Pretraining Probabilistic Models for Scalable Precision Agriculture

Reviewed on OpenReview:

Abstract

Genomic predictions can help breeders select desirable plant traits, improving crop yields and resilience. However, data collection for developing these prediction models is expensive. Using low-cost auxiliary data that exhibit correlation with desired traits can reduce costs and ultimately improve prediction accuracy. Although such data are abundant, identifying meaningful auxiliary candidates is time-consuming. To this end, we propose a transfer learning mechanism on Gaussian processes to search for potential good candidates via Bayesian optimization. Our results demonstrate promising transferability, paving a new way for efficient searching with a parsimonious sample size.

Keywords: Transfer Learning, Precision Agriculture

1 Introduction

Genomic predictions enable biologists to identify plants with desirable traits early in the growing season, eliminating the need to wait for trait observation. Precise selections can boost crop yield and strengthen resilience to climate change and pest infestations. Collecting data for these predictive models is a costly endeavor, often requiring scientists to bring individual plants into a lab. To reduce costs, high throughput data are used in place of measuring the trait of interest. For example, scientists favor plants with better water absorption. It is known that the degree of water absorption is correlated with the wavelength spectrum of crops (Roberts et al., 2018). Hence, an efficient way to collect data is to fly a drone mounted with a hyperspectral imaging sensor to measure the wavelengths emitted from each crop.

Quantifying the appropriate wavelength spectra that correlate with the desired genetics hinges on two factors (Fernandes et al., 2023). Firstly, understanding the correlation between the wavelength(s) and the desired trait: $r(w, t)$. Secondly, assessing the amount of variance in the wavelength(s) w and the trait t of the crop cohorts attributed to genetics: $h(w)$ and $h(t)$. Unraveling these relationships is non-trivial because they fluctuate depending on the current crop cohort. Each of these factors can be amalgamated into a metric called "co-heritability." To address the expense associated with sorting through various wavelengths, we take Bayesian optimization approach (Azam et al., 2023), which can effectively reduce the number of wavelength(s) considered.

In this paper, we focus on identifying wavelength ratios that have the highest co-heritablity with the desired trait. As shown in Figure 1, we observe that different traits can have similar co-heritability properties. To this end, we can reduce the search cost by transferring the knowledge learned from one set of traits and the corresponding wavelength

