Statistical and Computational Challenges in Whole Genome Prediction and Genome-Wide Association Analyses for Plant and Animal Breeding*

Symposium on Frontiers in Big Data
September 23, 2016
University of Illinois, Urbana-Champaign

Robert J.Tempelman
Michigan State University

MICHIGAN STATE UNIVERSITY
AgBioResearch

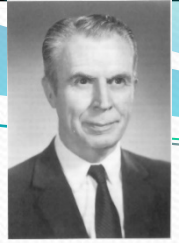
*See also December 2015 special issue of *Journal of Agricultural, Biological, and Environmental Statistics*

Who am I?

- I am proud to be known as a “animal breeder” or “animal breeding scientist”
  - That may conjure up some romantic images

  - But really one of the first “big data” agricultural scientists.
    - Also seminal users of mixed model analyses.

Mixed models



Charles R. Henderson (1911-1989)

Helped developed mixed model analyses for animal breeders before they became more widely popular later.

With O. Kempthorne, S. R. Searle, and C. M. von Krosigk. The estimation of environmental and genetic trends from records subject to culling. 1959, Biometric 15:192-218.

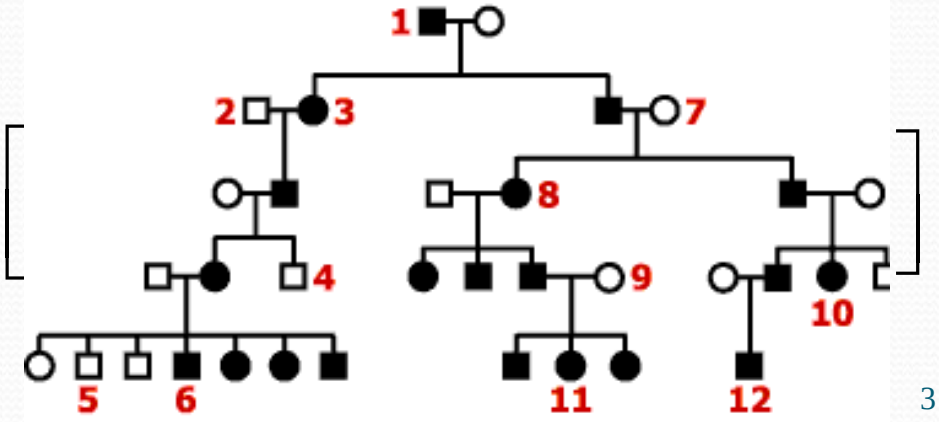
$$\mathbf{y}_{n\text{x}1} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u}_{q\text{x}1} + \mathbf{e}$$

data — **fixed effects (e.g. age herd)** — **genetic (polygenic) effects** — **residual**

$\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma^2_u)$

$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_e)$

A: genetic relationships (pedigree/kinship)



u

Resource constrained computing (in the good old days)

- My Masters Thesis (University of Guelph) 1986-1988.
  - Maximum memory request on Guelph IBM mainframe computer was (wait for it!): 8MB
  - q > 2*13,722 (additive and dominance genetic effects for each of 13,722 Holstein cattle including ancestors,
    - n = 8,329 cows with phenotypes
    - >MME: 180 MB (Full-store) without fixed effects!
  - Solution?.. sparse matrix storage techniques
    - MME in animal breeding are > 99% zeroes (whew!)..so only save non-zeroes
    - Thank you Karin Meyer! (DFREML)

Mixed models and animal breeding

- Mixed Models on Steroids.
  - A subtle "$q > n$" problem: (q = number of animal effects, n = number of records)
  - Let's go back in time (**1988**):

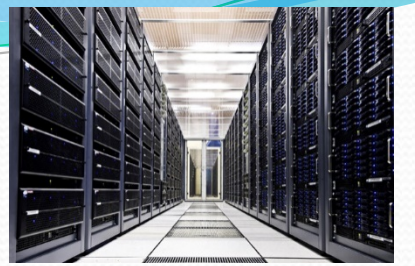
  - already both q and $n >$ **10M** cattle back then for USDA Holstein national genetic evaluation (Wiggans, 1988)
    - Hmm… even sparse matrix software can't save you there!
    - Solution? Save MME on hard disk; use Gauss-Seidel (now precondition conjugate gradient) iteration on data.
      - Now (2015) $q >$ 70M cattle, $n >$ 130M or > 650M depending on definition of phenotypes.

