Validation of Genomic selection in pigs:
a small data set with medium-dense marker coverage

Luc Janss, Vivi Gregersen,
Christian Bendixen, Mogens Lund

A A R H U S U N I V E R S I T E T

Dept. of Genetics and Biotechnology
Statistical genetics and bioinformatics

Outline

- **Material**
  - Small pig data, 6K SNP chip
- **Methods**
  - 2 Bayesian models
  - Estimation of hyper parameters
  - 10X cross validation
- **Results etc**
  - Model fit, prediction, ...

Material: a small pig data set for a "case study"

- **Originally 169 genotyped boars**
- **Using 6K Illumina porcine SNP array**
- **After combining with phenotypes and marker edits:**
  - 127 genotyped boars with reasonable progeny groups
  - 3463 good polymorphic SNP markers
- **Trait:**
  - Boar EBV for growth

Methods: association model

with scaling factors for variance modelling

All markers included

Allele design matrix

Allele additive effects

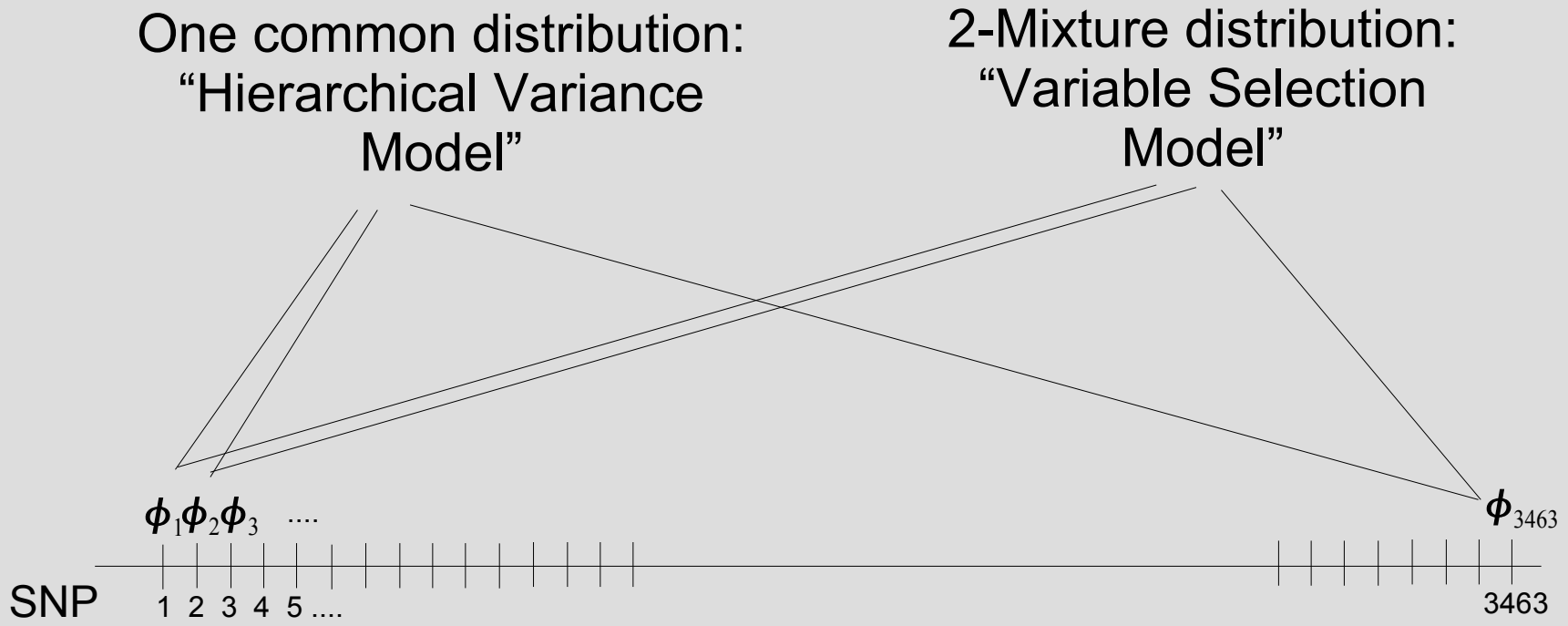
$$\mathbf{y} = \boldsymbol{\mu} + \sum_{i=1}^{M} \phi_i \mathbf{X}_i \mathbf{b}_i + \mathbf{e}$$

$$\mathbf{b}_i \sim N(0,1)$$

ϕ_i Models the variance

Priors for variance terms

or how to model 3463 marker variances on 127 observations



AARHUS UNIVERSITET

Dept. of Genetics and Biotechnology
Statistical genetics and bioinformatics

Luc Janss
EAAP Vilnius Aug '08

Prior distributions

- 1 common distribution
- $\phi_i \sim N(0, \sigma^2_H)$
- σ^2_H is ~ variance per marker and is estimated
-
-

- 2-mixture distribution
- $\phi_i \sim \pi_0 N(0, \sigma^2_{s_0}) + \pi_1 N(0, \sigma^2_{s_1})$
- $\sigma^2_{S_1}$ is ~ variance per "on" marker and is estimated
- $\sigma^2_{S_0}$ is ~ variance per "off" marker and is set small (e.g. 1% total)
- Proportions "on" and "off" are set and determine peakedness of profile

Further model details:

- **MCMC based on Gibbs samplers**
  - Normal for mean, allele effects, scaling factors
  - Inverse chi-square for 2 variance components: hyper variance for markers, residual variance
- **Add functions of parameters:**
  - total genomic values
  - genomic variance

