Scalable non-parametric pre-screening method for searching higher-order genetic interactions underlying quantitative traits

Juho A. J. Kontio and Mikko J. Sillanpää

Supplement A

1 Comparison between the PH-E model with and without including the random term for correlated responses

The proposed dimension reduction procedure is based on the following model

$$\log((Y_i - Y_j)^{-2}) = \log(\varrho) + \sum_{k=1}^{p}(-\rho_k)\|X_{ik} - X_{jk}\|^{\gamma} + \log(\varepsilon_{i,j}) \qquad (1)$$

with a defect that the pseudo-observations are incorrectly assumed to be mutually independent. However, with a cost of increasing computational complexity, dependencies between the pseudo-observations could be accounted for by additional random effect terms $u_{i,j}$ such that $u_{i,j} \perp\!\!\!\perp \log(\varepsilon_{i,j})$ and

$$\log((Y_i - Y_j)^{-2}) = \log(\varrho) + \sum_{k=1}^{p}(-\rho_k)\|X_{ik} - X_{jk}\|^{\gamma} + u_{i,j} + \log(\varepsilon_{i,j}). \qquad (2)$$

The random effect vector $\mathbf{u} \in \mathbb{R}^{n(n-1)/2}$ is assumed to follow a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{G})$ where $\mathbf{G} \in \mathbb{R}^{n(n-1)/2 \times n(n-1)/2}$ is a known expression covariance matrix between the pseudo-observations. With reference to the methods where genomic relationship matrices are estimated from molecular markers located across the genome (see e.g. VanRaden 2008) the matrix $\mathbf{G}$ could be estimated from the model matrix $\mathbf{Z} \in \mathbb{R}^{n(n-1)/2 \times p}$ consisting of all pseudo-variables $\|X_{i1} - X_{j1}\|^{\gamma}, \ldots, \|X_{ip} - X_{jp}\|^{\gamma}$ such that

$$\hat{\mathbf{G}} = (\mathbf{Z} - \hat{\boldsymbol{z}}_{\text{row}}^{\mathrm{T}}\mathbf{1}_p)(\mathbf{Z} - \hat{\boldsymbol{z}}_{\text{row}}^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}}/(p-1)$$

where $\hat{\boldsymbol{z}}_{\text{row}} \in \mathbb{R}^{n(n-1)/2}$ denotes a vector of the row means of the model matrix $\mathbf{Z}$ and $\mathbf{1}_p \in \mathbb{R}^p$ is a constant vector of ones. In practice the matrix G can be simply calculated for instance in R as cov(t($\mathbf{Z}$)).

However, it can be shown that the results do not differ substantially between the models (1) and (2). Especially in quantitative trait locus analyses multilocus association models have been shown repeatedly to perform well without including any polygenic term to account for residual dependencies in the model (see e.g. Setakis et al. 2006; Pikkuhookana and Sillanpää 2009; Kärkkäinen and Sillanpää 2012; Würschum and Kraft 2015; Toosi et al. 2018). In these studies, multilocus models are assumed to be consisting of q loci with significant effects and $p - q$ loci with negligible effects on the trait. The sum over the effects of these $p - q$ loci is perceived as a finite locus approximation to the polygenic effects (see e.g. Pikkuhookana and Sillanpää 2009; Kärkkäinen and Sillanpää 2012). In other words, the effects of these q loci emulate cumulatively the excluded polygenic component such that the dependencies among individuals are modeled by the loci itself (Habier et al. 2007).

With reference to the finite polygenic approximations we partition the systematic part of the model (1) according to the magnitude of the parameters ρ_k ($k = 1, \dots, p$) such that

$$\sum_{k=1}^{p}(-\rho_k)\|X_{ik}-X_{jk}\|^{\gamma} = \sum_{k=1}^{q}(-\rho_k)\|X_{ik}-X_{jk}\|^{\gamma}+\sum_{k=q+1}^{p}(-\rho_k)\|X_{ik}-X_{jk}\|^{\gamma}, \quad (3)$$

where the first q pseudo-variables are assumed to be significantly associated with the dependent pseudo-variable $\log((Y_i - Y_j)^{-2})$. The random effects $u_{i,j}$ are then approximated by the sum $\sum_{k=q+1}^{p}(-\rho_k)\|X_{ik} - X_{jk}\|^{\gamma}$ in a sense that the expression covariances among the pseudo-observations are taken into account cumulatively by the effects of these $p - q$ pseudo-variables.

The role of the penalty parameter λ: The random effects $u_{i,j}$ are approximated cumulatively by the effects of the negligible $p-q$ pseudo-variables. However, once the Lasso/elastic net approach is used we are actually shrinking the effect sizes of these $p - q$ pseudo-variables towards zero. Larger penalty parameter λ values therefore provide less degrees of freedom to emulate the random effects $u_{i,j}$ causing more differences between the results obtained from the models (1) and (2).

The role of the γ-parameter: Since different γ-parameters in pseudo-variables yield different estimates of the expression covariance matrices $\mathbf{G}$ the similarities between the results obtained from the models (1) and (2) clearly depend on the γ-parameter. It appears (as will be later shown) that the proposed γ-parameter value 0.2 consistently implies highly equivalent results between the models (1) and (2) for both large and small penalty parameter λ values. The aberrations in the model estimates between the models (1) and (2) tend to increase more clearly with the γ-parameter value 1.0 and even more so with the value 2.0 as the penalty parameter λ values increase.

