Genomic prediction of longitudinal body weights in pigs using a neural network

Z. Ni[1], R.L Fernando[1], H. Yu[1], E.F. Knol[2], and J.C.M. Dekkers[1]

[1]Iowa State University, Ames, Iowa 50011, USA; [2]Topigs Norsvin, 6640 AA Beuningen, The Netherlands

Abstract

Neural networks have drawn considerable interest for their use in genomic prediction in recent years. As a flexible universal approximator, nonlinearity can be easily accommodated in a neural network, allowing non-additive genetic effects and genotype by environment interactions to be captured. However, SNP effects must be formulated as additive effects for use in genetic improvement. We developed a neural network model for genomic prediction for body weight of pigs. The model consists of two parts: 1) additive effects of SNPs using genomic prediction, and 2) one layer of a feedforward neural network to predict longitudinal body weights based on the SNP effects. By decoupling the prediction of additive SNP effects from the prediction of phenotype, the model is expected to better predict body weight across populations and environments.

Introduction

Genomic prediction has been shown to improve genetic progress by allowing more accurate estimated breeding values (EBV) at an early age (Meuwissen et al., 2001). However, the fundamental assumption of additive models for observed phenotypes, as applied in most genomic prediction models, remains debatable. Over the last decade, there have been multiple attempts to incorporate non-additive genetic effects and genotype by environment interactions (GxE) into genomic prediction. As a universal approximator, neural networks have been shown to improve the accuracy of phenotype prediction in several settings (Waldmann 2018; Gianola et al. 2011). Whether neural networks can be used for genomic prediction based on longitudinal trait data, such as body weight, which is a complex quantitative trait with moderate heritability, deserves investigation. Several studies have attempted to integrate non-linear mechanistic models into genetic evaluation of body weight in pigs (Doeschl-Wilson et al., 2007). These mechanistic models usually require specific assumptions of the relationship between latent variables and phenotypes, which may not be valid across populations and environments. Neural networks don't require such assumptions. Against this background, the objective of this study was to develop an assumption-free neural network model for genomic prediction of longitudinal body weights in pigs.

Materials & Methods

Neural network model for genomic prediction. Figure 1 illustrates the neural network model with 3 neurons that each consist of a pair of latent variables, b_{ni} and a_{ni}, that model the intercept and slope, respectively, of body weight against centered age. The latent variables are linked to the observed longitudinal body weight of pig_n using an one-layer feedforward neural network, with nonlinearity introduced through the activation function Relu (Stursa and Dolezel 2019). Similar to a random regression model (Jamrozik et al., 1997), each b_{ni}/a_{ni} pair is linked to SNP genotypes using an additive whole-genome prediction model, as linear combinations of an intercept (μ_{ani}/μ_{bni}), a fixed contemporary group effect (CG_{bni}/CG_{ani}), random SNP effects for the slope and intercept ($\delta_{jbni}/\delta_{jani}$), and a residual ($\varepsilon_{bni}/\varepsilon_{ani}$). The number of b_{ni}/a_{ni} pairs is a tunable parameter.

Increasing the number of pairs increases the expressive power of the model, but predictive ability may decrease if the size of the dataset is limited. The final predictions of body weight with age are obtained as a weighted sum of the predictions from each b_{ni}/a_{ni} pair, equivalent to an ensemble learning process (Hashem 1997).

