Bayesian hierarchical approach to longitudinal high-throughput plant phenotyping

Jeffrey Berry[1,1,1,1], Josh Sumner[2,2,2,2], and Noah Fahlgren[1,1,1,1]

[1]Donald Danforth Plant Science Center
[2]Washington University School of Medicine

November 30, 2022

Abstract

High-throughput plant phenotyping is increasingly implemented in a wide array of experimentation and presents challenges both logistically and analytically. Phenotype data are often longitudinal and proper modeling of plant growth requires sophisticated modeling techniques to account for the intra-plant correlations and changing variation over time (heteroskedasticity). For this reason, plant growth is often analyzed by comparing only single time points or the start and end points for inference with no regard for the trends themselves. Single time point analysis can be sufficient for simple biological comparisons, but modeling has the potential to unlock additional insights by utilizing all the information at hand. Current plant growth modeling strategies do account for intra-plant correlations but are still limited to constant variance assumptions and therefore perform sub-optimally. Here we propose a Bayesian hierarchical approach as an alternative method for plant growth modeling by demonstrating the utility of heteroskedastic sub-model parameterizations. We show that accounting for heteroskedasticity greatly improves model accuracy and subsequent inference. Additionally, Bayesian methodologies inherently lend themselves to near real-time model updating and we propose integration with Clowder to facilitate adaptive experimental designs. We show by example the utility of Bayesian updating and how it relates to experimental decision making.

Bayesian hierarchical approach to longitudinal high-throughput plant phenotyping

Josh Sumner[a,b], Noah Fahlgren[a], and Jeffrey C. Berry[a]

[a]Donald Danforth Plant Science Center, 975 N Warson Rd, St. Louis MO, USA 63132
[b]Washington University School of Medicine, 660 S Euclid Ave, St. Louis MO, USA 63110

ABSTRACT

High-throughput plant phenotyping is increasingly implemented in a wide array of experimentation and presents challenges both logistically and analytically. Phenotype data are often longitudinal and proper modeling of plant growth requires sophisticated modeling techniques to account for the intra-plant correlations and changing variation over time (heteroskedasticity). For this reason, plant growth is often analyzed by comparing only single time points or the start and end points for inference with no regard for the trends themselves. Single time point analysis can be sufficient for simple biological comparisons, but modeling has the potential to unlock additional insights by utilizing all the information at hand. Current plant growth modeling strategies do account for intra-plant correlations but are still limited to constant variance assumptions and therefore perform sub-optimally. Here we propose a Bayesian hierarchical approach as an alternative method for plant growth modeling by demonstrating the utility of heteroskedastic sub-model parameterizations. We show that accounting for heteroskedasticity greatly improves model accuracy and subsequent inference. Additionally, Bayesian methodologies inherently lend themselves to near real-time model updating and we propose integration with Clowder to facilitate adaptive experimental designs. We show by example the utility of Bayesian updating and how it relates to experimental decision making.

Keywords: Bayesian, Growth Modeling, Heteroskedasticity, Clowder, High-throughput Phenotyping, brms

1. INTRODUCTION

New technologies have propelled advancements in every facet of the life sciences. High-throughput image-based phenotyping is one advancement that has the potential to accelerate progress in plant science and agriculture, in basic research, breeding, and production.[1] High-throughput image-based phenotyping is paving the way for significant advancements by producing massive amounts of highly quantifiable data.[2] Amongst a large number of challenges this presents, one challenge is statistical modeling of highly multivariate and longitudinal data that often results from these technologies. Plant growth modeling strategies have successfully shown differences in plant growth under stress conditions and recovery by parameterizing growth in different ways and comparing coefficients across experimental conditions.[3–8]

Plant growth modeling requires sophisticated approaches to handle the correlation structure that accompanies repeated measures data. This often is handled by mixed-effect modeling using a random slope and intercept term for each individual which accounts for individual variability on the population-level estimations. The major challenge with plant growth modeling is that plant growth patterns are often non-linear. Common non-linear parameterizations include 3- and 4-component logistic and Gompertz growth models.[9] Moreover, random effects are underestimated when modeled as random slopes and intercepts due to the non-linear growth patterns. Alternatively, modeling can be avoided by limiting analysis to select days where the experimental design effects are the largest to use for statistical inference, which is sufficient in many circumstances but misses the opportunity to detect additional insights and ultimately undermines the potential of these technologies.

