Aster and envelopes

Daniel J. Eck

September 1, 2021

Aster models (Geyer, Wagenius, and Shaw, 2007) were developed for use in life history analyses (Shaw, Geyer, Wagenius, Hangelbroek, and Etterson, 2008). A specific application of aster models in life history analysis is the estimation of expected Darwinian fitness across covariates and trait values of interest where Darwinian fitness is the total offspring for a plant or animal over the course of its lifetime. The estimates of expected Darwinian fitness are plotted in a fitness landscape when covariates and trait values are continuous (Shaw and Geyer, 2010; Eck et al., 2015).

More formally, aster models (Geyer, Wagenius, and Shaw, 2007) are directed graphical models that satisfy the following five structural assumptions:

A1. The directed graph is acyclic.

A2. A node has, at most, one predecessor node.

A3. The joint distribution is the product of conditional distributions, one conditional distribution for each arrow in the aster graph.

A4. Predecessor is sample size.

A5. Conditional distributions for arrows are one-parameter exponential families. The exponential families across arrows are not required to be the same.

Assumptions A4 and A5 mean for an arrow $y_k \longrightarrow y_j$ that y_j is the sum of independent and identically distributed random variables from the exponential family for the arrow and there are y_k terms in the sum where the sum of zero terms is zero (Eck, Geyer, and Cook, 2017). The sum y_j must be over discrete exponential families when y_j is not a terminal node in the

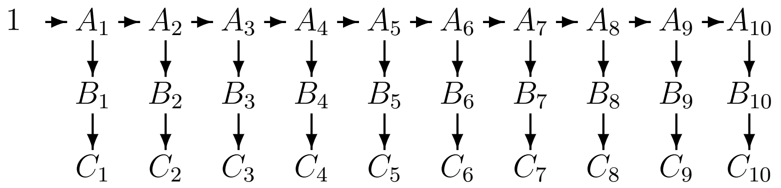

Figure 1: Graphical structure of the aster model for the simulated data in Eck, Geyer, and Cook (2017, Example 1). The top layer corresponds to survival; these random variables are Bernoulli. The middle layer corresponds to whether or not an individual reproduced; these random variables are also Bernoulli. The bottom layer corresponds to offspring count; these random variables are zero-truncated Poisson.

aster graph. These assumptions imply that the joint distribution of the aster model is an exponential family (Geyer, Wagenius, and Shaw, 2007, Section 2.3). In life history analyses, terminal nodes of the aster model correspond to offspring counts, while intermediate nodes represent important life stages in the plant or animal's life leading up to reproduction.

The aster model is a "generalized" generalized linear regression model (glm) for one parameter exponential families. All one parameter glms can be represented as an aster model. However, in aster, only the canonical link function can be used or the joint distribution of the aster model will cease to be an exponential family and inferential abilities of the aster model will be lost.

The log likelihood for the aster model in canonical form is

$$l(\beta) = \langle M^T Y, \beta \rangle - c(a + M\beta)$$

with canonical statistic $M^T Y$, $Y \in \mathbb{R}^m$ is the vector of responses consisting of one component for every node in the graph for every individual in the study, M is the model matrix assumed to have full column rank, a is a known offset vector, and β is the aster submodel canonical parameter vector.

One key difference between aster models and other methods in which envelope methodology is applied is that inference with respect to β is not desired. This is due in part to the fact that $M\beta_1$ and $M\beta_2$ can have the same value despite $\beta_1 \neq \beta_2$. Envelope methodology is not invariant to this type of non-uniqueness. To incorporate envelope methodology into the aster modeling framework, we focus on the aster submodel mean-vector parameter $\tau = E(M^T Y)$ which is a well-defined quantity in aster modeling (Geyer, 2010; Eck, Geyer, and Cook, 2017).

We now go through some details of Eck, Geyer, and Cook (2017, Example 1) to explore how envelope methodology fits within the context of aster modeling.

The dataset in this example is formed by generating data for 3000 organisms progressing through the lifecycle depicted in Figure 1. There is a known true envelope space in this example and it was shown that envelope methods in Eck, Geyer, and Cook (2017) find useful variance reduction at no cost to consistency. Darwinian fitness for this example is $\sum_{i=1}^{10} C_i$. There are two covariates (z_1, z_2) associated with Darwinian fitness and the aster model used in this analysis supposes that expected Darwinian fitness is a full quadratic model in z_1 and z_2.

