On models and challenges in quantitative genetics and genomics



Daniel Gianola

Sewall Wright Professor Emeritus of Animal Breeding and Genetics

University of Wisconsin Madison Dairy Science







ANIMAL AND PLANT BREEEDING LARGELY EMPLOY STATISTICAL ABSTRACTIONS

Coping with complexity

First assumption: there is a genetic signal and an environmental signal **Second assumption**: the joint effect translates into a phenotye **y**

Y = f(G, E)

For some **UNKNOWN** function *f*

Huge number of possibilities for *f*!

s? $\begin{cases} Y = G^{E}?\\ Y = E^{G}?\\ Y = G + E + GE?\\ Y = (G + E)^{GE}?\\ Y = G + E? \end{cases}$ Is an assumption (plant breeders very aware) Is an even a stronger assumption (animal breeders typically ignore it)

Choices?

BIG-BANG OF WHOLE-GENOME REGRESSIONS

Copyright © 2001 by the Genetics Society of America

Cast God: BLUP

Adam: Bayes A

Eve: Bayes B

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes[†] and M. E. Goddard^{†,‡}

*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, [†]Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and [‡]Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia

> Manuscript received August 17, 2000 Accepted for publication January 17, 2001

GENOMIC SELECTION: A DOMINANT RESEARCH AND DEVELOPMENT THEME



"The curse of the Bayesian Alphabet" Bayes A, B, Bayes BLUPC, C-pi, D, Fast-B, L, R, RC, RS, TA, TB, TC, RKHS, NN...





LOOKING AHEAD

(entirely personal and subjective "shopping list")

Area 1: Outlier detection, control and accommodation

Area 2: Genomic similarity matrix as estimand

Area 3: The Bayesian Alphabet marches on!

Area 4: On MCMC and "discovery"

Area 5: GE interactions and multi-omics

Area 6: Deep learners: experience so far

Area 1: Outlier detection, control and accommodation

The problem is of concern... (at least in Denmark and Finland)



J. Dairy Sci. 101:11159–11164 https://doi.org/10.3168/jds.2018-15123

© 2018, The Authors. Published by FASS Inc. and Elsevier Inc. on behalf of the American Dairy Science Association[®]. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Short communication: Multivariate outlier detection for routine Nordic dairy cattle genetic evaluation in the Nordic Holstein and Red population

H. Gao,^{*1} P. Madsen,^{*} J. Pösö,[†] G. P. Aamand,[‡] M. Lidauer,§ and J. Jensen^{*} *Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark †Faba Co-Op, FIN-01301 Vantaa, Finland ‡Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark §Natural Resources Institute Finland (Luke), FIN-31600 Jokioinen, Finland

9-trait model: milk, fat, protein in 3 lactations -Compute Mahalanobis intra year-lactation-DIM classes for each record. Cut-offs for edits discarding outliers of increased stringency

The results showed that, averaged over all scenarios, gains of 0.005 to 0.048 on prediction accuracy have been obtained by deleting the multivariate outliers. The improvements were more profound for progeny of young bulls compared with progeny of proven bulls. It is easy to implement this multivariate outlier-detection procedure in the routine genetic evaluation for different dairy cattle breeds; however, an optimal cutoff value for Mahalanobis distance needs to be defined to achieve

$N = 500 \text{ bivariate normals } N(0, \Sigma)$ $\Sigma = \begin{bmatrix} 0.8 & 0.7 \\ 0.7 & 0.8 \end{bmatrix} \text{ Prob(mutation)} = 0.08$ *Mutation adds* $\delta = (1, -1)'$ **Would not appear as outliers in x or y axis**









ACCOMODATING OUTLIERS CAN AUTOMATE?

- DISCARD DATA USING AD-HOC RULES (no account for exclusion uncertainty and arbitrariness in rules—GAO et al. 2018 recognized the issue)
- FIT ROBUST RESIDUAL DISTRIBUTION TO (Andrews and Mallows, 1974):
 →ATTENUATE ABERRANT (W.R. TO THE MODEL) OBSERVATIONS
- ANIMAL BREEDERS HAVE DONE IT FOR INFERENCE, NOT PREDICTION!

