A Principled Approach to Developing Chlorophyll Models for Remote Sensing

Erdem M. Karaköylü

October 16, 2018

Contents

1 Abstract

2 Introduction

2.1 Background

- Necessity for estimating chlorophyll
- Opportunities afforded by chlorophyll estimation through remote sensing
- State of current chlorophyll algorithms for remote sensing
  - OC4 for SeaWiFS
  - Hu & Franz 2012
  - Gholizadeh et al (2015?)
- Basic empirical form

$$log_{10}\left(chlor_a\right) = a_0 + \sum_{i=1}^{j} a_i log_{10}\left(\frac{max\left(Rrs\left(\lambda_{blue}\right)\right)}{Rrs\left(\lambda_{green}\right)}\right) \tag{1}$$

- Problems with current algorithms:
  - collinearity of inputs
  - poor performance in coastal
  - maximum likelihood estimation approach → risk of overfitting (lack of in-situ data vailability compared to satellite data makes it worse)

2.2 Proposed framework

Here, I use Bayesian regression models to predict chlorophyll concentration from apparent optical properties. There are a number of advantages for adopting this approach. The first of these is a problem agnostic set of steps detailed in the methods section. A second advantage is that model construction is transparent and assumptions are laid bare. This makes any model developed with this framework easily, and constructively, criticizable. A positive spillover of this is that the process requires that the model code be made available for model criticism to occur; i.e. the study in question must be reproducible to be effective. Another attractive feature of the Bayesian framework is that rather than point predictions, the output of the Bayesian model is the posterior distribution. The posterior distibution is a rich construct that can be used to develop insight in the process of interest, uncertainties around predictions, and a means to rank model performance. Finally, Bayesian regression models are inherently regularized through the stipulated priors. This has a 'calming' effect on the model, which is less likely to overfit an inappropriately scant data set; a common problem in bio-optical model development for marine remote sensing.

Succintly, this approach ensures:

- transparent construction of models with explicit formulation of assumptions,
- assumptions/background information codified as priors that are easy to criticize and modify,
- verifiable prior feasibility before data collection via prior predictive checks
- built-in regularization
- built-in structure for selecting relevant features,
- posterior distribution as rich information structure from which to estimate parameter uncertainty, output predictin uncertainty, and likelihood of model performance on out-of-sample data,
- assessment of predictive ability via posterior predictive checks,
- multiple model development, which avoids overemphasis on any particular formulation,

- evaluation/comparison between models using information theory

Because the usefullness of bayesian modeling relies on the reproducibility of the models and their results, the interested reader can obtain the model code from our github repository [here] and the data from the Open Science Framework (OSF) page associated with this projet, [here].

3 Methods

3.1 Model Development

3.1.1 Bayesian Linear Regression

- Order 1 regression for interpretable coefficients
- no interaction terms
- regularized horseshoe prior for feature selection

3.1.2 Bayesian Linear Regression with Interaction Terms

- generation of 1st order interaction terms
- allowing for both strong and weak heredity

3.1.3 Bayesian Neural Network

- Specific hierarchical structure for ARD
- HL1 4 NN with elu activation

3.1.4 Bayesian OC4 version as Baseline

3.2 Prior Predictive Checks

3.3 Data Acquisition/Exploration/Transformation

3.3.1 Data exploration and transformation

-

3.3.2 Basis reduction via PCA

- PCA of Rrs to reduce overlap of information between predictor variables

3.4 Model Fitting

3.5 Marginal Posterior of Coefficients $\rightarrow$ Feature Relevance Determination

3.6 Posterior Predictive Checks

3.7 Model Comparison Through Posterior Predictive Checks

4 Results