Partial envelope estimation is used in this example. The aster submodel mean-value parameter vector τ is partitioned into $(\gamma^T, \upsilon^T)^T$ where $\gamma \in \mathbb{R}^4$ are nuisance parameters and $\upsilon \in \mathbb{R}^5$ are relevant to the estimation of expected Darwinian fitness. Here, $\upsilon \in \mathbb{R}^5$ because our model is full quadratic in z_1 and z_2. We then estimate υ using both maximum likelihood estimation (the standard in aster) and envelope estimation. We use a weighted envelope estimator and a parametric bootstrap procedure suggested by Efron (2014) to assess its variability. For more details, see Eck, Geyer, and Cook (2017).

The contour plots of the ratios of estimated standard errors for estimated expected Darwinian fitness are displayed in Figure 2. The contours show that the envelope estimator of expected Darwinian fitness is less variable than the maximum likelihood estimator for the majority of the observed data. Most importantly, we see efficiency gains at the values of z_1 and z_2 that maximize estimated expected Darwinian fitness.

One consequence of this example is that lower variability in estimation, resulting from the incorporation of envelope methodology, allows researchers in life history analysis to narrow their search for candidate trait values and covariates associated with interesting aspects of expected Darwinian fitness. This is of importance in life history analysis (Eck et al., 2015; Shaw and Geyer, 2010). All estimates of expected Darwinian fitness have variability in estimation. As a result, there can be many trait values statistically indistinguishable from the trait value which maximizes expected Darwinian fitness. We can see that the combination of envelope methodology with aster models leads to a useful reduction in the number of traits that are statistically indistinguishable from the reported maximizer of estimated expected Darwinian fitness. When aster is used alone, there are 14 candidate maximizers of estimated expected Darwinian fitness as opposed to only 7 candidate maximizers of estimated expected Darwinian fitness when using envelope methodology.

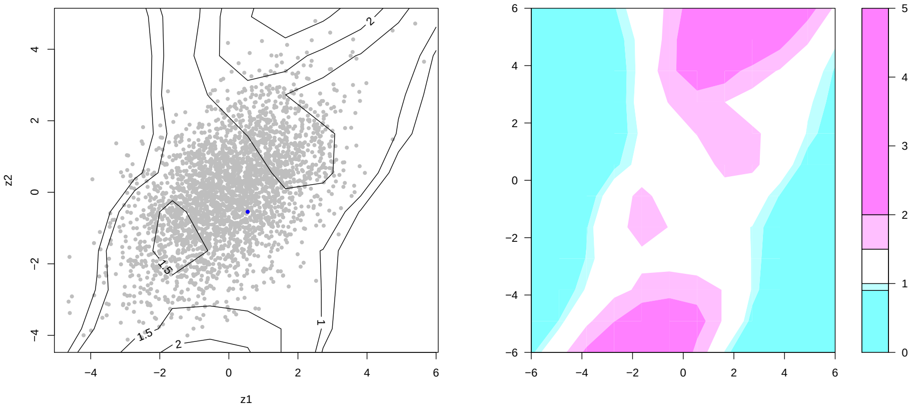

Figure 2: Contour plots for the ratios of bootstrapped standard errors for the model fit using the MLE to the ratios of bootstrapped standard errors for the model fit using the envelope estimator. The point in blue corresponds to the highest estimated expected Darwinian fitness value using envelope methodology and the MLE.

References

Eck, D. J., Shaw, R., Geyer, C. J., Kingsolver J. G. (2015). An Integrated Analysis of Phenotypic Selection on Insect Body Size and Development Time. *Evolution*, **69**: 2525-2532.

Eck, D. J., Geyer, C. J., and Cook, R. D. (2017). An Application of Envelope and Aster Models. *Submitted*.

Eck, D. J., Geyer, C. J., and Cook, R. D. (2016). Supporting Data Analysis for “An Application of Envelope Methodology and Aster Models.” `http://hdl.handle.net/11299/178384`.

Efron, B. (2014). Estimation and Accuracy After Model Selection. *JASA*, **109:507**: 991-1007.

Geyer, C. J., Wagenius, S., Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**: 415-426.

Geyer, C. J. (2010). A Philosophical Look at Aster Models. Technical Report No. 676. School of Statistics, University of Minnesota. `http://purl.umn.edu/57163.`

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H., and Etterson, J. R. (2008). Unifying life-history analyses for inference of fitness and population growth. *The American Naturalist*, **172**: E35-E47.

Shaw, R. G., Geyer, C. J. (2010). Inferring fitness landscapes. *Evolution*, **64**: 2510-2520.