Zaawansowane aspekty praktycznego zastosowania modeli mieszanych

Joanna Szyda, Tomasz Suchocki

1. Po co?
2. Predykcja wartości hodowlanych, a bioinformatyka
3. **Teaser**
4. **Przykładowy projekt**
5. Wykłady
6. Literatura
7. Kontakt

(1/7) **Po co?**

Po co ① ?

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{p} + \boldsymbol{e}$$

- $\boldsymbol{p}$ >>> $\boldsymbol{y}$
- większa liczba zmiennych niezależnych ($\boldsymbol{p}$) niż zależnych ($\boldsymbol{y}$)
- Bardzo częsty problem bioinformatyczny → bardzo duża liczba zmiennych objaśniających
- np. GWAS → więcej SNP (miliony) niż osobników (setki)
- **np. predykcja wartości hodowlanej → więcej osobników do predykcji (p) niż osobników z fenotypami (y)**

Copyright ©2023 Joanna Szyda

Po co 1 ?

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{p} + \boldsymbol{e}$$

- $\boldsymbol{p} >>> \boldsymbol{y}$
- Estymacja ($\boldsymbol{p}$) niemożliwa przy pomocy tradycyjnych modeli
- Feature pre-selection
- Shrinkage
  - — Ridge regression
  - — Lasso regression
  - — **Mixed models**

Copyright ©2023 Joanna Szyda

Po co ② ?

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{p} + \boldsymbol{e}$$

- $\boldsymbol{p} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{V})$
- Znana kowariancja pomiędzy $\boldsymbol{p}$
- np. GWAS → kowariancja między genami
- **np. predykcja wartości hodowlanej → kowariancja addytywmie genetyczna między osobnikami**

Copyright ©2023 Joanna Szyda

(2/7) Predykcja wartości hodowlanych, a bioinformatyka

Praktyczne zastosowanie modeli mieszanych – predykcja wartości hodowlanej

Copyright ©2023 Joanna Szyda

Predykcja wartości hodowlanych, a bioinformatyka

Cechy charakterystyczne danych hodowlanych

- Bardzo duże rozmiary
- Duża dynamika przepływu danych
- Wiele (skorelowanych) zmiennych objaśniających

Copyright ©2023 Joanna Szyda

Bardzo duże rozmiary danych

I) Data materials and computational requirements

- Genotyped population of German Holsteins for routine genomic evaluation April 2021
  - 949,636 genotyped Holstein animals including young candidates and culled animals
- Phenotype data of national cows and international bulls for milk production traits and SCS
  - Full data set of national test-day with bull MACE data
    - 12,432,940 cows as in official conventional evaluation with 242,121,126 test-day records
    - 138,770 bulls with deregressed MACE EBV added (ΔEDC >0), → 12,571,710 cows and bulls
  - Truncated data set following Interbull trend validation procedure
    - Removing **last 4 years test-day records** of cows and **last 4 birth years** of MACE bulls
    - Additionally daughters of validation bulls removed
    - 10,903,891 cows with 222,634,210 test-day records
    - 128,504 bulls with deregressed MACE EBV, → 11,032,395 cows and bulls
- Pedigree file for the single-step model evaluation
  - For youngest candidates 20 generations of ancestors selected
  - 20,461,400 animals in pedigree for the full evaluation and 177 phantom parent groups

Liu et al.

Bardzo duże rozmiary danych

I) Data materials and computational requirements

- Genotyped population of German Holsteins for routine genomic evaluation April 2023
  - 1,331,318 genotyped Holstein animals including young candidates and culled animals
- Phenotype data of national cows and international bulls for test-day protein yields
  - Full data set of national test-day with bull MACE data
    - 13,427,424 cows as in official conventional evaluation with 263,673,315 test-day records
    - 142,275 bulls with deregressed MACE EBV added (ΔEDC >0), → 13,569,699 cows and bulls
  - Truncated data set following Interbull trend validation procedure
    - Removing **last 4 years test-day records** of cows and **last 4 birth years** of MACE bulls
    - Additionally daughters of validation bulls removed
    - 11,898,375 cows with 244,186,399 test-day records
    - 132,009 bulls with deregressed MACE EBV, → 12,030,384 cows and bulls
- Pedigree file for the single-step model evaluation
  - For youngest candidates 20 generations of ancestors selected
  - 21,455,884 animals in pedigree for the full evaluation and 179 phantom parent groups

Liu et al.

