GBLUP and Kernel-based regression

...

Evangelina López de Maturana & Oscar González-Recio

Topics

Background → Non-parametric regressions → Reproducing Kernel Hilbert Spaces → RKHS-BLUP → RKHS-GBLUP → Single -Step

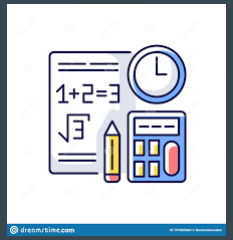
Background

Complex traits are likely to be influenced by many genomic regions, often interacting among them

Genome prediction of ‘total’ genetic effects is motivated by the non-linear relationship between outcomes and genotypes

State-of-the art statistical approaches (kernels based) capturing also non-additive effects, either parametrically or non-parametrically (Morota and Gianola, 2014)

Background



To find the relationships between a set of independent/predictor variables (inputs), and the set of dependent variables (responses/outputs):

$$y = f(x) + e$$

The goal is to estimate an unknown (desired) continuous and real valued function f(x)

Background

The parametric regression function has a rigid structure comprising a set of assumptions which may not be met in genomic selection problems

Sample size (n) is usually smaller than the number of predictors (*m*)--> large *m* small *n* problem ("the curse of dimensionality" (Bellman 1961))

Departures from linearity can be addressed by semiparametric approaches, such as Reproducing Kernel Hilbert Space (RKHS) regressions or neural networks

Background - Nonparametric regression

No explicit parameterization is given

$$y = f(x) + e$$

It does not assume major hard-to-satisfy hypotheses on the regression function

It makes minimal assumptions about the dependency of the outputs on the input variables

Generally speaking, all kernel methods differ from each other in the:

- Choice of f(x)
- Type of regularization to balance complexity and goodness of fit

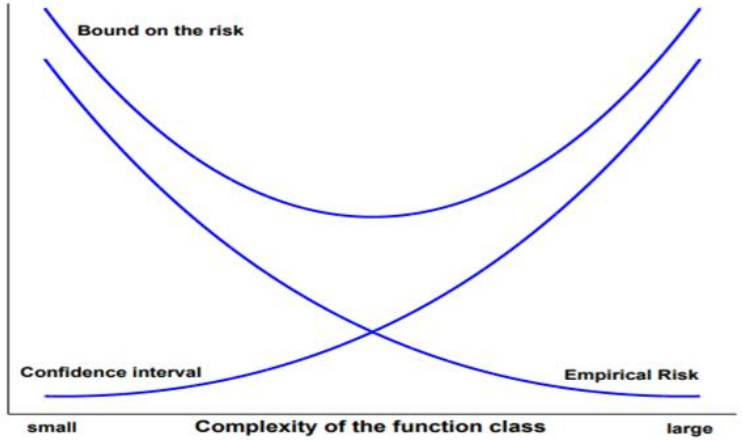


Fig. 2.2: Schematic illustration of bound on the risk as a summation of the empirical risk and of the confidence interval. Increasing the complexity of the function class decreases the empirical risk but upper bound of the risk is increased due to the increase of confidence interval. The smallest bound of the risk is a trade off between the empirical risk and complexity (confidence).

Reproducing Kernel Hilbert Spaces (RKHS)

A Hilbert space is inner product (type of normed vector) space which satisfies completeness (defined distance function)

$$d(x,y) = ||x-y|| = \sqrt{\langle x-y, x-y \rangle}$$

The aim of GP is approximating the true genetic signal (g) as an unknown function of genetic effects

$$g = \{g_i\}$$

The function of genetic effects may be viewed as the average phenotypic value of individuals with genotype x_i without restricting the form of $g(x_i)$

$$g(x_i) = E(y_i | x = x_i)$$

Morota and Gianola (2014)

RKHS

Procedure:

- Search of a function
- Loss function: residual sums of squares
- Penalty: squared norm of **g** under a Hilbert space

Objective function to be minimized with respect to **g**:

$$l(\mathbf{g}|\lambda) = \| y - g \|^2 + \lambda \| \mathbf{g} \|^2_{\mathbf{H}}$$

Penalization parameter

Hilbert space

RKHS- kernel ridge regression

Following Kimeldorf and Wahba (1971), the objective function $g(\mathbf{x})$ reduces to a linear function $K\alpha$ where:

K is an *n* x *n* kernel constructed from the observed data

α is an *n* x *1* vector of regression coefficients to be estimated by minimizing

$$l(\alpha|\lambda) = (\mathbf{y} - \mathbf{K}\alpha)'(\mathbf{y} - \mathbf{K}\alpha) + \boxed{\lambda\alpha'\mathbf{K}\alpha}$$

To regularize the α, to ensure that optimal solutions lie in a finite dimensional space

Minimizing (taking the derivative with respect to α and setting to 0)

$$\hat{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}y \qquad \hat{g} = K\hat{\alpha}$$

RKHS- kernel ridge regression

First introduced in QG by Gianola et al (2006), Gianola and van Kaam (2008) in the context of a Bayesian mixed effects modelling.

