Remote Sensing of Environment 132 (2013) 102–119

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection

Kaiguang Zhao [a,b,*], Denis Valle [b], Sorin Popescu [c], Xuesong Zhang [d], Bani Mallick [e]

[a] Center on Global Change, Duke University, Durham, NC, USA
[b] Nicholas School of the Environment, Duke University, Durham, NC, USA
[c] Dept. of Ecosystem Sciences and Management, Texas A&M University, College Station, USA
[d] Joint Global Change Research Institute, Pacific Northwest National Laboratory and University of Maryland, College Park, USA
[e] Dept. of Statistics, Texas A&M University, College Station, USA

ARTICLE INFO

Article history:
Received 23 November 2011
Received in revised form 30 December 2012
Accepted 30 December 2012
Available online 10 February 2013

Keywords:
Hyperspectral
Plant biochemistry
Leaf pigment
Chlorophyll
Carotenoid
Nitrogen
Carbon
Band selection
Bayesian model averaging
MCMC
Model misspecification
Model selection
Model uncertainty

ABSTRACT

Model specification remains challenging in spectroscopy of plant biochemistry, as exemplified by the availability of various spectral indices or band combinations for estimating the same biochemical. This lack of consensus in model choice across applications argues for a paradigm shift in hyperspectral methods to address model uncertainty and misspecification. We demonstrated one such method using Bayesian model averaging (BMA), which performs variable/band selection and quantifies the relative merits of many candidate models to synthesize a weighted average model with improved predictive performances. The utility of BMA was examined using a portfolio of 27 foliage spectral–chemical datasets representing over 80 species across the globe to estimate multiple biochemical properties, including nitrogen, hydrogen, carbon, cellulose, lignin, chlorophyll (a or b), carotenoid, polar and nonpolar extractives, leaf mass per area, and equivalent water thickness. We also compared BMA with partial least squares (PLS) and stepwise multiple regression (SMR). Results showed that all the biochemicals except carotenoid were accurately estimated from hyerspectral data with R^2 values > 0.80. Compared to PLS and SMR, BMA substantially reduced overfitting and enhanced model generalization; BMA also yielded error estimation better indicative of true uncertainties in predictions, when evaluated using a statistic called "prediction interval coverage probability". The relative band importance, which was quantified by band selection probability, differed markedly between BMA and SMR, cautioning the use of SMR for band selection. Computationally, the model calibration with datasets of moderate sizes (>100) was faster for BMA via a hybrid reversible-jump Monte Carlo Markov Chain sampler than for PLS via literal optimization of a cross-validation criterion. Our BMA scheme also provides a generic hierarchical Bayesian framework to assimilate prior knowledge of diverse forms, as illustrated by its use to account for nonlinearity in spectral–chemical relationships. We emphasize that BMA is a competitive, paradigm-shifting alternative to conventional statistical methods and it will find wide use as the virtue of Bayesian inference is increasingly appreciated by the remote sensing community.

1. Introduction

Biospheric functioning is mediated by both external abiotic forcings and internal vegetation physiology (Jackson et al., 2008), with the latter inherently linked to plant biochemistry. Measuring plant biochemicals therefore represents a critical component in quantifying how ecosystems function, as exemplified by the use of foliage nitrogen concentration derived from remote sensing as a proxy for photosynthetic capacity (Ustin et al., 2004). During the past few decades, the capabilities of remote sensing to determine plant biochemical status have been established, owing primarily to advances in spectroscopy from three platforms — ground, airborne and spaceborne (e.g., Asner & Martin, 2009; Milton et al., 2009; Townsend et al., 2003). Spectroscopic data typically contain hundreds to thousands of contiguous, narrow wavebands that allow detailed hyperspectral characterization of foliage absorption spectra. Physical linkages between foliage chemicals and their absorption characteristics provide an algorithmic basis for hyperspectral remote sensing of plant biochemicals (Knyazikhin et al., 2012), including water, carbon, nitrogen, cellulose, lignin, and foliage pigments such as chlorophyll, carotenoids, and anthocyanins (Asner & Martin, 2008; Gitelson et al., 2006; Serrano et al., 2002).

Hyperspectral methods to retrieve biochemical properties can be roughly grouped into two categories: physically- or empirically-based. Physically-based inversion of biochemicals was facilitated by progresses in radiative transfer modeling of leaf spectra, as demonstrated by the widespread use of the PROSPECT and LIBERTY models (e.g., Asner & Martin, 2008; Dawson et al., 1998; Feret et al., 2008; Jacquemoud & Baret, 1990). Typically, these leaf spectra models are

* Corresponding author at: Center on Global Change, Duke University, Durham, NC, USA. Tel.: +1 9797399981.
E-mail address: lidar.rs@gmail.com (K. Zhao).

0034-4257/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.rse.2012.12.026

scaled up with canopy reflectance models or further coupled with atmospheric radiative modules to simulate reflectance of vegetated landscapes as observed by airborne and spaceborne spectrometers (Zarco-Tejada et al., 2001). Such integrated simulations serve as forward physical models and generate look-up tables for inverting pigment concentrations (Zhang et al., 2008). However, inverting forward physical models is often ill-posed in nature, partly because the contributions of plant biochemicals to observed spectra are obscured by numerous extraneous factors, such as leaf anatomical structure, canopy architecture, viewing geometry, and atmospheric conditions. These confounding effects pose a practical limit to retrieval accuracies. Moreover, difficulties exist in explicitly quantifying absorption and scattering of some common biochemicals, such as nitrogen, phosphorus, lignin, cellulose, and phenols. Thus, the prevalent parameterization of leaf optical models is based not on these biochemicals but instead on foliage pigments such as chlorophyll and carotenoids. Also, the same biochemical such as nitrogen and phosphorus is disproportionally allocated in different foliage pigments (Kokaly et al., 2009). These constraints preclude direct inversions of forward physical models for retrieving many important biochemicals.

Empirically-based inversion encompasses a wide range of statistical methods to relate biochemical contents to hyperspectral indices or reflectance (e.g., Martin et al., 2008; Richardson et al., 2002). In particular, identifying several wavelengths to construct a spectral index is both conceptually and practically attractive, but this reductionist paradigm seemingly goes against harnessing the information-richness of full spectra. Spectral indices are often formulated heuristically through simple transforms, including difference, ratio, and high-order derivatives, in attempts to enhance desirable signals and subdue confounding factors (Garbulsky et al., 2011; Gitelson et al., 2006). Existing indices were proposed mostly based on small individual datasets of a limited number of species, contributing to the diverging forms of indices available in the literature even for the same biochemical (le Maire et al., 2004; Schlerf et al., 2010). Also, the utility of hyperspectral indices is affected by scales examined; thus, the direct transfer of indices between leaf and canopy scales is not always warranted (Zhang et al., 2008). Nevertheless, a few indices do manifest high levels of reliability and reproducibility. For example, the photochemical reflectance index responds to xanthophyll cycle pigment activities and is indicative of photosynthetic light use efficiency (Gamon et al., 1997) and relative levels of cholorophylls and carotenoids (Garbulsky et al., 2011).

In contrast to spectral indices, another important type of empirical method focuses on exploiting full spectra using regression techniques. Common examples of this type include stepwise multiple regression (SMR), ridge regression, principal component regression (PCR), partial least squares (PLS), and machine learning such as neural networks (NN) and Gaussian process (GP) (Asner et al., 2011; Pasolli et al., 2010; Zhao et al., 2008). When a large number of strongly correlated predictors are present, variable selection via SMR is recognizably vulnerable and unreliable. As an alternative to SMR, PLS becomes prevalent because it alleviates the problem of high dimensionality by seeking a parsimonious number of factors through a linear projection of original bands (Nguyen & Lee, 2006). PLS also outcompetes PCR because PCR accounts for only the variance of explanatory variables (e.g., band reflectance) without any bearing on response variable (e.g., biochemical content) whereas PLS accounts for both (Atzberger et al., 2010; Wold et al., 2001). Unlike linear regression, advanced methods such as NN and GP are flexible to approximate nonlinearity. Recently, such nonlinear regression has increasingly attracted interest from remote sensing practitioners as many machine learning tools reach maturity for practical purposes (e.g., Pasolli et al., 2010; Zhao et al., 2008). Kernel-based machines such as GP and support vector machine are particularly appealing for tackling high-dimensional problems; yet, machine learning tools in use for remote sensing were often criticized due to a high risk of over-fitting and a lack of physical interpretability (Liang, 2007; Zhao et al., 2011).

The past few decades have also witnessed a rapid re-awakening of interest in Bayesian statistical modeling. A notable example concerning spectroscopy is the recent adoption of Bayesian regression as an alternative to PLS and PCR for calibration problems in chemometrics (Brown et al., 1998; Chen & Martin, 2009). These problems are of the same nature as those of hyperspectral remote sensing considered here, both aiming to estimate chemical contents from high-dimensional curve data (Brown et al., 1998). The efficacy of Bayesian regression in this setting has been demonstrated in several chemometric studies with results superior to PLS (e.g., Chen & Martin, 2009). Bayesian regression also provides a natural framework for variable and model selection to efficiently explore an enormous model space (e.g., band combinations) via Monte Carlo Markov Chain (MCMC) sampling algorithms. This framework is particularly useful to tackle "$p > n$" problems wherein the number of spectral bands is larger than the training sample size, namely a case where traditional frequentist methods (i.e., non-Bayesian) often lack suitable procedures for efficient variable selection (Denison, 2002). More interestingly, with Bayesian variable and model selection, each model is given a posterior probability of being the true model, thereby offering an intuitive model-averaging mechanism to synthesize multiple competing models into inference and account for model uncertainty (Sloughter et al., 2007).

The purpose of this paper is to present a Bayesian variable selection and model averaging approach to estimating foliage biochemicals from hyperspectral data. The approach was formulated in a Bayesian hierarchical modeling framework and implemented using a hybrid MCMC sampler; one of its salient features is efficient variable and band selection. In stark contrast to the common practice of selecting only the single "best" model, this Bayesian model averaging (BMA) scheme seeks to leverage the many plausible models. Considerations of multiple models help to characterize model uncertainty, alleviate model misspecification, and improve predictive ability. The theoretical underpinning of BMA aligns with the fact that numerous hyperspectral indices or band combinations, though different, are all useful to some extent when estimating the same biochemical. We evaluated this Bayesian approach using a total of 27 spectral–chemical datasets from three sources representing 82 plant species to estimate a variety of biochemical properties, including nitrogen, hydrogen, carbon, cellulose, lignin, chlorophyll (a or b), carotenoid, polar and nonpolar extractives, leaf mass per area, and equivalent water thickness. We also compared BMA with two conventional methods, namely PLS and SMR.

2. Data

Leaf spectral–chemical data are compiled from three sources, namely the NASA's Accelerated Canopy Chemistry Program (ACCP, 1994), the Leaf Optical Properties Experiment 93 (LOPEX93) of the Joint Research Center (JRC — Italy) (Hosgood et al., 1994), and some existing field data collected of maize and maple (MM) from several individual projects (Gitelson et al., 1999, 2003, 2005, 2006). All these experiments aim to understand the relationships between plant biochemistry and spectra to support retrievals of biochemical constituents from hyperspectral data. In particular, ACCP and LOPEX data are available publically and have been previously examined for various purposes (e.g., Bolster et al., 1996; Jacquemoud et al., 1995). A complete list of the biochemicals we considered is summarized in Table 1, comprising a portfolio of 27 paired spectral–chemical datasets. Next, we briefly describe the three data sources.

2.1. ACCP

The ACCP data are an assemblage of measurements from five field sites across USA as well as from seedlings of Douglas-fir (*Pseudotsuga menziesii*) and bigleaf maple (*Acer macrophyllum*) grown in greenhouse, representing more than 30 deciduous and coniferous tree species. Only

leaf-level measurements were considered here. Foliage chemistry, including nitrogen, carbon, cellulose, lignin, polar and nonpolar extractives, and pigments such as chlorophyll (Chl) and beta-carotenoid (Carot), was determined through standard wet chemical analyses using the Perkin-Elmer CHN Elemental Analyzer and a sequential extraction/digest method. Not every leaf sample has a complete record of all the chemical concentrations; therefore, the number of leaf samples available for model calibration and validation varies from one biochemical to another (Table 1). Spectral reflectance of fresh leaf, dry leaf and sometimes ground leaf powder was measured in the wavelength range of 400–2500 nm at a 2-nm spectral sampling interval with either a NIRSystems Model 6250 or 6500 scanning monochromator (NIRSystems—Silver Springs, MD, USA). We discarded the ground foliage power data due to its small sample size, leaving 647 dry and 240 fresh leaf samples for our analyses.

2.2. LOPEX93

The LOPEX93 data include 70 leaf samples representative of more than 50 species of Monocotyledon, Dicotyledon or Gymnosperm collected in the vicinity of the Joint Research Center, Ispra, Italy. Each sample consists of multiple leaves taken from the same tree or plant. Five representative leaves from each sample were chosen for measuring reflectance and transmittance using a Perkin Elmer Lamdba 19 double-beam spectrophotometer equipped with a BaSo4 integrating sphere over the wavelength range of 400–2500 nm, at a sampling of 1.0–2.0 nm for visible/NIR (400–1000 nm) and 4–5 nm for SWIR (1000–2500 nm). The availability of five leaf spectra per sample allows accounting for leave-level spectral variations within each sample. The measured foliage properties include dry matter content (i.e., leaf matter per unit area — LMA), equivalent water thickness (EWT), chlorophyll a and b, carotenoid, nitrogen, carbon, and some biochemical compounds such as cellulose, lignin, and starch. The biochemical analyses were conducted by two independent laboratories, each using ~250 g of fresh leaves from each sample; the averages of the two estimates were used for our analysis to reduce measurement errors. Of the measured leaf properties, LMA and EWT were available for individual leaves, thus allowing us to quantify these two properties using leaf-level spectra. In contrast, all the other biochemicals were measured at the sample level; therefore, it is the average of five spectra of a sample, not the individual leaf spectrum, that will be related to these biochemicals for our modeling analyses. Moreover, we considered only the fresh leaf samples. The number of paired spectral–chemical measurements available for model calibration and validation also varies from one biochemical to another (Table 1).

2.3. Spectral–chemical data of maize and maple (MM)

We also compiled 30 Norway maple (*Acer platanoldes* L.) and 42 maize leaf samples that were collected in a park at Moscow State University, Russia, and at Mead Nebraska, USA, respectively (Gitelson et al., 2003, 2005, 2006). Adaxial reflectance spectra of these leaves were taken in the visible-near infrared range of 400 to 800 nm at a resolution of 2 nm. Specifically, a 150-20 Hitachi spectrophotometer equipped with a 150 mm integrating sphere was used for measuring Norway maple; and a clip with a 2.3-mm diameter bifurcated fiber-optic attached to both an Ocean Optics USB2000 radiometer and an Ocean optics LS-1 light source was used for maize. The biochemical measurements we considered are either total chlorophyll for both maple and maize or carotenoid and chlorophyll a and b for the maple samples only. In our subsequent analyses, we mainly combined the maple and maize samples, resulting in 191 wavebands common to both.

3. Bayesian regression with model averaging

3.1. Motivation and theoretical basis

Remote sensing practitioners, when implementing statistical inversion, are confronted with at least two questions — what predictors and what model forms should be used (Zhao & Popescu, 2009)? The use of different schemes to resolve these questions leads to a proliferation of model specifications for the same spectroscopic problem.

Table 1
A portfolio of 27 spectral–biochemical datasets compiled from three sources.