From Brown and Kass (2009)

"What is statistics?" The American Statistician (2009) 63: 105-110

- *"Physicists and engineers very often become immersed in the subject matter. In particular, they work hand in hand with neuroscientists and often become experimentalists themselves. Furthermore, engineers (and likewise computer scientists) are ambitious; when faced with problems, they tend to attack, sweeping aside impediments stemming from limited knowledge about the procedures that they apply"*
  - This has been the culture of animal breeding (for better or for worse)
  - Field of study is important….passion even more so.

It's not always about obtaining the biggest servers!



- ***Algorithm development!***
  - Example 1: MME requires $\mathbf{A}^{-1}$...not $\mathbf{A}$.

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad \lambda = \frac{\sigma_e^2}{\sigma_u^2}$$

    - Inverting **A** is not possible.
    - Henderson (again!) developed rules for computing $\mathbf{A}^{-1}$
      - computations linear in *q*! (numerically stable/sparse)
  - Example 2: Estimating variance components
    - REML. Standard algorithms in canned statistical packages unworkable
    - AI(Average Information)-REML is based on hybrid Newton-Raphson/Fisher scoring algorithm that exploits sparsity of MME.
      - (Gilmour et al., 1985; *Biometrics* 51: 1440)

Genetic improvement of livestock

- Has led to ***dramatic changes***
  - Since 1963, milk production / cow has doubled…>50% of that due to genetic trend (Garcia et al., 2016; PNAS 113:E3995)
  - Historically: Artificial insemination with frozen semen, embryo transfer/ estrus synchronization
    - Widespread global exchange of germplasm

  - ***Whole genome prediction (WGP)*** based on the use of 10s/100s of thousands "high" density single nucleotide polymorphism (SNP) marker genotypes on each cow has further increased selection intensity
    - Impact on higher accuracy of genetic merit and lower generation interval
    - Selection response should accelerate….