Gonzalez-Recio et al. (2008) first applied RKHS to genomic prediction.

de los Campos et al. (2010) developed efficient Gibbs sampling algorithms for RKHS regression

The basic idea underlying the RKHS approach to GS is to use the matrix of markers **X** to build a covariance structure among genetic values

Reproducing Kernel Hilbert Spaces

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\alpha} + \mathbf{e}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{K}_h \\ \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{X} & \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{K}_h + \frac{1}{\lambda^{-1}}\mathbf{K}_h \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\lambda,h} \\ \hat{\boldsymbol{\alpha}}_{\lambda,h} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

It measures the genomic similarity among pairs of individuals

Variance of non-parametric coefficients

Vector of non-parametric coefficients

Kernel

In non-parametric statistics, a kernel is a weighting function with the following characteristics:

- Symmetrical
- Area under the curve of the function must be equal to 1

It handles non-linear relationships between a pair of random variables

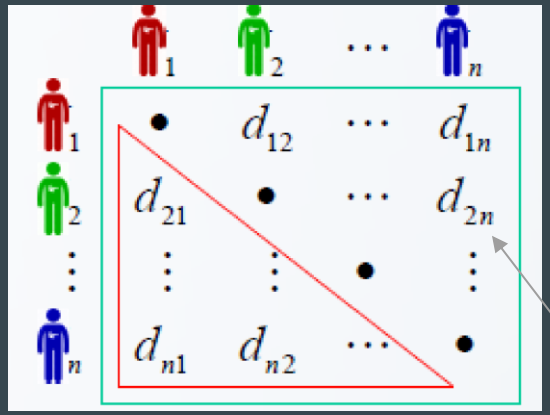
Kernel

Smoothing parameter to regularize the similarity

$$K_h(x_i, x_j) = f(h^{-1} dist(x_i, x_j))$$

Distance between the individuals

Kernel



$$K_h(x_i, x_j) = f(h^{-1} dist(x_i, x_j))$$

RKHS-BLUP

De los Campos et al (2009) brought up the important connection between RKHS regression and BLUP

$$\alpha \sim N(0, A\sigma_\alpha^2)$$

Model:

$$y = X\beta + I\alpha + \epsilon$$

$$\epsilon \sim N(0, I\sigma_\epsilon^2)$$

MME:

$$\begin{bmatrix} X^TX & X^T \\ X & I + A^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} X^Ty \\ y \end{bmatrix}$$

Numerator relationship matrix

RKHS-BLUP

Transforming the additive genetic effects as

$$\mathbf{GEBV} = \hat{\mathbf{u}} = \mathbf{K}\hat{\boldsymbol{\alpha}}$$

BLUP of additive effects can be viewed as a regression on pedigree or on additive genomic relationship kernels (Morota and Gianola, 2014)

RKHS - BLUP

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{K}_h \\ \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{X} & \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{K}_h + \frac{1}{\lambda^{-1}}\mathbf{K}_h \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\lambda,h} \\ \hat{\boldsymbol{\alpha}}_{\lambda,h} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{K}_h'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\lambda^{-1}}\mathbf{K}^{-1}{}_h \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\lambda,h} \\ \hat{\mathbf{u}}_{\lambda,h} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

(1) **(2)**

(1) It is used instead of (2) because inverting K may not be trivial

RKHS-GBLUP

$$y = g + e \; ; \; g \sim N\left(0, \; \boldsymbol{\Gamma}\sigma^2_g\right)$$

True genetic signal

True genomic relationship matrix among individuals (unknown)

g is approximated with a linear function $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

n ind x *m* markers

RKHS-GBLUP

$$\mathbf{V}_y = \mathbf{V}_g + \mathbf{V}_\epsilon$$
$$= \mathbf{X}\mathbf{X}^T\sigma_\beta^2 + \mathbf{I}\sigma_\epsilon^2$$

$$\mathbf{V}_g = \mathbf{X}\mathbf{X}^T\sigma_\beta^2$$

g has to be predicted with 2 conditions:

- $E(\hat{\mathbf{g}})=E(\mathbf{g})=0$
- $\mathrm{var}(\hat{g}_i - g_i)$ is minimum

(co)variance of marker genotypes, considering X is centered

$$BLUP(\hat{\mathbf{g}}) = \left[\mathbf{I} + \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{\sigma_\epsilon^2}{\sigma_\beta^2}\right]^{-1}\mathbf{y}$$

Morota and Gianola (2014)

