Supplementary information for "The benefits of selecting phenotype-specific variants for applications of mixed models in genomics"

Christoph Lippert, Gerald Quon, Eun Yong Kang, Carl M. Kadie, Jennifer Listgarten, and David Heckerman

In the following we demonstrate the well-known fact that, in expectation, misspecification of the genetic similarity matrix leads to reduced predictive accuracy, both in terms of an inflated squared error as well as a lower predictive log-likelihood.

Let $\boldsymbol{y}$ and $\boldsymbol{y}^\star$ denote training data and a test point, respectively. For the linear mixed model, $\boldsymbol{y}$ and $\boldsymbol{y}^\star$ are jointly distributed as

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}^*\end{bmatrix} \sim N\left(\boldsymbol{m};\ \sigma_g^2\begin{bmatrix}\boldsymbol{Z}\\ \boldsymbol{Z}^*\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}\\ \boldsymbol{Z}^*\end{bmatrix}^T + \sigma_e^2\boldsymbol{I}\right),$$

where, for the sake of simplicity and without loss of generality, we suppress the dependence on the number S of variants contained in $\boldsymbol{Z}$.

It follows that

$$\boldsymbol{y}^*|\,\boldsymbol{y} \sim N\left(\underbrace{\boldsymbol{m}+\sigma_g^2\boldsymbol{Z}^\star\boldsymbol{Z}^T\left(\sigma_g^2\boldsymbol{Z}\boldsymbol{Z}^T+\sigma_e^2\boldsymbol{I}\right)^{-1}(\boldsymbol{y}-\boldsymbol{m})}_{\boldsymbol{\mu}_{\boldsymbol{Z}}^\star};\ \underbrace{\sigma_g^2\boldsymbol{Z}^\star\boldsymbol{Z}^{\star T}+\sigma_e^2\boldsymbol{I}-\sigma_g^4\boldsymbol{Z}^\star\boldsymbol{Z}^T\left(\sigma_g^2\boldsymbol{Z}\boldsymbol{Z}^T+\sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{Z}\boldsymbol{Z}^{\star T}}_{\boldsymbol{\Sigma}^\star}\right).$$

The expected value of the squared error under the correct model equals

$$E\left(\sum_{i=1}^{N}\left(y_i^\star-\mu_{\boldsymbol{Z}_i}^\star\right)^2\right)$$

$$= E((\boldsymbol{y}^\star-\boldsymbol{\mu}_{\boldsymbol{Z}}^\star)^T(\boldsymbol{y}^\star-\boldsymbol{\mu}_{\boldsymbol{Z}}^\star))$$

$$= Trace[\boldsymbol{\Sigma}^\star].$$

Let the misspecified joint genomic covariance matrix be $\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{U}^*\end{bmatrix}\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{U}^*\end{bmatrix}^T$. Then the misspecified predictor is

$$\boldsymbol{\mu}_{\boldsymbol{U}}^\star = \boldsymbol{m}+\sigma_g^2\boldsymbol{U}^\star\boldsymbol{U}^T\left(\sigma_g^2\boldsymbol{U}\boldsymbol{U}^T+\sigma_e^2\boldsymbol{I}\right)^{-1}(\boldsymbol{y}-\boldsymbol{m}).$$

It follows that the squared error of the misspecified model equals

$$E\left(\sum_{i=1}^{N}\left(y_i^\star-\mu_{\boldsymbol{U}_i}^\star\right)^2\right)$$

$$= E((\boldsymbol{y}^\star-\boldsymbol{\mu}_{\boldsymbol{U}}^\star)^T(\boldsymbol{y}^\star-\boldsymbol{\mu}_{\boldsymbol{U}}^\star))$$

$$= \underbrace{Trace[\boldsymbol{\Sigma}^{\star}]}_{E\left(\sum_{i=1}^{N}\left(y_i^{\star}-\mu_{Z_i}^{\star}\right)^2\right)} + \underbrace{(\boldsymbol{\mu}_{\boldsymbol{U}}^{\star}-\boldsymbol{\mu}_{\boldsymbol{Z}}^{\star})^T(\boldsymbol{\mu}_{\boldsymbol{U}}^{\star}-\boldsymbol{\mu}_{\boldsymbol{Z}}^{\star})}_{\geq 0}.$$

Here, we used the well-known identity regarding the expectation of a squared form of a normally distributed variable (see, e.g. Section 0.5 in Roweis[1]).

Misspecification also leads to a lower predictive log likelihood. The expected difference between the correctly specified model and the incorrectly specified model is

$$E\left(\underbrace{\ln P(\boldsymbol{y}^{\star}|\boldsymbol{y})}_{correct\ model} - \underbrace{\ln P_{\boldsymbol{U}}(\boldsymbol{y}^{\star}|\boldsymbol{y})}_{incorrect\ model}\right)$$

$$= \int P(\boldsymbol{y}^{\star}|\boldsymbol{y}) \ln \frac{P(\boldsymbol{y}^{\star}|\boldsymbol{y})}{P_U(\boldsymbol{y}^{\star}|\boldsymbol{y})} d\boldsymbol{y}^{\star}$$

$$KL\big(P(\boldsymbol{y}^{\star}|\boldsymbol{y}); P_{\boldsymbol{U}}(\boldsymbol{y}^{\star}|\boldsymbol{y})\big)$$

$$\geq 0,$$

where the KL divergence is zero if and only if the distributions are equal, namely $P_{\boldsymbol{U}}(\boldsymbol{y}^{\star}|\boldsymbol{y}) = P(\boldsymbol{y}^{\star}|\boldsymbol{y})$. It follows that, in expectation, the correctly specified model achieves a higher predictive log likelihood than the misspecified model.

References

1. Roweis, S. *Gaussian identities*. (1999, http://www.cs.nyu.edu/~roweis/notes/gaussid.pdf).