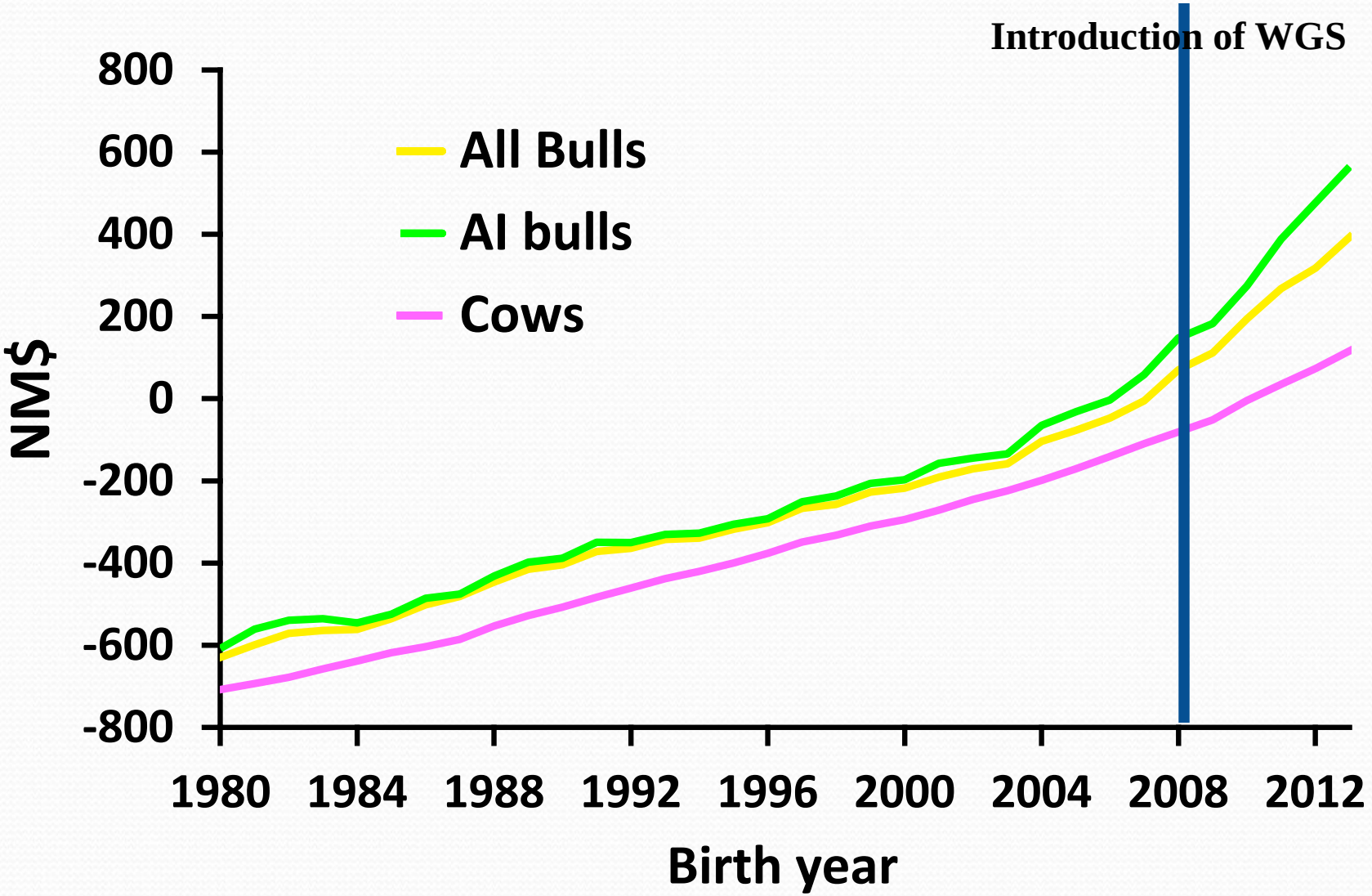
Largest genomic databases (courtesy, Paul Van Raden, USDA-AGIL; August, 2016)

	Ancestry.com	23andMe	CDCB / USDA
Genotypes	>2 million	>1 million	1.4 million
Species	Human	Human	Cattle
Countries	?	>55	53
Genotyping cost	$99	$199	$37–135
Delivery (weeks)	6–8	6–8	1–2
DNA generations	Few	Few	>10
EBV reliability	NA	Low	High