Further author information: (Send correspondence to J. C. Berry)
J. Sumner: E-mail: jsumner@danforthcenter.org, Telephone: (314) 587-1000, ORCiD: 0000-0002-3399-9063
N. Fahlgren: E-mail: nfahlgren@danforthcenter.org, Telephone: (314) 587-1000
J. C. Berry: E-mail: jberry@danforthcenter.org, Telephone: (314) 587-1000

Here, we investigated the use of Bayesian hierarchical models (BHMs) in both simulated and real high-throughput image-based phenotyping datasets from the Bellwether Phenotyping Facility at Donald Danforth Plant Science Center. Using a Bayesian approach, prior distributions naturally reflect the uncertainly in estimates before the experiment begins, and the posterior predictive distributions reflect the same uncertainly at the end. BHMs provide a natural environment to include any parameterization of population-level and family-specific effects, including heteroskedastic sub-models, by specifying prior distributions on all parameters and estimating the posteriors using Monte-Carlo Markov Chain (MCMC) computational methods. This work shows how different sub-models of heteroskedasticity affect the posterior predictions and subsequent inference. Additionally, this work explores Bayesian updating as a utility function of high-throughput phenotyping that can provide near real-time decision making in an ongoing experiment.

2. METHODS

2.1 Software and Data Availability

All analyses were conducted in R,[10] models were created and implemented using R package `brms`,[11] and graphics were created using `ggplot2`[12] and `patchwork`.[13] All code and data to recreate figures and results can be found at our GitHub repository[14] that has GPLv3 licensing.

2.2 Creating Simulated Data

3-parameter logistic growth data was simulated for plant size (area) in two treatments (a and b) each containing twenty individuals over the course of 25 days. The logistic growth model, the three parameters (asymptote, inflection point, and inflection rate) for each treatment with their respective simulated noise, and graphical representation are as follows:

3-Parameter Logistic Growth
Area $\sim \phi_1/(1 + e^{(\phi_2 - \text{Time})/\phi_3})$

Treatment a
Asymptote $:= \phi_1 \sim N(\mu = 200, \sigma^2 = 25)$
Inflection Point $:= \phi_2 \sim N(\mu = 13, \sigma^2 = 1)$
Inflection Rate $:= \phi_3 \sim N(\mu = 3, \sigma^2 = 0.2)$

Treatment b
Asymptote $:= \phi_1 \sim N(\mu = 160, \sigma^2 = 25)$
Inflection Point $:= \phi_2 \sim N(\mu = 13, \sigma^2 = 1)$
Inflection Rate $:= \phi_3 \sim N(\mu = 3.5, \sigma^2 = 0.2)$

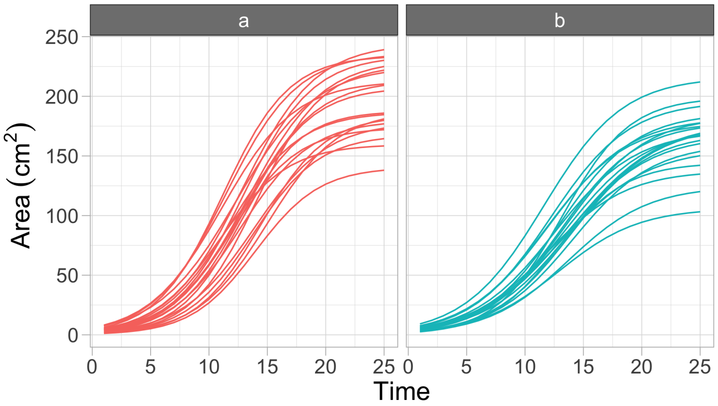

Figure 1: 3-Parameter logistic growth simulation and trend lines.