Data source	Biochemical	Leaf status	Dataset acronym	Number	Min	Mean	Max	Unit
ACCP	Nitrogen	Fresh	N-f-ACCP	206	0.68	2.18	5.25	% dry weight
		Dry	N-d-ACCP	644	0.68	1.85	3.51	% dry weight
	Cellulose	Fresh	Cell-f-ACCP	26	25.50	37.21	59.24	% dry weight
		Dry	Cell-d-ACCP	555	23.69	37.91	67.57	% dry weight
	Lignin	Fresh	Lign-f-ACCP	26	14.18	18.47	25.97	% dry weight
		Dry	Lign-d-ACCP	555	12.42	22.59	33.70	% dry weight
	Chlorophyll	Fresh	Chl-f-ACCP	231	1.16	6.47	18.60	% dry weight
		Dry	Chl-d-ACCP	89	1.20	3.89	6.96	% dry weight
	Chlorophyll-a	Fresh	Chl.a-f-ACCP	140	0.83	6.24	13.64	% dry weight
	Chlorophyll-b	Fresh	Chl.b-f-ACCP	140	0.36	1.92	5.23	% dry weight
	Carbon	Dry	C-d-ACCP	555	43.87	50.30	53.12	% dry weight
		Fresh	C-f-ACCP	30	43.17	47.60	51.07	% dry weight
	Hydrogen	Dry	H-d-ACCP	555	5.76	6.57	7.25	% dry weight
	Polar	Dry	polar-d-ACCP	555	9.91	34.59	55.78	% dry weight
	Nonpolar	Dry	nonpolar-d-ACCP	555	1.00	4.92	12.60	% dry weight
LOPEX	LMA	Fresh	LMA-f-LOPEX	335	1.71	5.28	15.73	mg/cm^2
	EWT	Fresh	EWT-f-LOPEX	335	0.29	11.53	65.53	mg/cm^2
	Nitrogen	Fresh	N-f-LOPEX	83	7.40	19.68	35.55	% dry weight
	Cellulose	Fresh	Cell-f-LOPEX	83	6.66	18.12	31.47	% dry weight
	Lignin	Fresh	Lign-f-LOPEX	83	1.37	9.16	24.16	% dry weight
	Chlorophyll-a	Fresh	Chl.a-f-LOPEX	64	0.27	7.29	14.82	% dry weight
	Chlorophyll-b	Fresh	Chl.b-f-LOPEX	64	0.14	2.33	5.24	% dry weight
	Carotenoid	Fresh	Carot-f-LOPEX	64	0.85	2.05	3.69	% dry weight
MM	Chlorophyll-a	Fresh	Chl.a-f-MM	30	0.44	127.59	312.72	mg/m^2
	Chlorophyll-b	Fresh	Chl.b-f- MM	30	0.31	51.61	145.43	mg/m^2
	Chlorophyll	Fresh	Chl-f- MM	72	0.75	373.51	918.74	mg/m^2
	Carotenoid	Fresh	Carot-f- MM	30	1.62	7.20	12.30	mg/m^2

Conventional methods such as PLS and SMR tend to seek an optimal model based on certain selection criteria while discarding the other models. In particular, PLS implicitly incorporates all wavebands in a linear fashion and finds a single equation typically via leave-one-out cross-validation (LOOCV) (Atzberger et al., 2010; Wold et al., 2001). This single best model paradigm is subject to potential limitations, such as vulnerability to model misspecifications, an understating of model uncertainty, and a lack of flexibility in model diagnostics (Chen & Martin, 2009; Denison, 2002; Grossman et al., 1996). Such weaknesses can be alleviated by switching to a Bayesian inferential paradigm, which allows combining multiple competing models via model averaging to account for model uncertainty in the model selection process (Brown et al., 1998; Chen & Wang, 2010).

As a further illustration, we consider Bayesian model averaging and selection in a linear regression context. The aim is to uncover a linear relationship between y and $\mathbf{x}$ from a calibration/training dataset of n observations $\mathcal{D}=\{\mathbf{x}_i, y_i\}_{i=1,\cdots,n}$. The variable y can be any biochemical of interest. The covariate vector $\mathbf{x}$ comprises p covariates, examples of which include reflectance, transmittance, and hyperspectral indices. A subset of covariates chosen out of $\mathbf{x}$ uniquely determines a model form or configuration; correspondingly the combinatorics of all the p covariates in $\mathbf{x}$ gives rise to a space of 2^p possible model configurations $\mathcal{M}=\{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_{2^p}\}$. The number of covariates selected into model $\mathcal{M}_i$ is denoted by $p_{\mathcal{M}_i}$, with $0 \le p_{\mathcal{M}_i} \le p$. Each model configuration $\mathcal{M}_i$ can be represented either by the set of the $p_{\mathcal{M}_i}$ selected covariates or by a $n \times \left(1+p_{\mathcal{M}_i}\right)$ design matrix $\mathbf{X}_{\mathcal{M}_i}$, with its first column being all ones and the remaining columns corresponding to the $p_{\mathcal{M}_i}$ selected covariates. Accordingly, the linear equation of model $\mathcal{M}_i$ becomes

$$\mathbf{y}=\mathbf{X}_{\mathcal{M}_i}\boldsymbol{\beta}_{\mathcal{M}_i}+\varepsilon, i=1,\cdots,2^p \tag{1}$$

where $\mathbf{y}$ is the $n\times 1$ vector of observed responses in $\mathcal{D}$; $\boldsymbol{\beta}_{\mathcal{M}_i}$ is the $\left(1+p_{\mathcal{M}_i}\right)\times 1$ vector of coefficients, with its first element being the intercept and the other $p_{\mathcal{M}_i}$ elements being slopes for the selected covariates of model $\mathcal{M}_i$; and ε is the $n\times 1$ vector of independent normal error, with $var(\varepsilon)=\sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. The ordinary least-square solution to Eq. (1) is $\hat{\boldsymbol{\beta}}_{\mathcal{M}_i}=\left(\mathbf{X}_{\mathcal{M}_i}^T\mathbf{X}_{\mathcal{M}_i}\right)^{-1}\mathbf{X}_{\mathcal{M}_i}^T\mathbf{y}$.

Traditional regression relies on criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and adjusted R^2, to search for the single "best" model $\mathcal{M}_{best}$ out of the model space $\mathcal{M}$ elicited above. For spectroscopic applications, this searching is stymied or even prohibited for at least two reasons (Denison, 2002): (1) The computation required for enumerating all the 2^p models in $\mathcal{M}$ can be prohibitive, even for a moderately large number of wavebands. For example, an exhaustive evaluation of 2^{200} models (i.e., p = 200 bands) will take ~5.1×10^{44} centuries even if using a computer that processes 1×10^6 models per second. (2) The inversion of $\mathbf{X}_{\mathcal{M}_i}^T\mathbf{X}_{\mathcal{M}_i}$, as required for calculating $\hat{\boldsymbol{\beta}}_{\mathcal{M}_i}$ of Eq. (1), will fail due to rank deficiency if there are more covariates than observations (i.e., $p_{\mathcal{M}_i} > n$) or if the covariates in $\mathbf{X}_{\mathcal{M}_i}$ are highly correlated, thus making it impossible to numerically evaluate model configuration $\mathcal{M}_i$.

As a remedy and a conceptually appealing alternative to traditional criteria-based methods, Bayesian regression with model averaging (i.e., BMA) does not attempt to find a single optimal model but instead admits the relevance of all the 2^p models of $\mathcal{M}$ to the inference. BMA probabilistically quantifies the usefulness of all the models in $\mathcal{M}$ and synthesizes them into an average model. This averaging helps to alleviate model misspecification and address model uncertainty. Specifically, to formulate a BMA model, a prior probability distribution on $\mathcal{M}$, $\pi(\mathcal{M}_i)$, is first elicited to encode our prior belief in the truthfulness of individual models $\mathcal{M}_i$. Then, this prior is updated in light of calibration data $\mathcal{D}$ to generate the posterior $p(\mathcal{M}_i|\mathcal{D})$ according to Bayes' theorem:

$$p(\mathcal{M}_i|\mathcal{D})=\frac{p(\mathcal{D}|\mathcal{M}_i)\pi(\mathcal{M}_i)}{\sum_{k=1}^{2^p}p(\mathcal{D}|\mathcal{M}_k)\pi(\mathcal{M}_k)},\ i=1,\cdots,2^p \tag{2}$$

where $p(\mathcal{D}|\mathcal{M}_i)$ is the marginal likelihood of model $\mathcal{M}_i$, defined as the probability of observing $\mathcal{D}$ given $\mathcal{M}_i$.

The posterior $p(\mathcal{M}_i|\mathcal{D})$ in Eq. (2) denotes the probability of model $\mathcal{M}_i$ being the true one, given the information and evidence of the training data $\mathcal{D}$. Equivalently speaking, $p(\mathcal{M}_i|\mathcal{D})$ embodies the degree of our posterior belief in the usefulness of model $\mathcal{M}_i$ for explaining the unknown functional pattern underlying $\mathcal{D}$. Hence, the posteriors $p(\mathcal{M}_i|\mathcal{D})$ are natural choices as weights to apportion the contributions of individual models $\mathcal{M}_i$ when synthesizing an average model. More important, $p(\mathcal{M}_i|\mathcal{D})$ provides a probabilistic measure to guide variable/model selection and to draw random samples of models using MCMC procedures. As detailed in Section 3.3, an MCMC sample of a given length N, $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,N}$, is a finite realization of the posterior $p(\mathcal{M}_i|\mathcal{D}), i=1,\cdots,2^p$ and therefore, can substitute $p(\mathcal{M}_i|\mathcal{D})$ for making posterior inference and prediction (Fig. 1b). Typical choices of the MCMC sample size N are far smaller than the size of model space 2^p, thereby obviating the computational difficulty in enumerating $\mathcal{M}$. Moreover, unlike traditional regression that treats $\boldsymbol{\beta}_{\mathcal{M}_i}$ and σ^2 as unknown constants, the Bayesian paradigm treats the model parameters as random and hence assigns prior distributions on them. Mathematically speaking, these priors on $\boldsymbol{\beta}_{\mathcal{M}_i}$ and σ^2 play a regularization role and help to avoid inverting $\mathbf{X}_{\mathcal{M}_i}^T\mathbf{X}_{\mathcal{M}_i}$ directly, thus circumventing the rank deficiency problem as in the traditional methods for evaluation of $\hat{\boldsymbol{\beta}}_{\mathcal{M}_i}$.

To be self-contained, important technical specifics of our Bayesian method are presented in Sections 3.2–3.5. Readers may request the Matlab code of our method from the primary author.

3.2. Formulation of Bayesian model averaging (BMA)

We implemented the above BMA framework based on the model form $\mathbf{y}=\mathbf{X}_{\mathcal{M}_i}\boldsymbol{\beta}_{\mathcal{M}_i}+\varepsilon$ as in Eq. (1). Our inferential interest lies primarily in coefficients $\boldsymbol{\beta}_{\mathcal{M}_i}$, noise variance σ^2 of ε, and model configuration $\mathcal{M}_i$ (i.e., selected covariates). Fundamental to our BMA regression is the formulation of the posterior distribution $p(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i|\mathcal{D})$, $i=1,\cdots,2^p$, which provides all the information essential for model inference and prediction analysis. The posterior $p(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i|\mathcal{D})$ denotes a compromise between a likelihood model $p(\mathcal{D}|\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i)$ and a prior model $\pi(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i)$ according to

$$p\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\middle|\mathcal{D}\right)\propto p\left(\mathcal{D}\middle|\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right)\pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right) \tag{3}$$

where $p\left(\mathcal{D}\middle|\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right)$ can also be written as $p\left(\mathbf{y}\middle|\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i; \mathbf{x}\right)$. The likelihood is a data model, representing the conditional distribution of $\mathbf{y}$ given the model parameters $\left\{\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right\}$; its specific formulation is governed by the linear form of model $\mathbf{y}=\mathbf{X}_{\mathcal{M}_i}\boldsymbol{\beta}_{\mathcal{M}_i}+\varepsilon$ and therefore is simply Gaussian due to the normality of error ε:

$$p\left(\mathcal{D}\middle|\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right)=\mathbb{N}\left(\mathbf{y}; \mathbf{X}_{\mathcal{M}_i}\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2\mathbf{I}\right) \tag{4}$$

where $\mathbb{N}(;)$ denotes the multivariate Gaussian distribution with a mean vector of $\mathbf{X}_{\mathcal{M}_i}\boldsymbol{\beta}_{\mathcal{M}_i}$ and covariance matrix of $\sigma^2\mathbf{I}$. The prior distribution $\pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right)$ in Eq. (3) provides an avenue to encode our empirical knowledge as constraints on model parameters. In particular, we

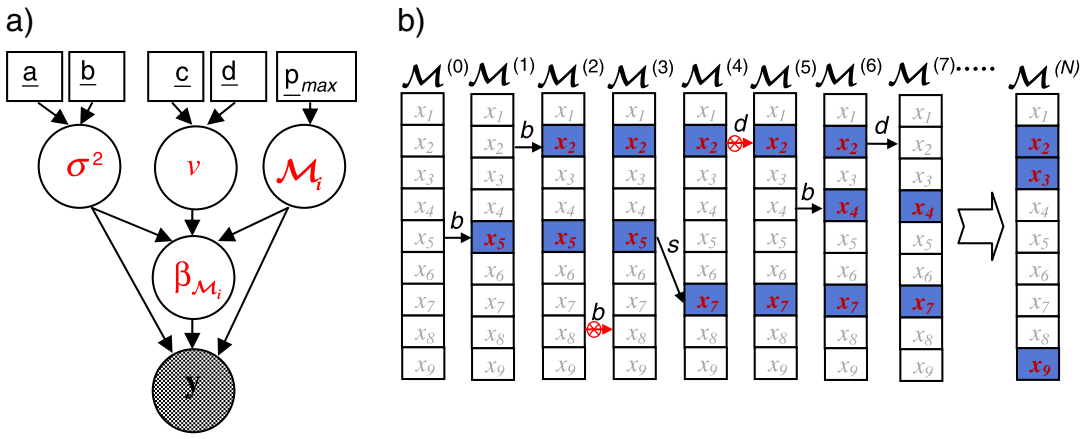


Fig. 1. (a) The hierarchical structure of our Bayesian regression model: The five underlined variables in the upper boxes are hyperparameters that specify priors and that should be prescribed a prior; the four red-highlighted variables in the circles are model parameters that are considered random and are of inferential interest. These model parameters are drawn using a hybrid Gibbs algorithm, with $\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2$, and v sampled from standard densities and the model configuration $\mathcal{M}_i$ sampled by a reversible-jump Monte Carlo Markov Chain (RJ-MCMC) sampler. (b) A simple hypothetical example using nine candidate covariates to illustrate the use of RJ-MCMC to draw model configuration $\mathcal{M}_i$: A MCMC chain is iterated N times, starting from a null model $\mathcal{M}^{(0)}$ (i.e., no covariates are chosen) and then making random local jumps from one model to the next by randomly applying one of three types of model proposal steps (i.e., b — birth, d — dead, and s — swap). Proposed models are accepted only with some probabilities and otherwise are rejected. For example, to generate $\mathcal{M}^{(3)}$ from $\mathcal{M}^{(2)}$, a "birth" step has been randomly chosen to add a randomly selected covariate x_8 into $\mathcal{M}^{(2)}$ to propose a new candidate, but this proposed candidate model was rejected (e.g., depicted by the red arrow with a crossed-circle); therefore, $\mathcal{M}^{(3)}$ is still the same as $\mathcal{M}^{(2)}$. In the chain, each sampled model $\mathcal{M}^{(t)}$ is represented by a column where the chosen covariates in $\mathcal{M}^{(t)}$ are blue-filled. Unlike conventional regression, Bayesian model averaging (BMA) uses all the sampled models of the chain for inference.

assumed $\pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right) = \pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2 \middle| \mathcal{M}_i\right)\pi(\mathcal{M}_i)$; therefore, it suffices to elicit separately the two conditional priors $\pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2 \middle| \mathcal{M}_i\right)$ and $\pi(\mathcal{M}_i)$, as detailed below.