Reference: http://genomemag.com/davies-23andme/#.VdY722zosY1

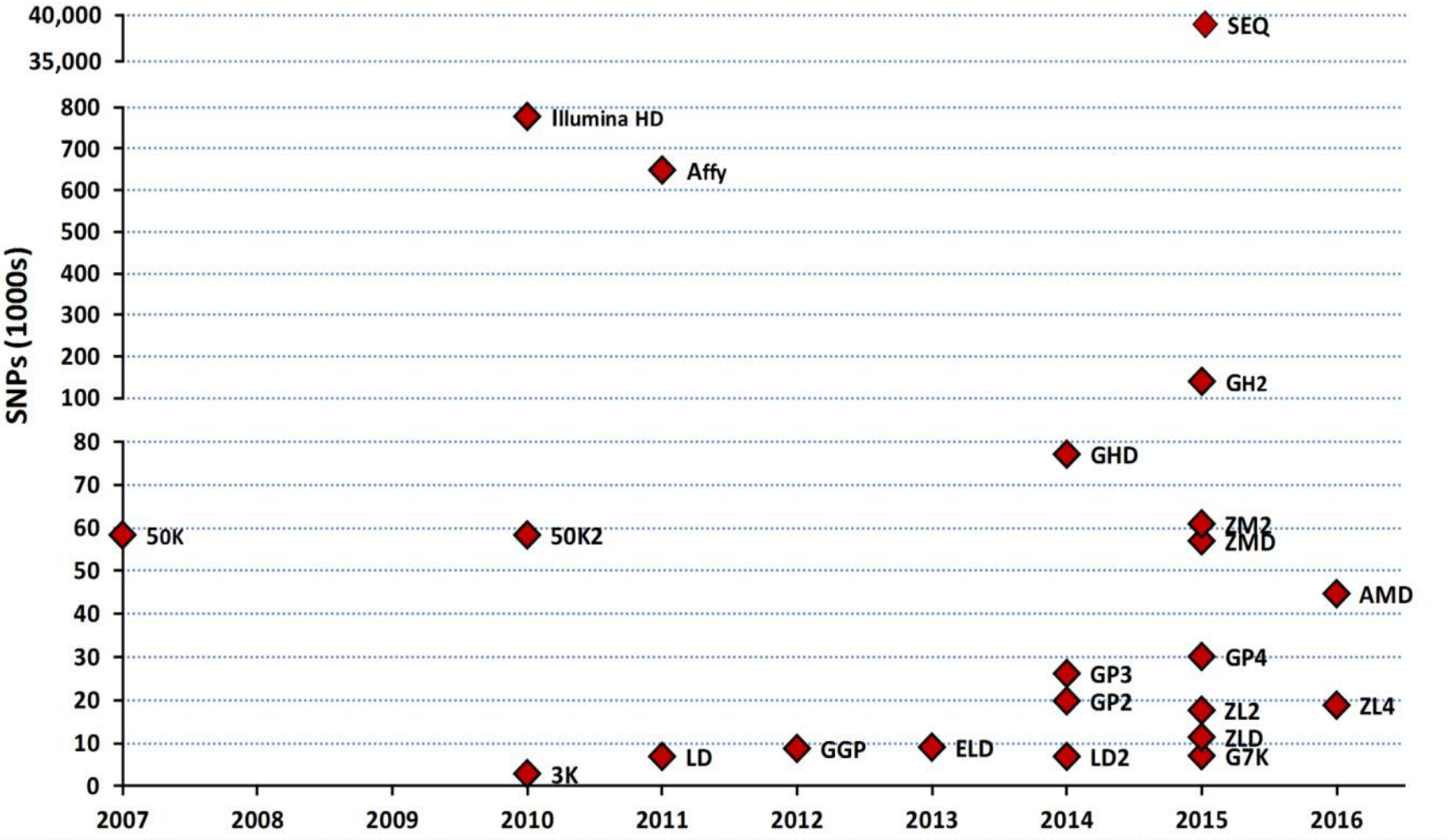
Web sites: https://www.23andme.com/
http://dna.ancestry.com/
https://www.cdcb.us/
http://aipl.arsusda.gov/Main/site_main.htm

Genetic trend for Dairy Net Merit $



Courtesy: USDA-AGIL (Animal Genomics Improvement Laboratory)

Evolution of SNP marker panels in cattle breeding



Courtesy: USDA-AGIL (Animal Genomics Improvement Laboratory)

Whole Genome Prediction (WGP)

Polygenic Effects ("optional")

$$\mathbf{u}=\{u_i\}_{i=1}^{n} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_u^2\right)$$

- Model: $\mathbf{y}_{nx1} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{W}\mathbf{u}_{qx1} + \mathbf{e}$

SNP allelic substitution effects: $\mathbf{g} = [g_1 \quad g_2 \quad g_3 \quad \dots \quad g_m]'$

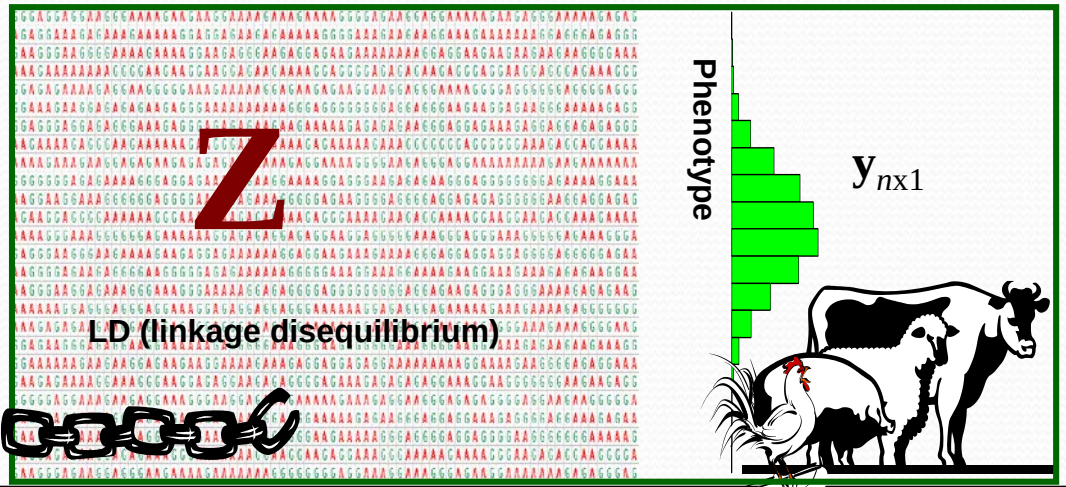
$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_q' \end{bmatrix}$$

