Experimental design for genome-wide association studies

Christoph Lippert, Oliver Stegle, Hannes Nickisch, Karsten Borgwardt, Detlef Weigel

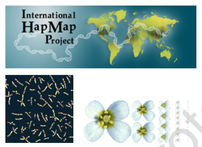
Max Planck Institutes, Tübingen

MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

Genome-wide association studies

- Genome-wide association studies (GWA) aim at mapping the **phenotypes** of samples to polymorphic **genetic loci** that explain the phenotypic variation.
- **Thousands of genotypes** are available on the web.
  The *human* HapMap [The International HapMap Consortium, 2007]
- **Hundreds of full genomes** are being sequenced.
  1000 *human* genomes project [1000 Genomes Project]
  1001 *A. thaliana* genomes project [1001 Genomes]



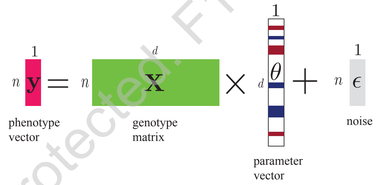
Measuring phenotypes is often expensive or time-consuming.

- Can phenotype 1000 plants at a time
- About 10 replica per genotype
  = **phenotypes for 100 genotypes**

⇒ How to **optimally select** the set of **genotypes** that should be **phenotyped**?

GWA via the Bayesian LASSO

- The Bayesian LASSO is a **sparse linear regression** model.
- Weights $\boldsymbol{\theta}$ determine the *strength* and *direction* of regulatory effects.
- In this model *multiple loci* can have an *additive effect* on a phenotype.



n: number of individuals
d: number of SNPs
$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

The corresponding **Gaussian likelihood**, defined by the conditional distribution of observed phenotypes $\mathbf{y}$ given genotypes $\mathbf{X}$ and model weights $\boldsymbol{\theta}$, follows as

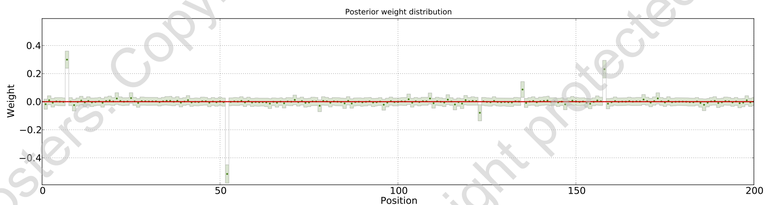
$$\mathtt{P}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}; \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}). \tag{1}$$

The Bayesian framework characterises the *uncertainty* in the weights $\boldsymbol{\theta}$ by defining a *posterior probability* given the data. Only a small subset of the measured loci influence the phenotype, motivating a *sparse* weight vector $\boldsymbol{\theta}$. *Sparsity* is encouraged by employing a **Laplace prior** on $\boldsymbol{\theta}$,

$$\mathtt{P}(\boldsymbol{\theta}|\sigma, \tau) \propto \prod_{i=1}^{d} \exp\left[-\tau\left|\frac{\theta_i}{\sigma}\right|\right]. \tag{2}$$

For the specified model, the **posterior distribution** over the weights $\boldsymbol{\theta}$ is given by Bayes' rule as

$$\mathtt{P}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}; \sigma^2, \tau) \propto \mathtt{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}; \sigma^2)\mathtt{P}(\boldsymbol{\theta}|\sigma, \tau). \tag{3}$$

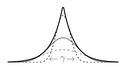


Approximate Bayesian learning

We follow the ideas developed in [Nickisch and Seeger, 2009] and approximate the intractable exact posterior, $\mathtt{P}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}; \sigma^2, \tau)$, by a **Gaussian approximation** of the form

$$\mathcal{Q}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}; \sigma^2, \tau) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}) \prod_{i=1}^{d} t_i(\theta_i, \gamma_i), \tag{4}$$

with variational parameters $\boldsymbol{\gamma} \in \mathbb{R}_+^d$. $t_i(\theta_i, \gamma_i) \propto \mathcal{N}(0, \boldsymbol{\gamma}_i)$ are Gaussian site functions that **lower bound** the exact Laplace sites. The product of the likelihood term and the approximate sites in Equation 4 is tractable and can be written as



$$\mathcal{Q}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}; \sigma^2, \tau) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{5}$$

with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2\mathbf{A}^{-1}, \qquad \mathbf{A} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathtt{diag}(\boldsymbol{\gamma})^{-1}\right) \tag{6}$$

Under this approximation, the prediction at an unseen test input $\mathbf{x}_\star$ again has a Gaussian distribution,

$$\mathbf{y}_\star \sim \mathcal{N}(\boldsymbol{\mu}_\star, \sigma_\star^2), \qquad \boldsymbol{\mu}_\star = \mathbf{x}_\star\boldsymbol{\mu}_{\boldsymbol{\theta}}, \qquad \sigma_\star^2 = \mathbf{x}_\star^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{x}_\star. \tag{7}$$

The variational parameters $\boldsymbol{\gamma}$ are determined such that $\mathcal{Q}_{\boldsymbol{\gamma}}$ approximates the exact posterior well. We developed an algorithm that minimises the convex relaxation of the KL-divergence between the exact and approximate posterior from [Nickisch and Seeger, 2009] in runtime that is **linear** in the number of SNPs, $\mathbf{O}(dn^3)$.

Experimental design

We perform **greedy blockwise experimental design** using alternative selection criteria.