First, the conditional prior of model coefficients $\boldsymbol{\beta}_{\mathcal{M}_i}$ and variance σ^2 on model structure $\mathcal{M}_i$, $\pi(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2|\mathcal{M}_i)$, was assigned a normal-inverse gamma distribution:

$$\begin{aligned}\pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2 \middle| \mathcal{M}_i\right) &= \pi_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_{\mathcal{M}_i} \middle| \sigma^2, \mathcal{M}_i\right)\pi_{\sigma^2}\left(\sigma^2\right) \\ &= \mathbb{N}\left(\boldsymbol{\beta}_{\mathcal{M}_i}; \mathbf{m}_{\mathcal{M}_i}, \sigma^2\mathbf{V}_{\mathcal{M}_i}\right)\mathbb{IG}\left(\sigma^2; \underline{a}, \underline{b}\right)\end{aligned} \tag{5}$$

where the conditional prior $\pi_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_{\mathcal{M}_i} \middle| \sigma^2, \mathcal{M}_i\right)$ is a normal density $\mathbb{N}\left(\boldsymbol{\beta}_{\mathcal{M}_i}; \cdot, \cdot\right)$, which depends on the model configuration $\mathcal{M}_i$ because the dimension of coefficient vector $\boldsymbol{\beta}_{\mathcal{M}_i}$, $\left(1 + p_{\mathcal{M}_i}\right) \times 1$, needs to be consistent with the number of covariates in $\mathcal{M}_i$; the prior $\pi_{\sigma^2}\left(\sigma^2\right)$ is an inverse-gamma density $\mathbb{IG}\left(\sigma^2; \underline{a}, \underline{b}\right)$ that is independent of $\mathcal{M}_i$ and is specified by two scalar hyperparameters $\underline{a}$ and $\underline{b}$. To parameterize the Gaussian prior $\pi_{\boldsymbol{\beta}}(\cdot) = \mathbb{N}\left(\boldsymbol{\beta}_{\mathcal{M}_i}; \mathbf{m}_{\mathcal{M}_i}, \sigma^2\mathbf{V}_{\mathcal{M}_i}\right)$, the prior mean $\mathbf{m}_{\mathcal{M}_i}$ is simply set to zeros, a justifiable choice if training data are centered to have zero means. The prior covariance $\sigma^2\mathbf{V}_{\mathcal{M}_i}$ needs to be further elicited with regard to the $\left(1 + p_{\mathcal{M}_i}\right) \times \left(1 + p_{\mathcal{M}_i}\right)$ matrix $\mathbf{V}_{\mathcal{M}_i}$; two choices for $\mathbf{V}_{\mathcal{M}_i}$ are often considered: ridge prior $\mathbf{V}_{\mathcal{M}_i} = v\mathbf{I}_{\mathcal{M}_i}$ and g-prior $\mathbf{V}_{\mathcal{M}_i} = v\left(\mathbf{X}_{\mathcal{M}_i}^T\mathbf{X}_{\mathcal{M}_i}\right)^{-1}$. In both cases, v is a scalar hyperparameter. Similar to Eq. (1), the g-prior $v\left(\mathbf{X}_{\mathcal{M}_i}^T\mathbf{X}_{\mathcal{M}_i}\right)^{-1}$ is problematic when the number of covariates in $\mathcal{M}_i$ is greater than the training sample size ($p_{\mathcal{M}_i} > \mathrm{n}$) because of the possible failure in inverting $\mathbf{X}_{\mathcal{M}_i}^T\mathbf{X}_{\mathcal{M}_i}$. Therefore, we chose the ridge prior $\mathbf{V}_{\mathcal{M}_i} = v\mathbf{I}_{\mathcal{M}_i}$. More interestingly, due to a lack of judicious value, the hyperparameter v in $\mathbf{V}_{\mathcal{M}_i}$ was also treated as random and was assigned an inverse-gamma prior $\pi_v(v) = \mathbb{IG}\left(v; \underline{c}, \underline{d}\right)$ with two hyperparameters $\underline{c}$ and $\underline{d}$. The prior $\pi_v(v)$ is called hyperprior because it is elicited at a level deeper than $\boldsymbol{\beta}_{\mathcal{M}_i}$ (Fig. 1). Subsequently, the full conditional prior of Eq. (5) can be re-expressed as

$$\pi\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, v \middle| \mathcal{M}_i; \underline{a}, \underline{b}, \underline{c}, \underline{d}\right) = \pi_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_{\mathcal{M}_i} \middle| \sigma^2, v, \mathcal{M}_i\right)\pi_{\sigma^2}\left(\sigma^2 \middle| \underline{a}, \underline{b}\right)\pi_v\left(v \middle| \underline{c}, \underline{d}\right) \tag{6}$$

where the hyperparameters $\underline{a}, \underline{b}, \underline{c}$ and $\underline{d}$ are underlined and made explicit for the respective priors.

Second, the model prior $\pi(\mathcal{M}_i)$ over the space $\mathcal{M}$, $i = 1, \cdots, 2^{\mathrm{p}}$, was chosen to be vague to reflect a lack of prior knowledge on which predictors or model configurations are useful, although modelers may elicit informative priors when such knowledge is available. In our Bayesian treatment, $\pi(\mathcal{M}_i)$ was specified with respect to the number of covariates selected into model $\mathcal{M}_i$, $p_{\mathcal{M}_i}$, which is a measure of model complexity or model dimension. Foremost, to preclude over-complicated models, we discarded those models with the number of selected covariates greater than a prescribed value $\underline{\mathrm{p}}_{\max}$; this preclusion is mathematically expressed as $\pi(\mathcal{M}_i) = 0$ if $p_{\mathcal{M}_i} > \underline{\mathrm{p}}_{\max}$. As a result, there are only $\left(\underline{\mathrm{p}}_{\max} + 1\right)$ model dimensions permissible, $p_{\mathcal{M}_i} \in \left\{0, 1, 2, \ldots, \underline{\mathrm{p}}_{\max}\right\}$, with $p_{\mathcal{M}_i} = 0$ indicating the null model (i.e., no covariate selected). For a given model dimension $p_{\mathcal{M}_i}$, there are totally $\binom{\mathrm{p}}{p_{\mathcal{M}_i}}$ possible models where p again is the total number of covariates in $\mathbf{x}$. We assumed that each of the $\left(\underline{\mathrm{p}}_{\max} + 1\right)$ allowable model dimension is equally possible a priori and that all the $\binom{p}{p_{\mathcal{M}_i}}$ models of the same dimension are equally likely a priori. Put together, our vague prior $\pi(\mathcal{M}_i)$ takes a discrete form of

$$\pi\left(\mathcal{M}_i \middle| \underline{\mathrm{p}}_{\max}\right) = \begin{cases} \dfrac{1}{\left(\underline{\mathrm{p}}_{\max} + 1\right) \cdot \binom{p}{p_{\mathcal{M}_i}}} & \text{if } p_{\mathcal{M}_i} \le \underline{\mathrm{p}}_{\max} \\ 0 & \text{if } p_{\mathcal{M}_i} > \underline{\mathrm{p}}_{\max} \end{cases}, \quad \mathcal{M}_i \in \mathcal{M}. \tag{7}$$

The maximum number of covariates allowed in models, $\underline{\mathrm{p}}_{\max}$, is a hyperparameter that should be pre-specified.

Finally, as a recap of the Bayesian model, the likelihood Eq. (4) and the priors in Eqs. (5) and (7) combine to reach the full formulation of the posterior according to Eq. (3):

$$p\left(\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,v,\mathcal{M}_i|\mathcal{D}\right) \propto p\left(\mathcal{D}\middle|\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,\mathcal{M}_i\right)\pi_\beta\left(\boldsymbol{\beta}_{\mathcal{M}_i}\middle|\sigma^2,v,\mathcal{M}_i\right)\pi_{\sigma^2}\left(\sigma^2\middle|\underline{a},\underline{b}\right)\pi_v\left(v\middle|\underline{c},\underline{d}\right)\pi\left(\mathcal{M}_i\middle|\underline{p}_{\max}\right). \tag{8}$$

The model parameters $\{\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,v,\mathcal{M}_i\}$ are of inferential interest and are all considered random. In contrast, the five hyperparameters $\left\{\underline{a},\underline{b},\underline{c},\underline{d},\underline{p}_{\max}\right\}$ are some fixed values and should be pre-specified empirically, although it is also permissible to treat them as random variables by further eliciting hyperprior distributions for them at higher levels in a manner similar to the treatment of v. No general rules exist on how to prescribe values of the hyperparameters. Our setup was chosen as $\underline{a}=\underline{b}=0.01$, $\underline{c}=\underline{d}=0.02$, and $\underline{p}_{\max}=\max(2n,m/10)$ with n and m being the numbers of observations and spectral bands, respectively. Such choices for the inverse gamma priors $\pi_{\sigma^2}(\sigma^2|\underline{a},\underline{b})$ and $\pi_v(v|\underline{c},\underline{d})$ are practically equivalent to non-informative priors, reflecting our vague knowledge on σ^2 and v a priori. Preliminary trials with various datasets suggest that the resulting predictive performances are insensitive to the settings of these hyperparameters as long as $\underline{a}$, $\underline{b}$, $\underline{c}$ and $\underline{d}$ take small values, $\underline{p}_{\max}$ assumes a moderately large value (e.g., >100), and data are standardized beforehand.

3.3. Extended model space for nonlinear regression

The Bayesian linear regression formulated above can be extended to model nonlinear relationships by expanding the set of covariates and augmenting the model space through the use of customized nonlinear transformations of raw spectral measurements. The transformations can take any forms and, by analogy to spectral indices, may involve a varying number of bands. Specifically, given a spectrum with m contiguous bands $\mathbf{r}=[r_1,r_2,\cdots,r_m]^T$, we can derive a new set of covariates $\mathbf{x}=[x_1,x_2,\cdots,x_p]^T$ wherein each elemental covariate x can be either the original spectral reflectance r itself or some transformed variables. Simple examples of transformations, in the form of hyperspectral indices, include $x=\log(1/r_i)$, $x=r_i/r_j$, $x=(r_i-r_j)/(r_i+r_j)$, $x=(r_i-r_k)/(r_i-r_k)$, and $x=r_i^\alpha$. Herein, α is a prescribed power parameter, and the band indices i, j, or k can be any out of the m wavebands of spectrum $\mathbf{r}$ that make transformations numerically valid. The total number of newly derived covariates in $\mathbf{x}$ can be much larger than the number of the original spectral bands (i.e., p > m); p may even be infinity in extreme cases, e.g., when the power parameter α in r_i^α takes continuous values. As a result, the augmented space of model configuration spanned by $\mathbf{x}$, $\mathcal{M}$, is enormously huge or even infinite in size, which cannot be tackled with conventional methods such as SMR and PLS.

The aforementioned use of various transformations to derive new covariates is a common strategy adopted in generalized additive models to account for nonlinearity. Apart from hyperspectral indices, the transformations for expanding model space in a more general setting can be any explicit or implicit functions, even including complex nonlinear functions such as multivariate adaptive regression spline and neural networks. Moreover, the transformations can involve far more than a few original bands; for example, full spectra can be analyzed with 1-D signal processing techniques such as wavelets to derive new covariates. Because our primary purpose is to compare BMA against PLS and SMR, we restricted our analyses to the linear model space of the original reflectance bands in most of the subsequent experiments, except in Section 5.6 where we did examine six possible transformations in the form of hyperspectral indices to illustrate the flexibility of our Bayesian method.

3.4. Monte Carlo implementation

The mathematical intractability of our Bayesian model precludes an analytical treatment to the posterior inference of $\{\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,v,\mathcal{M}_i\}$ in Eq. (8). Instead, these model parameters were estimated by sampling the posterior $p(\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,v,\mathcal{M}_i|\mathcal{D})$ using MCMC procedures. The MCMC algorithm we employed is a hybrid sampler that embeds a reversible-jump MCMC sampler (RJ-MCMC) into a Gibbs sampling framework. To apply the Gibbs sampling, the full posterior $p(\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,v,\mathcal{M}_i|\mathcal{D})$ of Eq. (8) is decomposed into three component conditional distributions that are separately sampled in three sequential Gibbs steps:

$$\begin{aligned}
&\text{Step 1. } p(\mathcal{M}=\mathcal{M}_i|v,\mathcal{D})\propto p(\mathcal{D}|v,\mathcal{M}_i)\pi\left(\mathcal{M}_i\middle|\underline{p}_{\max}\right),\ i=1,\cdots,2^p\\
&\text{Step 2. } p\left(\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2|v,\mathcal{M}_i,\mathcal{D}\right)\\
&\quad=\mathbb{N}\left(\boldsymbol{\beta}_{\mathcal{M}_i};\mathbf{V}^*_{\mathcal{M}_i}\mathbf{X}^T_{\mathcal{M}_i}\mathbf{y},\sigma^2\mathbf{V}^*_{\mathcal{M}_i}\right)\cdot\mathbb{IG}\left(\sigma^2;\underline{a}+\frac{n}{2},\underline{b}+\left[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}_{\mathcal{M}_i}\mathbf{V}^*_{\mathcal{M}_i}\mathbf{X}^T_{\mathcal{M}_i}\mathbf{y}\right]/2\right);\\
&\text{and}\\
&\text{Step 3. } p\left(v\middle|\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,\mathcal{M}_i,\mathcal{D}\right)=\mathbb{IG}\left(v;\underline{c}+\frac{p_{\mathcal{M}_i}}{2},\underline{d}+\frac{\sum_{k=1}^{p_{\mathcal{M}_i}}\beta^2_{k,\mathcal{M}_i}}{2}\right).
\end{aligned} \tag{9}$$

In Eq. (9), $\mathbf{V}^*_{\mathcal{M}_i}=\left(v^{-1}\mathbf{I}_{\mathcal{M}_i}+\mathbf{X}^T_{\mathcal{M}_i}\mathbf{X}_{\mathcal{M}_i}\right)^{-1}$; $\beta_{k,\mathcal{M}_i}$ is the kth element of $\boldsymbol{\beta}_{\mathcal{M}_i}$; and $p(\mathcal{D}|v,\mathcal{M}_i)$ is the conditional marginal likelihood that, due to the conjugacy of the prior $\pi(\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2|\mathcal{M}_i)$, has an analytical form as given in Denison (2002).

In the Gibbs algorithm of Eq. (9), the three conditional posteriors are sampled sequentially for a total of N iterations to generate a MCMC chain of samples $\left\{\mathcal{M}^{(t)},\boldsymbol{\beta}^{(t)}_{\mathcal{M}},\sigma^{2(t)},v^{(t)}\right\}_{t=1,\cdots,N}$. Within each Gibbs iteration t, Steps 2 and 3 for sampling $\left\{\boldsymbol{\beta}^{(t)}_{\mathcal{M}},\sigma^{2(t)},v^{(t)}\right\}$ are straightforward because the two densities $p(\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2|v,\mathcal{M}_i,\mathcal{D})$ and $p(v|\boldsymbol{\beta}_{\mathcal{M}_i},\sigma^2,\mathcal{M}_i,\mathcal{D})$ are standard distributions. However, Step 1 for sampling $\mathcal{M}^{(t)}$ is difficult for two reasons: (1) $p(\mathcal{M}=\mathcal{M}_i|v,\mathcal{D})$ is defined only up to an unknown proportionality constant and (2) the number of covariates in $\mathcal{M}_i$ can vary from one model to another. These difficulties are tackled using RJ-MCMC for sampling $p(\mathcal{M}=\mathcal{M}_i|v,\mathcal{D})$ in Step 1 of the Gibbs algorithm, as briefly described below.

To sample $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,N}$ from $p(\mathcal{M}=\mathcal{M}_i|v,\mathcal{D})$ for Step 1 of Eq. (9), the RJ-MCMC algorithm traverses the model space $\mathcal{M}$ by jumping locally from one model to another via a proposal move (Fig. 1b). Assuming that the current iteration is t-th, the proposed move from the current model $\mathcal{M}^{(t)}$ yields a candidate model $\mathcal{M}^{(t)}_{prop}$ to be accepted as the model of the next iteration $\mathcal{M}^{(t+1)}$ with a designated probability ρ so that all the so-traversed models $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,N}$ constitute a random sample from $p(\mathcal{M}|v,\mathcal{D})$. The candidate model $\mathcal{M}^{(t)}_{prop}$ is proposed from $\mathcal{M}^{(t)}$ by randomly applying one of the following three moves (Fig. 1b) – "birth", "death", or "swap" – each chosen with an equal probability of 1/3:

- the "birth" move randomly adds a covariate that is absent from $\mathcal{M}^{(t)}$;
- the "death" move randomly removes an existing covariate from $\mathcal{M}^{(t)}$;
- the "swap" move randomly replaces an existing covariate of $\mathcal{M}^{(t)}$ with another randomly selected covariate absent from $\mathcal{M}^{(t)}$.