Matrix of genotypes

$$\mathbf{z}_i' = [z_{i1} \quad z_{i2} \quad z_{i3} \quad \dots \quad z_{im}]$$

values = 0, 1 or 2 (# of copies of reference allele)

Genotypes



Typical distributional assumption

$$\mathbf{g}=\{g_j\}_{j=1}^{n} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_g^2\right)$$

LD generates "signal dependence" (Chen and Storey, 2006)…multicollinearity

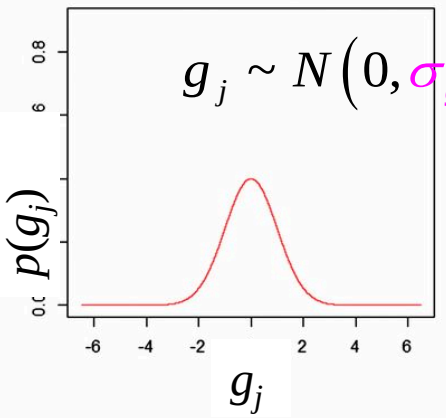
Big Data keeps getting bigger!

- m > 50,000 SNP markers (some imputation from lower density panels)
- q > 70M animals
  - (>1.4 M of which are genotyped…so how do you deal with that?…see later)
- n > 130M records.
- Most research studies involve far smaller q and n.
  - Greater recognition that if most SNP markers are NOT in tight LD with genes (QTL), then normality assumption is tenuous.
- With ↑ m and/or ↑ q for same n, "effective" sample size actually ↓…."big data" can be a misnomer.

Some alternative candidate priors for **g** (Meuwissen et al., 2001; de los Campos et al., 2013)

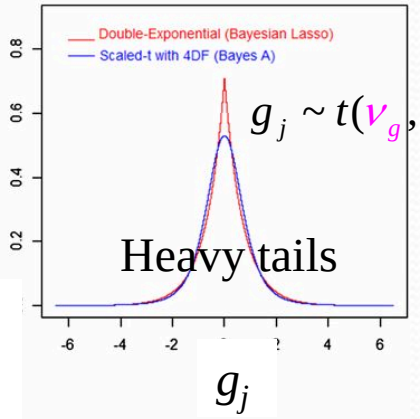
rrBLUP

$$g_j \sim N\left(0, \sigma_g^2\right)$$



BayesA

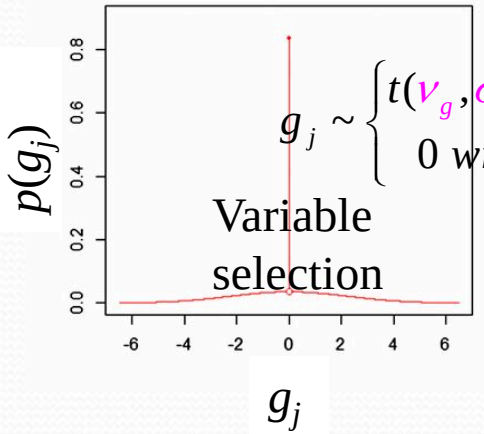
$$g_j \sim t(\nu_g, \sigma_g^2) \begin{cases} g_j \sim N\left(0, \sigma_g^2 \phi_j\right) \\ \phi_j \sim \chi^{-2}\left(\nu_g, \nu_g\right) \end{cases}$$