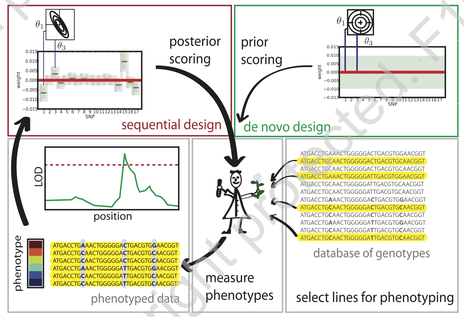
Mean marginal entropy

We use the expected reduction in mean marginal entropy of the posterior as a design criterion.

$$\Delta\mathrm{H}_{\mathrm{mm}} = \mathrm{H}_{\mathrm{mm}} - \mathrm{H}_{\mathrm{mm}}^{\mathcal{C}}. \tag{8}$$

The reduction in mean marginal entropy $\Delta\mathrm{H}_{\mathrm{mm}}$ of a fully trained model, defined by the approximate posterior of $\boldsymbol{\theta}$, is

$$\frac{1}{2}\sum_{d=1}^{D}\log\boldsymbol{\Sigma}_{\boldsymbol{\theta}}[d,d] - \frac{1}{2}\sum_{d=1}^{D}\log\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathcal{C}}[d,d]. \tag{9}$$



Population posterior variance

Another experimental design criterion is the expected posterior variance evaluated on a **target population** $\mathcal{P}$.

$$\frac{1}{|\mathcal{P}|}\sum_{\mathbf{x}_\mathcal{P}\in\mathcal{P}}\mathbf{x}_\mathcal{P}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathcal{C}}\mathbf{x}_\mathcal{P}^{\mathrm{T}}. \tag{10}$$

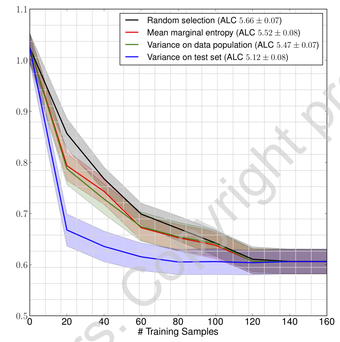
Note that $\mathcal{P}$ usually is either the set of *all genotypes* or in a transductive setting an *independent test set*.

In both criteria the **posterior covariance matrix** $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathcal{C}}$ including a number of candidate instances $\mathbf{X}_\mathcal{C}$ is approximated by

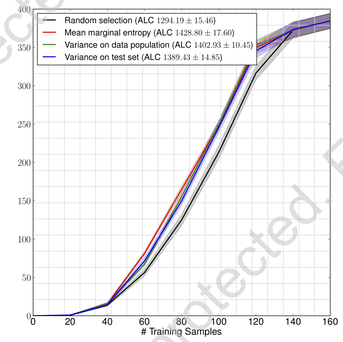
$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathcal{C}} = \sigma^2\left(\mathbf{X}_\mathcal{C}^{\mathrm{T}}\mathbf{X}_\mathcal{C} + \mathbf{A}\right)^{-1}. \tag{11}$$

Study of flowering-time in A. thaliana [Atwell et al., 2010]

- 166 accessions
- 216,509 confident SNP measurements from a 250k SNP chip
- 10 random splits in 133 candidates and 33 test individuals
- Hyperparameters σ and τ estimated via cross-validation
- Likelihood-ratio tests on each SNP j, comparing $\mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ and $\mathcal{N}(\mathbf{y}|\mathbf{X}_{-j}\boldsymbol{\mu}_{\boldsymbol{\theta}_{-j}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{-j}})$



RMSE on independent test set



number of associations at FDR ≤ 1%

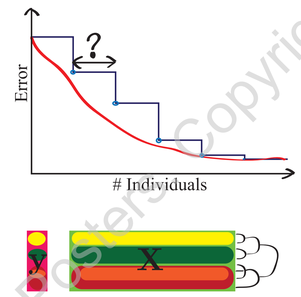
Discussion and Outlook

Discussion

- Sequential experimental design significantly improves GWA results compared to random selection.
- By doing **greedy blockwise** experimental design it is feasable to include *multiple individuals* per iteration.

Outlook

- Experimental design on 1200 recently genotyped *A. thaliana* accessions to extend the study by [Atwell et al., 2010].
- *Optimal solution* of blockwise selection
- How to determine the *optimal block size*?
- Extend to **multiple phenotypes**
- Correct for **population structure**
  Non-i.i.d. data
  Remove spurious associations



References

1000 Genomes Project. A deep catalog of human genetic variation. World Wide Web electronic publication, 2009. URL http://www.1000genomes.org.

1001 Genomes. A catalog of *Arabidopsis thaliana* genetic variation, 2009. URL http://www.1001genomes.org.

S. Atwell, Y. Huang, B. Vilhálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. Tarone, T. Hu, R. Jiang, N. Muliyati, X. Zhang, M. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. Ecker, N. Faure, J. Kniskern, J. Jones, T. Michael, A. Nemri, F. Roux, D. Salt, C. Tang, M. Todesco, M. Traw, D. Weigel, P. Marjoram, J. Borevitz, J. Bergelson, and M. Nordborg. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, 2010.

H. Nickisch and M. Seeger. Convex variational bayesian inference for large scale generalized linear models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 761–768, 2009.

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449: 851–861, 2007.