RESEARCH ARTICLE

Lost in translation: On the impact of data coding on penalized regression methods with interactions

Johannes WR Martini[??,??*†], Francisco Rosales[??†], Ngoc-Thuy Ha[??†], Thomas Kneib[??], Johannes Heise[??], Valentin Wimmer[??] and Henner Simianer[??]

*Correspondence:
jmartin2@gwdg.de
[??]University of Goettingen, Department of Animal Breeding and Genetics, Germany
[??]KWS SAAT SE, Einbeck, Germany
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background Penalized regression approaches are standard tools in quantitative genetics. It is known that the fit of an *ordinary least squares* (OLS) regression is independent of certain transformations of the coding of the predictor variables, and that the standard mixed model *ridge regression best linear unbiased prediction* (RRBLUP) is neither affected by translations of the variable coding, nor by global scaling. However, it has been reported that an extended version of this mixed model, which incorporates interactions by products of markers as additional predictor variables, indeed is affected by translations of the marker coding.

Results In this work, we identify the cause of this loss of invariance in a general context of mixed models defined on polynomials in the predictor variables. We show that in most cases, translating the coding of the predictor variables has an impact on penalized regressions, with the exception of the situation in which only the size of the coefficients of monomials of highest degree are penalized. The invariance of RRBLUP can thus be considered as a special case of this setting, that is as a polynomial of degree 1, where the size of the fixed effect (degree 0) is not penalized but all coefficients of monomials of degree 1 are. The extended RRBLUP which includes interactions is not invariant to translations, since it does not only penalize interactions (degree 2), but also additive effects (degree 1). Finally, we investigate the impact of changes of the coding on estimated effect sizes in a pair epistasis model on a publicly available wheat data set.

Conclusion Our results give a general insight into the behavior of penalized regressions. The fact that coding translations alter the estimates of interaction effects, provides an additional reason to interpret the biological meaning of these interactions with caution. Moreover, this problem does not only apply to gene by gene interactions, but also to other types of interactions modeled in mixed models with Hadamard products of covariance matrices (for instance gene by environment interactions).

Keywords: epistasis; extended GBLUP; coding-dependence

Background

Genomic prediction, that is the prediction of properties of individuals from their genetic data, is a crucial ingredient of modern breeding programs [**? ? ? ? ?**]. The traditional quantitative genetics theory is built upon linear models in which allele effects are usually modeled additively [**?**]. In particular, the usual model to represent the effect of the genotype on the phenotype is given by

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of the phenotypic observations of n individuals and $\mathbf{1}_n$ an $n \times 1$ vector with each entry equal to 1. Moreover, μ is the y-intercept, and $\mathbf{M}$ the $n \times p$ matrix describing the marker states of n individuals at p loci. Dealing with single nucleotide polymorphisms (SNPs) and a diploid species, the entries $M_{i,j}$ can for instance be coded as 0 (**aa**), 1 (**aA** or **Aa**) or 2 (**AA**) counting the occurrence of the reference allele **A**. The $p \times 1$ vector $\boldsymbol{\beta}$ represents the allele substitution effects of the p loci, and $\boldsymbol{\epsilon}$ the $n \times 1$ error vector. For single marker regression, which may for instance be used in *genome wide association studies* (GWAS), we can apply ordinary least square regression to determine β. However, in approaches of genomic prediction, we model the effects of many different loci simultaneously and the number of markers p is usually much larger than the number of observations n. To reduce overfitting and to deal with a large number of predictor variables, different methods have been applied in the last decades, among which *ridge regression best linear unbiased prediction* (RRBLUP) is the most popular [**?**]. RRBLUP penalizes the squared ℓ_2 norm of $\boldsymbol{\beta}$ and is built on the additional model specifications of μ being a fixed unknown parameter, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$ with $\mathbf{I}$ the identity matrix. With an approach of maximizing a certain density, these assumptions allow to derive the optimal penalty factor as the ratio of the variance components $\lambda := \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$ [**? ? ?**]. In practice these variance components are usually estimated from the data, albeit the theory to derive the optimal penalty is based on the assumption of σ_β^2 and σ_ϵ^2 being known. Note here that the fixed effect μ is not penalized in RRBLUP, which means that this method is not a pure ridge regression but actually a mixed model in which the size of μ is not penalized but the entries of $\boldsymbol{\beta}$ are. This mixed model RRBLUP is also called *genomic best linear unbiased prediction* (GBLUP) when the model is reformulated with $\mathbf{g} := \mathbf{M}\boldsymbol{\beta}$, and thus $\mathbf{g} \sim \mathcal{N}(0, \sigma_\beta^2 \mathbf{M}'\mathbf{M})$.