Assuming that $\mathcal{M}^{(t)}$ contains a total of $p^{(t)}$ covariates, the three types of moves will result in $\mathcal{M}^{(t)}_{prop}$ with the number of covariates being $p^{(t)}+1$, $p^{(t)}-1$, and $p^{(t)}$, respectively. The proposal $\mathcal{M}^{(t)}_{prop}$ is

accepted as the model of the next iteration $\mathcal{M}^{(t+1)}$ with a probability of

$$\rho = \min\left(1, \frac{p\left(\mathcal{D}|v^{(t)}, \mathcal{M}_{prop}^{(t)}\right)}{p\left(\mathcal{D}|v^{(t)}, \mathcal{M}^{(t)}\right)}\right); \tag{10}$$

and in case of a rejection that occurs with a probability of $1-\rho$, the current model $\mathcal{M}^{(t)}$ will be carried over to the next iteration as $\mathcal{M}^{(t+1)}$. That is,

$$\mathcal{M}^{(t+1)} = \begin{cases} \mathcal{M}_{prop}^{(t)} \text{ with a probability } \rho \\ \mathcal{M}^{(t)} \text{ otherwise} \end{cases}. \tag{11}$$

Two caveats should be noted when executing the RJ-MCMC algorithm. First, the "birth" move is forbidden when $p^{(t)} = \mathrm{p}_{\max}$, and the "death" and "swap" moves are forbidden when $p^{(t)} = 0$; in such cases, the other viable moves may be used instead. Second, the probabilities of choosing the three proposal moves do not necessarily all equal 1/3; if they take different values, the acceptance probability ρ of Eq. (10) needs to be redressed to reflect such differences. More details on RJ-MCMC can be found in Fan and Sisson (2010).

The reliability in using a MCMC chain of samples $\left\{\mathcal{M}^{(t)}, \boldsymbol{\beta}_{\mathcal{M}}^{(t)}, \sigma^{2(t)}, v^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$ for inference is contingent largely on how well the chain converges to its stationary distribution $p(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, v, \mathcal{M}_i|\mathcal{D})$. Monitoring the convergence with simple diagnostic statistics is often difficult. As a practical remedy, we ran the hybrid MCMC sampler sufficiently long to safeguard against instability. The length of individual chains in our experiments was set to 60,000 iterations with the first 10,000 being discarded as burn-in samples. We also thinned chains by retaining only every fifth sample. Moreover, a total of five such chains were run in parallel, thinned, and lastly merged into a final chain of samples $\left\{\mathcal{M}^{(t)}, \boldsymbol{\beta}_{\mathcal{M}}^{(t)}, \sigma^{2(t)}, v^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$. In the iterating processes, chains all started from the null model (i.e., no covariate being selected), although other choices are allowable. Graphical diagnostics of some preliminary trials using data from various sources have suggested that the chains mixed rapidly and in most cases reached stability after only a few thousand iterations. Therefore, our use of five parallel chains, each with 60,000 iterations, is sufficient to ensure convergence, and the long length of the chains also helps to obliterate the effect of initial model choices on the chains.

3.5. Posterior inference and prediction

An MCMC chain $\left\{\mathcal{M}^{(t)}, \boldsymbol{\beta}_{\mathcal{M}}^{(t)}, \sigma^{2(t)}, v^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$ generated by the preceding hybrid sampler contains all information such as covariates chosen and model coefficients that are necessary for statistical inference and predictive analysis. Apparently, the final inference and analysis using BMA are based not on a particular model $\mathcal{M}^{(t)}$ but instead on all the sampled models $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$. The predicted response y^* for a new spectrum $\mathbf{r}^*$ is the average of individual predictions over the sampled models $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$:

$$\begin{aligned} \hat{y}^* &= \frac{\sum_{i=1}^{\mathrm{q}} \int y^* p\left(y^* \middle| \mathbf{r}^*, \boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i\right) p\left(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \mathcal{M}_i \middle| v, \mathcal{D}\right) p\left(v \middle| \mathcal{D}\right) dy^* d\boldsymbol{\beta}_{\mathcal{M}_i} d\left(\sigma^2\right) dv}{2^{\mathrm{q}}} \\ &\approx \frac{\sum_{t=1}^{\mathrm{N}} \left(\mathbf{x}^{*(t)}\right)^{\mathrm{T}} \left[\left(v^{(t)}\right)^{-1} \mathbf{I}_{\mathcal{M}^{(t)}} + \left(\mathbf{X}_{\mathcal{M}^{(t)}}^{(t)}\right)^T \mathbf{X}_{\mathcal{M}^{(t)}}^{(t)}\right]^{-1} \mathbf{X}_{\mathcal{M}^{(t)}}^{(t)} \mathbf{y}}{\mathrm{N}} \end{aligned} \tag{12}$$

along with its associated prediction error

$$\hat{\sigma}_{\hat{y}^*}^2 \approx \frac{\sum_{t=1}^{\mathrm{N}} \left[\left(\mathbf{x}^{*(t)}\right)^T \boldsymbol{\beta}_{\mathcal{M}}^{(t)} - \hat{y}^*\right]^2}{\mathrm{N}} + \frac{\sum_{t=1}^{\mathrm{N}} \sigma^{2(t)}}{\mathrm{N}} \tag{13}$$

where $\mathbf{x}^*(t)$ is the vector of predictors constructed from the reflectance spectrum $\mathbf{r}^*$ according to the model configuration $\mathcal{M}^{(t)}$. The prediction equation of Eq. (12) corresponds to the weighted average model inferred by our BMA scheme.

The MCMC chain of sampled models $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$ also encapsulates valuable information for diagnosing model structures, especially for quantifying the importance of each waveband for explaining variations in observed responses (Fig. 1). Because each $\mathcal{M}^{(t)}$ represents a subset of covariates selected as predictors, one possible indicator of the relative variable/band importance is the marginal probability of a covariate x_i being included into BMA, which can be obtained by counting the relative occurrence frequency of x_i in $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$:

$$p(x_i \in \mathcal{M}|\mathcal{D}) \approx \frac{\textit{Number of } \mathcal{M}^{(t)} \text{ that includes } x_i}{\mathrm{N}}, \mathrm{i} = 1, 2, \cdots, \mathrm{p}. \tag{14}$$

Intuitively, the larger the inclusion probability $p(x_i \in \mathcal{M}|\mathcal{D})$ is, the more important role the covariate x_i is likely to play in predicting the response variable. Additionally, the mean number of covariates included in the BMA model provides a measure of model complexity and it can be simply obtained by averaging the numbers of chosen covariates across $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$, namely $\sum_{\mathrm{t}=1}^{\mathrm{N}} p^{(t)}/\mathrm{N}$.

The individual contributions of bands to a model's overall predictive power have also been assessed by comparing the relative magnitudes of regression coefficients. For example, the magnitudes of fitted coefficients in PLS models have been treated as crude indicators to quantify the relative importance of bands, and the so-identified band importance was reported to accord roughly with the absorption signatures of chemicals examined (Bolster et al., 1996). In our BMA regression, the sampled coefficient $\boldsymbol{\beta}_{\mathcal{M}}^{(t)}$ for a sampled model $\mathcal{M}^{(t)}$ is $(p^{(t)}+1)\times 1$ in size, and it can be extended to a fullsize $(\mathrm{m}+1)\times 1$ vector $\boldsymbol{\beta}_{full}^{(t)}$ by zero-padding the positions at those bands not chosen by $\mathcal{M}^{(t)}$. Then, the averaging of $\boldsymbol{\beta}_{full}^{(t)}$ over $\left\{\mathcal{M}^{(t)}\right\}_{t=1,\cdots,\mathrm{N}}$ results in the coefficient $\hat{\boldsymbol{\beta}}_{BMA}$ for the final BMA prediction model:

$$\hat{\boldsymbol{\beta}}_{BMA} = \frac{\sum_{\mathrm{t}=1}^{\mathrm{N}} \boldsymbol{\beta}_{full}^{(t)}}{\mathrm{N}} \tag{15}$$

which can be plotted graphically as a function of wavelength to reveal the band features important for explaining the response variable in question.

4. Benchmark analyses

4.1. Partial least squares (PLS) and stepwise multiple regression (SMR)

For comparison with the BMA regression, we examined two alternative methods — PLS and SMR. Unlike BMA, these two traditional methods employ some stopping rules to find a single best model configuration. In PLS, the number of factors entering into a final model needs to be carefully chosen to avoid excessive overfitting; we employed a LOOCV scheme to tune the number of chosen factors by minimizing the predicted residual sums of squares (PRESS). One practical difficulty with the use of PLS is a lack of simple analytical methods to estimate prediction errors, and we tackled this via a bootstrapping method. Specifically, prediction errors with PLS are assumed to comprise two independent components, one

associated with model uncertainty in $\boldsymbol{\beta}$ and another with the noise error ε:

$$\hat{\sigma}^2_{\text{pls}} = \hat{\sigma}^2_{\boldsymbol{\beta}} + \hat{\sigma}^2_{\varepsilon} \tag{16}$$

where $\hat{\sigma}^2_{\varepsilon}$ was computed directly from residuals of the fitted PLS model, and $\hat{\sigma}^2_{\boldsymbol{\beta}}$ was estimated as the sample variance of 1000 bootstrapped predictions that were generated by re-fitting PLS models upon 1000 bootstrap samples of the original training data. In contrast, procedures of variable selection and error estimation for SMR have long become standardized and thus are not described here.

4.2. Statistical measures for model evaluation

Predictive performances of BMA, PLS and SMR were assessed with respect to three statistics, including coefficients of determination (i.e., R^2, interpreted as the percentage of variance in the test data explained by predictions), root mean squared error (RMSE), and prediction interval coverage probability (PICP). In our analyses of Section 5, these three model evaluation statistics were computed from out-of-sample validation data not from training data. Of the three statistics, R^2 and RMSE are common criteria for evaluating point estimation $\hat{y}_i$; in contrast, PICP is a measure for evaluating uncertainty estimation $\hat{\sigma}_i$, namely how well the estimated standard deviation $\hat{\sigma}_i$ reflects the true observed uncertainty in $\hat{y}_i$. The use of PICP is rare in remote sensing literature probably because evaluations of uncertainty estimation are not frequently practiced in remote sensing research.

By definition, PICP represents the probability that prediction intervals $[\hat{y}_i - t_{1-\alpha/2}\hat{\sigma}_i, \hat{y}_i + t_{1-\alpha/2}\hat{\sigma}_i]$ cover actual observations y_i. It is simply computed according to

$$\text{PICP}_{1-\alpha} = \sum_{i=1}^{n} \text{I}\left(|y_i - \hat{y}_i| < t_{1-\alpha/2}\hat{\sigma}_i\right) \Big/ \text{n} \tag{17}$$

where n is the number of observations in the validation data, $\hat{y}_i$ is the prediction of the ith observation y_i with $\hat{\sigma}_i$ being its estimated error, $t_{1-\alpha/2}$ is the $(1-\alpha/2)100\%$ percentile of a t-distribution with a proper degree of freedom, and $\text{I}(\cdot)$ is the indicator function that equals 1 if the condition is satisfied and 0 otherwise. A $\text{PICP}_{1-\alpha}$ value closer to its nominal value $(1-\alpha)100\%$ implies more realistic estimation of $\hat{\sigma}_i$. If $\text{PICP}_{1-\alpha}$ is plotted as a function of $(1-\alpha)100\%$, the resultant curve is expected to coincide with the 1:1 line in an ideal situation. We mainly reported the PICP statistic associated with a significance level of 95% (i.e., $\alpha = 0.05$), as denoted by PICP95.

5. Experiments and results

To evaluate the utility of BMA for estimating foliage biochemicals, we designed seven experiments using either an artificial dataset (Section 5.1) or the 27 spectral–chemical datasets (Sections 5.2–5.7). Acronyms of the 27 datasets are summarized in Table 1. The emphasis of these experiments is on comparing BMA against PLS and SMR while assessing the extent to which hyperspectral data can be used to estimate multiple biochemicals. We did not utilize all the data for every experiment, due to the relatively small sample sizes for some biochemicals and the disparities in data acquisition protocols among the three sources or simply due to a space limit on reporting all results.

5.1. An ideal experiment with artificial data

We first synthesized an artificial dataset for model comparison. To generate the synthetic data, 200 curves of Gaussian processes were simulated using a Fourier transform-based algorithm with an exponential covariance function and were discretized evenly at 500 locations, mimicking 200 observations of 500 correlated covariates.

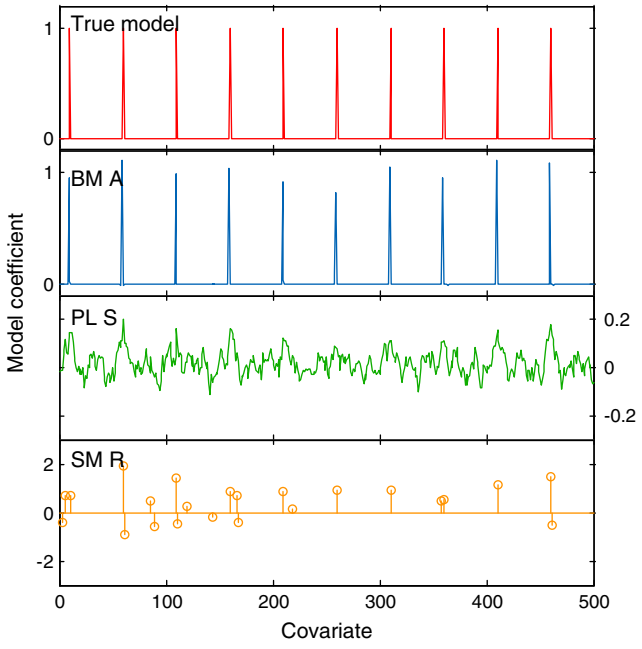


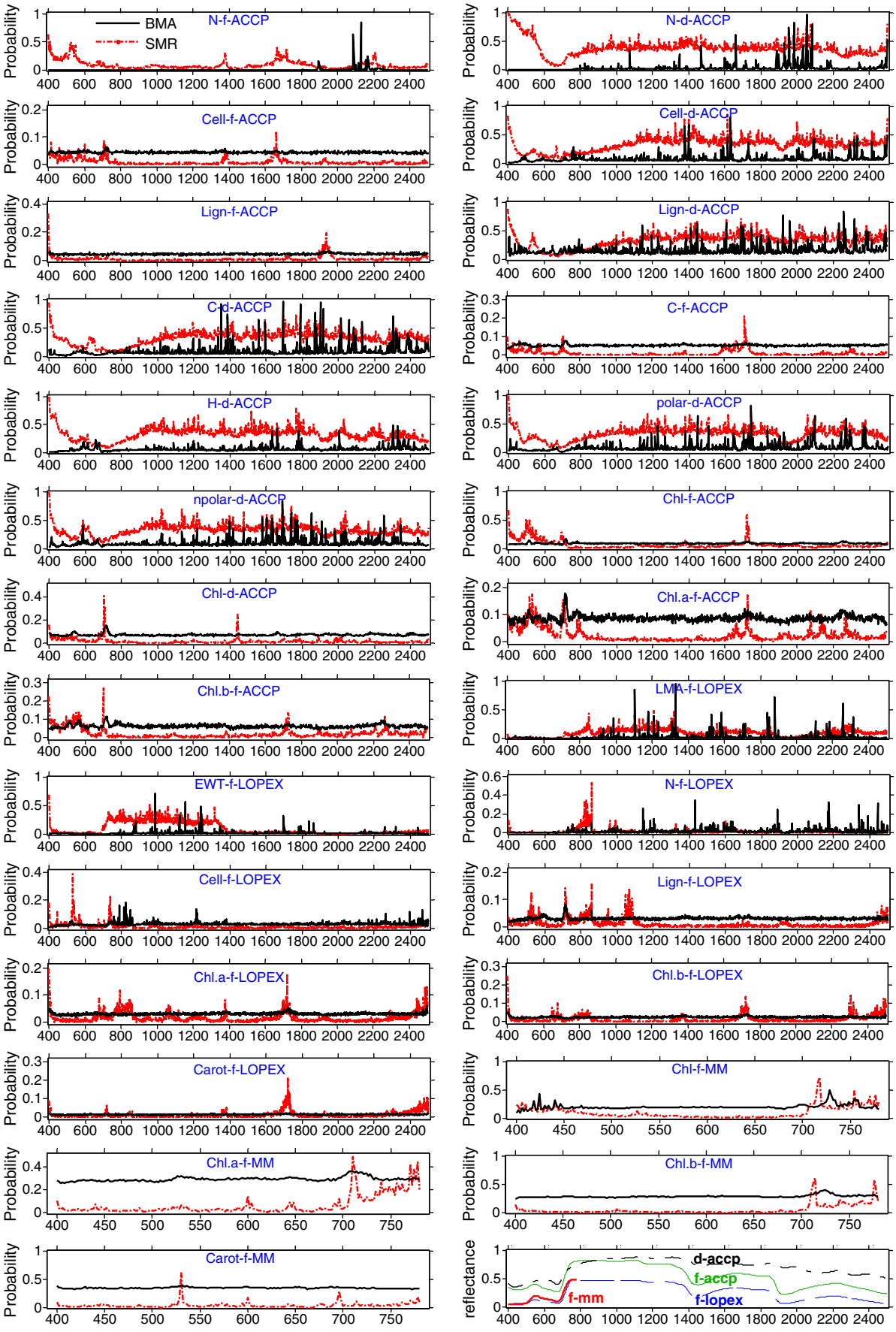
Fig. 2. Comparisons of BMA, PLS and SMR models fitted to the synthetic data of Section 5.1 in reference to the true model (top) (i.e., $y = 1.0 \cdot x_1 + 1.0 \cdot x_{51} + \cdots + 1.0 \cdot x_{451} + \varepsilon$). Models are depicted here as graphs of coefficients of the 500 covariates. Only BMA uncovered the true model whereas PLS and SMR, though with good predictive performances, essentially inferred "wrong" models.