More flexible priors generally require a Bayesian approach

BayesB

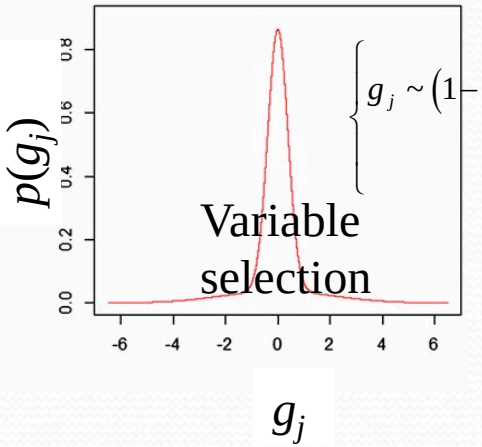
$$g_j \sim \begin{cases} t(\nu_g, \sigma_g^2) \text{ with prob } \pi \\ 0 \text{ with prob } (1\text{-}\pi) \end{cases}$$



SSVS

$$\begin{cases} g_j \sim (1-\phi_j) N\left(0, \frac{\sigma_g^2}{c}\right) + \phi_j N\left(0, \sigma_g^2\right) \\ \phi_j \sim Bernoulli(\pi) \end{cases}$$

$c >> 1$



Bayesian alphabet models (Gianola)

ϕ_j: augmented variables (computational convenience)

hyperparameters should be estimated..they define genetic architecture

Bayesian analyses

- More ambitious priors typically require use of **MCMC** (Markov Chain Monte Carlo) methods.
  - Sample from the posterior distribution of all unknowns.
  - Don't be at the "*mercy of the prior!*" (Dan Gianola)…**estimate hyperparameters**
- Research reproducibility issue:
  - Many researchers (animal breeders and others) may not draw enough MCMC samples from the posterior density
    - Autocorrelated draws
    - Issue potentially more dramatic for computing Bayesian credible bounds than, say, posterior means.
    - A much greater issue also for denser SNP chips (or $m >> n$).
      - Ongoing research to improve "mixing" (↓ autocorrelation)

"Exact" inference and computational efficiency

- MCMC: Real-time posterior densities (updates) very difficult to get with "big data"
  - no "memory" from previous analyses…need to rerun
- Analytical approximations
  - Expectation-maximization (EM) often referred to as "big data Bayes" (Allenby et al., 2014)
  - However, extreme sensitive to starting values (Chen and Tempelman, 2015) especially with large *m* relative to *n* (multimodal joint posterior densities more likely)
- Normal prior on **g**: Pragmatic alternative for large national genetic evaluations.
  - Yet limited effectiveness for genome wide association studies relative to more flexible priors.

Reducing dimensionality with equivalent models (if $q<<m$)

Assume $q = n$ (one record per animal)

SNP effects model (rrBLUP)

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zg} + \mathbf{e}$$

$$\mathbf{g} = \{g_j\}_{j=1}^{m} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_g^2\right)$$

Genomic animal effects model (GBLUP)

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} + \mathbf{e}$$

$$\mathbf{u}_{qx1} = \mathbf{Zg}$$

$$\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{ZZ'}\sigma_g^2\right)$$

G = **ZZ**' is the *genomic relationship matrix* (need its inverse for MME for GBLUP).

Can easily backsolve for **g** from **u** in GBLUP (Stranden and Garrick, 2008).

The latest and greatest

- Using biological information
  - having a disciplinary passion is useful for a data scientist!
  - Assign higher prior probability to SNP in coding or regulatory genomic regions (BayesRC; McLeod et al., 2016 *BMC Genomics* 17:144)
- Addressing industry concerns: Combining data on genotyped and non-genotyped animals.
  - "single" step procedures (Aguilar et al., 2010) based on blending **G** (on genotyped animals) with **A** (on non-genotyped animals)
  - APY (Mizstal, 2016 *Genetics* 202:401) based on "sparsifying" $\mathbf{G}^{-1}$...numerically stable!