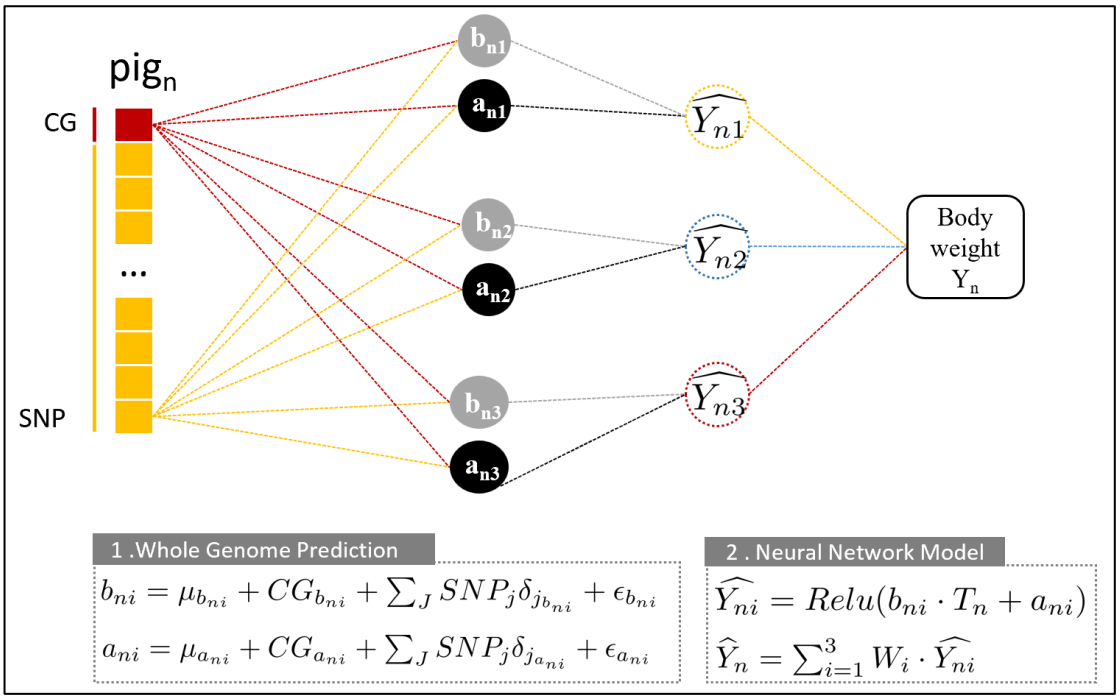

Figure 1. Structure of the neural network model for genomic prediction of body weights based on 3 neurons, each consisting of a pair of latent variables (g and t).

The neural network was built using the Flux.jl package (Innes et al. 2018) and trained for 5,000 iterations using the minibatch gradient descent method (Goodfellow et al., 2016) and the ADAM optimizer (Kingma and Ba, 2014). A mean square error loss function that incorporates the L2 regularization (Friedman et al., 2001) was used, which avoids overfitting, since the number of SNPs is typically much larger than the number of records. The model has 4 hyperparameters: the number of neurons, the L2 regularization coefficient λ, the size of the minibatch, and the learning rate of the ADAM optimizer. The first step in applying the model is tuning the hyperparameters, which was based on a grid search by training the model for a given set of hyperparameters and evaluating its prediction accuracy in validation data to identify the set with the highest accuracy.

Application. The model was applied to a dataset of daily body weight records from 79 to 197 days of age from 3934 purebred boars in 97 contemporary groups. Evaluation of predictions was by forward cross-validation by using the first-born 2000 boars as the training set and the remaining boars as the test set. To avoid information leakage, tuning of the hyperparameters was conducted by forward cross-validation within the 2000 training animals, by fitting the model to the first 1000 of the 2000 training set animals, and computing the accuracy of predictions using data from the second set of 1000 animals. For this purpose, the accuracy of predictions for a given set of hyperparameters was computed for each day of age based on the correlation of predictions with observed body weights adjusted for contemporary group effects for that day of age. Choice of the

best set of hyperparameters was based on the average accuracy across days 103 to 153, which is when most pigs had records. The optimal set of hyperparameters was then used to train the model on the full training set of 2000 and applied to forward-predict the test set. The accuracy of predictions in the test set was computed for each day of age as for tuning. The bias of predictions was computed for each day of age based on the regression coefficient of adjusted body weights on predictions and would be 1 if there's no bias.

Results

The grid search identified the optimal hyperparameters to be 5 neurons, a batch size of 1,000, λ = 240,000, and a learning rate of 1.0 e^{-5}. Example predictions in the test set at example days are in Figure 2, compared to predictions based on a hierarchical Bayesian model that fits the Gompertz function, similar to Cai et al. (2012). The cross-validation accuracy and bias of predictions of body weights for each day are shown in Figure 3. Results from the Bayesian hierarchical model are provided for comparison. The average accuracy across days was slightly higher for the neural network model (0.17) than for the Bayesian hierarchical model (0.16). The neural network model also had a lower bias of 0.39 compared to 0.33 for the Bayesian hierarchical model.

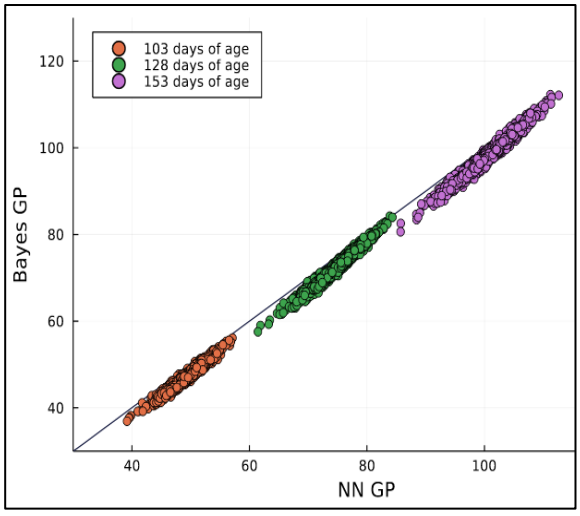

Figure 2. Predictions of body weight (kg) using a Neural Network Model (NN GP) versus the Bayesian method (Bayes GP) at 103, 128, and 153 days of age.

