Fast and accurate multi-environment genomic prediction using penalized factorial regression

Willem Kruijer[1] , Emilie Millet[1,2] , Aalt-Jan van Dijk[1,3], Daniela Bustos-Korts[1], Vahe Avagyan[1], Jip Ramakers[1], Martin Boer[1] and Fred van Eeuwijk[1]

[1] Biometris, Wageningen University and Research, Wageningen, The Netherlands

[2] INRA Montpellier, Montpellier, France

[3] Laboratory of Bioinformatics, Wageningen University and Research, Wageningen, The Netherlands

E-mail for correspondence: willem.kruijer@wur.nl

Abstract:

Genomic prediction has become an important tool in applications ranging from genomic selection to personalized medicine. While methodology for univariate genomic prediction is well-established, prediction for new environments remains a notorious challenge, which is however of great interest in plant breeding, where new varieties need to be adapted to a range of increasingly extreme conditions. At least in theory, phenotypes in new environments can be predicted using environmental covariates (ECs) such as temperature, which quantify both tested and untested environments.

Millet et al (2019) recently showed that this is indeed possible using factorial regression models, in which Genotype-by-Environment (GxE) interactions are modelled using genotype-specific sensitivities to ECs. Their approach however relied on 3 predefined ECs, driven by biological knowledge. Such knowledge is however often unavailable, with potentially hundreds of ECs to choose from. Here we consider various ways to penalize factorial regression models, and simultaneously regularize SNP, EC and GxE effects, implicitly assuming different causal models. We show that for the most appropriate of these models, penalized factorial regression leads to accurate predictions for new genotypes in new environments. For simulated and real data with medium to high heritabilities, the within-environment accuracy is on average r = 0.68, outperforming a state-of-the-art deep-learning approach (Khaki and Wang 2019), as well as a popular Bayesian method (Jarquin et al 2014).

Key words: Multi-environment genomic prediction; penalized regression; deep learning;

Jarquin D, Crossa J, Lacaze X et al (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics: 127(3): 595–607.

Khaki S, and Wang L. (2019) Crop Yield Prediction Using Deep Neural Networks. Frontiers in Plant Science : https://doi.org/10.3389/fpls.2019.00621

Millet, EJ, Kruijer, W, Coupel-Ledru, A et al (2019). Genomic prediction of maize yield across European environmental conditions. Nature Genetics: 51: 952-956.