Analysis and validation

- **Step 1: hyper parameters estimated using all data (”REML”)**
- **Step 2: 10X cross validation with hyper parameters set fixed (“BLUP”)**
  - Data randomly divided in 10 groups
  - In 10 analyses, data in 1 group was left out and predicted based on other 9 groups
  - Predictions collected for observation from analysis where observation was left out

Results: estimated variances

Model	Prior value π_1	σ^2_E	σ^2_G	$\sigma^2_{S_1}, \sigma^2_H$	σ^2_{TOT}	BF
VSM	10	429	368	16.3	797	-470.3
VSM	20	356	441	4.00	797	-458.5
VSM	40	253	529	0.88	782	-452.8
HVM	100	215	598	0.20	813	-433.1

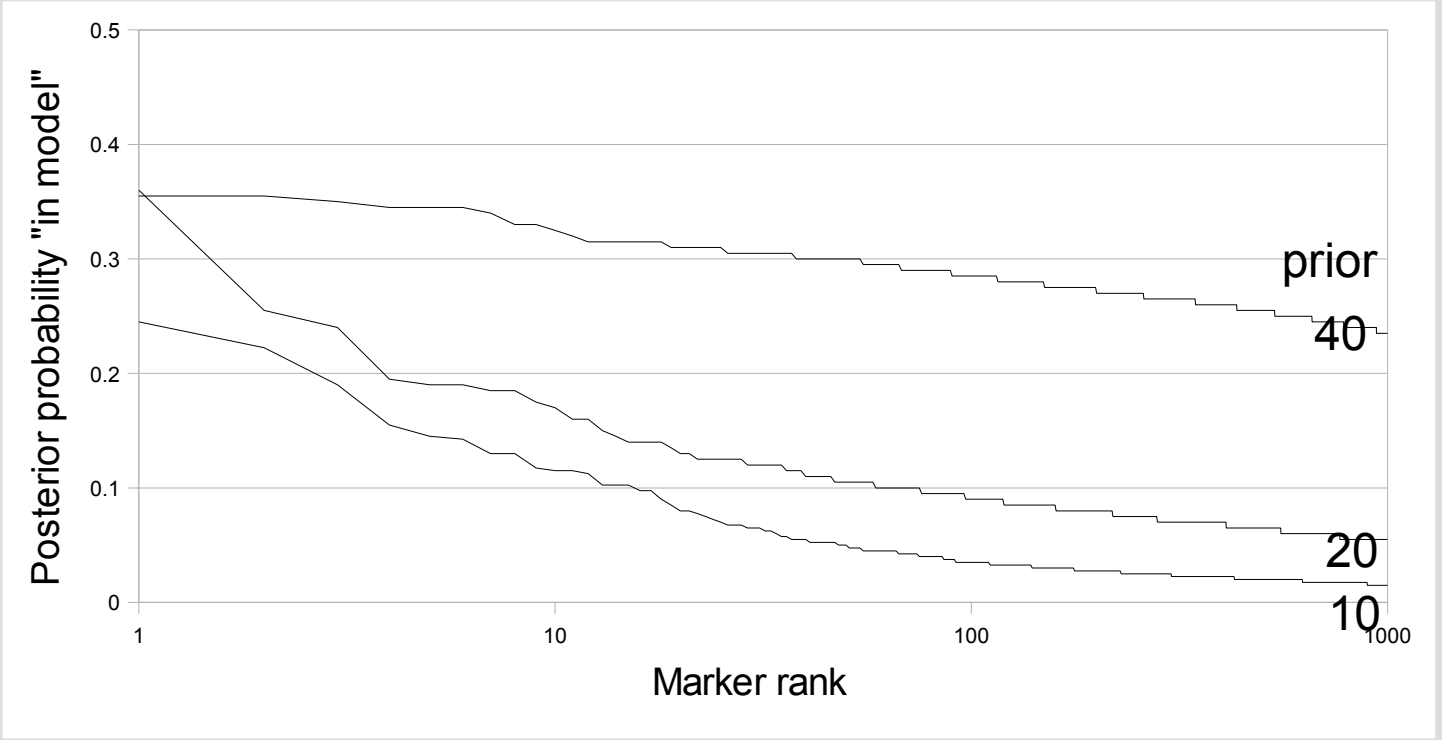
HVM fits best: lowest residual variance, highest genomic variance, highest Bayes Factor, although some overestimation in total variance (raw variance 776).

Results: prediction

Model	Prior value π_1	Estimated values $\sigma^2_{S_1}, \sigma^2_H$	For predicted EBVs correlation	regression
VSM	10	16.3	0.39	0.98
VSM	20	4	0.46	1.05
VSM	40	0.88	0.48	1.01
HVM	100	0.2	0.5	0.99

HVM predicts best but VSM with large proportion markers “on” comes close.
Predictions are all (close to) unbiased.

Posterior probabilities for VSM



Conclusions

- **Bayesian models fit and predict well**
  - Despite having 30x more predictors than observations
  - Only slight overfit in variance ? (2-4%)
  - Unbiased predictions with hyper parameters estimated
    - Setting hyper parameters away from estimates gave biased predictions
  - Behaves like we're used from BLUP for predicting
    - explained variance ~75%, prediction correlation ~50%

Conclusions

- **A model with 1 common distribution for variance terms (HVM) performed best**
  - Signs that in this small data identification of associated markers was difficult
    - But even then sensible predictions can be made, mostly based on "genomic relationship"
- **Even in small data and with medium-dense markers Genomic Predictions work and behave well.**