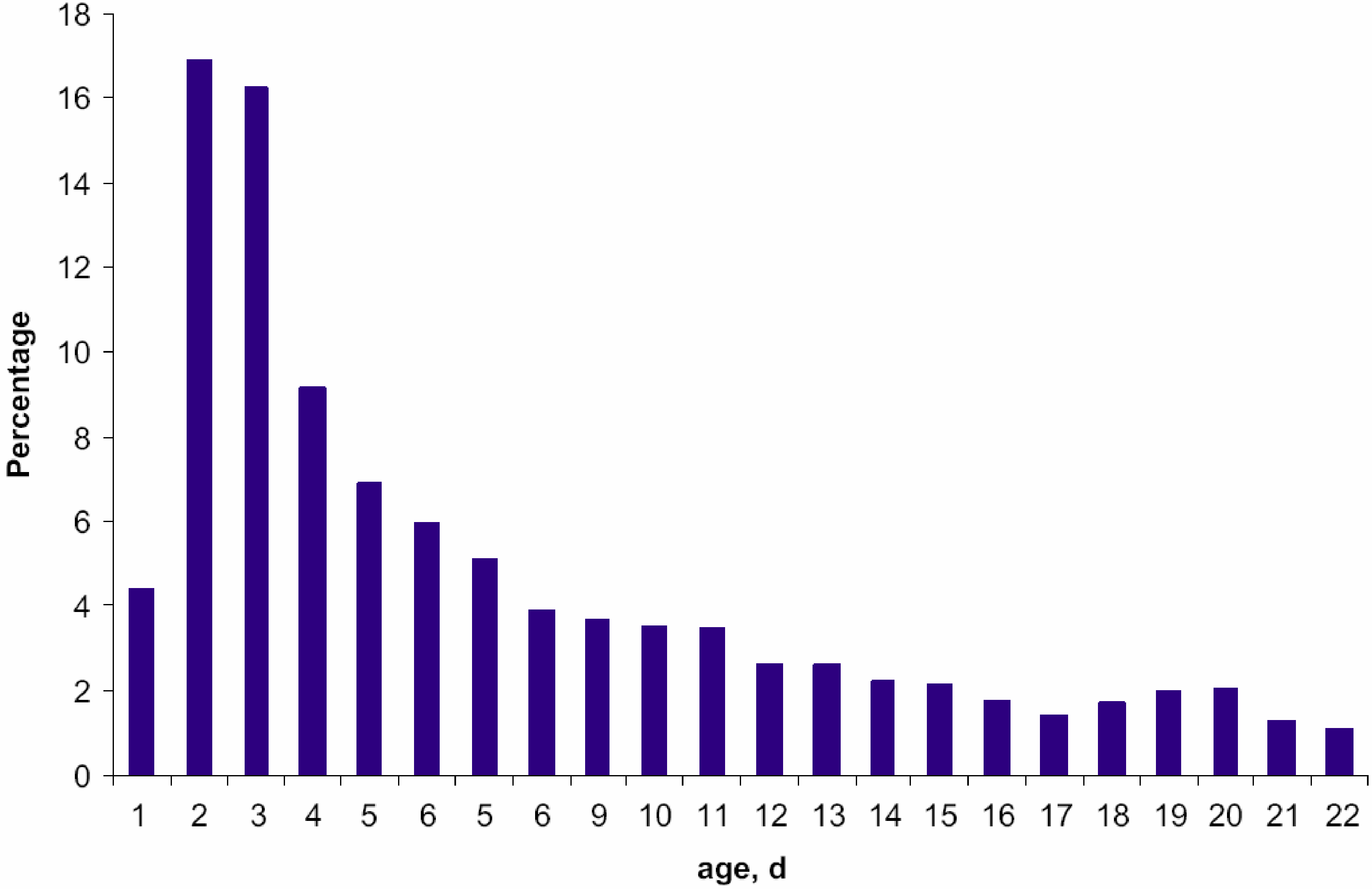
[a] 306 Department of Animal and Dairy Science, the University of Georgia, Athens, GA 30602-2771, USA

[b] Smithfield Premium Genetics, Roanoke Rapids, NC 27870, USA

Received 23 June 2005; received in revised form 10 November 2005; accepted 24 November 2005

Traits

- Number of stillborn
- Mortality as f(days)
- Birth weight (not for stillborn)


Percentage
0
2
4
6
8
10
12
14
16
18
1
2
3
4
5
6
5
6
9
10
11
12
13
14
15
16
17
18
19
20
21
22
age, d