Loosely speaking, this is due to the fact that the pseudo-variables with the γ-parameter value 0.2 are more favourably distributed to approximate normally distributed random effects $u_{i,j}$ than the ones corresponding to the γ-parameter values 1.0 and 2.0. In Figure S1 we have provided a simple example based on the DREAM-challenge dataset described in the simulation section of the article.

Let us denote the pseudo-variables of the PH-E model for each $\gamma \in \{0.2, 1.0, 2.0\}$ by $\mathbf{X}(\gamma)$. To generate "negligible" effects we simulated a known inverse bandwidth parameter vector $\boldsymbol{\rho}$ from the multivariate normal distribution $\mathcal{N}(0, 0.3^2\mathbf{I})$. Then the histograms of $\mathbf{X}(\gamma)\boldsymbol{\rho}$-values as well as the values of the randomly chosen pseudo-variable (among $X_1(\gamma)\ldots, X_p(\gamma)$) are plotted in Figure S1 for each $\gamma \in \{0.2, 1.0, 2.0\}$.

——————— INSERT FIGURE S1 ABOUT HERE ———————

It can be seen that especially the histograms of $\mathbf{X}(\gamma)\boldsymbol{\rho}$-values corresponding to the γ-parameter value 0.2 is quite well normally distributed. However, for the γ-parameter values 1.0 and 2.0 the tails of the histograms of $\mathbf{X}(\gamma)\boldsymbol{\rho}$-values get evidently heavier and longer as can be expected by comparing the histograms of $\mathbf{X}(\gamma)$-values between different γ-parameter values. Since the random effects $u_{i,j}$ are assumed to be normally distributed, the sum $\sum_{k=q+1}^{p}(-\rho_k)\|X_{ik} - X_{jk}\|^{\gamma}$ with $\gamma = 0.2$ can be therefore expected to emulate the random effects $u_{i,j}$ more effectively than with $\gamma = 1.0$ or $\gamma = 2.0$.

1.1 Simulated examples

We provide a comparison between the proposed PH-E model with and without including the random effect term to the model by using the DREAM-challenge gene-expression dataset described in the simulation section of the article. However, we used only the first 120 individuals in both analyses since the estimation process for the model with the random effect term is extremely time- and memory-consuming. One three-way interaction term (dissembled at genes 21, 105, 207 indexes starting from the ACTB gene) was simulated without the main effects such that

$$Y = X_{21}X_{105}X_{207} + \varepsilon, \tag{4}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We simulated a single replicate of the phenotype vector and analyzed the same replicate by the both models through the grid $\{0.2, 1.0, 2.0\}$ of different γ-parameter values. The intercept was set to zero and normal residual variance σ^2 was chosen to be relatively small ($\sigma^2 = 0.1^2$) due to the small sample size.

——————— INSERT FIGURE S2 ABOUT HERE ———————

The model coefficients were estimated with the elastic net estimator using the *glmnet* R-package (Friedman et al. 2010) for the model (1) and with the *ggmix* R-package (Bhatnagar et al. 2019) for the mixed model (2). In both cases we used the α-parameter value of 1/3 in the elastic net estimator for each $\gamma \in \{0.2, 1.0, 2.0\}$. We note that the model estimates for the models (1) and (2) are not perfectly aligned by the same λ-values. We have therefore presented in Figure S2 the regularization paths for the estimated inverse bandwidth parameters against the ℓ_1-norm of the whole inverse bandwidth parameter vector as λ varies. We note that the mixed elastic net algorithm sometimes showed convergence problems for some fixed penalty parameter λ values. The used R-code and the simulated phenotype replicate are available at the supplementary materials B.

It is evident from Figure S2 that the results are overall equivalent between the models (1) and (2) for each $\gamma \in \{0.2, 1.0, 2.0\}$ at least in terms of dimension reduction. In concordance with the finite polygenic approximation studies (e.g. Pikkuhookana and Sillanpää 2009; Kärkkäinen and Sillanpää 2012) we are inclined to believe that also in the case of our PH-E model the benefits of the incorporated random effects $u_{i,j}$ can be emulated by the pseudo-variables itself accurately enough for practical purposes.

More precisely, Figure S3 displays the estimates of the inverse bandwidth parameters ρ_k $k = 1, \dots, p$ for two sets of fixed penalty parameter λ values. The top six panels (A1-3 and B1-3) represent the estimates of the model (2) and the estimates the model (1) (B1-3) for a relative small penalty parameter λ value 0.001 (representing the right edges in the regularization path figures). Different panel columns separate the estimates produced by different kernels parameters: $\gamma = 0.2$ (panels A1 and B1), $\gamma = 1$ (the exponential kernel, panels A2 and B2) and $\gamma = 2$ (the Gaussian kernel, panels A3 and B3). Moreover, since the results start to deviate between the models (1) and (2) as the penalty parameter λ increases (see Figure S2) we have also plotted the estimates of the inverse bandwidth parameters for larger values of λ. In the bottom six panels we chose the penalty parameter λ value separately for each panel such the number of non-zero inverse bandwidth parameters was approximately 60 in each case (blue lines in Figure S3 represent the associated ℓ_1-norm values).