The response variable was set to the sum of every fiftieth covariate plus some corrupting noises with $\sigma^2_{\varepsilon} = 1$; hence, the true model can be written as $y = x_1 + x_{51} + x_{101} \cdots + x_{451} + \varepsilon$, with coefficients being 1's for the ten selected covariates and being zeros for the rest (Fig. 2). The 200 curves were randomly halved into a training and a validation subset. Such well-controlled data provide a true statistical relationship that allows us to confidently evaluate and compare the inferential power of BMA, PLS and SMR.

All the three regression methods predicted the simulated response variable well, with R^2 values all greater than 0.98 when tested upon the validation data. However, PLS and SMR greatly overfitted the data. Their estimated errors $\hat{\sigma}^2_{\varepsilon}$ were 0.10 and 0.00, respectively, compared to 0.976 estimated by BMA; only BMA estimated the true value $\sigma^2_{\varepsilon} = 1$ well. Hence, BMA yielded reasonable estimation of prediction errors with a PICP95 value of 93%. In contrast, the respective PICP95 values for PLS and SMR were 58% and 0%, both of which deviate markedly from their nominal value of 95%. Further, a closer examination of the fitted model coefficients reveals that only BMA approximated the true model with high fidelity (Fig. 2). Despite the good predictive performances of PLS and SMR (i.e., $R2 > 0.98$), their fitted models were essentially "wrong" in reference to the true underlying model (Fig. 2). SMR chose 23 covariates as predictors; PLS incorporated essentially all the 500 covariates; and BMA chose an average of 10.3 predictors, close to the true number of 10.

5.2. Band selection probability

Model diagnostics for BMA and SMR were analyzed and compared in terms of identifying important bands for predicting each biochemical. In this experiment, raw reflectances of the spectral–chemical datasets served as candidate predictors without any band transformation. Because no model validation was required here, the entire samples in each of the 27 spectral–chemical datasets were used for model calibration. Band selection probabilities for BMA were estimated using Eq. (14) from MCMC-sampled model structures. Band selection by SMR is sensitive to the randomness associated with the compilation of training data. To account for such variability as well


BMA
SMR
N-f-ACCP
N-d-ACCP
Cell-f-ACCP
Cell-d-ACCP
Lign-f-ACCP
Lign-d-ACCP
C-d-ACCP
C-f-ACCP
H-d-ACCP
polar-d-ACCP
npolar-d-ACCP
Chl-f-ACCP
Chl-d-ACCP
Chl.a-f-ACCP
Chl.b-f-ACCP
LMA-f-LOPEX
EWT-f-LOPEX
N-f-LOPEX
Cell-f-LOPEX
Lign-f-LOPEX
Chl.a-f-LOPEX
Chl.b-f-LOPEX
Carot-f-LOPEX
Chl-f-MM
Chl.a-f-MM
Chl.b-f-MM
Carot-f-MM
d-accp
f-accp
f-mm
f-lopex
Probability
reflectance

as to develop a method to estimate band selection probability for SMR, we bootstrapped 10,000 samples for each spectral–chemical dataset and accordingly replicated the fitting of SMR models 10,000 times. The frequency of a band occurring in these 10,000 bootstrapped SMR models was counted and deemed as a heuristic probability measure to quantify the importance of the band as gauged by SMR.

The band importance, which is quantified as the probability of each band being chosen, differed strikingly between SMR and BMA (Fig. 3). Important bands generally correspond to local spikes in the probability graphs of Fig 3. In some cases, the identified local-spike bands coincided between BMA and SMR; for example, 712 nm peaked for both models when estimating lignin with the LOPEX data (Lign-f-LOPEX), but the associated probabilities differed slightly, being 0.08 for BMA and 0.14 for SMR. Although some common bands were identified for estimating the same chemical using the different datasets, discrepancies in band importance were frequently observed among different datasets. For example, when estimating nitrogen from spectra of fresh leaves, the useful bands identified by BMA for the ACCP data (i.e., N-f-ACCP) mostly fell within 1850–2200 nm, including 2128, 2086, 2166, 2172, 2156, 2116, 2098, and 1892 nm whereas those identified for the LOPEX data (N-f-LOPEX) within this range included 2178, 2175, 2170, 2178, 2155, 2061, 2065, 2094,2104,2122, 2155, 1970, 1897, 1876 and 1889 nm. Beyond this spectral range, a few extra bands were also identified as useful for the N-f-LOPEX data, including 1431, 1143, 1204, 1185, 860, and 753 nm. Moreover, important bands selected by BMA or SMR depend on the water status of leaves. For example, when estimating nitrogen, cellulose and lignin, BMA identified more bands for spectra of dry leaves than for spectra of fresh leaves (Fig. 3).

It may come as a surprise to find that the band selection probabilities obtained by stepwise regression were nonzero for all bands in the probability graphs of Fig. 3 for all the 27 spectral–chemical datasets; that is, every band was likely to be chosen by SMR as useful predictor for the same problem if we consider the randomness associated with the training data. Even more surprisingly, in our results for the biochemicals with training datasets larger than 100 in size, the associated 10,000 SMR models based on the bootstrap samples were all unique in terms of the included bands. These observations suggest that caution should be exercised when interpreting a stepwise regression model in the context of hyperspectral applications. In addition, these SMR results reinforce our motivation of using BMA: All the models are relevant and useful to a varying degree for estimating biochemicals.

5.3. Band importance as evidenced in model coefficients

We further evaluated band importance in terms of the magnitudes of model coefficients for the three regression methods. As described in Section 3.5, prediction equations of PLS, BMA and SMR all reduce to a final form of $\hat{y} = \mathbf{r}^{T} \cdot \hat{\boldsymbol{\beta}}$ when reflectance spectra $\mathbf{r}$ are used as predictors. Using the entire samples of each dataset, we fitted the three types of regression models and graphed the respective model coefficients $\hat{\boldsymbol{\beta}}$ as a function of wavelength. Prior to model fitting, data have been standardized to have zero means and unit variances at each band so that the intercept term in the models is always zero and the magnitudes of the fitted coefficients $\hat{\boldsymbol{\beta}}$ are more indicative of the relative usefulness of each band. The closer to zero a coefficient is, the less useful the associated band is. To avoid a proliferation of figures, only eight representative graphs are depicted in Fig. 4.

In most of our results, the curves of coefficients $\hat{\boldsymbol{\beta}}$ versus wavelengths diverged substantially among the three regression methods (Fig. 4). The curve of SMR $\hat{\boldsymbol{\beta}}_{SMR}$ differed inherently from those of BMA and PLS because the elements of $\hat{\boldsymbol{\beta}}_{SMR}$ are nonzero only at the selected bands whereas the curves of $\hat{\boldsymbol{\beta}}_{BMA}$ and $\hat{\boldsymbol{\beta}}_{PLS}$ are more or less continuous with respect to wavelength. Moreover, the curves of BMA $\hat{\boldsymbol{\beta}}_{BMA}$ were generally less smoother than those of PLS $\hat{\boldsymbol{\beta}}_{PLS}$. This contrast partially suggests that BMA was more parsimonious than PLS in terms of the effective number of selected predictors, although the actual number of factors selected by PLS was often smaller than the mean number of bands selected by BMA (Fig. 4). We also calculated the sum of absolute values of model coefficients over all bands $\left\|\hat{\boldsymbol{\beta}}\right\|$, defined here as the L-1 norm of $\hat{\boldsymbol{\beta}}$. In most cases, BMA had the smallest L-1 norm whereas SMR had the largest one. Using the three models fitted upon the Lign-f-ACCP data as an example, the L-1 norm of $\hat{\boldsymbol{\beta}}$ was 82.6, 18.1 and 1.9 for SMR, PLS and BMA, respectively, although the number of chosen bands or factors was nine for both SMR and PLS and was on average 48 for BMA. The smallest $\left\|\hat{\boldsymbol{\beta}}\right\|$ observed for BMA is consistent with the less continuity of the curve $\hat{\boldsymbol{\beta}}_{BMA}$ and, in a broad sense, aligns with the tenet of many regression techniques such as Lasso that penalize large coefficients to improve model generality. Our results also reveal a good correspondence between BMA and PLS in terms of bands of peak magnitudes (Fig. 4).

5.4. Prediction of biochemicals

We tested the performances of BMA in predicting various biochemicals, compared to PLS and SMR. In this test, raw reflectance spectra without any band transformations were again used as predictors. Each spectral–chemical dataset was randomly partitioned into two subsets with a splitting ratio of 2:1 — the two-thirds subset for model calibration, and the one-third subset for validation. To avoid any fortuitous results associated with a particular splitting, the partition of each dataset was randomly repeated 50 times. Then, each resultant partition was used for model fitting and validation. This replication correspondingly generated 50 values for a model evaluation criterion (e.g., R^2) and therefore permitted checking the statistical significance of the improvement achieved by one method over the other via a paired t-test. This test procedure, though less theoretically sound, offers a pragmatic expediency because other well-established techniques for comparing R^2 or RMSE between modeling approaches are still lacking.

Table 2 summarizes the three validation statistics (i.e. R^2, RMSE, and PICP95) for all the 27 spectral–chemical datasets. These reported statistics represent the averages over the 50 runs of random splitting for each dataset. According to R^2 and RMSE, both BMA and PLS markedly outperformed SMR; BMA in most cases was superior to PLS and in a few cases was at least as good as PLS (Fig. 5). Only in the case of Lign-d-ACCP did PLS perform better than BMA, as indicated by the paired t-test comparing the respective R^2 values of 0.883 vs. 0.874 or RMSE values of 1.53% vs. 1.58% (p-value$<$0.001); however, such a difference may be of little practical significance. As shown for predicting the same biochemical from different spectral samples of the ACCP data (Table 2), higher accuracies were usually achieved using dry leaf spectra than fresh leaf spectra. With regard to predictive powers for individual biochemicals, results show that nitrogen, chlorophyll, LMA and EWT could be estimated by linear models with reasonable accuracies but the estimation of lignin, cellulose and carotenoid from fresh leaf spectra were much less accurate or even unsuccessful, regardless of the regression methods. For example,

Fig. 3. Waveband selection probability was obtained as a quantitative measure of band importance for SMR (red dashed line) and BMA (black solid line); the selection probability (y axis) is plotted here as a function of wavelength (x axis) for all the spectral–chemical datasets considered (see Table 1 for acronyms of the spectral–chemical data such as C-f-ACCP and Chl-f-MM). Note that the last subfigure at the bottom right depicts the mean spectra of data from the three sources (d: dry leaves, and f: fresh leaves).

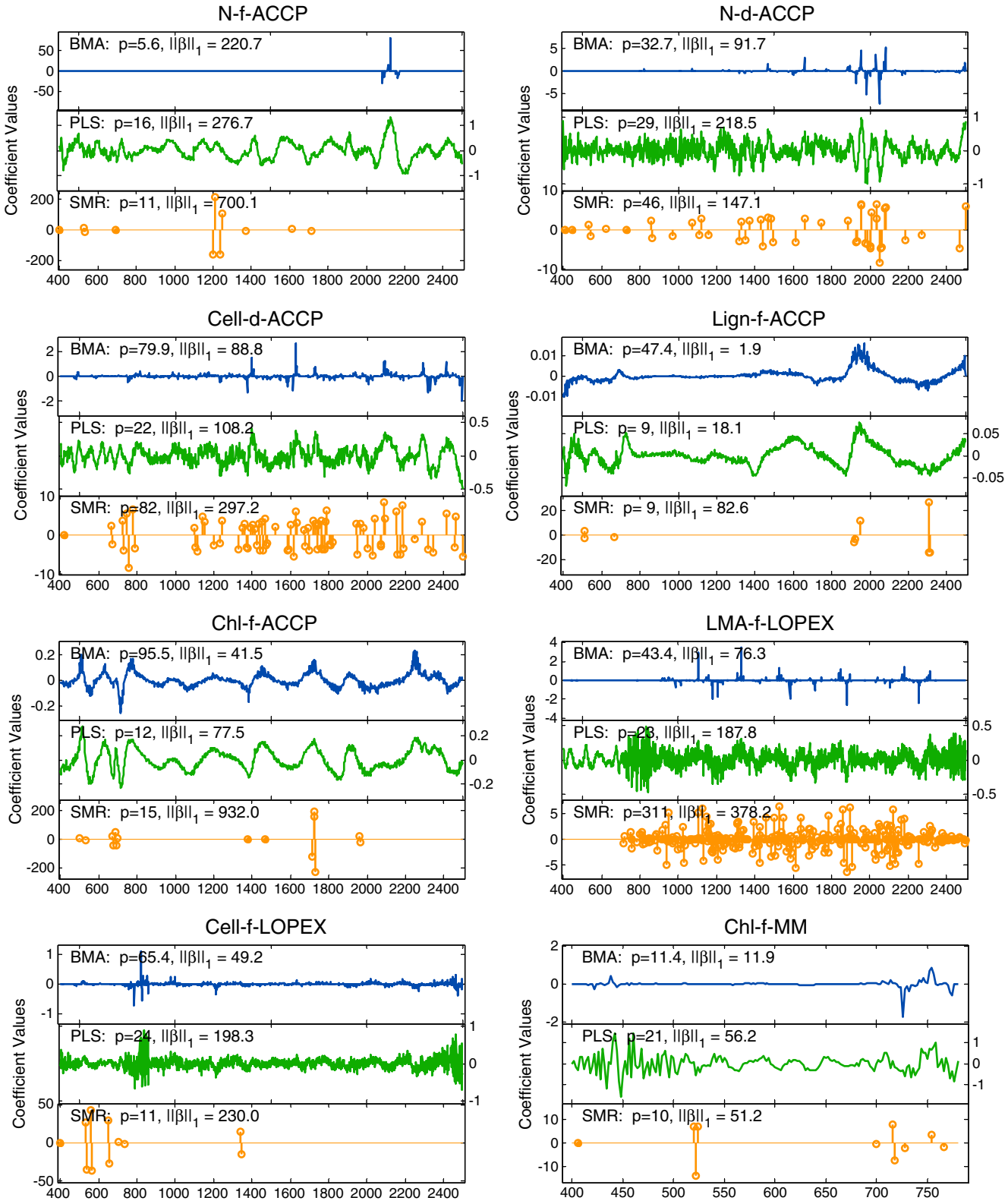


Fig. 4. Model coefficients β fitted to eight spectral–chemical datasets are plotted as functions of wavelength (x axis) for BMA (blue lines), PLS (green), and SMR (orange): Coefficients of SMR are nonzero only at selected waveband whereas those of BMA and PLS are more or less continuous across the spectrum. In each panel, "p" denotes either the number or average number of selected bands (i.e., for BMA and SMR) or factors (i.e., for PLS); ||β|| is the L1-norm of coefficients, defined as the sum of absolute values of coefficients. The value of ||β|| for BMA is often the smallest among the three regression methods, partially explaining why BMA is the least likely to overfit data.

all the three regression methods failed to relate spectra to carotenoid based on the Carot-f-LOPEX dataset, with R^2 values all less than 0.03 when tested on the validation data.