It is known that translations of the marker coding, that is subtracting a constant p_i from the i-th column of $\mathbf{M}$ does not change the predictions $\hat{\mathbf{y}}$ of an OLS regression (provided it is well-defined). This invariance also holds for RRBLUP, when the variance components remain unchanged. However, when modeling interactions by products of two predictor variables, that is when fitting the coefficients of a polynomial of degree two to the data, OLS predictions are not affected by translations of the marker coding, but the predictions of its penalized regression analogue *extended genomic best linear unbiased prediction* (EGBLUP) indeed are sensible to a translation of the coding [**? ?**].

In this work we will address the question of why the penalized regression method is affected by translations of the marker coding when a polynomial function of higher degree is used. We will start with a short recapitulation of the different methods.

Theory: Specification of regression methods

In the following we specify the relevant models and regressions to answer the research question previously stated. If an expression includes an inverse of a matrix, we implicitly assume that the matrix is invertible for the respective statement, also if not mentioned explicitly. Analogously, some statements for OLS may implicitly assume that a unique estimate exists, which for instance implicitly restricts to cases of $n > p$ for OLS.

Additive effect regression

The additive effect model has already been presented in Eq. (1)

OLS The ordinary least squares approach is to determine $\boldsymbol{\beta}$ by minimizing the squared residuals:

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{\text{OLS}} := \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{M}_{i,\bullet}\boldsymbol{\beta} - \mu)^2 \tag{2}$$

$\mathbf{M}_{i,\bullet}$ denotes here the i-th row of $\mathbf{M}$, that is the genomic data of individual i. The solution to the minimization problem of Eq. (2) is given by the well-known OLS estimate

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{\text{OLS}} = \left(\begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \mathbf{y} \tag{3}$$

provided that the required inverse exists, which in particular also means that n has to be greater than p.

In problems of statistical genetics, we often deal with a high number of loci and a relatively low number of observations. In this situation of $p+1 > n$, the solution to Eq. (2) is not unique but a vector subspace of which each point minimizes Eq. (2) to zero. Due to this overfit, the quality of predictions $\hat{\mathbf{y}}$ for genotypes which have not been used to estimate the parameter $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ are usually poor. An approach to overcome this problem is RRBLUP.

RRBLUP / GBLUP minimizes

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{\text{RR}_\lambda} := \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{M}\boldsymbol{\beta} - \mu)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

for a penalty factor $\lambda > 0$. Using an approach of maximizing the density of the joint distribution of $(\mathbf{y}, \boldsymbol{\beta})$, the model specifications of $\beta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\beta^2)$ and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ allow to determine the penalty factor as ratio of the variance components as $\lambda := \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$. We stress that Eq. (4) is not a pure ridge regression (RR), as the name RRBLUP might suggest, but a mixed model which treats μ and $\boldsymbol{\beta}$ differently by not penalizing the size of μ. This is the version which is most frequently used in the context of genomic prediction (often with additional fixed effects).

The corresponding solution is given by

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{\text{RR}_\lambda} = \left(\begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix} + \lambda \begin{pmatrix} 0 & \mathbf{0}_p^t \\ \mathbf{0}_p & \mathbf{I}_p \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \mathbf{y}. \tag{5}$$

Here, $\mathbf{0}_p$ denotes the $p \times 1$ vector of zeros and $\mathbf{I}_p$ the p-dimensional identity matrix. The effect of the introduction of the penalization term $\lambda \sum_{i=1}^{p} \beta_i^2$ is that for the minimization of Eq. (4), we have a trade-off between fitting the data optimally and shrinking the square effects to 0. The method will only "decide" to increase the estimate $\hat{\beta}_j$, if the gain from improving the fit is greater than the penalized loss generated by the increase of $\hat{\beta}_j$.