Predykcja wartości hodowlanych, a bioinformatyka

Najważniejsze cechy statystyczne i numeryczne

- Modelowanie cech stałych i zmiennych w czasie
- Mieszanka efektów stałych i (wielu) efektów losowych
- Duże rozmiary danych → brak możliwości zastosowania standardowych metod estymacji efektów modeli

Copyright ©2023 Joanna Szyda

Najważniejsze aspekty obliczeniowe

- Predykcja rutynowa → stabilne, bardzo wydajne obliczeniowo oprogramowanie
- Potok (pipeline) wielu programów
  - Edycja danych
  - Formatowanie danych
  - Predykcja / estymacja parametrów modeli
  - Formatowanie danych wyjściowych
- Pipeline setup → Bash, Nextflow, UC4, Snakemake
- Edycja danych, estymacja efektów → **Fortran**, Python, R, SAS
- Estymacja, obliczenia rónoległe → openMP in Fortran

Copyright ©2024 Joanna Szyda

pipeline

```
VERSUS/mnt/dysk/DYSK/GBV:ls -l
total 72
drwxrwxrwx 12 tomek users  4096 2015-10-26 19:00 DATA
drwxrwxrwx  2 tomek users  4096 2013-10-01 11:11 dt_edu
drwxrwxrwx  2 tomek users  4096 2015-10-28 12:01 EBV
drwxrwxrwx  2 szyda users  4096 2014-12-02 08:56 EG
drwxrwxrwx 31 tomek users  4096 2012-06-28
drwxrwxrwx 16 szyda users  4096 2014-03-29
drwxrwxrwx 12 tomek users  4096 2015-10-26
drwxrwxrwx 10 tomek users  4096 2013-07-28
drwxrwxrwx  2 szyda users  4096 2011-10-03
drwxrwxrwx 11 tomek users  4096 2015-11-07
drwxrwxrwx 15 tomek users  4096 2015-10-27
drwxrwxrwx 10 tomek users  4096 2015-08-09
-rwxrwxrwx  1 tomek users 11109 2009-08-02
drwxrwxrwx  2 tomek users 12288 2015-11-08
VERSUS/mnt/dysk/DYSK/GBV:
```

```
-rwxrwxrwx 1 suchy users  17052 2013-08-09 23:39 CombDGVEBVmu_VALEAAP.f90
-rwxrwxrwx 1 suchy users  16847 2013-08-09 23:39 REL4DGVxPH_VALEAAP.f90
-rwxrwxrwx 1 szyda users  19288 2013-08-11 21:22 makefile~
-rwxrwxrwx 1 suchy users  17056 2013-08-11 22:18 CombDGVEBV_VALEAAP.f90
-rwxrwxrwx 1 suchy users  17056 2013-08-12 16:03 CombDGVEBVEAAP.f90
-rwxrwxrwx 1 suchy users   1289 2014-03-12 18:36 BullDatSUR.sas
-rwxrwxrwx 1 suchy users   6395 2014-08-28 14:58 ChooseDGVPI.sas
-rwxrwxrwx 1 suchy users   3811 2014-08-28 20:08 ChooseDGVPI2.sas
-rwxrwxrwx 1 szyda users  20509 2014-09-06 09:26 deregressREL10.f90~
-rwxrwxrwx 1 szyda users  19470 2014-09-06 09:27 makefile
-rwxrwxrwx 1 szyda users  20487 2014-09-06 09:28 deregressREL10.f90
-rwxrwxrwx 1 szyda users  11406 2014-09-09 07:57 Val4ITB732.sas.old
-rwxrwxrwx 1 szyda users  11406 2014-09-09 07:57 Val4ITB732.sas~
-rwxrwxrwx 1 szyda users   3160 2014-09-10 19:12 cmpDPRF2runs.sas~
-rwxrwxrwx 1 szyda users   3160 2014-09-10 19:13 cmpDPRF2runs.sas
-rwxrwxrwx 1 szyda users 113914 2014-09-10 19:13 cmpDPRF2runs.lst
-rwxrwxrwx 1 szyda users  11716 2014-09-10 19:13 cmpDPRF2runs.log
-rwxrwxrwx 1 szyda users   4052 2014-09-10 20:07 cmpDGV2runsNatID.sas~
-rwxrwxrwx 1 szyda users   4010 2014-09-10 20:09 cmpDGV2runsNatID.sas
-rwxrwxrwx 1 szyda users  58350 2014-09-10 20:09 cmpDGV2runsNatID.lst
-rwxrwxrwx 1 szyda users  17565 2014-09-10 20:09 cmpDGV2runsNatID.log
-rwxrwxrwx 1 szyda users  11406 2014-09-11 18:34 Val4ITB732.sas
-rwxrwxrwx 1 szyda users   3656 2015-07-19 11:18 CombDGVPI.sas
VERSUS/mnt/dysk/DYSK/GBV/PROGRAMS/SOURCE:
```