Diagnostic statistics of BMA generated during the model training phase, including band selection probability, R^2, and RMSE, were found to serve as more reliable measures of model capabilities,

Table 2
Statistics of R^2, RMSE, and PICP95 for evaluating predicted biochemicals from the 27 foliage spectral–chemical datasets of Table 1 using three regression methods, including Bayesian model averaging (BMA), partial least squares (PLS), and stepwise multiple regression (SMR): PICP95 is a statistic called "95% prediction interval coverage probability" for assessing the reliability of error estimation; the closer it is to 95%, the more reliable the error estimation. All values reported here were calculated based on test data and they represent the average over 50 runs in which data were randomly split into training and test sets. R-squares followed by ** and by * indicate that the starred method was better than the other method at a significance level of 0.001 and 0.01, respectively, according to a paired-t test comparing the 50 R^2 values between BMA and PLS. Unstarred rows suggest that BMA and PLS performed similarly. Statistics are bolded to identify the methods of the best performance with regard to the statistics.

Acronym	R^2			RMSE			PICP95		
	BMA	PLS	SMR	BMA	PLS	SMR	BMA	PLS	SMR
N-f-ACCP	**0.791****	0.573	0.509	**0.42**	0.64	0.72	**95.7%**	80.7%	84.7%
N-d-ACCP	**0.968****	0.963	0.955	**0.12**	0.13	0.15	**94.3%**	91.1%	78.3%
Cell-f-ACCP	**0.052**	0.014	0.002	**9.85**	10.38	11.08	**90.0%**	82.5%	77.5%
Cell-d-ACCP	**0.902****	0.891	0.843	**2.08**	2.19	2.65	**96.1%**	92.5%	59.3%
Lign-f-ACCP	**0.245**	0.224	0.190	**2.56**	3.77	3.44	**95.0%**	82.5%	72.5%
Lign-d-ACCP	**0.874**	0.882*	0.758	1.58	**1.53**	2.29	**95.5%**	86.2%	54.2%
Chl-f-ACCP	**0.755****	0.705	0.654	**1.59**	1.73	1.93	92.7%	**94.1%**	85.1%
Chl-d-ACCP	**0.764****	0.736	0.704	**0.65**	0.70	0.72	**93.0%**	85.0%	90.3%
Chl.a-f-ACCP	**0.643***	0.621	0.197	**1.45**	1.55	2.26	**94.0%**	88.3%	89.6%
Chl.b-f-ACCP	**0.513***	0.479	0.008	**0.49**	0.55	0.75	**94.9%**	88.3%	88.9%
C-d-ACCP	**0.937***	0.930	0.850	**0.37**	0.39	0.62	**95.4%**	83.6%	31.2%
C-f-ACCP	**0.284***	0.247	0.375	**2.00**	2.21	1.92	**89.0%**	**89.0%**	73.0%
H-d-ACCP	**0.845***	0.829	0.725	**0.12**	1.33	0.17	**94.2%**	89.2%	60.8%
polar-d-ACCP	**0.950***	0.947	0.867	**1.98**	2.02	3.49	**96.7%**	93.9%	24.4%
npolar-d-ACCP	**0.858**	0.856	0.690	**0.89**	0.90	1.30	**95.5%**	88.8%	45.7%
LMA-f-LOPEX	**0.942**	0.934	0.844	**0.63**	**0.63**	1.08	**93.8%**	73.1%	0.0%
EWT-f-LOPEX	**0.950***	0.943	0.881	**1.87**	1.93	2.81	**94.3%**	92.8%	0.0%
N-f-LOPEX	**0.712***	0.667	0.436	**3.07**	4.28	6.01	**92.1%**	85.7%	46.1%
Cellu-f-LOPEX	0.340	**0.409**	0.280	**5.04**	5.28	6.42	**96.4%**	80.4%	69.6%
Lign-f-LOPEX	0.244	**0.267**	0.231	5.18	**4.92**	5.13	**95.4%**	87.1%	83.6%
Chl.a-f-LOPEX	**0.324****	0.054	0.029	**2.50**	3.00	3.18	**97.6%**	92.4%	76.7%
Chl.b-f-LOPEX	**0.129***	0.104	0.120	**0.88**	0.99	**0.88**	**93.3%**	90.5%	79.1%
Carot-f-LOPEX	**0.027****	0.019	0.018	**0.66**	0.74	0.81	**94.3%**	91.0%	71.9%
Chl-f-MM	**0.957***	0.929	0.956	**53.4**	69.7	53.9	**94.6%**	89.6%	87.5%
Chl.a-f-MM	**0.909***	0.895	0.869	**26.0**	29.6	31.6	**99.0%**	91.0%	68.0%
Chl.b-f-MM	0.740	**0.804**	0.668	24.1	**20.3**	26.9	91.9%	90.7%	**92.6%**
Carot-f-MM	**0.420***	0.295	0.114	**1.91**	2.60	3.38	**95.0%**	89.0%	74.0%

compared to those of PLS and SMR. Using the Carot-f-LOPEX data as an example, the graphs of band selection probability showed that BMA found no bands useful for predicting carotenoid whereas SMR mis-identified several "important" bands with 1722 nm being the most probable one (Fig. 2). The R^2 values obtained in the calibration/training phase by PLS and BMA for Carot-f-LOPEX were 0.54 and 0.23, respectively, whereas the counterpart values in the validation phase were 0.019 and 0.027, respectively (Table 2). Only the diagnostic information provided by BMA during the calibration phase was consistent with the actual poor predictive relationships. Also, the smaller difference in R^2 between calibration and validation for BMA suggests that it is less prone to overfitting than PLS. Thus, the statistical measures calculated by BMA during the training phase have more fidelity in projecting models' generalization abilities.

Uncertainty estimation of BMA was more reliable than that of PLS and SMR, as evaluated in terms of the PICP95 statistic (Table 2). Fig. 5 presents a specific example showing that only BMA captured the large uncertainty in predicted nitrogen for a selected spectrum (i.e., the red-filled point). Across our results, BMA frequently yielded a PICP95 statistic closest to the expected nominal value of 95% whereas the PICP95 statistic of SMR in many cases deviated largely from 95%. For example, when predicting LMA and EWT from the LOPEX data,

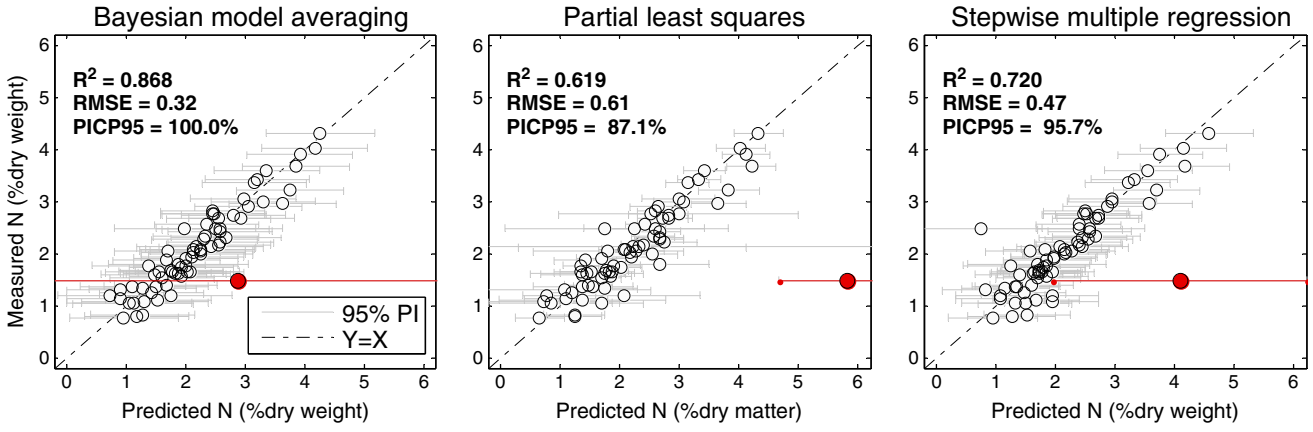


Fig. 5. Measured versus predicted foliage nitrogen concentrations (N) for three regression models (i.e., BMA, PLS, and SMR). Scatterplots shown here are the validation results for one random splitting of the N-f-ACCP dataset of Table 1 with a split ratio of 2:1 for training and validation. Light-gray horizontal bars represent 95% prediction intervals. The red-filled data point singled out provides a case where only BMA estimated the error interval reasonably.

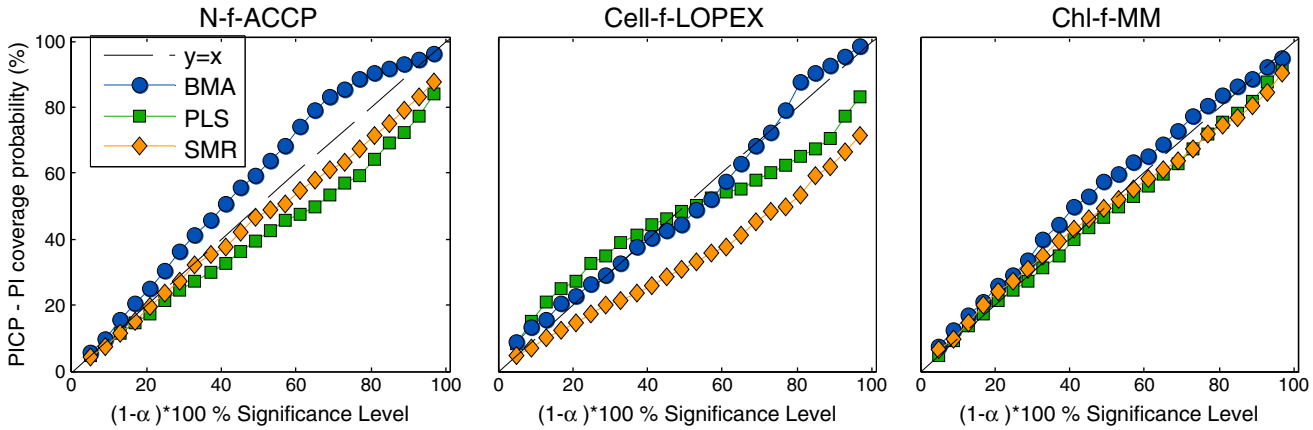


Fig. 6. Use of a statistical measure called prediction interval coverage probability (PICP) to evaluate and compare the reliability of error interval estimation for BMA, PLS and SMR regression. Only three spectral–chemical datasets (i.e., N-f-ACCP, Cell-f-LOPEX, and Chl-f-MM) are depicted here as examples. A PICP curve more closely following the 1:1 line indicates higher trustworthiness of error estimation.

the PICP95 values for SMR degenerated to 0%; these extreme and wrong values were attributed to overly small estimates of $\hat{\sigma}_\varepsilon^2$ by the SMR models as a result of substantial overfitting (i.e., almost perfect fitting). We further plotted the $\text{PICP}_{1-\alpha}$ curve as a function of uncertainty significance level $(1-\alpha)100\%$ for three selected spectral–biochemical datasets (Fig. 6). These curves indicate that the PICP estimates of BMA either followed close to or deviated upward from the ideal line 1:1 while those of PLS and SMR tended to deviate downward. In practice, the upward deviation of $\text{PICP}_{1-\alpha}$ relative to $(1-\alpha)$ 100% is advantageous over the downward deviation because the former suggests wider error intervals than should be and because on a safe side, wider error estimates will give more conservative error budgets.

5.5. Incorporation of customized spectral indices into BMA

This experiment aims mainly to demonstrate the flexibility of the BMA method in accounting for nonlinearity in spectral–chemical relationships. We chose four datasets, LMA-f-LOPEX, EWT-f-LOPEX, Chl-f-MM, and Carot-f-MM, to illustrate how prior knowledge, though vague or useless in a classical sense, can be encoded and assimilated into BMA to improve the inference of spectral–chemical relationships. In particular, the use of spectral indices as surrogates to biochemical contents is common but with differing choices of wavelengths in literature. Sheer forms of spectral index provide vague knowledge to guide transforming raw reflectances into new covariates for the BMA regression, as elaborated in Section 3.3. For a given type of spectral index, BMA can incorporate all valid indices of that type (e.g., $\sim 1.8\times 10^8$ possible three-band indices out of 1000 bands as a result of combinatorics) and also gauges their relative usefulness in terms of marginal probabilities of being selected into the BMA model. In this experiment, we considered separately six types of indices that involved either two, three or four bands (Table 3). For each type of index, we fitted BMA models upon the four selected spectral–chemical datasets. As in Section 5.3, each of the four datasets was randomly split into a two-thirds training and a one-third test subset to calibrate and validate the models, with a repetition of 50 times. PLS and SMR were not considered because they are unable to tackle the enormous set of all band combinations associated with a given type of index type.

This experiment also aims to examine the benefits gleaned from the model averaging paradigm compared to the use of a single "optimal" model. Therefore, in addition to BMA, we searched for the particular band combination that yields the best correlation with a given biochemical for each type of spectral index; we then used this identified single best index (SBI) as the only predictor to build a simple linear model, which is termed here as SBI model. The searching of these optimal indices was done using either exhaustive enumeration for the two-band indices (i.e., $(r_{\lambda_1}-r_{\lambda_2})/(r_{\lambda_1}+r_{\lambda_2})$, $r_{\lambda_1}/r_{\lambda_2}$, and $1/r_{\lambda_1}-1/r_{\lambda_2}$) or a Genetic algorithm optimizer for the three- and four-band indices (i.e., $(1/r_{\lambda_1}-1/r_{\lambda_2})r_{\lambda_3}$, $(r_{\lambda_1}-r_{\lambda_2})/(r_{\lambda_1}+r_{\lambda_3})$, and $(r_{\lambda_1}-r_{\lambda_2})/(r_{\lambda_3}+r_{\lambda_4})$). These SBI models were fitted and tested upon the same splitting of data as that of the BMA models.

In our fitted simple linear models using SBI as predictor, the wavelengths selected in the optimal indices for Chl and Carot were consistent with those of previous studies (e.g., le Maire et al., 2004), but not for LMA and EWT. The discrepancies for LMA and EWT were attributed possibly to our consideration of the whole spectrum whereas previous studies focused on some selected ranges of spectra (Sims & Gamon, 2003). To predict a given biochemical, spectral indices manifested better predictive power for both BMA and SBI models if the indices involved more bands (Table 3). The BMA models outperformed the SBI models for predicting LMA, EWT and Carot, except for Chl. The improvement achieved by BMA was particularly evident for LMA and Carot (Table 3). For Chl and Carot, the BMA models fitted in this experiment with spectral indices as predictors outcompeted those BMA models fitted in Section 5.3 with raw reflectance as predictors, but for LMA and EWT, the improvement was marginal or absent. In terms of PICP95, both the BMA and SBI models yielded realistic estimation of prediction errors. The PICP95 values of SBI models were much closer to 95% than those of linear models fitted with SMR in Section 5.3; this improvement is because SBI models with a single predictor are the most parsimonious and are unlikely to overfit data whereas SMR usually included dozens of predictors. Overall, the extra benefits achieved by BMA over SBI were more pronounced when correlations between the biochemicals and the optimal indices in the SBI models were lower, which reaffirms the usefulness of model averaging to overcome model misspecification.

5.6. Effect of training sample size on prediction accuracy

Model inference and prediction are expected to improve as the training sample size increases. We demonstrated and assessed such effects using three representative datasets, including N-d-ACCP, Cell-d-ACCP and Lign-d-ACCP. Each dataset was halved into two parts, one reserved for training and another for validation. From the reserved training data only, we further selected a series of subsets with a progression of sample size to calibrate models, but the calibrated models were always validated upon the other reserved half of the full data.