As could be seen already in Figure S2, the top six panels (A1-3 and B1-3) show that the estimates of the inverse bandwidth parameters are almost identical when the penalty parameter λ is small for each $\gamma \in \{0.2, 1.0, 2.0\}$. As the penalty parameter λ values increase, we get less degrees of freedom to emulate the random effects by the pseudo-variables itself and the differences between models (1) and (2) start to expose. It is clear from the bottom six panels (A1-3 and B1-3) that smaller γ-parameter values provide better approximation for the random effects as expected from the results of the first example (Figure S1). Especially, the differences between the results obtained from the models (1) and (2) remained practically equivalent with the γ-parameter value 0.2.

——————— INSERT FIGURE S3 ABOUT HERE ———————

However, while not incorporating the random effect terms into the model does not seem to cause any substantial drawbacks especially with the proposed γ-parameter value 0.2 there are significant differences in estimation times in favor of the model (1). We illustrate the amount of computational alleviation achieved by using the model (1) without included random effect terms with respect to the increasing number of individuals (n) and random variables (p). First, multiple datasets are simulated such that the number of random variables is fixed to 100 and the number of individuals is changed (25, 50, 75, 100 and 125). Subsequently, the number of individuals is fixed to 100 and the number of random variables is changed (125, 250, 500, 750 and 1000). The random variables were simulated independently from each other such that each variable was assigned to follow the standard normal distribution.

——————— INSERT TABLE S1 ABOUT HERE ———————

The estimation times for both methods are listed in Table 1. As can be seen, when the random effects are not included into the model the estimation process was extremely fast with respect to the increasing number of individuals and random variables. Yet, if the random effects are incorporated into the model the estimation time increases relatively quickly as the number of individuals or random variables increases. It therefore seems that the benefits of using the model where the dependencies among pseudo-observations are properly accounted for are insignificant relative to the disadvantages. However, we still suggest to use the model (2) whenever possible (in terms of computational capacity) for better theoretical justification.

2 Proofs for the propositions

Proposition 1 *The powered exponential kernel function $K_\gamma(\boldsymbol{X}_i, \boldsymbol{X}_j; \rho)$ with $0 < \gamma \leq 2$ has the infinite series representation that contains all possible product terms $\prod_{r \in \mathcal{M}_s} \rho_r \phi_{\mathcal{M}_s}(\boldsymbol{X}_i)\phi_{\mathcal{M}_s}(\boldsymbol{X}_j)$ with respect to all possible subsets $\mathcal{M}_s$ of indices $\{k_1, \ldots, k_s\} \subset \{1, \ldots, p\}$ of size s for all $1 \leq s \leq p$, where $\phi_{\mathcal{M}_s} : \mathbb{R}^p \longrightarrow \mathbb{R}$ is a mapping $\phi_{\mathcal{M}_s}(X) = \prod_{k \in \mathcal{M}_s} X_k$.*

Proof. *This representation is a consequence of several series expansions (c.f. Cotter et al. 2011 for $\gamma = 2$) of the powered exponential kernel $K_\gamma(\boldsymbol{X}_i, \boldsymbol{X}_j; \rho)$. Let us begin by re-writing the powered exponential kernel as*

$$\exp\left(-\sum_{k=1}^{p} \rho_k \|X_{ik} - X_{jk}\|^\gamma\right) = \exp\left(-\sum_{k=1}^{p} \rho_k ((X_{ik} - X_{jk})^2)^{\gamma/2}\right) \quad (5)$$

$$= \exp\left(-\sum_{k=1}^{p} \rho_k (X_{ik}^2 - 2Z_{(i,j)k} + (\frac{Z_{(i,j)k}}{X_{ik}})^2)^{\gamma/2}\right) \quad (6)$$

where $Z_{(i,j)k} = X_{ik}X_{jk}$. Now the terms $\rho_k(X_{ik}^2 - 2Z_{(i,j)k} + (\frac{Z_{(i,j)k}}{X_{ik}})^2)^{\gamma/2}$ for each $k = 1, \ldots, p$ can be represented with the Taylor series expansion around $Z_{(i,j)k} = 0$ in the form of

$$\rho_k(X_{ik}^2 - 2Z_{(i,j)k} + (\frac{Z_{(i,j)k}}{X_{ik}})^2)^{\gamma/2} = \sum_{l=1}^{m} Q_{l,\gamma,X_{ik}} \rho_k (Z_{(i,j)k})^l + \mathcal{O}(Z_{(i,j)k}^{m+1}), \quad (7)$$

where the terms $Q_{l,\gamma,X_{ik}}$ depend on γ, l and X_{ik} (only on the ith observation). By using the equation (3), the powered exponential kernel (2) can be written as *the following power series:*

$$K_\gamma(X_i, X_j; \rho) = \sum_{q=0}^{\infty} \frac{1}{q!} \left(-\sum_{k=1}^{p} \sum_{l=1}^{m} Q_{l,\gamma,X_{ik}} \rho_k (Z_{(i,j)k})^l + \mathcal{O}(Z_{(i,j)k}^{m+1}) \right)^q \qquad (8)$$

$$\approx \sum_{q=0}^{\infty} \frac{1}{q!} \left(-\sum_{k=1}^{p} \sum_{l=1}^{m} Q_{l,\gamma,X_{ik}} \rho_k (Z_{(i,j)k})^l \right)^q = 1 - \sum_{k=1}^{p} \sum_{l=1}^{m} Q_{l,\gamma,X_{ik}} \rho_k (Z_{(i,j)k})^l$$

$$+\frac{1}{2} \sum_{k_1,k_2} \sum_{l_1,l_2} Q_{l_1,\gamma,X_{ik_1}} Q_{l_2,\gamma,X_{ik_2}} \rho_{k_1} \rho_{k_2} (Z_{(i,j)k})^{l_1} (Z_{(i,j)r})^{l_2}$$