-STRANDEN AND GIANOLA (1998, 1999) -ROSA ET AL. (2003, 2004) -KIZILKAYA ET A. (2003) -CARDOSO ET AL. (2006) -CARDOSO ET AL. (2006)

-GIANOLA ET AL. (2018) → t and Laplace distributions: "Iterative GBLUP"

OUTLIERS IN TMAP: weights assigned automatically

d-values vs test-day yield (h2guess=0.05) t4=black t8=red t12=blue t16=green



Bootstrap distribution (b=15,000 samples) of predictive mean-squared error (PMSE) and predictive correlation (PCOR) for GBLUP, TMAP (df=4) and LMAP at selected genomic heritability values (guesses of 0.05 and 0.50 produced MINQUE estimates of 0.07 and 0.15, respectively): test-day milk yield in Brown-Swiss cows.



LMAP "BEST" FOLLOWED BY TMAP4 AND THEN BY GBLUP

EXTENSION TO MULTIVARIATE OUTLIERS NEEDED

-FOUNDATION: STRANDEN (1996) FOR T-DISTRIBUTION (MCMC)
-WORK IN PROGRESS FOR NON-MCMC T AND LAPLACE DISTRIBUTIONS

Area 2: Learning similarity matrices among individuals (G) as an estimation problem

Van Raden proposed marker-based similarity matrix (acts as hype-parameter): $G_{VR} = \frac{XX'}{\sum_{i=1}^{p} 2p_i(1-p_i)}$

Inferring similarity among individuals from molecular markers and phenotypes with Bayesian regression

Daniel Gianola a,b,c and Rohan L. Fernando

Model specific genomic variance

An unknown matrix

 $Var(\mathbf{g}|\mathbf{X}) = \mathbf{X}\mathbf{D}\mathbf{X}' = \mathbf{G}\sigma_{g^{\mathbf{X}}}^2$ A diagonal matrix

p(G|X) = prior distribution based on markersy = vector of phenotypes (single or multiple - traits)p(G|X, y) = posterior distribution informed by markers and phenotypes

HAVE DEVELOPED A FORMALISM FOR BAYES A, B, C, Cπ, R



EXTENT TO TO WHICH PHENOTYPES INFORM ABOUT SIMILARITY:

<u>Frobenius distances between TRUE QTL G</u>

and their prior and posteriors

10 chromosomes- 2000 SNP (100 QTL)- Bayes Cπ







Pinus taeda: N=807 p=4828 Bayes Cπ bivariate analysis (only 100 pairs of individuals)



Frobenius Distance

Area 3: The Bayesian Alphabet marches on!

A multiple-trait Bayesian Lasso for genome-enabled association and prediction of complex traits

Daniel Gianola
 a,b,c,d Rohan L. Fernando and Chris-Carolin Schön
 d

- ^a Department of Animal Sciences, University of Wisconsin-Madison, USA;
 - ^b Department of Dairy Science, University of Wisconsin-Madison, USA;

^cDepartment of Animal Science, Iowa State University, USA

^d Department of Plant Sciences, Technical University of Munich,

TUM School of Life Sciences, Germany.

2.1.2 Multivariate Laplace prior distribution (MLAP) for marker effects

Independent *T*-variate Laplace prior distributions with a null mean vector will be assigned to each of the $T \times 1$ vectors β_j (j = 1, 2, ..., p). Gómez et al. (2007) presented a multi-dimensional version of the power exponential (PE) family of distributions; one special case is the multivariate Laplace distribution (MLAP). The density of the MLAP distribution with a zero-mean vector used here has form

$$p(\boldsymbol{\beta}_{j}|\boldsymbol{\Sigma}) = \frac{T\Gamma\left(\frac{T}{2}\right)}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} \pi^{\frac{T}{2}}\Gamma\left(1+T\right)2^{(1+T)}} \exp\left(-\frac{1}{2}\sqrt{\boldsymbol{\beta}_{j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_{j}}\right); \ j = 1, 2, ..., p,$$
(8)

where $\Sigma = \{\Sigma_{kk'}\}$ is a $T \times T$ positive-definite scale matrix. The variance-covariance matrix of the distribution is

$$Var(\boldsymbol{\beta}_{j}|\boldsymbol{\Sigma}) = 4\left(T+1\right)\boldsymbol{\Sigma} = \mathbf{B};$$
(9)