Table 3
Comparisons of three validation statistics, R^2, RMSE, and PICP95, between Bayesian model averaging (BMA) and single-best-index (SBI) models for six types of spectral indices when predicting leaf matter per area (LMA-f-ACCP), equivalent water thickness (EWT-f-ACCP), Chl (Chl-f-MM) and Carot(Carot-f-MM). Note that R^2, RSME and PICP95 were evaluated upon test data and their reported values represent averages over 50 random runs. In the "single-best-index" column, the subscripts of r's indicate the optimal wavelengths selected in each type of SBI model for predicting the four biochemicals, with units being nm.

Index type	Chemical	Single best index (SBI)	R2		RMSE		PICP95	
			BMA	SBI	BMA	SBI	BMA	SBI
$\frac{r_{\lambda_1}-r_{\lambda_2}}{r_{\lambda_1}+r_{\lambda_2}}$	*LMA*	$(r_{1869}-r_{2279})/(r_{1869}+r_{2279})$	**0.942**	0.534	**0.58**	1.64	**94.3%**	93.1%
	EWT	$(r_{1148}-r_{1127})/(r_{1148}+r_{1127})$	**0.945**	0.917	**1.81**	2.23	**93.6%**	97.2%
	Chl	$(r_{744}-r_{732})/(r_{744}+r_{732})$	0.970	0.970	45.4	**45.0**	**95.8%**	96.3%
	Carot	$(r_{494}-r_{470})/(r_{494}+r_{470})$	**0.840**	0.708	**1.01**	1.41	**92.0%**	89.0%
$\frac{r_{\lambda_1}}{r_{\lambda_2}}$	*LMA*	r_{1870}/r_{2292}	**0.947**	0.689	**0.56**	1.32	94.8%	94.9%
	EWT	r_{1148}/r_{1128}	**0.946**	0.913	**1.79**	2.29	92.0%	**97.5%**
	Chl	r_{744}/r_{728}	0.972	**0.975**	43.8	**41.0**	92.9%	**93.7%**
	Carot	r_{470}/r_{494}	**0.864**	0.724	**0.96**	1.37	**91.0%**	90.0%
$\frac{1}{r_{\lambda_1}}-\frac{1}{r_{\lambda_2}}$	*LMA*	$1/r_{1875}-1/r_{2295}$	**0.909**	0.732	**0.73**	1.25	**93.9%**	93.2%
	EWT	$1/r_{1138}-1/r_{1155}$	**0.865**	0.852	**2.89**	2.95	**95.0%**	98.2%
	Chl	$1/r_{730}-1/r_{738}$	0.966	**0.971**	48.3	**45.0**	**92.5%**	91.2%
	Carot	$1/r_{494}-1/r_{466}$	**0.881**	0.751	**0.91**	1.29	90.0%	**97.0%**
$\left(\frac{1}{r_{\lambda_1}}-\frac{1}{r_{\lambda_2}}\right)r_{\lambda_3}$	*LMA*	$(1/r_{2272}-1/r_{1873})r_{1105}$	**0.942**	0.810	**0.58**	1.05	**93.9%**	93.8%
	EWT	$(1/r_{1156}-1/r_{1134})r_{732}$	**0.946**	0.926	1.**80**	2.09	92.4%	**95.4%**
	Chl	$(1/r_{686}-1/r_{736})r_{722}$	0.973	0.**976**	43.07	**40.9**	93.3%	**94.6%**
	Carot	$(1/r_{446}-1/r_{692})r_{624}$	**0.867**	0.818	**0.95**	1.10	**95.0%**	89.0%
$\frac{r_{\lambda_1}-r_{\lambda_2}}{r_{\lambda_1}+r_{\lambda_3}}$	*LMA*	$(r_{2279}-r_{1869})/(r_{2279}+r_{2302})$	**0.948**	0.701	**0.55**	1.32	94.8%	**95.0%**
	EWT	$(r_{1145}-r_{1127})/(r_{1145}+r_{1405})$	**0.950**	0.929	**1.73**	2.07	92.1%	**97.1%**
	Chl	$(r_{728}-r_{744})/(r_{728}+r_{418})$	0.972	**0.976**	43.7	**40.2**	94.2%	**94.5%**
	Carot	$(r_{462}-r_{512})/(r_{462}+r_{600})$	**0.858**	0.819	**0.97**	1.10	94.0%	94.0%
$\frac{r_{\lambda_1}-r_{\lambda_2}}{r_{\lambda_1}+r_{\lambda_4}}$	*LMA*	$(r_{1718}-r_{1649})/(r_{2280}+r_{2134})$	**0.949**	0.741	**0.54**	1.23	**94.8%**	93.0%
	EWT	$(r_{1148}-r_{1127})/(r_{2443}+r_{1354})$	**0.952**	0.926	**1.71**	2.13	92.0%	**96.7%**
	Chl	$(r_{728}-r_{744})/(r_{726}+r_{418})$	0.974	**0.976**	41.6	**40.1**	**95.8%**	94.5%
	Carot	$(r_{466}-r_{514})/(r_{596}+r_{570})$	**0.842**	0.828	**1.02**	1.07	**94.0%**	93.0%

Results show that SMR was much more sensitive to training sample size than BMA and PLS (Fig. 7). In terms of R^2 and RSME, the prediction accuracies of BMA and PLS indeed showed an increasing trend with training sample size whereas SMR showed sizable variations in performances. The curves of performance versus sample size obtained from this experiment can help to determine the minimum training samples required to capture patterns underlying spectra and biochemicals (Fig. 3). The predictive performances of all three models improved pronouncedly around sample sizes between 30 and 80, and gradually leveled off after a training sample size around 120, except for SMR. Overall, BMA and PLS fitted better linear models than SMR regardless of the training sample size, although under extreme and rare situations, SMR might fortuitously yield favorable results (e.g., using only 14 training data points for the Cell-d-ACCP as in Fig. 7). The inferior and instable performances of SMR suggest that more efforts and care are required in applying stepwise regression to predict biochemicals from spectra.

5.7. Computation complexity

Lastly we evaluated the computation complexity associated with training BMA and PLS. SMR was excluded because it rarely constitutes a computational concern. The Markov chain length of BMA runs was set to 60,000 iterations. PLS was literally implemented by progressively adding high-order factors and enumerating all the possible factor numbers to optimize the PRESS statistic, a procedure widely known as PLS-PRESS. In this evaluation, we used the N-d-ACCP dataset as an example and tested a series of training sample sizes ranging from 33 to 580.

Contrary to what might be expected, the training of BMA was found much faster than PLS-PRESS for a moderate training sample size larger than 100. PLS-PRESS had a computational advantage only for cases with small training sample sizes less than 50 (Fig. 8). The training computation for BMA depends to a large degree on the average number of selected bands and therefore appeared less sensitive to

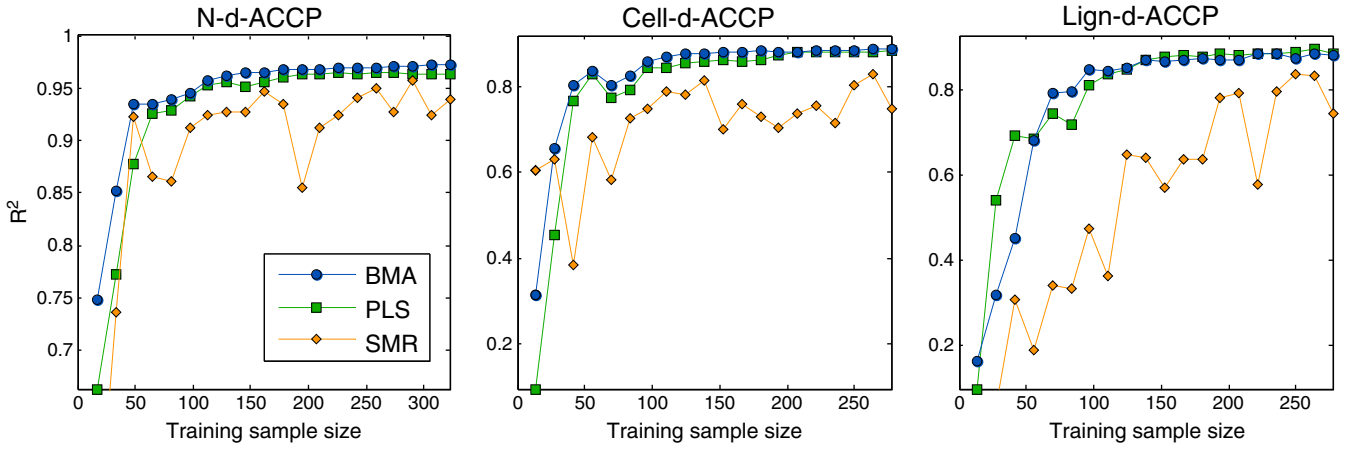


Fig. 7. Effects of training sample size on prediction performances of BMA, PLS and SMR as evaluated by R^2 values computed upon validation data: Only three examples using the ACCP data are depicted here for predicting dry-leaf Nitrogen (N-d-ACCP), Cellulose (Cell-d-ACCP) and Lignin (Lign-d-ACCP), respectively.

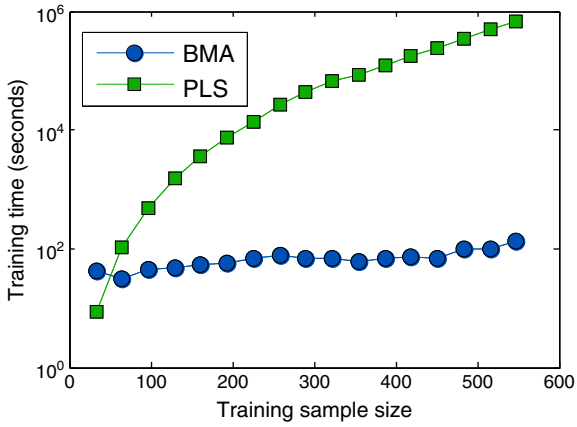


Fig. 8. Computational costs of training BMA (circles) and PLS-PRESS (square) models at a series of sample sizes. Y axis is in a logarithmic scale.

sample size over the range we considered. For example, all the BMA training experiments were completed within 150 s. In contrast, the computation for PLS-PRESS scaled exponentially with training sample size and, for example, took ~189 h for a training sample size of 550. The daunting computational costs render it infeasible to tune the factor number upon large datasets using PLS-PRESS, thus necessitating a modification of calibration scheme for PLS as has been explored by some researchers (e.g., Li et al., 2007).

6. Discussion

Along with many recent studies, our findings accumulate further empirical evidence suggesting that it is difficult, if not impossible, to establish a generic statistical framework in hyperspectral remote sensing of plant biochemistry for diverse vegetation communities across the globe (Atzberger et al., 2010; Bolster et al., 1996; Grossman et al., 1996; le Maire et al., 2004; Schlerf et al., 2010). This difficulty is directly manifested by a lack of commonality in the chosen wavebands and spectral indices among prior research for estimating the same biochemical of interest (Feret et al., 2011; Grossman et al., 1996), though with few exceptions (e.g., Martin et al., 2008). Efforts aiming to reveal and reconcile such a discrepancy seem enlightening but might turn inconsequential because many factors exist to confound the spectral–chemical linkage. These factors are not always within our observation capabilities or under direct controls, although their confounding effects are identifiable via physical modeling in conceivable manners. Another important type of uncertainty inherent in an inferred predictive relationship, especially in terms of variable and band selection, stems from the statistical nature of regression techniques and the randomness in collecting and preparing calibration data (Atzberger et al., 2010; Brown et al., 1998), as evident in our results for stepwise regression. All these ramifications lead to an ensemble of competing models. Alternatively speaking, model misspecification is essentially inevitable when establishing predictive spectral–chemical relationships for spectroscopy of plant biochemistry.

This research relied on a Bayesian version of model averaging to address model misspecification and uncertainty for hyperspectral applications. In BMA, the merit of each candidate model is explicitly quantified in a probabilistic manner. BMA clearly has theoretical advantages over standard regression. To date, BMA has become an acclaimed mechanism to exploit the values of multiple competing hypotheses and models in many disciplines, such as the statistics, ecology, social sciences, economics, hydrology, geophysics, and climate sciences (e.g., Denison, 2002; Gelman, 2004; Hoeting et al., 1999; Wintle et al., 2003). For example, Sloughter et al. (2007) proposed a flexible BMA scheme to improve predictive skills of rainfall forecast by integrating an ensemble of weather models. Zhang and Zhao (2012) applied BMA to account for the uncertainty in specifying neural network models for streamflow prediction. In most cases, the synthesis of multiple models not just enables quantifying model uncertainty but also improves model prediction power. An example pertaining to image classification is the combination of multiple weak classifiers to create a strong classifier. Our results also suggest that the BMA regression was generally superior to PLS and SMR, consistent with the results of previous chemometric studies (Brown et al., 1998; Chen & Martin, 2009; Chen & Wang, 2010).

The BMA regression approach offers a competitive alternative to PLS for hyperspectral estimation of plant biochemistry. The thrusts behind BMA and PLS are similar in that both tend to make the most use of full spectra (Atzberger et al., 2010), but their inference procedures differ fundamentally, resulting in predictive equations that are entirely different, though sometimes with similar predictive performances. Unlike BMA, PLS is unable to perform variable and model selection (Wold et al., 2001). The use of PLS coefficient magnitudes or loading matrices as ad-hoc proxies for variable importance, though sometimes suggestive (Bolster et al., 1996), is not truly a variable selection scheme. In particular, given an extremely large or infinite number of candidate predictors, BMA can still construct a useful model via variable and model selection (Denison, 2002), but PLS cannot (Li et al., 2007). Furthermore, BMA gains flexibilities to capture nonlinear spectral–chemical relationships by incorporating nonlinear basis terms, such as spectral indices, and complex transformations like wavelet and neural networks (Denison, 2002). These flexibilities with BMA have been demonstrated in this study through the use of six customized spectral indices for improving predictions of Chl and Carot over the use of the original reflectance predictors. In contrast, PLS can consider nonlinear terms only in a pre-defined manner, and it lacks an inference mechanism to automatically discriminate among a large set of nonlinear transformations of spectral reflectance (Li et al., 2007). Another demonstrated advantage of BMA is that its model calibration upon moderately large training datasets (e.g., >100) was computationally faster than that for PLS.

The Bayesian method introduced here outcompeted the traditional regression for dealing with the "$p>n$" problems in which the number of candidate predictors is larger than the training sample size. The greedy-searching scheme of stepwise regression for such problems is notoriously problematic (Grossman et al., 1996; Wold et al., 2001). As in our bootstrapping experiment, SMR frequently chose fortuitous predictors only to honor some spurious pattern in the observations. Grossman et al. (1996) also criticized SMR by noting that the selected hyperspectral bands did not accord well with the absorption features of chemicals under investigation. Other model selection criteria such AIC and BIC are more likely to guide finding better predictive models than does SMR. A benign property of AIC and BIC is their explicit penalty on complex models, which mitigates over-fitting (Hastie et al., 2001). Although our BMA formulation does not expressly penalize model complexity, a mechanism known as the Occam's razor naturally comes into play to discriminate against complex models in our Bayesian method (Denison, 2002). This implicit penalty on complexity helps to explain the low sensitivity of the BMA performances to the setup of the hyperparameter $\underline{p}_{max}$ (i.e., maximum number of covariates allowed in a model) as long as $\underline{p}_{max}$ is set to a relatively large number. Computationally, sweeping over all band combinations to optimize AIC/BIC is formidable, although combinatorial optimizers are sometimes introduced to help to pinpoint suboptimal models at reasonable costs. BMA circumvents this computation issue by resorting to MCMC sampling (Brown et al., 1998; Chen & Martin, 2009; Denison, 2002), which is typical of Bayesian modeling approaches.