Copyright ©2023 Joanna Szyda

Predykcja wartości hodowlanych, a bioinformatyka

Wartość hodowlana

- Podstawowe narzędzie przemysłu hodowlanego
- Definicja popularna: „Jakość genetyczna” osobników
- Definicja genetyczna: sumaryczna wartość **addytywnych** efektów **wszystkich** genów warunkujących daną cechę
- Najważniejsza wielkość pozwalająca na ustalenie rankingu zwierząt hodowlanych
- Ludzki ekwiwalent → polygenic risk scores

Copyright ©2024 Joanna Szyda

Predykcja wartości hodowlanych, a bioinformatyka

Pojęcia

- EBV: **E**stimated **B**reeding **V**alue
- PTA: **P**redicted **T**ransmitting **A**bility → ½ EBV
- DRP: **D**eregressed **P**roof → EBV poddane deregresji
- GEBV: **G**enomically **E**nhanced **EBV** → EBV z wykorzystaniem SNP
- PA: **P**arent **A**verage → EBV młodego osobnika obliczona na podstawie EBV rodziców / ojca i dziadka
- MACE: **M**ultiple **A**cross **C**ountry **E**valuation → międzynarodowa ocena wartości hodowlanej

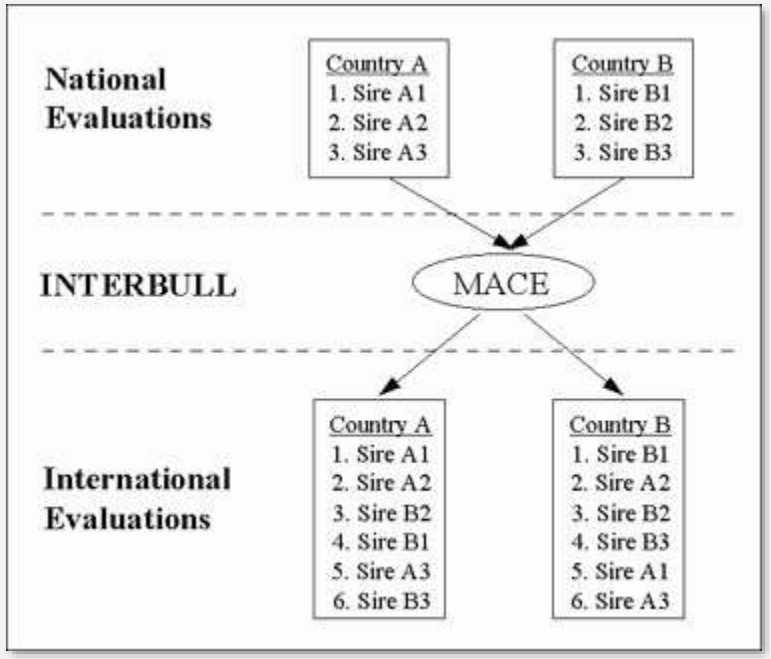
Copyright ©2024 Joanna Szyda

(3/7) **Teaser**

Pierwszy model

Międzynarodowa ocena wartości hodowlanej bydła - schemat

Różne kraje: EBV, dokładność, pochodzenie



EBV, dokładność

Copyright ©2024 Joanna Szyda

Pierwszy model

Międzynarodowa ocena wartości hodowlanej bydła - model

$$\mathbf{y}_i = 1\boldsymbol{\mu}_i + \mathbf{Z}_i\mathbf{Q}\mathbf{w}_i + \mathbf{Z}_i\mathbf{a}_i + \mathbf{e}_i$$

- $\boldsymbol{y}_i$ wektor DRP osobników z kraju *i*
- $\boldsymbol{\mu}_i$ średnia dla kraju *i*
- $\boldsymbol{w}_i$ grupa genetyczna *i* - $\boldsymbol{w} \sim N\left(\mathbf{0}, \boldsymbol{A}_{\boldsymbol{pp}} \cdot g\right)$
- $\boldsymbol{a}_i$ wektor EBV dla kraju i w skali międzynarodowej $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{A}_{\boldsymbol{nn}} \cdot g)$
- $\boldsymbol{e}$ wektor błędów losowych
- $\boldsymbol{Z}_i$ macierz wystąpień → przyporządkowuje $\boldsymbol{y}$ do $\boldsymbol{a}$
- $\boldsymbol{Q}$ macierz wystąpień → przyporządkowuje $\boldsymbol{y}$ do $\boldsymbol{w}$

Copyright ©2023 Joanna Szyda

Międzynarodowa ocena wartości hodowlanej bydła – struktura kowariancji

$$\mathbf{y}_i = 1\boldsymbol{\mu}_i + \mathbf{Z}_i\mathbf{Q}\mathbf{w}_i + \mathbf{Z}_i\mathbf{a}_i + \mathbf{e}_i$$