Analysis of wheat data set

- n=599 inbred lines
- p=1279 DaRT markers
- Traits: grain yield in environments 1 and 2
- Bivariate GBLUP (maximum likelihood for variance components)
- Bivariate Lasso
- Six chains of 1500 iteration each (evaluate MCMC)
- Gelman's R-statistics+ Geweke's batching for convergence assessment
- Six chains of 2000 samples post-burn in: 12000 samples used for inference



Bivariate GBLUP vs Bivariate LASSO marker effects: yield 1

Bivariate GBLUP vs Bivariate LASSO marker effects: yield 2



Bivariate GBLUP vs Bivariate LASSO absolute marker effects: yield 1



Bivariate GBLUP vs Bivariate LASSO absolute marker effects: yield 2



MUMA PLOTS (MULTIVARIATE MANHATTAN)





Area 4: On MCMC and "discovery"

MCMC-Based Inference in the Era of Big Data: A Fundamental Analysis of the Convergence Complexity of High-Dimensional Chains

> Bala Rajaratnam and Doug Sparks Stanford University

> > August 28, 2015

Variational distance between <u>posterior</u> π and <u>estimate</u> of (distribution) at iterate k

$$d_{\mathrm{TV}}(P_{x_0}^k, \Pi) \leq M_{x_0}r^k$$

"Starting distance"



Ranges in 0-1 1 when the two distributions differ Convergence rate

AUTHORS:

In summary, our theoretical and numerical analysis above indicates that regardless of the type or form of regression (standard regression, lasso, elastic net, or spike-and-slab), there is a universal geometric convergence rate of the form r = p/(n + p - 2).

ITERATIONS REQUIRED TO REDUCE STARTING VARIATIONAL DISTANCE TO 10^{-7}

n	р	$ {\bf Parameterization \ in} \ n \\$
1000	50000	23
	800000	23
	$20 imes10^{6}$	23
5000	50000	23
	800000	23
	$20 imes 10^{6}$	23

Parameterization in n Parameterization in p \mathbf{n} \mathbf{p} 20×10^{6} 20×10^{6}

MARKER EFFECTS

GENOTYPIC VALUES

MARKER EFFECTS

n	р	Parameterization in n	Parameterization in p	Parameterization in $4p$
1000	50000	23	816	3235
	800000	23	12886	51986
	$20 imes10^{6}$	23	322350	$1.6118 imes10^6$
5000	50000	23	169	653
	800000	23	2587	10324
	$20 imes10^{6}$	23	64664	2.6863×10^{5}

Area 5: Multi-omics and GE

HIGHLIGHTED ARTICLE GENETICS | GENOMIC SELECTION

Genetic Epidemiology OFFICIAL JOURNAL INTERNATIONAL GENETIC Poly-Omic Prediction of Complex Traits: OmicKriging PIDEMIOLOGY SOCIETY www.geneticegi.org

Heather E. Wheeler,¹ Keston Aguino-Michaels,² Eric R. Gamazon,² Vassily V. Trubetskoy,² M. Eileen Dolan,¹ R. Stephanie Huang,¹ Nancy J. Cox,² and Hae Kyung Im³*

Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America, "3Department of Health Studies, University of Chicago, Chicago, Illinois, United States of America

Received 26 November 2013; Revised 11 March 2014; accepted revised manuscript 12 March 2014. Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21808