First order epistasis: Polynomials of degree two

An extension of the additive model of Eq. (1) is a first order epistasis model given by a polynomial of degree two in the marker data [**? ? ?**]

$$y_i = \mathbf{1}_n \mu + \mathbf{M}_{i,\bullet} \boldsymbol{\beta} + \sum_{k=1}^{p} \sum_{j=k+1}^{p} h_{j,k} M_{i,j} M_{i,k} + \boldsymbol{\epsilon} \tag{6}$$

OLS Since the model is still linear in the coefficients, Eq. (3) represents the OLS solution, but with a modified matrix $\mathbf{M}$ including the products of markers as additional predictor variables.

eRRBLUP The extended RRBLUP is based on the additional assumption of $h_{j,k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_h^2)$. Also here the solution is given by an analogon of Eq. (5), but with two different penalty factors $\lambda_1 := \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$ and $\lambda_2 := \frac{\sigma_\varepsilon^2}{\sigma_h^2}$.

Translations of the marker coding

In quantitative genetics, often allele frequencies are subtracted from the original 0, 1, 2 coding of $\mathbf{M}$ to use $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}_n \mathbf{P}'$ with $\mathbf{P}$ the vector of column means of $\mathbf{M}$ such that

$$\sum_{i=1,\ldots,n} \tilde{M}_{i,j} = 0 \; \forall j \in \{1, \ldots, p\}.$$

However, also other types of translations, for instance a symmetric $\{-1, 0, 1\}$ coding or a genotype-frequency centered coding [**?**] can be found in quantitative genetics literature. Thus, the question occurs whether this has an impact on the estimates of the marker effects or on the prediction of new genotypes.

The answer is that for the additive setup of Eq. (1), a shift from $\mathbf{M}$ to $\tilde{\mathbf{M}}$ will change $\hat{\mu}$ but not $\hat{\boldsymbol{\beta}}$ and any prediction $\hat{\mathbf{y}}$ will not be affected, neither for OLS, nor for RRBLUP (provided that λ is not changed). This invariance we observe for the additive model does not hold for the extended RRBLUP method.

We will give an example and discuss the effect of translations of the marker coding in a more general way afterwards.

Example 1 (Translations of the marker coding) *Let the marker data of five individuals with two markers be given:*

$$\mathbf{y} = (-0.72, 2.34, 0.08, -0.89, 0.86)^t \qquad \mathbf{M} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ 2 & 0 \\ 2 & 1 \\ 1 & 0 \end{pmatrix}$$

Moreover, let us use the original matrix $\mathbf{M}$*, and the by allele frequencies centered matrix* $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}\underbrace{(1.6, 1)}_{=:\mathbf{P}^t}$*. We consider the first order epistasis model*

$$y_i := \mu + \beta_1 M_{i,1} + \beta_2 M_{i,2} + h_{1,2} M_{i,1} M_{i,2} + \varepsilon_i.$$

Then, we obtain the corresponding estimates based on i) an OLS model, ii) a the mixed model RRBLUP1 of Eq. (5) with $\lambda = 1$*, and iii) a mixed model RRBLUP2 of EQ. (5) with* $\lambda = 1$ *when* β_1 *and* β_2 *are non-penalized. The results are reported in table 1. The columns labeled "centred" correspond to the usage of design matrix* $\mathbf{M}$*, and the columns labeled "uncentered" correspond to the usage of design matrix* $\tilde{\mathbf{M}}$*.*

Table 1: Results from Example 1

	Non-Centred ($\mathbf{M}$)			Centred ($\tilde{\mathbf{M}}$)		
Coeffs.	OLS	RRBLUP1	RRBLUP2	OLS	RRBLUP1	RRBLUP2
μ	1.83	1.81	2.69	0.33	0.33	0.33
β_1	-0.97	-0.89	-1.54	-2.11	-1.15	-2.11
β_2	1.88	0.71	1.03	0.06	0.09	0.11
$h_{1,2}$	-1.14	-0.48	-0.57	-1.14	-0.56	-0.56