Przykład dla 2 krajów (1 i 2)

$$\mathrm{var}\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{pp}g_{11} & \mathbf{A}_{pp}g_{12} & \mathbf{A}_{pn}g_{11} & \mathbf{A}_{pn}g_{12} & 0 & 0 \\ \mathbf{A}_{pp}g_{21} & \mathbf{A}_{pp}g_{22} & \mathbf{A}_{pn}g_{21} & \mathbf{A}_{pn}g_{22} & 0 & 0 \\ \mathbf{A}_{np}g_{11} & \mathbf{A}_{np}g_{12} & \mathbf{A}_{nn}g_{11} & \mathbf{A}_{nn}g_{12} & 0 & 0 \\ \mathbf{A}_{np}g_{21} & \mathbf{A}_{np}g_{22} & \mathbf{A}_{nn}g_{21} & \mathbf{A}_{nn}g_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{D}_1\sigma^2_{e1} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{D}_2\sigma^2_{e2} \end{pmatrix}$$

- $\boldsymbol{A}$ macierz spokrewnień addytywnie genetycznych między osobnikami / grupami genetycznymi
- $\boldsymbol{G} = \begin{bmatrix} \boldsymbol{g_{11}} & \boldsymbol{g_{12}} \\ \boldsymbol{g_{12}} & \boldsymbol{g_{22}} \end{bmatrix}$ wartości (ko)wariancji genetycznej $\boldsymbol{D}$ macierz (ko)wariancji efektów błędu

Copyright ©2023 Joanna Szyda

Pierwszy model

Międzynarodowa ocena wartości hodowlanej bydła

	Country 1			Country 2		
Sire	EDC	BV	DRP	EDC	BV	DRP
1	58	9.0	9.7229	90	13.5	14.5088
2	150	10.1	9.9717	65	7.6	7.7594
3	20	15.8	19.2651	–	–	–
4	25	–4.7	–8.5711	–	–	–
5	–	–	–	30	19.6	23.9672
6	–	–	–	55	–5.3	–9.6226

EDC, effective daughter contribution; BV, breeding value; DRP, deregressed proof.

	Solutions			
Effects	Country 1		Country 2	
Country effect	7.268		9.036	
Animal/group	A	B	A	B
1	2.604	9.871	2.661	11.697
2	2.176	9.444	0.403	9.439
3	8.059	15.327	5.001	14.037
4	–9.865	–2.597	–5.605	3.431
5	13.634	20.902	9.728	18.764
6	–18.086	–10.818	–13.203	–4.167
7	4.310	11.578	3.071	12.106
8	7.015	14.283	4.489	13.525
9	–6.299	0.969	–5.059	3.977
G1	0.174	7.442	–0.092	8.944
G2	–0.124	7.144	0.126	9.162
G3	–0.071	7.197	0.264	9.300
G4	0.087	7.355	–0.288	8.748
G5	–0.067	7.201	–0.010	9.026

A = solutions for animals and groups from the MME; B = solutions for animals and groups expressed in each country scale.

Copyright ©2023 Joanna Szyda

(4/7) **Przykładowy projekt**

Projekt THETA

A comparison of **genomically enhanced breeding values** predicted by different **single-step** approaches

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{W_G}\,\boldsymbol{a} + \boldsymbol{e}$$

- $\boldsymbol{y}$ fenotypy osobników
- $\boldsymbol{b}$ efekty stałe
- $\boldsymbol{a}$ efekty losowe = GEBV
  - $\boldsymbol{a} \sim N\left(0, \boldsymbol{H_G} \cdot \sigma_a^2\right)$
- $\boldsymbol{e}$ błąd, $\boldsymbol{e} \sim N\left(0, \boldsymbol{D} \cdot \sigma_e^2\right)$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{W_G}\,\boldsymbol{a} + \boldsymbol{e}$$

- $\boldsymbol{y}$ fenotypy osobników
- $\boldsymbol{b}$ efekty stałe
- $\boldsymbol{a}$ efekty losowe, $\boldsymbol{a} = \boldsymbol{Z}\boldsymbol{g} + \boldsymbol{u}$
  - $\boldsymbol{g}$ efekty SNP
  - $\boldsymbol{u}$ efekt poligeniczny
- $\boldsymbol{e}$ błąd, $\boldsymbol{e} \sim N\left(0, \boldsymbol{D} \cdot \sigma_e^2\right)$

Copyright ©2024 Joanna Szyda

Projekt THETA

A comparison of **genomically enhanced breeding values** predicted by different **single-step** approaches