$$-\frac{1}{3!} \sum_{k_1,k_2,k_3} \sum_{l_1,l_2,l_3} Q_{l_1,\gamma,X_{ik_1}} Q_{l_2,\gamma,X_{ik_2}} Q_{l_3,\gamma,X_{ik_3}} \rho_{k_1} \rho_{k_2} \rho_{k_3} (Z_{(i,j)k_1})^{l_1} (Z_{(i,j)k_2})^{l_2} (Z_{(i,j)k_3})^{l_3}$$

$$+\cdots+\frac{(-1)^s}{s!} \sum_{k_1 \dots k_s} \sum_{l_1 \dots l_s} Q_{l_1,\gamma,X_{ik_1}} \dots Q_{l_s,\gamma,X_{ik_s}} \rho_{k_1} \cdots \rho_{k_s} (Z_{(i,j)k_1})^{l_1} \cdots (Z_{(i,j)k_s})^{l_s},$$

where $1 \leq k_1, \dots, k_s \leq p$ and $1 \leq l_1, \dots, l_s \leq m$ for some $m \in \mathbb{N}$. This representation contains all possible product terms $\prod_{r \in \mathcal{M}_s} Q_{1,\gamma,X_{ir}} \rho_r (Z_{(i,j)r})$ with respect to all possible subsets $\mathcal{M}_s$ of indices $\{k_1, \dots, k_s\} \subset \{1, \dots, p\}$ of size s for all $1 \leq s \leq p$.

By applying a simple linear approximation around a point $(Q_{1,\gamma,X_{ir}}, \rho_r(Z_{(i,j)r})) = (a, 0)$ we have for some constant $C \in \mathbb{R}$ that

$$\prod_{r \in \mathcal{M}_s} Q_{1,\gamma,X_{ir}} \rho_r (Z_{(i,j)r}) \approx C + a \prod_{r \in \mathcal{M}_s} \rho_r (Z_{(i,j)}) \qquad (9)$$

$$= C + a \prod_{r \in \mathcal{M}_s} \rho_r \phi_{\mathcal{M}_s}(\boldsymbol{X}_i) \phi_{\mathcal{M}_s}(\boldsymbol{X}_j) \qquad (10)$$

from which it can be seen that the powered exponential kernel implicitly enumerates all possible product terms $\prod_{r \in \mathcal{M}_s} \rho_r \phi_{\mathcal{M}_s}(\boldsymbol{X}_i) \phi_{\mathcal{M}_s}(\boldsymbol{X}_j)$ with respect to all possible subsets $\mathcal{M}_s$ of indices $\{k_1, \dots, k_s\} \subset \{1, \dots, p\}$ of size s for all $1 \leq s \leq p$, as stated in the proposition.

Proposition 2 *Let us consider a random vector $\boldsymbol{X}_i \in \mathbb{R}^p$, the corresponding phenotype $Y_i \in \mathbb{R}$ and independent and identically distributed copies $\boldsymbol{X}_j \in \mathbb{R}^p$ and $Y_j \in \mathbb{R}$. Then for any Borel-measurable function T we have that*

$$Y_i \perp\!\!\!\perp \prod_{k \in \mathcal{M}_s} X_{ik} \quad \text{if and only if} \quad T(Y_i - Y_j) \perp\!\!\!\perp \phi(\boldsymbol{X}_i) \phi(\boldsymbol{X}_j) \qquad (11)$$

where $\phi \colon \mathbb{R}^p \longrightarrow \mathbb{R}$ is a mapping $\phi(\boldsymbol{X}) = \prod_{k \in \mathcal{M}_s} X_k$ for some fixed subset $\mathcal{M}_s$ of indices $\{l_1, \dots, l_s\} \subset \{1, \dots, p\}$ of size $1 \leq s \leq p$.

Proof. We begin by assuming that the phenotype Y_i is independent on the term $\phi(\mathbf{X}_i) = \prod_{k\in\mathcal{M}_s} X_{ik}$ for some integer $1 \leq s \leq p$ and show that it implies the right-hand side of the proposition. Let us consider an independent copy $\mathbf{Z}' = (Y_j, \phi(\mathbf{X}_j))$ of the random vector $\mathbf{Z} = (Y_i, \phi(\mathbf{X}_i))$ so that $Z_t \perp\!\!\!\perp Z_h^{'}$ for $1 \leq t, h \leq 2$. Let $\mathcal{D}_Y$, $\mathcal{D}_{\phi(\mathbf{X}_i)}$ and $\mathcal{D}_{\phi(\mathbf{X}_j)}$ denote the $\sigma-$algebras generated by the variables Y_i, $\phi(\mathbf{X}_i)$ and $\phi(\mathbf{X}_j)$, respectively. Now we can consider the product $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ as $\mathcal{D}_{\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)}$-measurable where $\mathcal{D}_{\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)}$ is the σ-algebra generated by the $\sigma-$algebras $\mathcal{D}_{\phi(\mathbf{X}_i)}$ and $\mathcal{D}_{\phi(\mathbf{X}_j)}$ since $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ is a Borel-measurable function of $\phi(\mathbf{X}_i)$ and $\phi(\mathbf{X}_j)$.