Attempts and Problems

- Issues
  - Some traits binary
  - Some traits direct (BW), some maternal, some uncertain (stillbirth)

- Initial plans
  - All traits altogether
  - Mortality as continuous

- Convergence problems
  - Limited data with binary variables, also low incidences

- Final choices
  - Mortality treated as one, early, or late
  - Several models with maternal effects

Models by trait combination for individual piglet birth weight, preweaning mortality, and litter farrowing mortality

Model	Trait combination[a]						
	BW	TM	SB	ELM	EM	LM	NBD
1	✓	✓	–	–	–	–	–
2	✓	–	✓	✓	–	–	–
3	–	–	✓	–	✓	✓	–
4	✓	–	–	✓	–	–	✓

[a] BW=piglet birth weight; TM=total preweaning mortality including stillbirth; SB=stillbirth; ELM=preweaning mortality; EM=early preweaning mortality; LM=late preweaning mortality; NBD=number of piglet born dead at litter level.

Methodology

- Program THRGIBBSF90 (Lee and Tsuruta)
  - Any number of categorical and linear traits
  - Flat priors
  - Optimizations:
    - Block sampling by traits and direct-maternal
    - Single-trait left-hand side in pieces created once

  - Up to 250,000 samples required for some models
  - Computing - up to a few days

Estimation of genetic correlation between purebreds and crossbreds

- Purebreds - only paternal lines
- Crossbreds: dams identified but no pedigrees

Parameters of crossbred model

(Zumbach et al., 2006)

Terminal cross model by Lo et al. (1997):

$$\begin{array}{llll} y_A = .. + Z_A & u_A & & ...+ e_A \\ & & & \\ y_C = .. + Z_{AC} & u_{AC} & + \ Z_D u_D & ...+ e_C \end{array}$$

A –purebred
C – crossbred
y – trait value
u – additive effects; d – dam effects

Computing

- AI REML
- If maternal effects fit for purebreds:
  - Convergence 5 → 200 rounds
  - Very small variance for maternal

- Old guideline for estimation of maternal effects (Quaas): enough MGS (or dams) with own records

Analyzes of number of born alive and dead

- First parity only
  - Backfat
  - Days to reach 113.5 kg
- Three parities
  - Number of born alive
  - Number of born dead
- Born dead treated as categorical
- All traits too many!

Analyzes (Arango et al., 2005)

- First model
  - Number of born alive in first parity
  - Number of born dead in first parity
  - Backfat
  - Days to reach 113.5 kg
- Second model - 3 parities
  - Number of born alive
  - Number of born dead (categorical)
- No major computing problems - large data sets

Survival for sows

- Many reasons for disposal

- Why sow disposed?
  - Genes (QTLs) for each reason separately?
  - General poor fitness?

- Few general categories for disposal
  - Reproduction, disease, other

Can all be analyzed jointly?

Traits combinations

Parity at Disposal

Repro	Disease	Other
2	2+	2+
3+	**3**	3+
1+	1+	**1**

One trait observed, others censored

Computing (Arango et al., 2005)