> HIGHLIGHTED ARTICLE GENOMIC SELECTION GENETICS |

Prediction of Plant Height in Arabidopsis thaliana Using DNA Methylation Data

Yaodong Hu, *.1 Gota Morota, * Guilherme J. M. Rosa, *.* and Daniel Gianola *.*.8 *Department of Animal Sciences, [‡]Department of Biostatistics and Medical Informatics, and [§]Department of Dairy Science, University of Wisconsin, Madison, Wisconsin 53706, and [†]Department of Animal Science, University of Nebraska, Lincoln, Nebraska 68583

Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome **Multiomic Profiles**

Ana I. Vazquez,*.' Yogasudha Veturi,' Michael Behring,*.⁵ Sadeep Shrestha,^{\$} Matias Kirst,**.'' Marcio F. R. Resende, Jr., **.** and Gustavo de los Campos*.**

* Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan 48824, *Biostatistics Department, *Comprehensive Cancer Center, and [§]Department of Epidemiology, University of Alabama at Birmingham, Alabama 35294, **School of Forest Resources and Conservation and ⁺⁺University of Florida Genetics Institute, University of Florida, Gainesville, Florida 32611, and **Statistics Department, Michigan State University, East Lansing, Michigan 48824

ABSTRACT Whole-genome multiomic profiles hold valuable information for the analysis and prediction of disease risk and progression. However, integrating high-dimensional multilayer omic data into risk-assessment models is statistically and computationally challenging. We describe a statistical framework, the Bayesian generalized additive model ((BGAM), and present software for integrating multilayer high-dimensional inputs into risk-assessment models. We used BGAM and data from The Cancer Genome Atlas for the analysis and prediction of survival after diagnosis of breast cancer. We developed a sequence of studies to (1) compare predictions based on single omics with those based on clinical covariates commonly used for the assessment of breast cancer patients (COV), (2) evaluate the benefits of combining COV and omics, (3) compare models based on (a) COV and gene expression profiles from oncogenes with (b) COV and wholegenome gene expression (WGGE) profiles, and (4) evaluate the impacts of combining multiple omics and their interactions. We report that (1) WGGE profiles and whole-genome methylation (METH) profiles offer more predictive power than any of the COV commonly used in clinical practice (e.g., subtype and stage), (2) adding WGGE or METH profiles to COV increases prediction accuracy, (3) the predictive power of WGGE profiles is considerably higher than that based on expression from large-effect oncogenes, and (4) the gain in prediction accuracy when combining multiple omics is consistent. Our results show the feasibility of omic integration and highlight the importance of WGGE and METH profiles in breast cancer, achieving gains of up to 7 points area under the curve (AUC) over the COV in some cases.

Theor Appl Genet (2014) 127:595-607 DOI 10.1007/s00122-013-2243-1

ORIGINAL PAPER

A reaction norm model for genomic selection using high-dimensional genomic and environmental data

Diego Jarquín · José Crossa · Xavier Lacaze · Philippe Du Chevron · Joëlle Daucourt · Josiane Lorgeou · François Piraux · Laurent Guerreiro · Paulino Pérez · Mario Calus · Juan Burgueño · Gustavo de los Campos

RKHS FOR GENOTYPE X ENVIRONMENT INTERACTION



J. Dairy Sci. 100:2042–2056 https://doi.org/10.3168/jds.2016-11543 © American Dairy Science Association[®], 2017.

Genotype by environment (climate) interaction improves genomic prediction for production traits in US Holstein cattle

F. Tiezzi,*¹ G. de los Campos,† K. L. Parker Gaddis,‡ and C. Maltecca* *Department of Animal Science, North Carolina State University, Raleigh 27695 †Department of Epidemiology and Biostatistics, Michigan State University, East Lansing 48828 ‡Council on Dairy Cattle Breeding, Bowie, MD 20716



Figure 3. Accuracy of prediction for the new bulls and incomplete progeny test cross-validation scenarios. Models included genomic effect (G), geographical region effect (R), and effect (H), latitude and longitude effects (L), herd management effects (M), herd fertility effects (F), service-sire choice effects (S), culling descriptors effects (C), and elimate variables effects (W). Histograms (error bars) report the average (SD) over the 4 folds used in the cross-validation.