RKHS-GBLUP

Assuming that gi is parameterized as $g_i = \sum x_{ij}\beta_j$, where x and B are treated as random and independent, then, under HWE:

$$E(x_{ij}) = 2p_j \qquad Var(x_{ij}) = 2p_j(1-p_j)$$

Considering that all markers have the same variance (homogeneous marker variance):

$$\sigma^2_\beta = \frac{\sigma^2_g}{2\sum_{j=1}^{p} p_j(1-p_j)}$$

p_i is the minor allele frequency

RKHS-GBLUP

Substituting $\sigma_\beta^2 = \frac{\sigma_g^2}{2\sum_j p_j(1-p_j)}$ in $BLUP(\hat{\mathbf{g}}) = \left[\mathbf{I} + \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \frac{\sigma_\epsilon^2}{\sigma_\beta^2}\right]^{-1} \mathbf{y}$

Results in:

$$BLUP(\hat{\mathbf{g}}) = \left[\mathbf{I} + \mathbf{G}^{-1} \frac{\sigma_\epsilon^2}{\sigma_g^2}\right]^{-1} \mathbf{y}$$

where $\mathbf{G} = \frac{\mathbf{X}\mathbf{X}^T}{2\sum_j p_j(1-p_j)}$ (VanRaden, 2008)

BLUP estimates of the markers

GBLUP is linked to BLUP of marker regression coefficients

Suppose that the phenotype-genotype mapping function is $\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}$

and the genetic effect is $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$

genotypes

Allele substitution effects

$$BLUP(\boldsymbol{\beta}) = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1} BLUP(\mathbf{g})$$

See Morota and Gianola (2014) for more details

Backsolving of marker effects from GBLUP estimates

BLUP of marker coefficients once $\hat{\mathbf{g}}$ is obtained from GBLUP (backsolving):

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\hat{\mathbf{g}}$$

Ridge regression with markers treated as random effects is mathematically equivalent to BLUP (Ruppert et al, 2003)

However, predictive ability differed when applied to real data (Morota and Gianola, 2014)

Single step (Misztal et al, 2009; Legarra et al, 2009)

It is considered a genomic relationship-based method

Not all animals may be genotyped

Genomic relationships are identical by state (IBS) because they account for the probability that two alleles randomly picked from each individual are identical, independently of origin.

Pedigree relationships are identical by descent (IBD) because they consider that the shared alleles come from the same ancestor in a base population

Single step (Misztal et al, 2009; Legarra et al, 2009)

H matrix is a relationship matrix constructed with SNP markers and pedigree, where the SNP information is projected to the individuals that are not genotyped (subscript 1 for non-genotyped animals, 2 otherwise)

$$\mathrm{H} = \begin{bmatrix} \mathrm{var}\left(\mathrm{u}_2\right) & \mathrm{cov}\left(\mathrm{u}_1, \mathrm{u}_2\right) \\ \mathrm{cov}\left(\mathrm{u}_2, \mathrm{u}_1\right) & \mathrm{var}\left(\mathrm{u}_2\right) \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{A}_{11} + \mathrm{A}_{12}\mathrm{A}_{22}^{-1}\left(\mathrm{G} - \mathrm{A}_{22}\right)\mathrm{A}_{22}^{-1}\mathrm{A}_{22} & \mathrm{A}_{22}\mathrm{A}_{22}^{-1}\mathrm{G} \\ \mathrm{G}\mathrm{A}_{22}^{-1}\mathrm{A}_{21} & \mathrm{G} \end{bmatrix},$$

Single step

Aguilar et al (2010); Christensen and Lund (2010) derived a method to directly construct the inverse of H

H properties include being always semi-positive definite and being positive definite and invertible if G is invertible

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

to avoid double-counting of pedigree information for genotyped animals

Single step

G and A_{22} should be compatible (genomic and genetic bases should be at the same level):

- Compatibility can be understood as both matrices referring to the same genetic base and to the same genetic variance
- genomic relationships can be biased if G is constructed based on allele frequencies other than the ones from the base population (VanRaden, 2008)

Single-step

In a typical breeding program, average elements in A_{22} are > than in G when computed with current allelic frequency (it does not account for past selection)

Possible solutions:

- G should be calculated using the same allelic frequencies as in the base population of the pedigree (difficult)
- Align G and A_{22} (average of the diagonal and off-diagonal elements) (Meyer et al 2018)

Single-step

A^{-1} is often built ignoring the inbreeding, although algorithms to built G^{-1} and A_{22}^{-1} considered it; to avoid convergence problems due to the unbalance between A^{22} and A_{22}^{-1}

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau\mathbf{G}^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix}$$

ω controls inflation due to incompleteness of pedigree (unknown parent groups)

τ controls additive genetic variance

ω 0.7 for beef and dairy cattle ssGBLUP evaluations, from 0.5 to 0.8 for pig evaluations (Lourenco et al, 2020)

Single-step

alfa=0.95,beta= 0.05, tau=1, and omega=0.70 (Lourenco et al, 2014)

it accounts for the fact that genotyped animals are more related through A than G

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix}$$

BLUPf90 family of programs (http://nce.ads.uga.edu/html/projects/programs)

Topics

Background → Non-parametric regressions → Reproducing Kernel Hilbert Spaces → RKHS-BLUP → RKHS-GBLUP → Single -Step