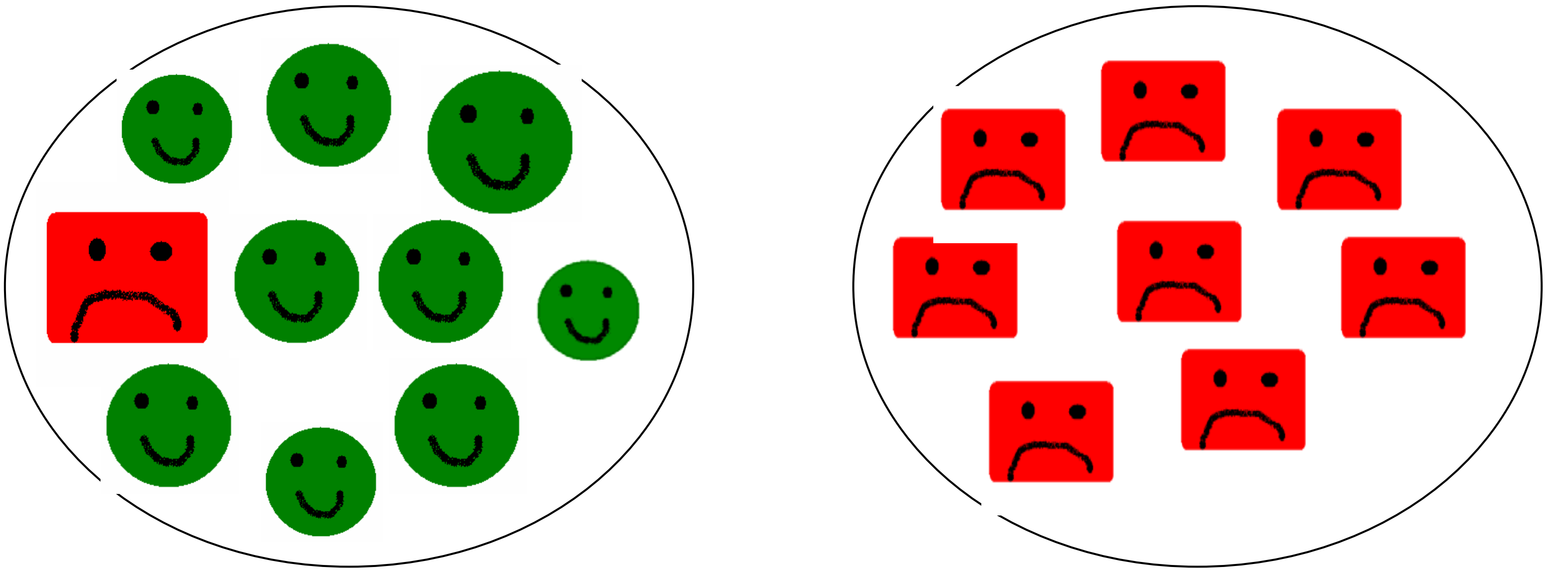
Rework of existing Gibbs samplers to censored

Traits categorical – no convergence

Traits linear – slow mixing but convergence

Competition effects

(Muir and Schinkel, 2002)



Competition model

d_i, c_i -- direct and competitive effects for animal i

$y = \ldots\ d_i + \sum c_j\ \ldots\ +e$ $\qquad \mathrm{var}(c,d) = G_0 \otimes A$

All c's in $\sum c_j$ contribute to same effect

Issues and experiences

(Arango et al., 2005)

- Implementation with simple MCMC and REML programs without modifications – optimized versions did not work!
- Good results with simulated data
- Real data set – convergence problems due to very flat likelihood
- Corr(d,c) dropped, L computed for several values of var(c) - similar to DF REML
- Model not realistic -- expression of competitiveness not linear but categorical

Fertility in Uruguayan Herefords (Urioste et a., 2006)

- Traits
  - days from exposure to bull to conception
  - 3 parities
  - some cows do not calve in some years

- Ways of treating missing calvings
  - Penalized (missing = max+20d)
  - Censored
  - 2 traits: days + calving success (binary)

Analyses and results

- Some 5000 cows with data
- 3-6 traits analyzes

- Problems until fixed effects refined

- Split data correlations of EBV for days
  - Penalized: 0.4
  - Censored: 0.48
  - Threshold-linear: 0.65

- Censored model less correct?

Changes in genetic parameters over time

Genetic parameters assumed constant in time

prediction of correlated responses

Is this correct?