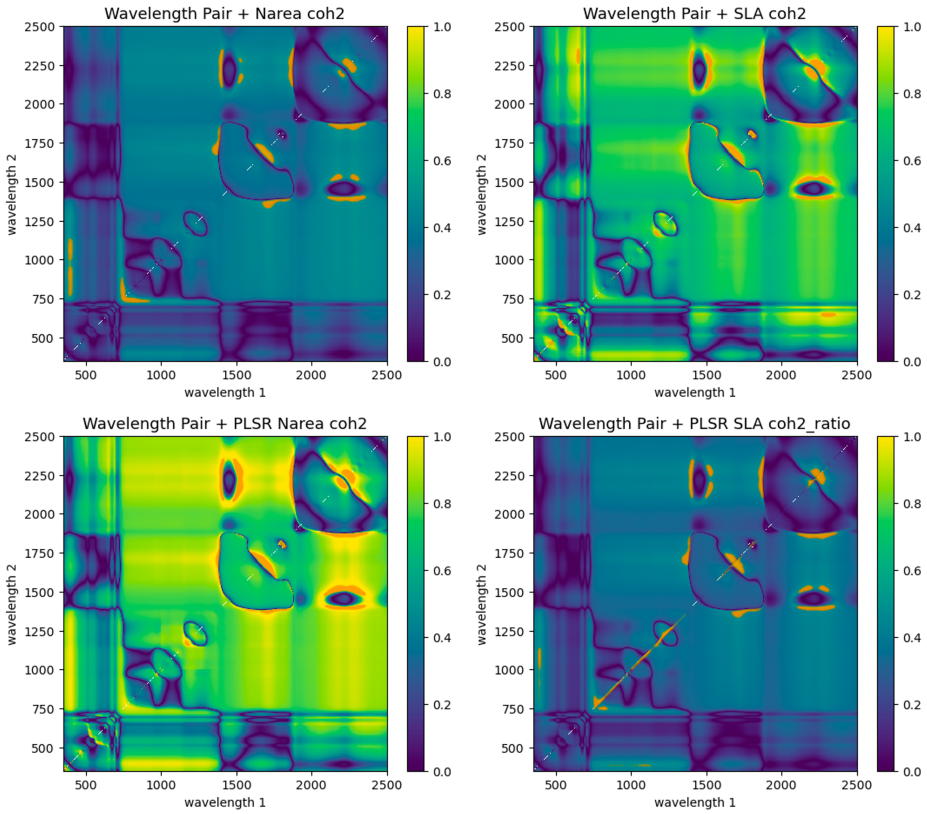

Figure 1: An illustration of the search space of co-heritabilities of nitrogen area (upper left), specific leaf area (upper right), plsr nitrogren area (lower-left), and plsr specific leaf area (lower-right) for each wavelength ratio $\frac{w_1}{w_2}$. The orange highlights the top-1% of the search space in the crop Sorghum.

spectrum to another set. This work presents preliminary results showing that transfer learning of good wavelength ratios is possible for different phenotypes. Additionally, we address the issue of scalability of such models.

2 Preliminary

We begin with formally introducing the concept of co-heritability, Bayesian optimization, and transfer learning.

Co-heritability Heritability $\mathbf{h} : \mathcal{T} \to \mathbb{R}$ is a function that measures the portion of the population variance of a trait $t \in \mathcal{T}$ explained by genetic factors (Hill and Mackay, 2004), opposed to environmental factors. Traits with high heritability can have their genetic components predicted with greater reliability, this measure is desirable for plant biologists during genomic selection. Co-heritability $\mathbf{f} : \mathcal{W} \times \mathcal{T} \to \mathbb{R}$ is a measure that combines the heritability of two traits and their Pearson correlation r between the two traits (Janssens, 1979; Fernandes et al., 2023):

$$\mathbf{f}(w,t) = \mathbf{h}(w) \times \mathbf{h}(t) \times \sqrt{r(w,t)}. \tag{1}$$

For the context of this paper, we fix the desired trait and vary ratios of hyperspectral reflectance, namely the ratio of the wavelength.

Bayesian Optimization (BO) BO is a technique for global optimization of expensive, black-box functions. The goal of such methods is to solve the following optimization problem:

$$x^* = \arg\max_{x \in X} f(x). \tag{2}$$

BO utilizes a probabilistic surrogate model to estimate f, then adaptively select the next data point, ensuring that we extract the maximum information from each experiment.

In this work, we wish to find the spectra with the optimal co-heritability. The derivative over the co-heritability function is non-trivial. Furthermore, the computation of each co-heritability value demands ~ 0.2 seconds, a potentially sluggish process contingent upon the granularity and scale of the hyperspectral data. Given these challenges, treating co-heritability as a black-box within a BO framework becomes a pragmatic approach.

Transfer Learning Transfer learning is a technique where a model trained on one task is repurposed or fine-tuned for another related task. It is particularly helpful if data is expensive or scarce, as it leverages knowledge gained from previous task(s) to improve performance on the target task. From Figure 1, we observe the similarities between the co-heritabilities for two target phenotypes. Although the peaks (in orange) do not exactly line up, the search spaces show similarity in shape. We wish to leverage such similarities.

3 Experiment

In this section, we attempt to train a multi-task model for predicting co-heritability of wavelength ratio jointly on four different tasks. This experiment will prove that positive transfer is possible in predicting co-heritability of wavelength ratios.

Our dataset comprises 869 Sorghum Lines from two growouts near the University of Illinois Urbana-Champaign. In our experiment, we consider four target traits 1) Nitrogren Area (narea) 2) Specific Leaf area (sla) 3) PLSR Nitrogren Area (pn) 4) PLSR Specific Lead area (ps). The auxiliary trait consists of wavelength ratios (w_1/w_2) spectrography ranging from 350nm–2500nm. Using these traits, co-heritability can be calculated between the target and wavelength ratios using a Linear Mixed Effects (LMER) model using the lme4 package in R.