2.3 Source of Real High-throughput Phenotyping Data

Data was generated using the Bellwether Phenotyping Facility at Donald Danforth Plant Science Center.[15] In short, the experiment was on a single variety of sorghum designed to test the interaction of added microbes and different watering regimes on overall plant growth and development. The dataset was subset to include only the drought treatment and only two microbe inoculates (SynCom A and SynCom B). Full details about the experiment that produced the dataset, including outlier removal, were described previously.[15]

2.4 Bayesian Hierarchical Models

Robust estimation of the logistic growth model parameters were done using the heavy-tailed t-distribution.[16] Priors for ϕ_1, ϕ_2, ϕ_3, and ν are all weak and strictly positive and individual variation was accounted for with auto-regressive moving average (ARMA) covariance. Complete parameterizations of the considered heteroskedastic sub-models and their respective weak t-distributed hyperpriors are shown here in tabular form. All modeling is performed using the following logistic growth model, priors, heteroskedastic sub-models, and hyperpriors:

Plant Growth Model: $\forall i \in Treatment_i$

$$\text{Area} \sim T(\phi_{1,i}/(1+e^{(\phi_{2,i}-\text{Time})/\phi_{3,i}}) + Z_{k,i},\ \sigma+\sigma_i,\ \nu)$$

$$Z_{k,i} \sim ARMA_k(1,1)\ \forall k \in \{1,2,3,...,20\}$$

Priors: $\forall i \in Treatment_i$

$\phi_{1,i} \sim Lognormal(\mu = 130, \epsilon = 2.5)$
$\phi_{2,i} \sim Lognormal(\mu = 12, \epsilon = 2.5)$
$\phi_{3,i} \sim Lognormal(\mu = 3, \epsilon = 2.5)$
$\sigma \sim T(\mu = 0, \epsilon = 3, \omega = 5)$
$\nu \sim Gamma(\alpha = 2, \beta = 0.1)$

Heteroskedastic Sub-models: $\forall i \in Treatment_i$

Homoskedastic	$\sigma_i \sim \beta_0$
Linear	$\sigma_i \sim \beta_0 + \beta_1 time + \beta_{2,i} treatment_i : time$
Exponential	$\sigma_i \sim \beta_0 + \alpha * e^{\beta_1 time} + \beta_{2,i} treatment_i : time$
Splines	$\sigma_i \sim \frac{1}{n}\sum_{j=1}^{n}(Y_j - f(x_j)_i)^2 + \lambda \int (f''(x)_i)^2 dx$
Quadratic	$\sigma_i \sim \beta_0 + \beta_1 time + \beta_2 time^2 + \beta_{3,i} treatment_i : time + \beta_{4,i} treatment_i : time^2$

Hyperpriors: $\forall i \in Treatment_i$

$\forall \beta_* \in \sigma_i(\beta_*) \sim T(\mu = 0, \epsilon = 3, \omega = 5)$

Figure 2: Parameterization of logistic growth, priors of estimates, and heteroskedastic sub-models.

3. RESULTS

3.1 Heteroskedastic sub-model parameterization influences posterior predictions

Logistic growth data was simulated for two treatment groups (Figure 1, See methods: Creating Simulated Data). Multiple Bayesian hierarchical models were fit to recapitulate the simulated model parameters using different heteroskedastic sub-models (Figure 2, See methods: Bayesian Hierarchical Models), and Bayesian credible intervals were estimated to visualize model predictions. Finally, the different sub-models were evaluated for model fit using leave-one-out information criterion (LOO IC). For efficiency, LOO is approximated via Pareto-smoothed importance sampling (PSIS)[17] in `brms`, then the expected log pointwise predictive density (elpd) is calculated and multiplied by negative two to yield a model's LOO IC.[18]