Multitrait random regression model

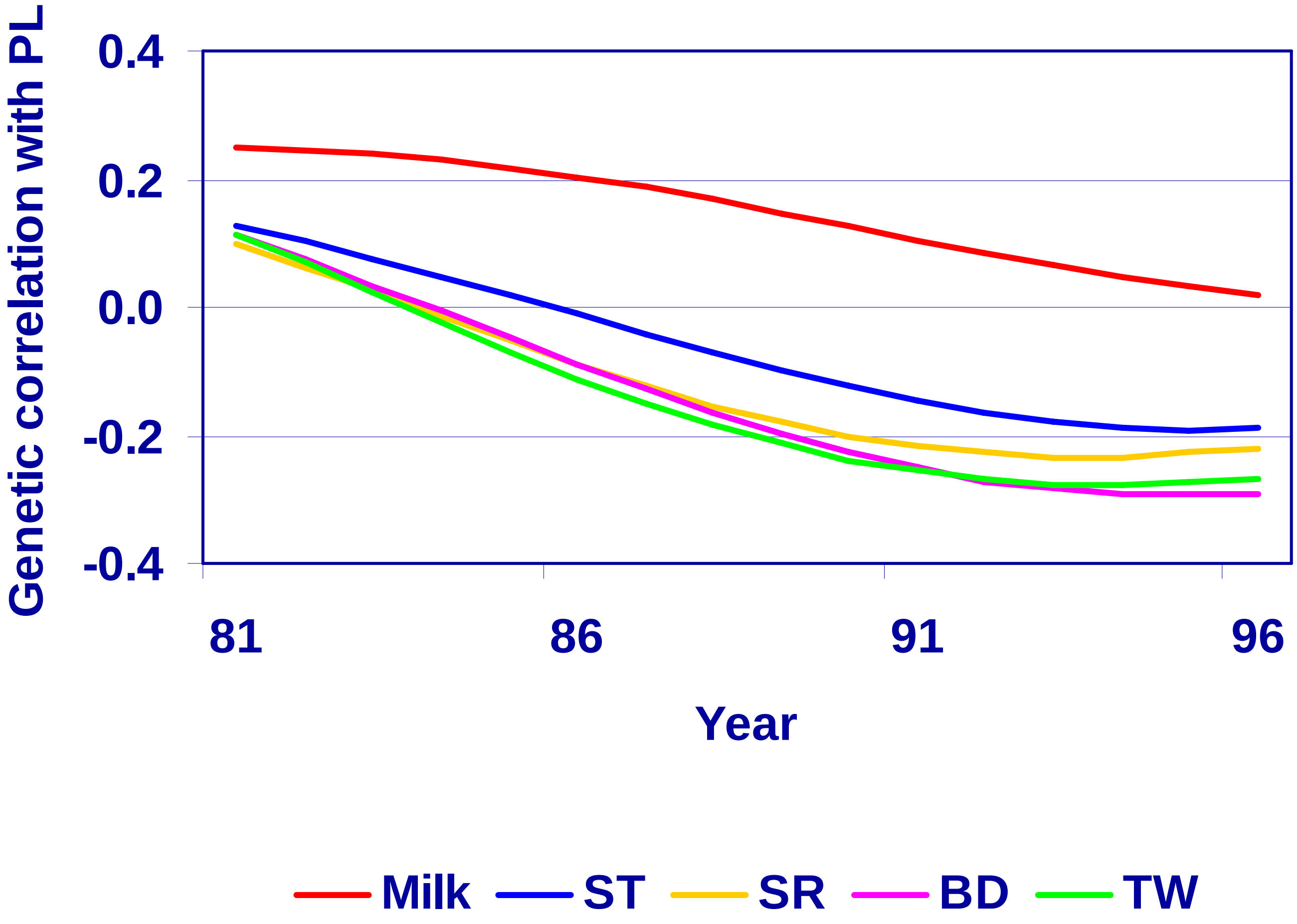
(Tsuruta et al.)

$$y_{ijknp} = \text{"}fixed\text{"} + \sum_{m=0}^{1 or 2} a_{mp} z_{m} + e_{ijjnp}$$

z = first order Legendre polynomial on year of birth

a = additive genetic effect with RR on year

18 type,3 production, somatic cells, days open + productive life (PL)