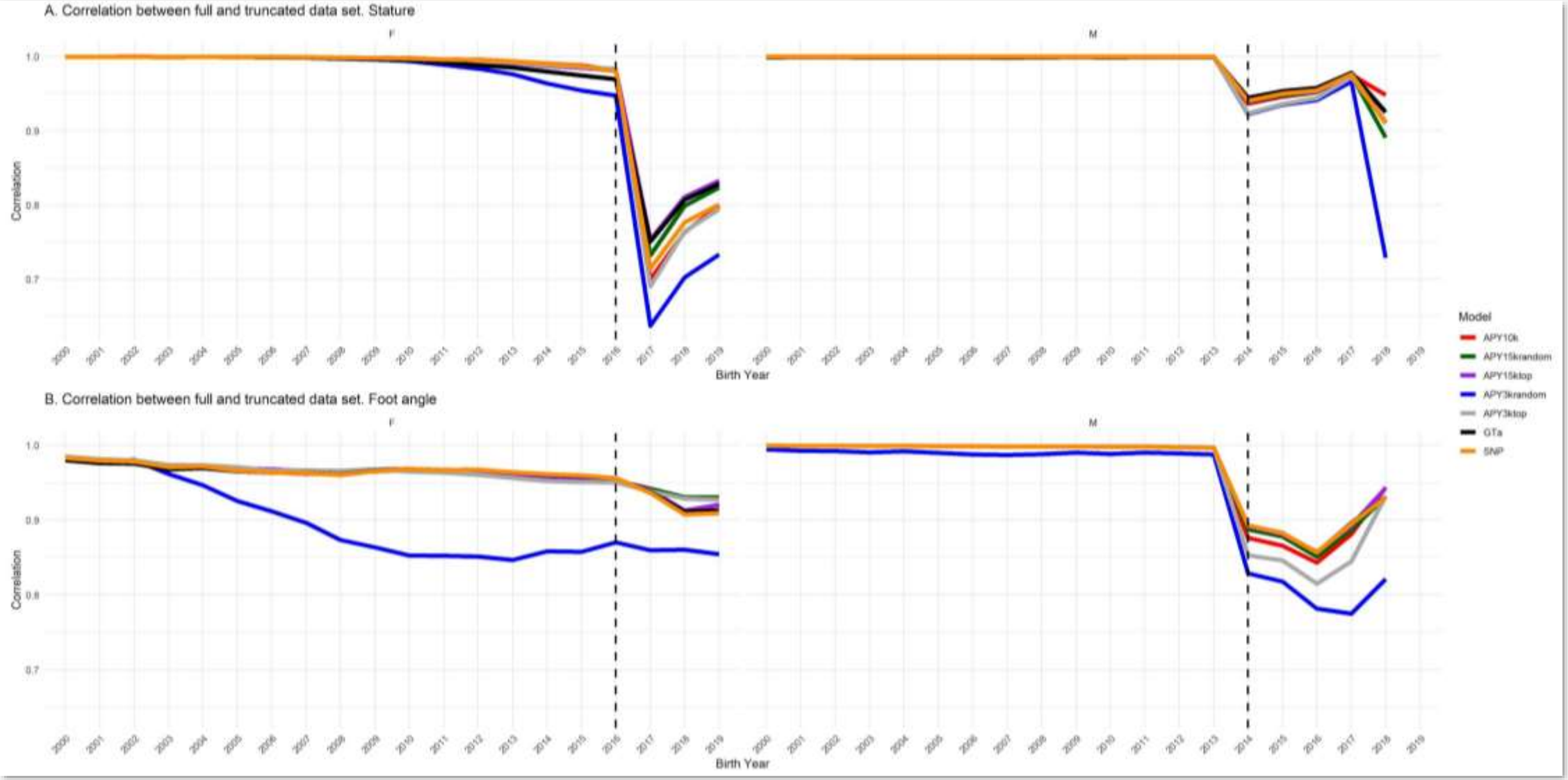
Copyright ©2024 Joanna Szyda

Porównane metody estymacji

Copyright ©2024 Joanna Szyda

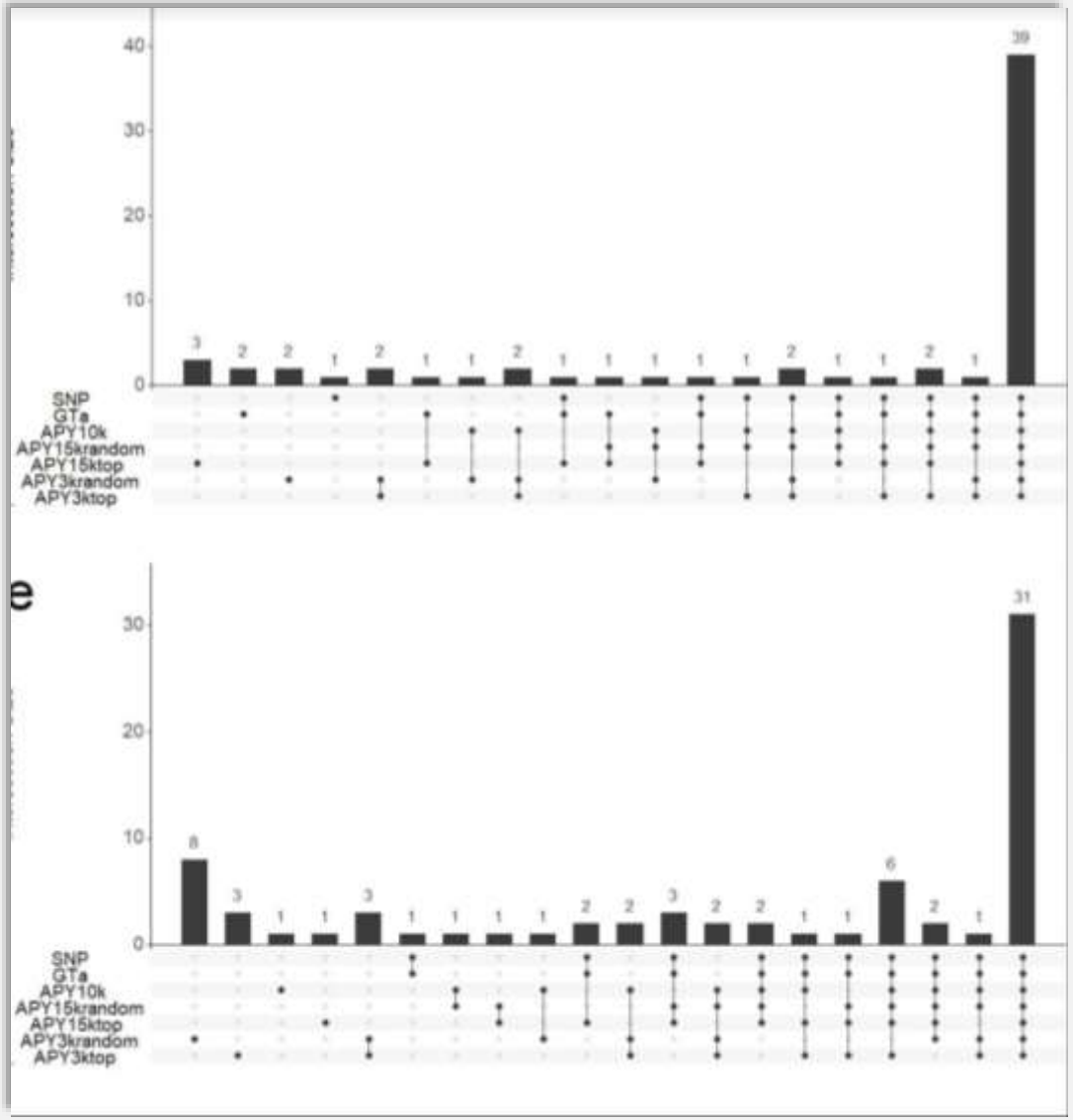
Projekt THETA

Korelacje między GEBV



Copyright ©2024 Joanna Szyda

Liczba wspólnych osobników top-50



Copyright ©2024 Joanna Szyda

(domowe) Przykłady zastosowania modeli mieszanych

Short Communication | Open Access | Published: 16 December 2022

Application of mixed linear models for the estimation of functional effects on bovine stature based on SNP summary statistics from a whole-genome association study

Krzysztof Kotlarz, Barbara Kosinska-Selbi, Zexi Cai, Goutam Sahana & Joanna Szyda

Genetics Selection Evolution **54**, Article number: 80 (2022) | Cite this article

851 Accesses | Metrics

Copyright ©2023 Joanna Szyda

(domowe) Przykłady zastosowania modeli mieszanych

RESEARCH | VOLUME 103, ISSUE 5, P4483-4494, MAY 2020

PD

Two-stage genome-wide association study for the identification of causal variants underlying hoof disorders in cattle

T. Suchocki • Ch. Egger-Danner • H. Schwarzenbacher • J. Szyda

Open Archive • Published: March 27, 2020 • DOI: https://doi.org/10.3168/jds.2019-17542

Copyright ©2023 Joanna Szyda

(domowe) Przykłady zastosowania modeli mieszanych

Exploring the Potential Genetic Heterogeneity in the Incidence of Hoof Disorders in Austrian Fleckvieh and Braunvieh Cattle

Barbara Kosińska-Selbi[1], Tomasz Suchocki[1,2], Christa Egger-Danner[3], Hermann Schwarzenbacher[3], Magdalena Frąszczak[1] and Joanna Szyda[1,2*]