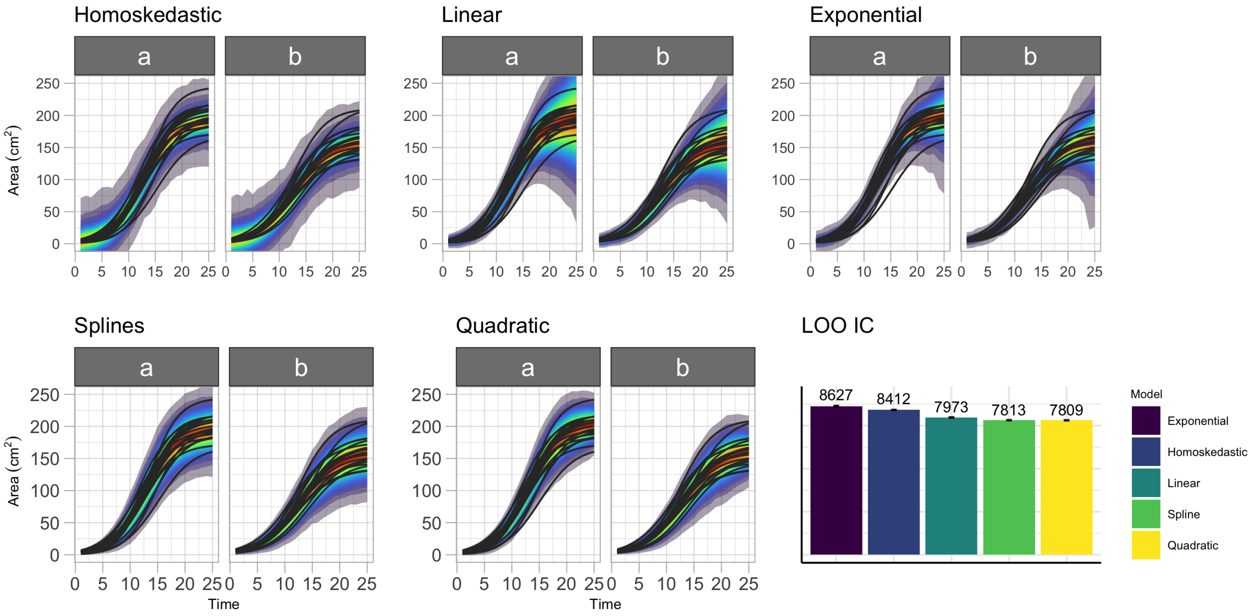

Figure 3: Bayesian credible intervals of different heteroskedastic sub-models and LOO IC model comparison.

A homoskedastic (constant variance) model and four heteroskedastic models were created: linear, exponential, spline fit, and quadratic relationship with time (See Methods: Bayesian Hierarchical Models). Shown in each sub-model panel are the trend lines of each individual in a solid black line, which are on top of Bayesian credible intervals as predicted from the model. The outermost edges represent the [1,99]% interval and colors gradually converge to red which is the mean trend line. In the homoskedastic model, the credible interval drastically overestimates the beginning and middle of development, whereas linear and exponential both overestimate the variance at the end of development. In contrast, splines and quadratic heteroskedastic models appear to fully contain the simulated data and do not appear to over or under estimate the variance at any time point. The splines and quadratic models also had the lowest LOO IC, which indicates they are the best fit models to the data. Given that the posterior credible intervals are highly influenced by model choice, it is obvious that choosing one that most reflects the data will more accurately reflect the true nature of the data and boost statistical power by not over estimating the variance. Spline heteroskedasticity is not constrained to a family of functions as with the others evaluated here and is therefore the most flexible type of heteroskedastic sub-model.

