Experimental design for genome-wide association studies

Christoph Lippert, Oliver Stegle, Hannes Nickisch, Karsten Borgwardt, Detlef Weigel
Max Planck Institutes, Tübingen

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.
- Many thousands of phenotypes are available on the web.
- Hundreds of full genomes are being sequenced.

Measuring phenotypes is often expensive or time-consuming.

- Can phenotype 100 plants at a time.
- About 10 replicates per genotype.
- Phenotypes for 100 genotypes.

GWA via the Bayesian LASSO

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

Approximate Bayesian learning

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

\[
Q_\rho(\theta, \mathbf{y}, \sigma^2, \tau) = \mathcal{N}(\mathbf{y}, \mathbf{X}^T \mathbf{X}\sigma^2 + \tau^{-1}) \prod_{i=1}^{d} \mathcal{N}(\theta_i | \gamma_i, \sigma^2) .
\] (4)

with variational parameters \( \gamma \in \mathbb{R}^d \) and \( t = (\gamma, \tau) \). 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

\[
Q_\rho(\theta, \mathbf{y}, \sigma^2, \tau) = \mathcal{N}(\mathbf{y}, \mathbf{X}^T \mathbf{X}\sigma^2 + \tau^{-1}) \prod_{i=1}^{d} \mathcal{N}(\theta_i | \gamma_i, \sigma^2) .
\] (5)

with mean \( \mu_p \) and covariance matrix \( \Sigma_p \).

\[
\mu_p = A^T \mathbf{y}, \quad \Sigma_p = \sigma^2 A^{-1}, \quad A = (\mathbf{X}^T \mathbf{X} + \tau \mathcal{N}(\sigma^2))^{-1} .
\] (6)

Under this approximation, the prediction at an unseen test input \( \mathbf{x}_* \) again has a Gaussian distribution.

\[
y_* \sim N(\mu_*, \sigma^2_*), \quad \mu_* = \mathbf{x}_*^T \mathbf{X} \mathbf{\mu}, \quad \sigma^2_* = \mathbf{x}_*^T \Sigma \mathbf{x}_*. \] (7)

The variational parameters \( \gamma \) are determined such that \( Q_\rho \) approximates the exact posterior well.

We developed an algorithm that minimizes 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. (O(\#))

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.

\[
\Delta H_{\text{mean}} = H_{\text{mean}}(\theta) - H_{\text{mean}}(\theta|\mathbf{y}, \mathbf{x}, \sigma^2, \tau) .
\] (8)

The reduction in mean marginal entropy, \( \Delta H_{\text{mean}} \), of a fully trained model, defined by the approximate posterior of \( \theta \), is

\[
\Delta H_{\text{mean}} = \frac{1}{2} \sum_{i=1}^{d} \Sigma_i^{\gamma_i} d_i + \frac{1}{2} \sum_{i=1}^{d} \Sigma_i^{\gamma_i} d_i .
\] (9)

Population posterior variance

Another experimental design criterion is the expected posterior variance evaluated on a target population \( P \).

\[
\text{PRMSE} = \frac{1}{|P|} \sum_{\mathbf{x} \in P} \text{var} (\mathbf{y} | \mathbf{x}) .
\] (10)

Note that \( P \) usually is either the set of all genotypes or in a transductive setting an independent test set.

Population posterior variance

In both criteria the posterior covariance matrix \( \Sigma_p \) including a number of candidate instances \( \mathbf{x}_C \) is approximated by

\[
\Sigma_p = \sigma^2 (\mathbf{X}_C \mathbf{X}_C + \mathbf{A})^{-1} .
\] (11)

References