We summarize our observations from the reported results as follows:

- *Comparing the two OLS models, the estimated effects* μ*,* β_1 *and* β_2 *change, but the estimated interaction* $\hat{h}_{1,2}$ *as well as* $\hat{\mathbf{y}}$ *remain unchanged.*
- *Comparing the two RRBLUP1 models, both methods give different estimates for all the parameters and the solutions produce different predictions* $\hat{\mathbf{y}}$*.*
- *Comparing the two RRBLUP2 models, both methods give different estimates for* μ*,* β_1 *and* β_2*, but the same for* $h_{1,2}$*, as well as* $\hat{\mathbf{y}}$*.*

The different cases presented in Example 1 have a certain systematic pattern, which we will discuss in the following section.

Results

The observations made in Example 1 are explained by following simple proposition which has several interesting implications.

Proposition 1 *Let* $\mathbf{M}_{i,\bullet}$ *be the* p *vector of the marker values of individual* i *and let* $f(\mathbf{M}_{i,\bullet}) : \mathbb{R}^p \to \mathbb{R}$ *be a polynomial of degree* D *in the marker data. Moreover, let* $\tilde{\mathbf{M}} :=$

$\mathbf{M} - \mathbf{1}\mathbf{P}^t$ *be a translation of the marker coding (as in Example 1) and let us define a polynomial* $\tilde{f}$ *in the translated variables* $\tilde{\mathbf{M}}$ *by* $\tilde{f}(\tilde{\mathbf{M}}_{i,\bullet}) := f(\tilde{\mathbf{M}}_{i,\bullet} + \mathbf{P}^t) = f(\mathbf{M}_{i,\bullet})$*. Then for any data* $\mathbf{y}$*, the sum of squared distances will be identical*

$$\sum_{i=1,\ldots,n} (y_i - f(\mathbf{M}_{i,\bullet}))^2 = \sum_{i=1,\ldots,n} (y_i - \tilde{f}(\tilde{\mathbf{M}}_{i,\bullet}))^2$$

and for any monomial m of degree D, the corresponding coefficient a_m *in* $f(\mathbf{M}_{i,\bullet})$ *and* $\tilde{a}_m$ *in* $\tilde{f}(\tilde{\mathbf{M}}_{i,\bullet})$ *will be identical:*

$$a_m = \tilde{a}_m.$$

□

Proposition 1 has the very simple statement that if we have a certain fit f based on $\mathbf{M}$, and we use the translated marker coding $\tilde{\mathbf{M}}$ in a second regression, the polynomial $\tilde{f}$ will fit the data with the same quadratic distance but also with the same predictions $\hat{\mathbf{y}}$ (due to the definition of $\tilde{f}$). Moreover, the coefficients of highest degree will be the same.

Since OLS is defined only by the minimal quadratic distance this also means that it is invariant to any translation of the coding, provided that the model structure allows the fit to adapt any f to the corresponding $\tilde{f}$ of Proposition 1. To allow this adaption, the possibility to adapt any coefficient of monomials of lower degree is required. We cannot adapt the regression completely if certain coefficients are forced to zero by the model structure. If a coefficient is equal to zero in f, it may be different from zero in $\tilde{f}$. We will illustrate this with an example.

Example 2 (Models without certain terms of intermediate degree) *Let us consider the data* $\mathbf{M}$ *and* $\mathbf{y}$ *of Example 1 but with the assumption that marker* 2 *does not have an additive effect. Then*

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 3.71 \\ -2.098 \\ -0.012 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \tilde{\mu} \\ \tilde{\beta}_1 \\ \tilde{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 0.334 \\ -2.11 \\ -1.162 \end{pmatrix}$$

and also the estimates $\hat{\mathbf{y}}$ *and* $\tilde{\mathbf{y}}$ *are different.*

Example 2 illustrates that “completeness" of the model is required to have the possibility to adapt to translations of the coding. More precisely, for any monomial of degree d, the model has to include all monomials of lower degree with these variables. If this is not the case, the adapted $\tilde{f}$ may not be a valid fit. Given that the model is “complete" in this sense, Proposition 1 has various implications. The following corollary explains the results observed in our examples and some additional properties of penalized regression methods.