3.2 Adaptive designs are accessible through Bayesian hierarchical frameworks

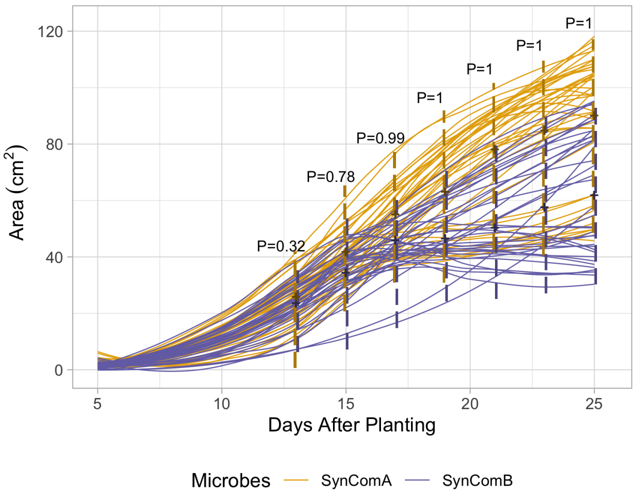

Figure 4: Pseudo Bayesian interim analysis for effect estimation and data-driven decision making.

Using a Bayesian hierarchical framework, conventional problems arising in continuous monitoring and optional stopping under a null hypothesis significance testing design can be alleviated.[19,20] Additionally, results of Bayesian hypothesis testing can be interpreted intuitively and implemented with excellent flexibility with `brms`. Adaptive designs are a data-driven decision making process wherein hypothesis tests are conducted on set days to decide whether or not there is sufficient evidence to trigger an event which commonly is to cease the experiment, this is known as interim analysis. The overarching goal of adaptive designs is to maximize efficiency through continuous monitoring.

Using published high-throughput phenotyping data,[15] logistic growth was modeled using a spline heteroskedastic sub-model for each of the two microbe inoculates. Pseudo-interim analysis shows that comparing the asymptotes of the two logistic curves with the hypothesis $P[\phi_{1a}/\phi_{1b} > 1]$ every two days starting on the 13th day indicates an increasing difference between the two microbe inoculates over time (Figure 4). This experiment was concluded after 25 days but there is sufficient evidence to support that there is an effect as early as day 17. Evaluating inoculate effects is one of many hypotheses that could be used for interim analysis in this dataset and future work will be done to determine optimal interim analysis strategies in this setting.

4. CONCLUSIONS

A Bayesian hierarchical framework offers several advantages for plant growth modeling and analysis over common linear and non-linear modeling approaches. The BHM framework is flexible, works with a variety of models, and can account for uneven variance, a common feature of growth curves, which results in better model fit. The BHM framework also produces models that can be updated when new observations are available, lending this approach to real-time monitoring of plant phenotypes and evidence-based decision making. In future work, we plan to integrate the BHM framework with the Clowder Framework.[21] Clowder can be used to manage the intake of image data from phenotyping instruments and analyze data in real-time using a catalog of data and metadata extractors. Interim analysis using Bayesian hypothesis testing could be done automatically and provide reports on experimental progress.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation (1835543) and the US Department of Energy (DE-SC0018072) (DE-SC0014395).

REFERENCES

[1] Casto, A. L., Schuhl, H., Tovar, J. C., Wang, Q., Bart, R. S., Fahlgren, N., and Gehan, M. A., "Picturing the future of food," *Plant phenome j.* **4** (Jan. 2021).

[2] Pieruschka, R. and Schurr, U., "Plant phenotyping: Past, present, and future," *Plant Phenomics* **2019**, 1–6 (Mar. 2019).

[3] Chen, D., Neumann, K., Friedel, S., Kilian, B., Chen, M., Altmann, T., and Klukas, C., "Dissecting the Phenotypic Components of Crop Plant Growth and Drought Responses Based on High-Throughput Image Analysis ," *The Plant Cell* **26**, 4636–4655 (12 2014).