In our experiment, we perform (vanilla) Gaussian process regression on 500–3000 data points. At every step, points are uniformly sampled from each task. The validation set comprises 1000 uniformly sampled points, distinct from the training set. The Gaussian process employs a mixed kernel. This kernel combines Matèrn-kernel(s) for the continuous variable (w_i), i.e., wavelength pairs and a categorical kernel(s) for discrete variables (t_i), i.e., the trait:

$$K((\mathbf{w}_1, \mathbf{t}_1), (\mathbf{w}_2, \mathbf{t}_2)) = K_{\text{mat}_1}(\mathbf{w}_1, \mathbf{w}_2) + K_{\text{cat}_1}(\mathbf{t}_1, \mathbf{t}_2) + K_{\text{mat}_2}(\mathbf{w}_1, \mathbf{w}_2) \cdot K_{\text{cat}_2}(\mathbf{t}_1, \mathbf{t}_2). \tag{3}$$

Each kernel consists of one or more hyperparameters. These hyperparameters are tuned by minimizing the marginal log-likelihood using the L-BFGS-B minimizer in sciPy.

In Figure 2, with 3000 data points, a consistent trend emerges: the validation loss steadily decreases across all tasks as the dataset size grows for each task. This reduction in mean squared error indicates that, on average, the training data from various tasks jointly improve the performance of all tasks. Furthermore, as depicted in Figure 3, we present the Gaussian Process model acquired after training on $N = 3000$ points. The resemblance of these plots to those in Figure 1 provides assurance that a robust Gaussian Process model is indeed being cultivated.

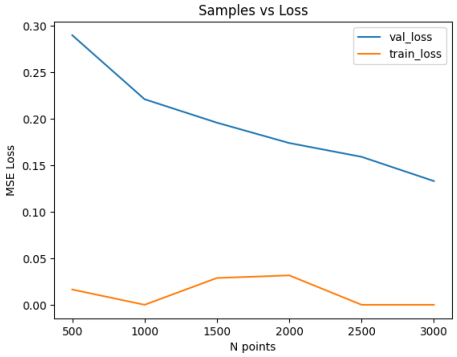

Figure 2: Vanilla GP trained on four target traits with increasing training points.

4 Conclusion & Future Work

We perform inference to find the posterior predictive distribution, conditioned on data from four different phenotypes. Our preliminary results suggest positive transfer when learning co-heritabilities of different target traits with hyperspectral traits. However, fully confirming this transfer effect necessitates a more careful investigation of the transfer between specific phenotypes. Subsequently, we wish to extend this testing to more phenotype(s) and more crops to further understand the utility and practicality of such methods.

Our current tests are on 3000 points. The challenge lies in scaling up this transfer effect to the potentially large-scale dataset, which contains up to 3 million data points. We plan to leverage a stochastic gradient method for the Gaussian process method (SDD)(Lin et al., 2023) to expedite the fitting procedure. This method is provably efficient and estimates the kernel regression parameters with the stochastic gradient descent by optimizing the dual objective. SDD requires domain knowledge to select the kernel's hyperparameter. To ensure the applicability of these hyperparameters across all phenotypes, we presume that an optimal hyperparameter obtained from Maximum Likelihood Estimation (MLE) remains stable with an increase of training data. Our proposed methodology involves randomly selecting a batch of 1000 points for each phenotype, and training a vanilla GP on a total of 4000 data points to ascertain the hyperparameters. We plan to run these experiments in the future.