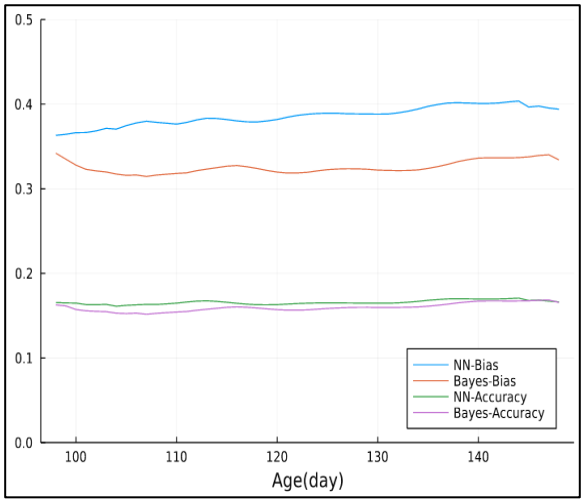

Figure 3. Comparison of bias and accuracy of the Neural Network Model (NN) and the Bayesian method (Bayes).

Discussion

We found that the neural network model obtained slightly higher accuracy and less bias than the Bayesian hierarchical model that was based on the Gompertz function as the underlying biological model. This may be because the Gompertz function models the S-shaped body weight curve over a longer lifespan, which contains much nonlinearity, compared to the fairly linear part of the curve that our data were obtained in. Nevertheless, the results suggest that the neural network can predict longitudinal data well when trained using properly tuned hyperparameters in the framework of random regression. And indeed, Sonoda and Murata (2017) previously proved that a neural network using Relu functions could act as a universal approximator. However, because a neural network model is intrinsically a data-driven method, extrapolating predictions, e.g. to other ages, remains problematic. This is less of an issue for mechanistic models.

Interpretability of a neural network model is also lacking under most circumstances. However, an appealing feature of the neural network model is that its computing time is much shorter than that of the Bayesian hierarchical model: less than 5 minutes on a Geforce Nvidia 3090 GPU given a specific set of hyperparameters, compared to at least 3.5 hours on an Intel Core i7-9800X CPU for the Bayesian model. But the current grid search requires numerous trials over different combinations of hyperparameters, which is time consuming. More sophisticated hyperparameter tuning methods such as heuristic algorithms need to be investigated to reduce tuning time.

In conclusion, the expressive power of a neural network can be incorporated into genomic prediction using a decoupling 2-step design, enabling the model to capture non-additive genetic effects and GxE, hence increasing the accuracy of predictions and reducing their bias. The predictive performance of the neural network model was comparable to a Bayesian hierarchical model that relies on a mechanistic biological model and associated assumptions. The neural network avoids the possible negative effects of improper assumptions of mechanistic models on predictions, especially under specific scenarios, which supports the broader application of predictions across different scenarios.

Acknowledgments
This research was funded by USDA-NIFA grant # 2020-67015-31031.

References
Cai, W., Kaiser, M. S., and Dekkers, J. C. M. (2012) J. Animal Sci. 90(1):127-141. https://doi.org/10.2527/jas.2011-4293
Doeschl-Wilson, A. B., P. W. Knap, B. P. Kinghorn, and H. A. M. Van der Steen. (2007). Animal 1(4): 489–99. https://doi.org/10.1017/S1751731107691848
Friedman, J., Hastie, T., Tibshirani, R., (2001). The elements of statistical learning. Springer Press, New York, USA.
Gianola, D., Okut, H., Weigel, K.A., Rosa, G.J., (2011). BMC Genet 12(87). https://doi.org/10.1186/1471-2156-12-87
Goodfellow, I., Bengio, Y., Courville, A., (2016). Deep Learning. MIT Press, Cambridge, USA.
Hashem S., (1997). Neural Networks.10(4):599-614. https://doi.org/10.1016/S0893-6080(96)00098-6
Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M.C., et al., (2018). Available at: http://arxiv.org/abs/1811.01457
Jamrozik, J., L. R. Schaeffer, and J. C. Dekkers., (1997). J. Dairy Sci. 80(6):1217–26. https://doi.org/10.3168/jds.S0022-0302(97)76050-8
Kingma, D.P., Ba, J., (2014). Available at: http://arxiv.org/abs/1412.6980
Meuwissen, T.H., Hayes, B.J., Goddard, M.E., (2001). Genetics (157):1819–1829. https://doi.org/10.1093/genetics/157.4.1819
Sonoda, S., Murata, N., (2017). Appl. Comput. Harmon. Anal. (43):233–268. https://doi.org/10.1016/j.acha.2015.12.005
Stursa, D., Dolezel, P., 2019. (2019). Proc. of 22nd International Conference on Process Control, Štrbské Pleso, Slovakia.
Waldmann, P., (2018). Genet. Sel. Evol.(50):70. https://doi.org/10.1186/s12711-018-0439-1
Yu, H., van Milgen, J., Knol, E., Fernando, R., and Dekkers, J. (2022) WCGALP 2022