Corollary 1 *For all statements it is assumed that penalty factors remain unchanged and that the model is complete in the sense that for any* f*, the corresponding* $\tilde{f}$ *is a valid fit.*

a) For a model of any degree D , the OLS estimates of the coefficients of highest degree as well as the predictions $\hat{\mathbf{y}}$ *are invariant with respect to translations of the marker coding.*

b) For a regression which only penalizes the coefficients of highest degree D, the estimates of the coefficients degree D as well as the predictions $\hat{\mathbf{y}}$ are invariant with respect to translations of the marker coding.

c) In particular, predictions $\hat{\mathbf{y}}$ of RRBLUP are invariant with respect to translations of the marker coding, since we are dealing with a model of degree 1 and a regression that does not penalize the fixed effect (degree 0).

d) An additive least absolute shrinkage and selection operator (LASSO) regression ℓ_1 penalizing the marker effects but not the intercept is invariant to translations of the markers coding.

Corollary 1 a) is a result of the OLS method being defined only by the sum of squares and explains why the OLS estimates $\hat{h}_{1,2}$ and $\tilde{h}_{1,2}$ of Example 1 are identical. Part b) is a results of the following observation: For each f, its corresponding $\tilde{f}$ will have the same sum of squared distances and the same coefficients of highest degree (with the translated marker coding). Thus, it will have the same value for the target function of Eq. (4) which we aim to minimize. Since this is true for any polynomial f, it is in particular true for the solution minimizing the target function. Corollary 1 b) applied to complete models of degree 1 gives the result of RRBLUP being invariant to translations of the marker coding which has previously for instance been proven using the mixed model equations (which is slightly more complicated and less general than the argumentation here). Part d) illustrates that these observations also transfer to other types of penalized regressions, for instance LASSO.

Before, we illustrate the impact on a publicly available data set, we give a small example highlighting cases which are not invariant to translations of the marker coding. We recommend to use the data of Example 1 to validate the statements.

Example 3

a) Pure ridge regression (with penalty on μ) is not invariant to translations.

b) RRBLUP with the fixed effect forced to zero is not invariant to translations of the marker coding.

c) An extended LASSO ℓ_1 penalizing additive effects and interactions is not in general invariant to translations of the coding.

We will now illustrate the practical relevance our observations on a well investigated publicly available wheat data set.

Results on a wheat data set

Data We use the well-investigated wheat data set published by Crossa *et al.* [**?**]. The data set provides the state of 1279 DArT markers of 599 genotyped wheat lines and records on their yield in four different environments. The provided coding of the marker data is a $0, 1$ coding. For more details on the data see Crossa *et al.* [**?**] or the `R` [**?**] package `BGLR` [**?**].

Calculating the interaction effects The assessment of the practical impact of translations of the marker coding on the effect estimates is difficult. Since in practice, the variance components and thus the penalty factors are estimated on the data, the translations of the marker coding may have an additional indirect effect of changing the penalty factors. Moreover, there may be numerical issues changing the interaction effects when a large number of interactions is included. Thus, it seems difficult to separate these superposed effects. For this reason, we decided to restrict our model to the 100 most important markers and their interactions. In more detail, for each environment we perform RRBLUP, chose the 100 markers with highest absolute effect size and built a model with all pairwise interactions between them. Thus, the corresponding eRRBLUP includes the fixed effect μ, 100 additive effects and 4450 interactions. We prefer this approach to an approach of randomly selecting 100 markers, since approaches of restricting interactions to markers with large additive effects can be found in literature (citation). Moreover, we estimated the variance components only for the allele-frequency centered coding and used the penalty factors also for the estimates with other codings. Analogously, the translational invariance of RRBLUP also holds when the penalty factor remains fixed. For the estimation of the variance components, we used the `regress` package [**?**].