It can be shown by the Dynkin $\pi-\lambda$ theorem (see *e.g.* Klenke 2014) that the σ-algebras $\mathcal{D}_{\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)}$ and $\mathcal{D}_{Y_i}$ are independent which implies that Y_i (and Y_j) and $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ are independent random variables. This is actually a special case of a more general case (see e.g. Pfeiffer 1990, Theorems 11.3.1. and 11.3.3).

Since $(Y_j, \phi(\mathbf{X}_j))$ is independent and identically distributed copy of $(Y_i, \phi(\mathbf{X}_i))$, it follows by the Kac's theorem (see *e.g.* Itô 1984) that $Y_i - Y_j \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$. Now that for every Borel measurable function $T : \mathcal{R}(Y_i - Y_j) \longrightarrow \mathbb{R}$ (where $\mathcal{R}(Y_i - Y_j)$ denotes the range of $Y_i - Y_j$) the $\sigma-$algebra generated by the mapping $T(Y_i - Y_j)$ is a sub-algebra of the σ-algebra generated by $Y_i - Y_j$ implies that $\sigma(T(Y_i - Y_j)) \perp\!\!\!\perp \mathcal{D}_{\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)}$. Therefore, for any Borel-measurable function $T : \mathcal{R}(Y_i - Y_j) \longrightarrow \mathbb{R}$ we have that

$$Y_i \perp\!\!\!\perp \prod_{k\in\mathcal{M}_s} X_{ik} \Longrightarrow T(Y_i - Y_j) \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j). \tag{12}$$

Let us now assume the independence $(Y_i - Y_j) \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$. This by the Kac's theorem implies that $Y_i \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ since Y_i and Y_j are independent. Since $\phi(\mathbf{X}_i)$, $\phi(\mathbf{X}_j)$, Y_i and Y_j as well as their products are assumed to be non-constant continuous random variables and that $\phi(\mathbf{X}_j)$ is independent on $(\phi(\mathbf{X}_i), Y_i)$ but similarly distributed as $\phi(\mathbf{X}_i)$ it must be that $Y_i \perp\!\!\!\perp \phi(\mathbf{X}_i)$. As an antithesis let us assume that Y_i and $\phi(\mathbf{X}_i)$ are not independent which implies that Y_j and $\phi(\mathbf{X}_j)$ are also dependent. Therefore $Y_i = h(\phi(\mathbf{X}_i)) + \varepsilon_i$ and $Y_j = h(\phi(\mathbf{X}_j)) + \varepsilon_j$ for some Borel-function h and independent error terms ε_i and ε_j.

Since Y_i, Y_j and $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ are independent we have by the Dynkin's theorem that $Y_iY_j \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$. Moreover, since Y_i and $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ are independent we have by the antithesis that

$$h(\phi(\mathbf{X}_i)) + \varepsilon_i \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j). \tag{13}$$

By the Kac's theorem $h(\phi(\mathbf{X}_i))$ must be independent on $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$. Since h is a Borel-function, $h(\phi(\mathbf{X}_i))$ and $h(\phi(\mathbf{X}_j))$ are independent so it follows by the Dynkin's theorem and the independency $Y_iY_j \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ that

$$h(\phi(\mathbf{X}_i))h(\phi(\mathbf{X}_j)) \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j) \tag{14}$$

which can be only true if h is degenerate *i.e.* $P(h(\phi(X_i) = c)) = 1$ for some constant $c \in \mathbb{R}$. As Y_i is a continuous random variable implies contradiction since now $Y_i = h(\phi(\mathbf{X}_i)) + \varepsilon_i = c + \varepsilon_i$ where $c + \varepsilon_i \perp\!\!\!\perp Y_i$. Therefore $Y_i \perp\!\!\!\perp \phi(\mathbf{X}_i)$ by the contradiction and

$$T(Y_i - Y_j) \perp\!\!\!\perp \phi(\mathbf{X}_i)\phi(\mathbf{X}_j) \Longrightarrow Y_i \perp\!\!\!\perp \prod_{k \in \mathcal{M}_s} X_{ik}. \tag{15}$$

Lemma 1 *Let us consider a random vector $\boldsymbol{X}_i \in \mathbb{R}^p$, the corresponding phenotype $Y_i \in \mathbb{R}$ and independent and identically distributed copies $\boldsymbol{X}_j \in \mathbb{R}^p$ and $Y_j \in \mathbb{R}$. Then we have that*

$$\mathrm{Cov}((Y_i - Y_j)^2, \phi(\boldsymbol{X}_i)\phi(\boldsymbol{X}_j)) \neq 0 \quad \text{if} \quad \mathrm{Cov}(Y_i, \prod_{k \in \mathcal{M}_s} X_{ik}) \neq 0.$$

where $\phi\colon \mathbb{R}^p \longrightarrow \mathbb{R}$ is a mapping $\phi(\boldsymbol{X}) = \prod_{k \in \mathcal{M}_s} X_k$ for some fixed subset $\mathcal{M}_s$ of indices $\{l_1, \ldots, l_s\} \subset \{1, \ldots, p\}$ of size $1 \leq s \leq p$. Moreover, the converse is also true under the assumption that all dependencies between Y and $\prod_{k \in \mathcal{M}_s} X_k$ are identifiable by the interaction model (2).