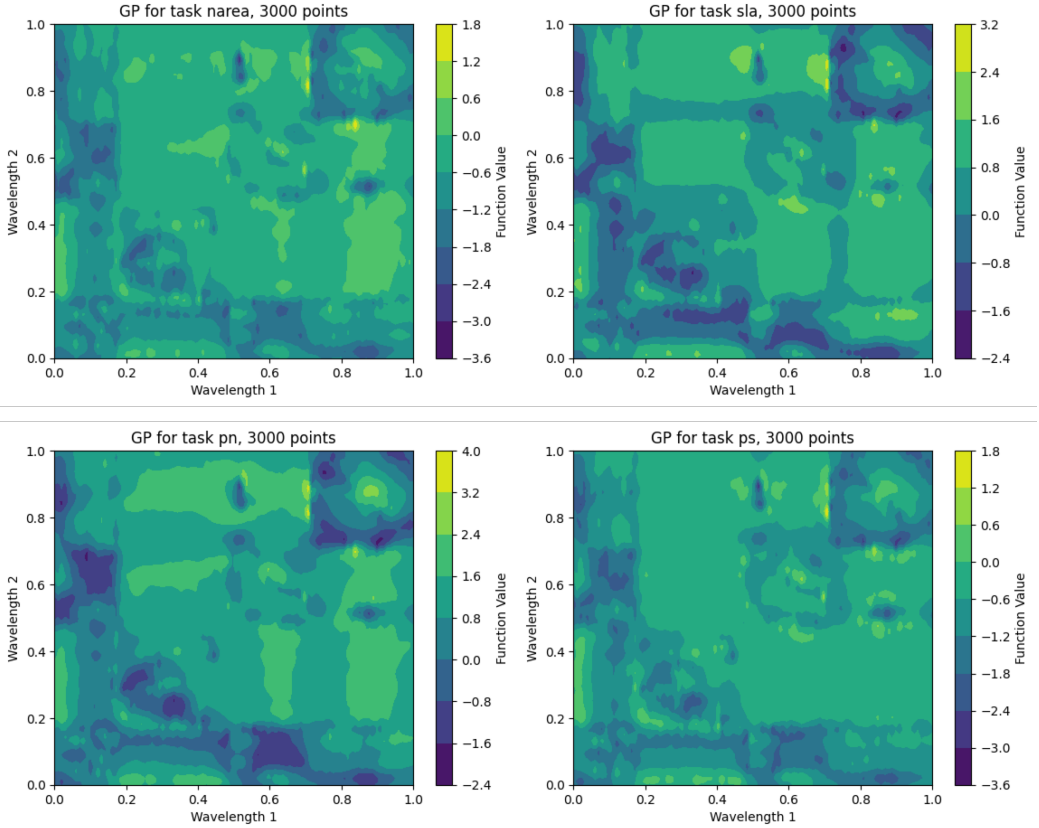

Figure 3: Multitask Gaussian process posteriors after training on $N = 3000$ points.

The next stage of this project aims at minimizing the number of wavelengths for which co-heritability is calculated. To accomplish this objective, we can employ our multi-task Gaussian process as a surrogate function in the Bayesian optimization process.

References

R. Azam, S. Koyejo, S. B. Fernandes, M. Kebir, A. Leakey, and A. Lipka. Rethinking bayesian optimization with gaussian processes: Insights from hyperspectral trait search. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL `https://openreview.net/forum?id=4ePJjCP14u`.

S. B. Fernandes, R. Azam, R. E. Paul, M. Yuan, M. El-Kebir, S. Koyejo, A. E. Lipka, and A. Leakey. Including high-throughput phenotyping derived traits in multi-trait genomic analysis. Presented at the CSSA: Translational Genomics Workshop, Plant and Animal Genome XXI Conference., 2023.

W. G. Hill and T. F. Mackay. Ds falconer and introduction to quantitative genetics. *Genetics*, 167(4):1529–1536, 2004.

M. Janssens. Co-heritability: its relation to correlated response, linkage, and pleiotropy in cases of polygenic inheritance. *Euphytica*, 28(3):601–608, 1979.

J. A. Lin, S. Padhy, J. Antorán, A. Tripp, A. Terenin, C. Szepesvári, J. M. Hernández-Lobato, and D. Janz. Stochastic gradient descent for gaussian processes done right, 2023.

D. A. Roberts, K. L. Roth, E. B. Wetherley, S. K. Meerdink, and R. L. Perroy. Hyperspectral vegetation indices. In *Hyperspectral indices and image classifications for agriculture and vegetation*, pages 3–26. CRC press, 2018.