[4] Neilson, E. H., Edwards, A. M., Blomstedt, C. K., Berger, B., Møller, B. L., and Gleadow, R. M., "Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a C4 cereal crop plant to nitrogen and water deficiency over time," *J. Exp. Bot.* **66**, 1817–1832 (Apr. 2015).

[5] Fahlgren, N., Feldman, M., Gehan, M. A., Wilson, M. S., Shyu, C., Bryant, D. W., Hill, S. T., McEntee, C. J., Warnasooriya, S. N., Kumar, I., Ficor, T., Turnipseed, S., Gilbert, K. B., Brutnell, T. P., Carrington, J. C., Mockler, T. C., and Baxter, I., "A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in setaria," *Mol. Plant* **8**, 1520–1535 (Oct. 2015).

[6] Vasseur, F., Bresson, J., Wang, G., Schwab, R., and Weigel, D., "Image-based methods for phenotyping growth dynamics and fitness components in arabidopsis thaliana," *Plant Methods* **14**, 63 (July 2018).

[7] Wang, R., Qiu, Y., Zhou, Y., Liang, Z., and Schnable, J. C., "A High-Throughput phenotyping pipeline for image processing and functional growth curve analysis," *Plant Phenomics* **2020**, 7481687 (July 2020).

[8] Brien, C., Jewell, N., Watts-Williams, S. J., Garnett, T., and Berger, B., "Smoothing and extraction of traits in the growth analysis of noninvasive phenotypic data," *Plant Methods* **16**, 36 (Mar. 2020).

[9] Paine, C. E. T., Marthews, T. R., Vogt, D. R., Purves, D., Rees, M., Hector, A., and Turnbull, L. A., "How to fit nonlinear plant growth models and calculate growth rates: An update for ecologists," *Methods in Ecology and Evolution* **3(2)**, 245–256 (2012).

[10] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2013).

[11] Bürkner, P.-C., "Advanced Bayesian multilevel modeling with the R package brms," *The R Journal* **10**(1), 395–411 (2018).

[12] Wickham, H., [*ggplot2: Elegant Graphics for Data Analysis*], Springer-Verlag New York (2016).

[13] Pedersen, T. L., *patchwork: The Composer of Plots* (2020). R package version 1.1.1.

[14] Sumner, J., Berry, J., and Schuhl, H., "danforthcenter/bayesian_growth: Nappn 2022 release," (Oct. 2021).

[15] Qi, M., Berry, J. C., Veley, K., O'Connor, L., Finkel, O. M., Salas-González, I., Kuhs, M., Jupe, J., Holcomb, E., del Rio, T. G., Creech, C., Liu, P., Tringe, S., Dangl, J. L., Schachtman, D., and Bart, R. S., "Identification of beneficial and detrimental bacteria that impact sorghum responses to drought using multi-scale and multi-system microbiome comparisons," *bioRxiv* (2021).

[16] Kruschke, J. K., "Bayesian estimation supersedes the t test," (2012).

[17] Vehtari, A., Gelman, A., and Gabry, J., "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing* **27**, 1413–1432 (2016).

[18] Bürkner, P., *Estimating Non-Linear Models with brms* (2021). R Vignette, URL https://cran.r-project.org/web/packages/brms/vignettes/brms_nonlinear.html.

[19] Deng, A., Lu, J., and Chen, S., "Continuous monitoring of a/b tests without pain: Optional stopping in bayesian testing," (2016).

[20] Wagenmakers, E.-J., Gronau, Q. F., and Vandekerckhove, J., "Five bayesian intuitions for the stopping rule principle," (Mar 2019).

[21] Marini, L., Gutierrez-Polo, I., Kooper, R., Satheesan, S. P., Burnette, M., Lee, J., Nicholson, T., Zhao, Y., and McHenry, K., "Clowder: Open source data management for long tail data," in [*Proceedings of the Practice and Experience on Advanced Research Computing*], *PEARC '18*, 1–8, Association for Computing Machinery, New York, NY, USA (July 2018).