Proof. Let us first assume that $\mathrm{Cov}(Y_i, \prod_{k \in \mathcal{M}_s} X_{ik}) \neq 0$. Since $(Y_j, \phi(\mathbf{X}_j))$ is independent and identically distributed copy of $(Y_i, \phi(\mathbf{X}_i))$ we can write that $\phi(\mathbf{X}_i) = aY_i + \varepsilon_i$ and $\phi(\mathbf{X}_j) = aY_j + \varepsilon_j$ for some non-zero constant $a \in \mathbb{R}$ and independent error terms ε_i and ε_j with $E(\varepsilon_i) = E(\varepsilon_j) = 0$. Now

$$\begin{gathered}
\mathrm{Cov}((Y_i - Y_j)^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) = \mathrm{Cov}((Y_i - Y_j)^2, (aY_i + \varepsilon_i)(aY_j + \varepsilon_j)) \\
= \mathrm{Cov}(Y_i^2 - 2Y_iY_j + Y_i^2, a^2Y_iY_j + aY_i\varepsilon_j + aY_j\varepsilon_i + \varepsilon_i\varepsilon_j) \\
= \mathrm{Cov}(Y_i^2 - 2Y_iY_j + Y_i^2, a^2Y_iY_j + aY_i\varepsilon_j + aY_j\varepsilon_i) = -2a^2\mathrm{Cov}(Y_iY_j, Y_iY_j) \neq 0.
\end{gathered}$$

Let us now assume that $\mathrm{Cov}((Y_i - Y_j)^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) \neq 0$ and that all dependencies between Y and $\prod_{k \in \mathcal{M}_s} X_k$ are identifiable by the interaction model (2). Then

$$\begin{gathered}
\mathrm{Cov}((Y_i - Y_j)^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) = \mathrm{Cov}(Y_i^2 - 2Y_iY_j + Y_j^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) \\
= \mathrm{Cov}(Y_i^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) - 2\mathrm{Cov}(Y_iY_j, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) + \mathrm{Cov}(Y_j^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) \\
= 2\mathrm{Cov}(Y_i^2, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) - 2\mathrm{Cov}(Y_iY_j, \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) \\
= E(Y_i^2\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) - E(Y_i^2)E(\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) - E(Y_iY_j\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)) \\
= E(\phi(\mathbf{X}_j)\left(E(Y_i^2\phi(\mathbf{X}_i)) - E(Y_i^2)E(\phi(\mathbf{X}_i)))\right) - E(Y_i\phi(\mathbf{X}_i))E(Y_j\phi(\mathbf{X}_j)).
\end{gathered}$$

Let us consider an antithesis $\mathrm{Cov}(Y_i, \phi(X_i)) = 0$ which further implies that

$$\begin{aligned}
E(\phi(\mathbf{X}_j))\left(E(Y_i^2\phi(\mathbf{X}_i)) - E(Y_i^2)E(\phi(\mathbf{X}_i)))\right) - E(Y_i\phi(\mathbf{X}_i))E(Y_j\phi(\mathbf{X}_j)) \quad &(16) \\
= E(\phi(\mathbf{X}_j))\left(E(Y_i^2\phi(\mathbf{X}_i)) - E(Y_i^2)E(\phi(\mathbf{X}_i)))\right) \neq 0, \quad &(17)
\end{aligned}$$

if and only if the inner term $E(Y_i^2\phi(\mathbf{X}_i)) - E(Y_i^2)E(\phi(\mathbf{X}_i))$ *i.e.* the covariance $\mathrm{Cov}(Y_i^2, \phi(X_i))$ is non-zero. This contradicts by the additional assumption that all dependencies between Y and $\prod_{k \in \mathcal{M}_s} X_k$ are identifiable by the interaction model (2) so it must be that $\mathrm{Cov}(Y_i, \phi(X_i)) \neq 0$.

Proposition 3 *Let us consider an interaction model (2) with the random vector* $\boldsymbol{X} \in \mathbb{R}^p$ *and the phenotype* $Y \in \mathbb{R}$*. If we have that*

$$\mathrm{Cov}(Y, \prod_{k\in\mathcal{M}_s} X_k) \neq 0 \quad \textit{for some} \quad \mathcal{M}_s = \{l_1, \ldots, l_s\} \subset \{1, \ldots, p\},$$

then the corresponding inverse bandwidth parameters ρ_k $(k \in \mathcal{M}_s)$ *estimated by the PH-E method tend to be non-zero.*

Proof: Let us assume that $\mathrm{Cov}(Y, \prod_{k\in\mathcal{M}_s} X_k)$ is non-zero for some set $\mathcal{M}_s = \{l_1, \ldots, l_s\} \subset \{1, \ldots, p\}$. Then by the lemma (1) the covariance between terms $(Y_i - Y_j)^2$ and $\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ is also non-zero. This implies by the proposition 1 that $\prod_{k\in\mathcal{M}_s\rho_k} \neq 0$ due to which the associated bandwidth parameters $\{\rho_k\}_{k\in\mathcal{M}_s}$ must be also non-zero.

References

[1] Bhatnagar, S. R., Y. Yang, T. Lu, E. Schurr, J. Loredo-Osti, M. Forest, K. Oualkacha, C. M. and Greenwood, 2019 Simultaneous snp selection and adjustment for population structure in high dimensional prediction models. bioRxiv. URL: https://www.biorxiv.org/content/early/2019/07/15/408484

[2] Cotter, A., J. Keshet, and N. Srebro, 2011 Explicit approximations of the Gaussian kernel. arXiv:1109.47603.

[3] Friedman J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33:** 1-22.

[4] Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics. **177:** 2389-2397.

[5] Itô, K., 1984 *An Introduction to Probability Theory.* University Press, Cambridge.