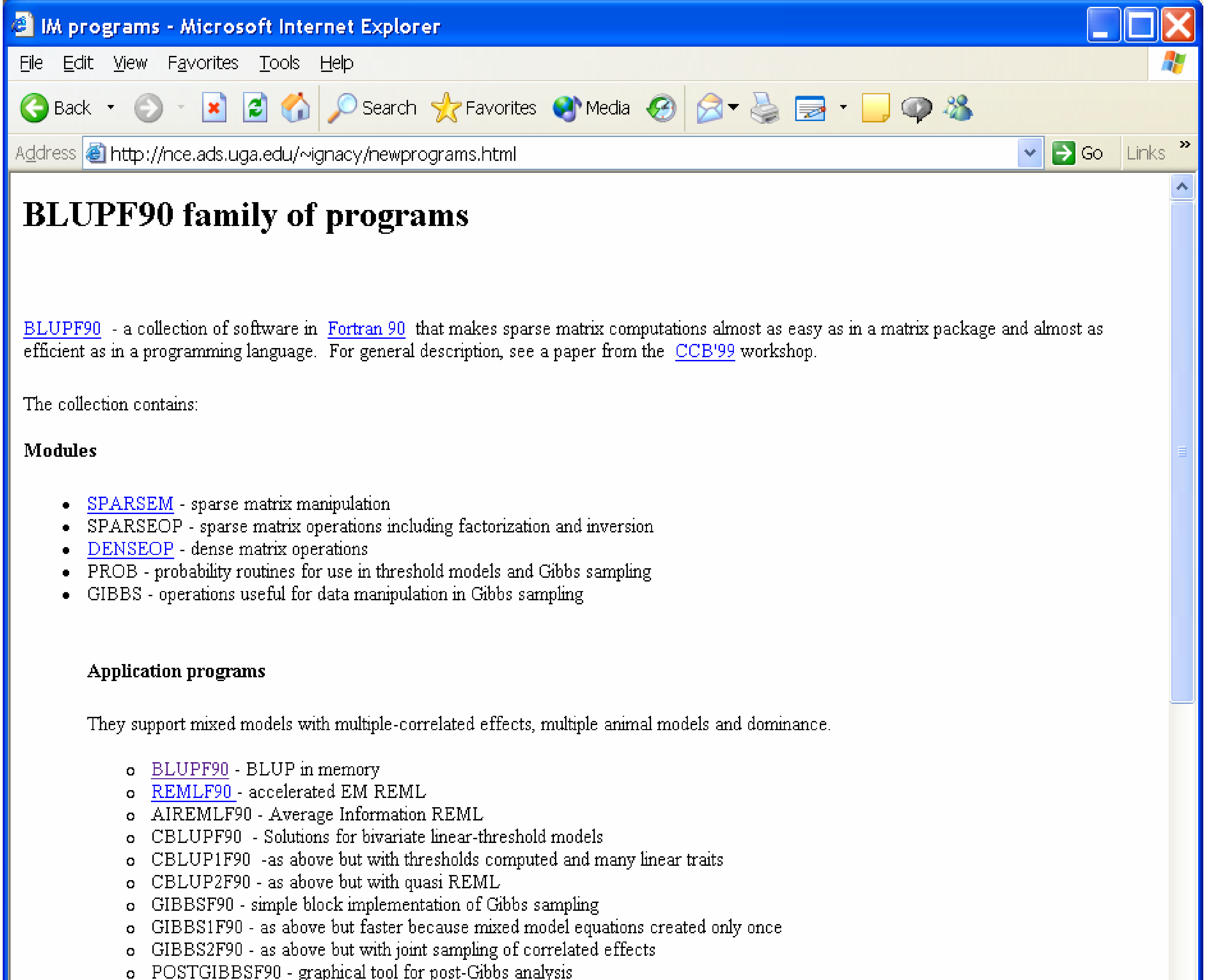


PL: productive life ST: stature SR: strength BD: body depth TW: thurl width

Methodology

Optimized sampler
1-3 months run time

Validation with MT



http://nce.ads.uga.edu/~ignacy/newprograms.html

BLUPF90 family of programs

BLUPF90 - a collection of software in Fortran 90 that makes sparse matrix computations almost as easy as in a matrix package and almost as efficient as in a programming language. For general description, see a paper from the CCB'99 workshop.

The collection contains:

Modules

- SPARSEM - sparse matrix manipulation
- SPARSEOP - sparse matrix operations including factorization and inversion
- DENSEOP - dense matrix operations
- PROB - probability routines for use in threshold models and Gibbs sampling
- GIBBS - operations useful for data manipulation in Gibbs sampling

Application programs

They support mixed models with multiple-correlated effects, multiple animal models and dominance.

- BLUPF90 - BLUP in memory
- REMLF90 - accelerated EM REML
- AIREMLF90 - Average Information REML
- CBLUPF90 - Solutions for bivariate linear-threshold models
- CBLUP1F90 -as above but with thresholds computed and many linear traits
- CBLUP2F90 - as above but with quasi REML
- GIBBSF90 - simple block implementation of Gibbs sampling
- GIBBS1F90 - as above but faster because mixed model equations created only once
- GIBBS2F90 - as above but with joint sampling of correlated effects
- POSTGIBBSF90 - graphical tool for post-Gibbs analysis

Choices in estimation of parameters

- Several versions of REML
- Several versions of MCMC

- Linear and nonlinear

- Modifications relatively simple

Which is the best model?

- Statistical tests
  - Show which models are better fitting
  - Do not show whether
    - differences important in practice
    - which important effects are missing

- Predictivity at WCGAL06
  - Blasco
  - “posterior predictive ability was the only criterion that ranked methods correctly” by .. & Sorensen

Performance of various test-day models

(Lopez-Romero and Carabano, 2002)

Model	**Relative BIC**	**Predictive ability**
Repeatability	2210	0.835
Legendre (3)	0	0.855
Legendre (4)	-255	0.857
Legendre (5)	-294	0.858

All models almost identical for ranking sires

Conclusions

- Different methodologies for different problems
- Balance of complexity and amount of data
- Caution with statistical criteria
- Importance of “fixed effects”

“All models are wrong, some are useful”