Copyright ©2023 Joanna Szyda

(5/7) **Wykłady**

Charakterystyka przedmiotu

Wykłady

1. Wykład wstępny
2. Algebra macierzowa
3. Best Linear Unbiased Prediction (BLUP): model jednocechowy z pojedynczym efektem losowym
4. Best Linear Unbiased Prediction (BLUP): model wielocechowy
5. Metody redukcji wymiarów modeli wielocechowych BLUP
6. Analiza danych powtarzalnych w czasie
7. Predykcja genomowej wartości hodowlanej
8. Model jednostopniowy
9. Aspekty obliczeń wielkoskalowych
10. - 15. Prezentacje zaliczeniowe

Copyright ©2024 Joanna Szyda

Charakterystyka przedmiotu

theta.edu.pl

Home Page | Contact Us | Teaching | Job | Login

Teaching

- Analiza danych pochodzących z sekwencjonowania następnej generacji
- Narzędzia bioinformatyczne w naukach biologicznych
- Pracownia Informatyczna (BL)
- Statystyka zaawansowana
- Zaawansowane aspekty praktycznego zastosowania modeli mieszanych

Przedmioty realizowane w semestrze zimowym 2024/2025.

- Alumni
- Collaboration
- Conferences
- Grants
- Projects
- Publications
- Seminars
- Statistics in medicine