[6] Klenke, A., 2014 *Probability Theory: A Comprehensive Course.* Second edition. Springer-Verlag, London.

[7] Kärkkäinen, H., and M. J. Sillanpää, 2012 Robustness of Bayesian multilocus association models to cryptic relatedness. Ann. Hum. Genet. **76:** 510-523.

[8] Pfeiffer, P. E., 1990 *Probability for Applications.* Springer-Verlag.

[9] Pikkuhookana, P., and M. J. Sillanpää, 2009 Correcting for relatedness in Bayesian models for genomic data association analysis. Heredity. **103:** 223-237.

[10] Setakis, E., H. Stirnadel, and D. J. Balding, 2006 Logistic regression protects against population structure in genetic association studies. Genome Res. **16:** 290-296.

[11] Toosi, A., R. L. Fernando, and J. C. M. Dekkers, 2018 Genome-wide mapping of quantitative trait loci in admixed populations using mixed linear model and Bayesian multiple regression analysis. Genet. Sel. Evol. **50:** 32 https://doi.org/10.1186/s12711-018-0402-1

[12] VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91:** 4414-4423.

[13] Würschum, T., and T. Kraft, 2015 Evaluation of multi-locus models for genome-wide association studies: a case study in sugar beet. Heredity. **114:** 281-290.

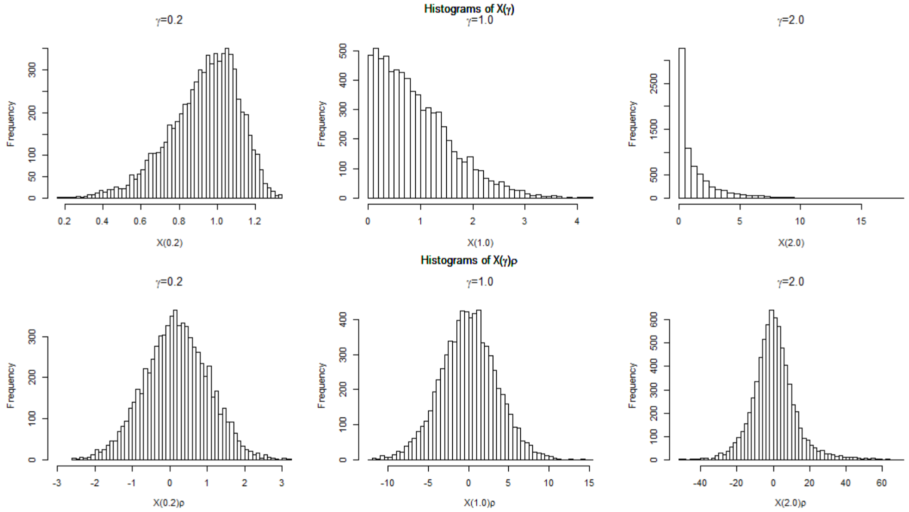

Figure S1: The top histograms represent how the values of the single randomly chosen pseudo-variable among the pseudo-variables $X_1(\gamma)\ldots,X_p(\gamma)$ are distributed for each $\gamma \in \{0.2, 1.0, 2.0\}$. The bottom panels in turn are the histograms of $\mathbf{X}(\gamma)\boldsymbol{\rho}$-values for each $\gamma \in \{0.2, 1.0, 2.0\}$ where the inverse bandwidth parameter vector $\boldsymbol{\rho}$ is generated from the multivariate normal distribution $\mathcal{N}(0, 0.3^2\mathbf{I})$. The panel columns separate the histograms for different kernels parameters.

Table S1: Comparison of computational times in seconds between the PH-E model without the random effect term (Model 1) and with the random effect term (Model 2). The top table represents the computational times with respect to the increasing number (n) of individuals with the the number of random variables fixed to 100. Respectively, the computational times with respect to the increasing number (p) of simulated random variables when the number of individuals is fixed to 100 are presented in the bottom table.

n	25	50	75	100	125
Model 1	0.01	0.01	0.02	0.03	0.05
Model 2	13.05	17.77	52.46	484.35	645.22

p	125	250	500	750	1000
Model 1	0.02	0.05	0.29	0.58	1.21
Model 2	26.49	95.66	2137.30	4718.11	5997.64