Worries keep coming

M.P.L. Calus, J. Vandenplas and J. Ten Napel
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, The Netherlands
E-mail: mario.calus@wur.nl

Journal of Animal Breeding and Genetics
J. Anim. Breed. Genet. ISSN 0931-2668

EDITORIAL

2015

Ever-growing data sets pose (new) challenges to genomic prediction models

- # of SNP markers increasing (10s of millions in sequencing).
  - Because of high LD, even greater multicollinearity creating instability, especially in MCMC inference.

Big data and sustainability

- "management systems & environments are changing more rapidly than animal populations can adapt to such changes through natural selection" (Hohenboken et al., 2005)..e.g.
  - Energy policy (corn distiller's grain)
  - More intensive management (larger farms)
  - Climate change
  - What are the implications for genetic improvement of livestock?
    - How should we prepare as statistical geneticists?
    - What kind of direction should we provide?

Scope of inference and livestock breeding

- **Scope of inference**
  - What is this in context of animal breeding?
  - **Broad scope**: Inferring upon average" additive genetic merit for animals *across all environments*.
    - Often is the primary focus
  - **Narrow scope**: Inferring upon (eventually adapting) genetic merit for animals *within a specific environment*….
  - To accommodate both objectives… need to create SNP*E terms
    - Curse of dimensionality intensifies!

Big Data Dairy Management

31st ADSA Discover Conference℠ on Food Animal Agriculture

ARPAS has assigned 16 ARPAS CEU's for ARPAS members participating in the 31st Discover Conference

Conference Objective

Across all industries, the availability of increasingly powerful computers and new technologies provides new business management opportunities. In the last few years, most large companies have embraced the concept of "big data" techniques as part of their management strategy. Definitions of big data vary. But, in general, the term refers to using large data sets for complex decisions where traditional data processing techniques may lack. The key components of big data are analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. Big data often involves using predictive analytics to analyze existing data sets in new ways. Another key characteristic of big data is merging data from multiple sources into cloud computing. For example, in the dairy industry, big data may involve combining DHI production records, financial records, precision dairy technology data, health records, milk cooperative records, historical weather data, genomic evaluations, ration and feeding management data, and human resource data into one large database. Combining this information helps to improve decision-making, operational efficiency, cost and revenue optimization, and risk management.

The dairy industry remains a perfect application of decision science and big data because: (1) it is characterized by considerable price, weather and biological variation, and uncertainty, (2) technologies, such as those that monitor dairy cow yield, physiology, and behavior are easily available, (3) and the primary output, fluid milk, is difficult to differentiate, increasing the need for alternative means of business differentiation. Big data represents a potential management

Conference Details

November 1-4, 2016

Hilton Chicago/Oak Brook Hills Resort & Conference Center

Hosted by the American Dairy Science Association

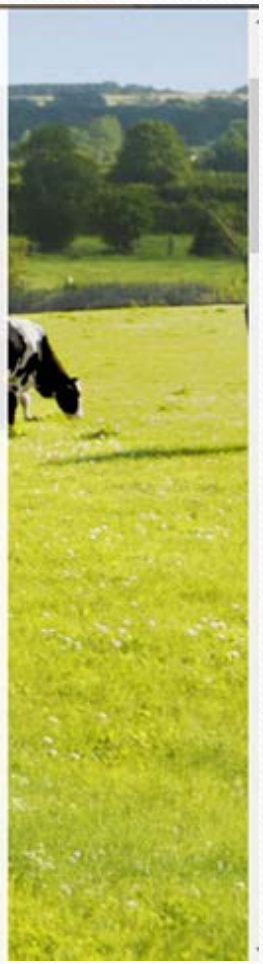
Conference Documents

- Program Announcement
- Conference Program
- Online Registration
- Printable Registration Form
- Cancellation Policy
- Invitation Letter Request Form
- Sponsorship Options

Conference Format

The Discover Conference Series is designed to provide a format and venue that encourages in-depth discussion of cutting-edge science.

ADSA Discover Conferences℠ focus on topics of importance to the science of food animal agriculture and are held in a relaxed



THANK YOU!

Funding by Agriculture and Food Research Initiative Competitive Grants # 2011-67015-30338 and 2011-68004-30340 is gratefully acknowledged.