Copyright ©2024 Joanna Szyda

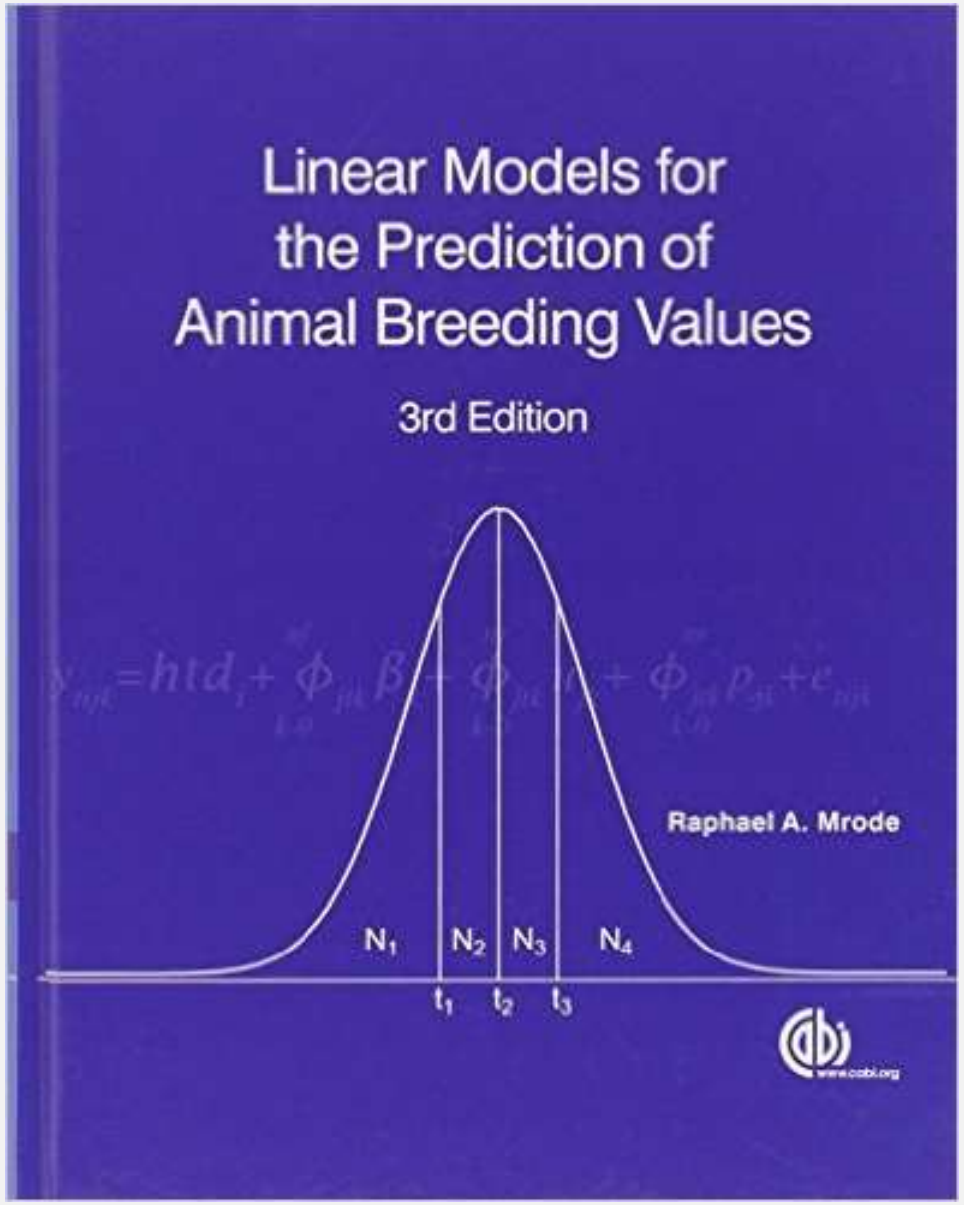
(6/7) **Literatura**

Literatura

1. Wykłady – literatura podstawowa
2. Literatura uzupłeniajaca – książki, publikacje

Copyright ©2023 Joanna Szyda

Literatura

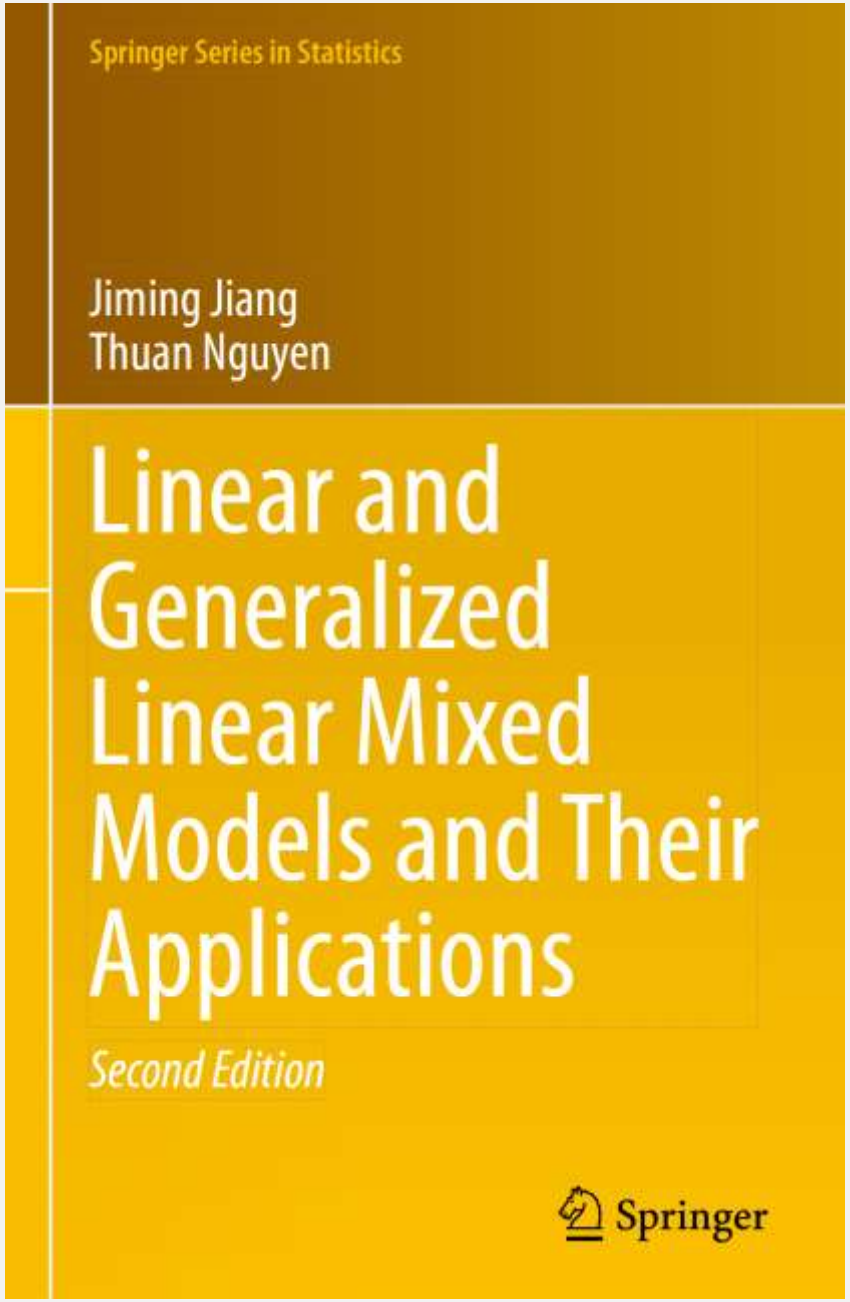


Copyright ©2023 Joanna Szyda

Literatura

Copyright ©2023 Joanna Szyda

Literatura



Copyright ©2024 Joanna Szyda

(7/7) **Kontakt**

Kontakt

Katedra Genetyki, Kożuchowska 7

Konsultacje

- termin ustalony indywidualnie
- online
- stacjonarne
- Email

Copyright ©2023 Joanna Szyda

1. Po co?
2. Predykcja wartości hodowlanych, a bioinformatyka
3. **Teaser**
4. **Przykładowy projekt**
5. Wykłady
6. Literatura
7. Kontakt

1. Na czym polega problem p >> n
2. Co to jest EBV?
3. Czym się charakteryzuje model mieszany?
4. Na czym polega Interbull MACE?