bioRxiv preprint first posted online Jul. 6, 2018; doi: http://dx.doi.org/10.1101/363309. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Predicting Growth and Carcass Traits in Swine Using Metagenomic Data and Machine Learning Algorithms

Christian Maltecca^{1,*,+}, Duc Lu^{1,+}, Costantino Schillebeeckx², Nathan P McNulty², Clint Schwab³, Caleb Schull³, and Francesco Tiezzi^{1,+}

Area 6: Deep learners: experience so far













Prediction of maize phenotype based on whole-genome single nucleotide polymorphisms using deep belief networks

H Rachmatia^{*}, W A Kusuma, and L S Hasibuan

Department of Computer Science, Bogor Agricultural University, Indonesia

Figure 1. The DBN architecture for genomic prediction

Trait anningument	Model ^a			
Iran-environment	RKHS	BL	BLUP	DBN
MFL – WW	0.607	0.790	b	0.295
MFL - SS	0.674	0.778	0.464	0.325
FFL – WW	0.588	0.781	b	0.270
FFL - SS	0.648	0.774	0.521	0.370
ASI – WW	0.547	0.513	0.469	0.559
ASI – SS	0.572	0.517	0.481	0.579
GY - WW	0.514	0.525	0.515	0.445
GY – SS	0.453	0.415	0.442	0.383

 Table 1.Cross-validation (cv) correlation between predicted and observed phenotypes

Four models were fitted to each trait (FFL, MFL, ASI, and GY) and environment (SS, severe drought stress; WW, well watered) combination.

^a Models were molecular marker (SNPs) using reproducing kernel Hilbert space (RKHS) regression, Bayesian LASSO (BL), best linear unbiased predictor (BLUP), and our proposed method deep belief network (DBN).

^b BLUP were not computed because the estimated genetic variances were negligible [9].

RESULTS: NO CONVINCING EVIDENCE OF SUPERIORITY OF DBN

Can Deep Learning Improve Genomic Prediction of Complex Human Traits?

Pau Bellot,*1 Gustavo de los Campos,** and Miguel Pérez-Enciso*52

*Centre for Research in Agricultural Genomics (CRAG), Consejo Superior de Investigaciones Científicas (CSIC) - Institut de Recerca i Tecnologies Agroalimentaries (IRTA) - Universitat Autónoma de Barcelona (UAB) - Universitat de Barcelona (UB) Consortium, 08193 Bellaterra, Barcelona, Spain, [†]Department of Epidemiology and Biostatistics, and [‡]Department of Statistics, Michigan State University, East Lansing, Michigan 48824, and [§]Institut Català de Recerca Avançada (CREA), 08010 Barcelona, Spain ORCID IDs: 0000-0001-9503-4710 (P.B.); 0000-0001-5692-7129 (G.d.); 0000-0003-3524-995X (M.P.-E.)



STATURE

Figure 4 Prediction performance across methods and SNP sets for height. Gray, green, blue, and magenta bars correspond to linear, MLP, one-hot encoding MLP, and CNN methods, respectively. Average SE of R's were ~3 × 10⁻³, BEST, set with the 10k or 50k top most-associated SNPs; BRR, Bayesian Ridge Regression; CNN, Convolutional Neural Network; MLP, Multilayer Perceptron; UNIF, set in which the genome was split in windows of equal physical length and the most-associated SNP within each window was chosen.



BONE HEEL MINERAL DENSITY

Figure 5 Tediction performance across methods and SNP sets for bone heel mineral density. Gray, green, blue, and magenta bars correspond to linear, MRP, one-hot encoding MLP, and CNN methods, respectively. Very low bar means method not converging. Average 52 of 1% verse - 33 × 10⁻⁹, BEST, set with the 10k or 50k top most-associated SNP; BRR, Bayesian Ridge Regression; CNN, Convolutional Neurork; MLP, Multilayer Perceptron; UNF, set in which the genome was split in windows of equal physical length and the most-associated SNP within each oxison.





CONCLUSION

Draw your own conclusions!