We compare three different codings: The original provided $0,1$ coding, a version translated by -0.5, that is a symmetric $-0.5,0.5$ coding, and a coding in which the mean of each column is subtracted. We will refer to these codings later as the *original* coding, the *symmetric* coding and *allele-frequency centered* coding.

For each of the environments, we compare the correlation of the estimated 4450 interaction effects for the three different codings. The results are summarized in Table 2. We see that the estimates are highly correlated, but not identical. In particular, the effects sizes seem to be more similar between the original $0,1$ coding and the ± 0.5 coding than compared to the allele-frequency centered version.

Table 2: Correlation of the estimates of the 4450 interactions with different marker coding. Colors indicate which data from which environment was used. Black: Environment 1; Red: Environment 2; Green: Environment 3; Blue: Environment 4.

		$\hat{\mathbf{h}}_{symm}$		$\hat{\mathbf{h}}_{centered}$
$\hat{\mathbf{h}}_{original}$	0.97	0.96	0.82	0.84
	0.95	0.95	0.80	0.83
$\hat{\mathbf{h}}_{symm}$		– –	0.86	0.87
		– –	0.85	0.88

Discussion

The illustrated problem of the coding having an impact on the estimates of interactions in penalizes regressions is essential for quantitative genetics, where Hadamard products are often used to model interaction such as epistasis or gene by environment interaction. In particular this shows once more that the size of these effects should be interpreted with caution, since a biological meaningfulness is not obvious, which is also reflected by the doubtful biologically mechanistic meaning of variance components [**?**].

Table 2 illustrates that the estimated effect sizes are different when the coding is altered, but also that the correlation of the estimates is relatively high. In particular, the correlation of effect sizes for the original coding and the symmetric coding are at least on the level of 0.95 for the wheat data set and for all environments. The correlation of the estimates of each of these codings and the allele-frequency centered coding drops below 0.90 with a minimum of 0.80. This also fits to previous observations where the predictive ability of different codings was compared and where the allele-frequency centered coding was found to be different from the codings in which each marker was coded identically [**?**]. There, it was also found that the symmetric coding seems to outperform other codings slightly (with respect to predictive ability).

It should be highlighted, that the problem does not seem to be a consequence of non-orthogonality of the predictor variables (marker values and their products), since these problems would not appear in an OLS regression (provided that it exists), where the variables have the same coding.

Finally, note that it was reported that Gaussian Kernel regression [**?**] can be interpreted as a limit of the polynomial regression with increasing degree (and all possible monomials) [**?**]. This raises the question of why the Gaussian Kernel regression is not affected by translations of the marker coding. It may be interesting from a theoretical point of view to reconsider the limit behavior.

Conclusion

We identified the cause of the coding-dependent performance of epistasis effects models. Our results were motivated by ridge regression, but do equally hold for many other types of penalized regressions, for instance for the ℓ_1 penalized LASSO. The fact that the estimated effect sizes depend on the coding in particular underlines again that estimated effect sizes should be treated with caution. Moreover, this problematic of coding is not only present for marker by marker interaction, but for any mixed model in which interactions are modeled by Hadamard products of covariance matrices, in particular also for gene by environment (G x E) models.

Appendix

Proposition 1 The fact that the goodness of fit remains the same results from the definition of the polynomials. To see that the coefficients of monomials of highest degree are identical, choose a monomial $m(M_{l_1}, M_{l_2}, ..., M_{l_D})$ of the loci $l_1, ..., l_D$ of degree D of f. Multiplying the factors of $m(\tilde{M}_{l_1} + P_{l_1}, \tilde{M}_{l_2} + P_{l_2}, ..., \tilde{M}_{l_D} + P_{l_D})$ gives the same monomial $m(\tilde{M}_{l_1}, \tilde{M}_{l_2}, ..., \tilde{M}_{l_D})$ as a summand of highest degree, plus additional monomials of lower degree. Thus, the coefficients of monomials of degree D remain the same. □

Author's contribution

JWRM: Proposed to consider the topic; derived the theoretical results; wrote the manuscript FR, NTH,JH: Verified the results All authors: Discussed the research

Acknowledgements

JWRM thanks KWS SAAT SE for financial support.