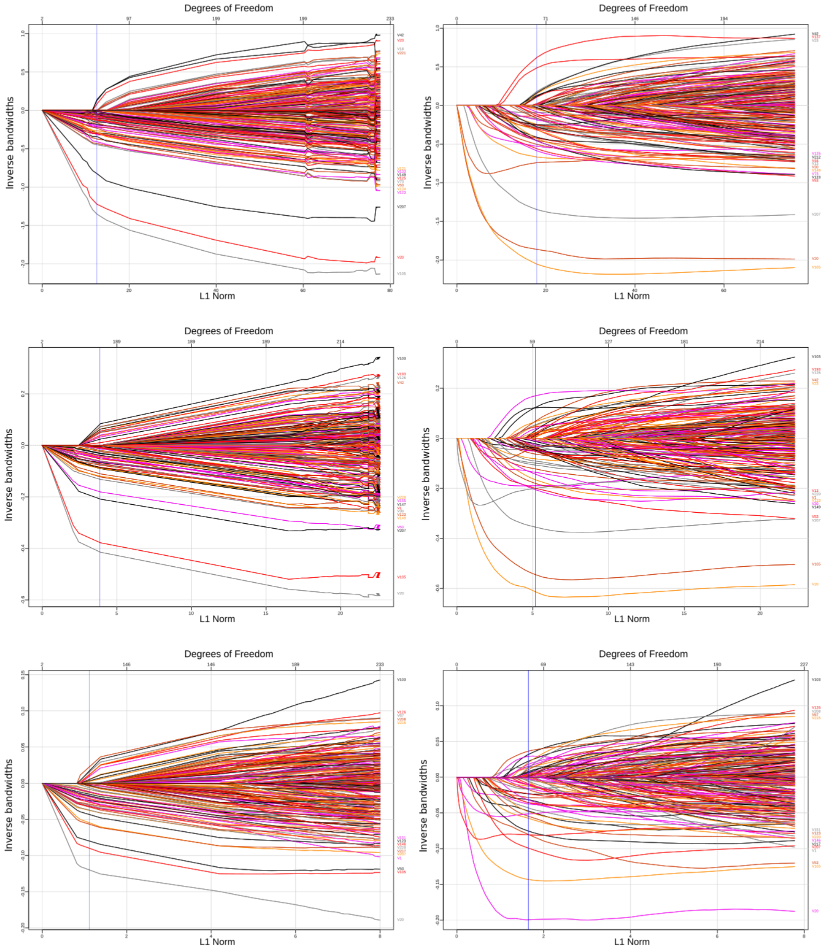

Figure S2: Regularization paths of the estimated bandwidth parameters against the ℓ_1-norm of the whole inverse bandwidth parameter vector as λ varies. Panel columns separate the regularization paths obtained by the PH-E method with (left panel) and without (right panel) including the random effect term. The panel rows represent the regularization paths produced by different kernels parameters: $\gamma = 0.2$ (panels A1 and B1), $\gamma = 1$ (the exponential kernel, panels A2 and B2) and $\gamma = 2$ (the Gaussian kernel, panels A3 and B3). Blue vertical lines denote the ℓ_1-norm values associated with the λ values that produced approximately 60 non-zero inverse bandwidth parameters.

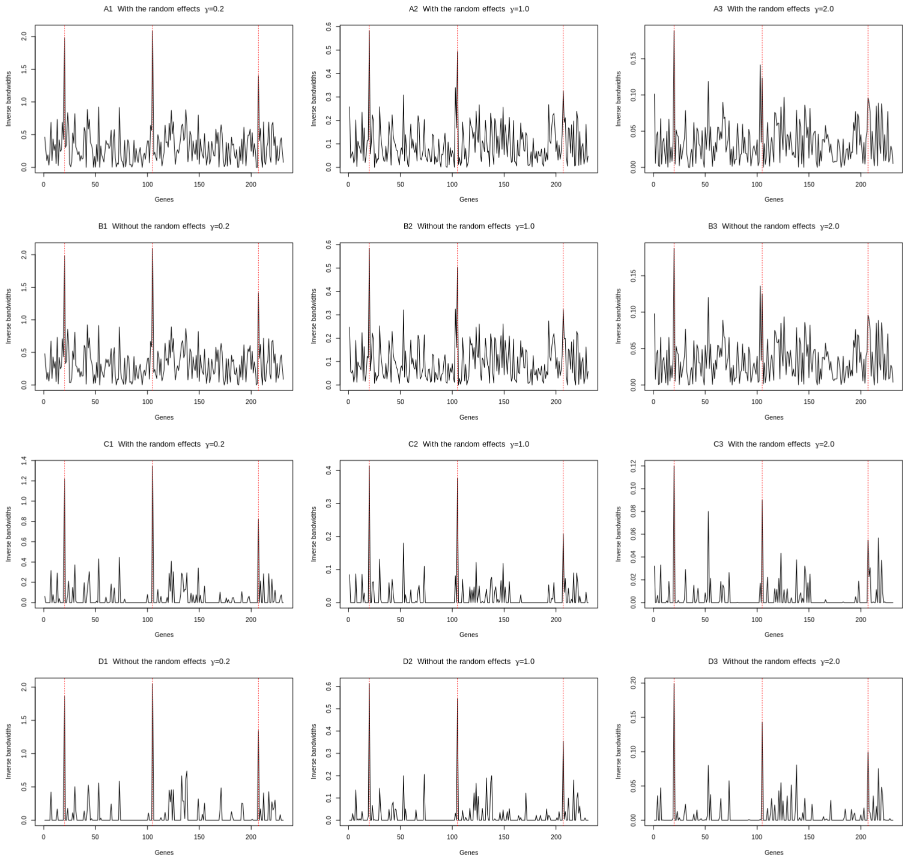

Figure S3: Panels represent the estimates of the inverse bandwidth parameters obtained by the PH-E method with (panels A1-3 and C1-3) and without (panels B1-3 and D1-3) including the random effect term in the model. The panel columns separate the estimates produced by different kernels parameters: $\gamma = 0.2$ (panels A1, B1, C1 and D1), $\gamma = 1$ (the exponential kernel, panels A2, B2, C2 and D2) and $\gamma = 2$ (the Gaussian kernel, panels A3, B3, C3 and D3). The same penalty parameter value $\lambda = 0.001$ was used in panels A1-3 and B1-3. In panels C1-3 and D1-3 the penalty parameter λ values where chosen such that the number of non-zero inverse bandwidth parameters was approximately 60 in each case. Black solid lines denote bandwidth parameter estimates and vertical red lines are the exact places of simulated phenotype associated genes.