Another important benefit derived from BMA is the improvement in error interval estimation (Hastie et al., 2001; Wintle et al., 2003), attributable particularly to the fact that model averaging incorporates model uncertainty that has been largely ignored by conventional

approaches. Although all statistical methods supposedly predict errors (i.e., interval estimate) to complement the evaluation of predicted means (i.e., point estimate), our experiments clearly show that the errors estimated by SMR and PLS are not always trustworthy when assessed in terms of PICP. Often enough, the major factor contributing to such unrealistic error estimation is the deviation of data from model assumptions. In addition, the reliability of error estimates and the degree of model overfitting are interlinked (Hastie et al., 2001). For example, the SMR models based on LMA-f-LOPEX and EWT-f-LOPEX were highly overfitted with R^2 values close to 1.0 in the training phase, which led to some extremely overconfident and useless estimates of prediction errors valued approximately at zero. These results signal a precaution to those practitioners who blindly rely on estimated model errors to report uncertainty without first evaluating their reliability.

Model interpretability has been improved through the use of BMA compared to conventional regression. This improvement is attributed primarily to the rich information inherent in the posterior distribution of model structure and model parameters $p(\boldsymbol{\beta}_{\mathcal{M}_i}, \sigma^2, \nu, \mathcal{M}_i|\mathcal{D})$ or its sampled version $\left\{\mathcal{M}^{(t)}, \boldsymbol{\beta}^{(t)}_{\mathcal{M}}, , \sigma^{2(t)}, , \nu^{(t)}\right\}_{t=1,\ldots,N}$. One salient feature of BMA that aids in model interpretation is the use of marginal band selection probability for quantifying band importance. Band selection probabilities derived by BMA bear some resemblance to absorption spectra of the examined biochemicals. In particular, the presence of many local spikes in the curve of band selection probability is consistent with the physical basis that the absorption features of a chemical bond comprises a series of vibrational or rotational lines attributable to multiple harmonics and overtones (Kokaly et al., 2009). Moreover, the graph of band selection probability for chlorophyll appears continuous over a selected range of spectrum. This continuity complies with the fact that biochemical compounds such as pigments absorb light continuously over wide spectral ranges.

As the average of individual linear models, the final predictive equation of a BMA model is still a linear one, but this average model is hard to be inferred directly with traditional linear regression (Denison, 2002). In particular, although the number of covariates chosen in each MCMC-sampled model is far less than the total number of spectral bands, the average of these individual models contains essentially all spectral bands as predictors. As emphasized throughout this paper, BMA synthesizes all models with their relative importance borne out of training data, thus helping to reduce the risk of excessive overfitting and derive a more robust predictive relationship by safeguarding against too severe model misspecification (Wintle et al., 2003). Such benefits are indirectly revealed in our comparisons of BMA against PLS and SMR based on the LOPEX data. For example, the spectral–chemical relationships for Cellu, Chl + a, Chl + b and Carot were all weak, which was reflected only in the training statistics of BMA but not those of PLS and SMR. In addition, the model evaluation criteria for PLS and SMR differed substantially between the training and validation phases, implying excessive overfitting in the PLS and SMR models. On a different note, we suspect that the primary reason for such weak spectral–chemical relationships for the four biochemicals is the data quality issue: Each spectrum in the LOPEX data represents the average over five leaves that are not necessarily the same ones as used for wet chemistry analysis. The weak relationships may also be contributed by the species diversity and the use of concentrations instead of contents for units of these biochemicals (Curran, 1989; Feret et al., 2011).

Similar to SMR, important bands identified by BMA do not perfectly match absorption peaks but instead fall near the peaks (Bolster et al., 1996; Grossman et al., 1996). One explanation for this is that absorption at peak wavelengths often saturates at low levels of pigment concentration so that the adjacent wavelengths with less absorption capacities exhibit more sensitivity to variations in chemical concentrations (Curran, 1989; Kokaly et al., 2009). In some cases, useful bands selected by BMA are irrelevant to absorption features of the chemical of interest but instead show some correspondence to those of other chemicals. Two reasons for this can be sought. First, concentrations of different chemicals are often intercorrelated, thus creating an indirect link between the chemical of interest and the absorption bands of other correlated chemicals (Schlerf et al., 2010). Second, the observed spectral variability at a band can be contributed by more than one chemical (Curran, 1989; Jacquemoud & Baret, 1990); the confounding effect of non-interested chemicals could be removed by the addition of their absorption bands to correct for the contribution of these non-interested chemicals to the absorption at the bands selected for predicting the chemical of interest. Overall, caution should be exercised when seeking physical justifications for the important bands selected by regression approaches (Grossman et al., 1996), because of the statistical nature of the calibration process.

Statistical inversion, despite its empirical nature, has played and will continue to play an indispensible role in forging a predictive science for hyperspectral remote sensing of vegetation (Asner & Martin, 2009; Feret et al., 2011), even given that recent advances in modeling leaf and canopy spectra have facilitated physically-based inversion (Feret et al., 2008). The choice between physical and statistical inversion is often a subjective matter, depending in part on data availability and modelers' expertise. The synergy of the two inversion paradigms remains largely untapped and has some new potential for algorithm improvements. Our BMA method offers one such possibility: The level of success achieved by physical inversion is generally related to four factors, including realistic forward models, high-fidelity spectral data, effective optimizers, and sufficient prior information as constraints on land surface variables (Darvishzadeh et al., 2008; Zhang et al., 2008); our BMA method could help in physical inversion with the optimization in at least two manners. First, the band selection probability from BMA can be used to weigh band contributions in the merit function of the optimization problem. Second, biochemical estimates from BMA may serve as informative guesses of initial values for the optimization. The actual extent to which this synergy can improve physically-based inversion needs to be determined in future research.

7. Summary

Enhancing the efficacy of spectroscopy for hyperspectral remote sensing of vegetation requires not just high-fidelity spectral measurements or ample field data but also flexible analytical methods. To improve predictive skills, we described a Bayesian regression method capable of variable selection and model averaging that is particularly useful for tackling high-dimensional problems. We demonstrated the capabilities of this method by applying it to estimate multiple foliage biochemicals from 27 spectral–chemical datasets. The major features of this method are highlighted as follows:

(1) Provides a rigorous band selection scheme to tackle high-dimensional problems, especially in cases that the number of spectral bands is larger than the observation number of training data;
(2) Accounts explicitly for model uncertainty by incorporating an ensemble of competing models into inference and prediction, which avoids choosing any particular fortuitous band combination and also reduces the risk of model misspecification;
(3) Avoids exhaustively exploring the space of all possible models through the use of MCMC sampling, which, though slower than the greedy search of stepwise regression, is much faster than PLS-PRESS for training data of moderate size larger than 100;
(4) Builds a final average model of superior generalization abilities, with less overfitting compared to many conventional regression techniques;

(5) Allows realistically and reliably estimating error intervals for predictions, due primarily to inclusion of multiple models and the treatment of model parameters as random;
(6) And adopts a generic Bayesian hierarchical framework that enable assimilating a variety of prior knowledge into the model processing, e.g., for inferring potential nonlinearity in the chemical–spectral relationship by considering spectral indices of various forms as potential predictors.

Mapping plant biochemicals over extensive regions will become more feasible as sensors of various types, especially those onboard aircraft, are increasingly available for data acquisitions. This feasibility is concomitantly complemented by advances in statistical learning that offer an expanding suite of tools to explore empirical relationships between spectra and canopy biochemical/biophysical properties (e.g., Gaussian processes regression as used in Chen et al., 2007; Zhao et al., 2008). Compared to traditional methods, Bayesian learning is generally more conducive to exploiting empirical knowledge, multiple modeling techniques, and information-rich data. Although the current use of Bayesian methods in remote sensing remains limited due possibly to a lack of professional training in Bayesian statistics, we envision their increased use in the foreseeable future. Bayesian methods, such as the BMA method of this study, hold great potential to boost the utility of spectroscopic data for characterizing chemical fingerprints of diverse species. In particular, the BMA method offers a very competitive alternative to standard conventional regression for hyperspectral estimation of biochemicals, as evidenced in our comparison results against PLS and SMR regression. Although the focus of this work is on estimation of foliage biochemistry from field spectroscopic data, our Bayesian regression method is a generic approach that, with no modification, is equally applicable for retrieving biogeophysical and biochemical variables from diverse remote sensing data and images, especially when a large set of inter-correlated remote sensing metrics is available.

Acknowledgment

Financial support to Kaiguang Zhao for this research came from a grant from the DOE-funded National Institute for Climate Change Research at Duke University to Rob Jackson. Xuesong Zhang received financial support from NASA under contract no. NNH12AU03I. We express our sincere gratitude to two anonymous reviewers for their insightful comments. We are also indebted to Dr. Anatoly Gitelson at the University of Nebraska-Lincoln who generously shared his maize and maple spectral–chemical data. Our Matlab code of the Bayesian model was a modified implementation of the algorithms for the book "Bayesian methods for nonlinear classification and regression" co-authored by Bani Mallick. Readers may request the Matlab code of this work from the primary author at lidar.rs@gmail.com.

References

ACCP (1994). *Accelerated canopy chemistry program final report to NASA-EOS-IWG.* Washington DC: National Aeronautics and Space Administration.
Asner, G. P., & Martin, R. E. (2008). Spectral and chemical analysis of tropical forests: Scaling from leaf to canopy levels. *Remote Sensing of Environment*, *112*, 3958–3970.
Asner, G. P., & Martin, R. E. (2009). Airborne spectranomics: Mapping canopy chemical and taxonomic diversity in tropical forests. *Frontiers in Ecology and the Environment*, *7*, 269–276.
Asner, G. P., Martin, R. E., Tupayachi, R., Emerson, R., Martinez, P., Sinca, F., et al. (2011). Taxonomy and remote sensing of leaf mass per area (LMA) in humid tropical forests. *Ecological Applications*, *21*, 85–98.
Atzberger, C., Guerif, M., Baret, F., & Werner, W. (2010). Comparative analysis of three chemometric techniques for the spectroradiometric assessment of canopy chlorophyll content in winter wheat. *Computers and Electronics in Agriculture*, *73*, 165–173.
Bolster, K. L., Martin, M. E., & Aber, J. D. (1996). Determination of carbon fraction and nitrogen concentration in tree foliage by near infrared reflectance: A comparison of statistical methods. *Canadian Journal of Forest Research—Revue Canadienne De Recherche Forestiere*, *26*, 590–600.
Brown, P. J., Vannucci, M., & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society Series B—Statistical Methodology*, *60*, 627–641.
Chen, T., & Martin, E. (2009). Bayesian linear regression and variable selection for spectroscopic calibration. *Analytica Chimica Acta*, *631*, 13–21.
Chen, T., Morris, J., & Martin, E. (2007). Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*, *87*, 59–67.
Chen, T., & Wang, B. (2010). Bayesian variable selection for Gaussian process regression: Application to chemometric calibration of spectrometers. *Neurocomputing*, *73*, 2718–2726.
Curran, P. J. (1989). Remote-sensing of foliar chemistry. *Remote Sensing of Environment*, *30*, 271–278.
Darvishzadeh, R., Skidmore, A., Schlerf, M., & Atzberger, C. (2008). Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland. *Remote Sensing of Environment*, *112*, 2592–2604.
Dawson, T. P., Curran, P. J., & Plummer, S. E. (1998). LIBERTY — Modeling the effects of leaf biochemical concentration on reflectance spectra. *Remote Sensing of Environment*, *65*, 50–60.
Denison, D. G. T. (2002). *Bayesian methods for nonlinear classification and regression.* Chichester: Wiley.
Fan, Y., & Sisson, S. A. (2010). *Reversible jump Markov chain Monte Carlo.* (Arxiv preprint arXiv:1001.2055).
Feret, J. B., Francois, C., Asner, G. P., Gitelson, A. A., Martin, R. E., Bidel, L. P. R., et al. (2008). PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. *Remote Sensing of Environment*, *112*, 3030–3043.
Feret, J. B., Francois, C., Gitelson, A., Asner, G. P., Barry, K. M., Panigada, C., et al. (2011). Optimizing spectral indices and chemometric analysis of leaf chemical properties using radiative transfer modeling. *Remote Sensing of Environment*, *115*, 2742–2750.
Gamon, J. A., Serrano, L., & Surfus, J. S. (1997). The photochemical reflectance index: An optical indicator of photosynthetic radiation use efficiency across species, functional types, and nutrient levels. *Oecologia*, *112*, 492–501.
Garbulsky, M. F., Penuelas, J., Gamon, J., Inoue, Y., & Filella, I. (2011). The photochemical reflectance index (PRI) and the remote sensing of leaf, canopy and ecosystem radiation use efficiencies. A review and meta-analysis. *Remote Sensing of Environment*, *115*, 281–297.
Gelman, A. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, Fla; London: Chapman & Hall/CRC.
Gitelson, A. A., Buschmann, C., & Lichtenthaler, H. K. (1999). The chlorophyll fluorescence ratio F-735/F-700 as an accurate measure of the chlorophyll content in plants. *Remote Sensing of Environment*, *69*, 296–302.
Gitelson, A. A., Gritz, Y., & Merzlyak, M. N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, *160*, 271–282.
Gitelson, A. A., Keydan, G. P., & Merzlyak, M. N. (2006). Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. *Geophysical Research Letters*, *33*.
Gitelson, A. A., Vina, A., Ciganda, V., Rundquist, D. C., & Arkebauer, T. J. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, *32*.
Grossman, Y. L., Ustin, S. L., Jacquemoud, S., Sanderson, E. W., Schmuck, G., & Verdebout, J. (1996). Critique of stepwise multiple linear regression for the extraction of leaf biochemistry information from leaf reflectance data. *Remote Sensing of Environment*, *56*, 182–193.
Hastie, T., Tibshirani, R., Friedman, J. H., & MyiLibrary (2001). The elements of statistical learning data mining, inference, and prediction: with 200 full-color illustrations. *Springer series in statistics*. New York: Springer (pp. xvi, 533 pp.).
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–401.
Hosgood, B., Jacquemoud, S., Andreoli, G., Verdebout, J., Pedrini, A., & Schmuck, G. (1994). *Leaf Optical Properties EXperiment 93 (LOPEX93)*. Ispra (Italy): European Commission — Joint Research Centre EUR 16095.
Jackson, R. B., Randerson, J. T., Canadell, J. G., Anderson, R. G., Avissar, R., Baldocchi, D. D., et al. (2008). Protecting climate with forests. *Environmental Research Letters*, *3*.
Jacquemoud, S., & Baret, F. (1990). Prospect — A model of leaf optical-properties spectra. *Remote Sensing of Environment*, *34*, 75–91.
Jacquemoud, S., Verdebout, J., Schmuck, G., Andreoli, G., & Hosgood, B. (1995). Investigation of leaf biochemistry by statistics. *Remote Sensing of Environment*, *54*, 180–188.
Knyazikhin, Y., Schull, M. A., Stenberg, P., Mõttus, M., Rautiainen, M., Yang, Y., et al. (2012). Hyperspectral remote sensing of foliar nitrogen content. *Proceedings of the National Academy of Sciences*.
Kokaly, R. F., Asner, G. P., Ollinger, S. V., Martin, M. E., & Wessman, C. A. (2009). Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sensing of Environment*, *113*, S78–S91.
le Maire, G., Francois, C., & Dufrene, E. (2004). Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements. *Remote Sensing of Environment*, *89*, 1–28.
Li, L., Ustin, S. L., & Riano, D. (2007). Retrieval of fresh leaf fuel moisture content using genetic algorithm partial least squares (GA-PLS) modeling. *IEEE Geoscience and Remote Sensing Letters*, *4*, 216–220.
Liang, S. L. (2007). Recent developments in estimating land surface biogeophysical variables from optical remote sensing. *Progress in Physical Geography*, *31*, 501–516.
Martin, M. E., Plourde, L. C., Ollinger, S. V., Smith, M. L., & McNeil, B. E. (2008). A generalizable method for remote sensing of canopy nitrogen across a wide range of forest ecosystems. *Remote Sensing of Environment*, *112*, 3511–3519.

Milton, E. J., Schaepman, M. E., Anderson, K., Kneubuhler, M., & Fox, N. (2009). Progress in field spectroscopy. *Remote Sensing of Environment*, *113*, S92–S109.

Nguyen, H. T., & Lee, B. W. (2006). Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. *European Journal of Agronomy*, *24*, 349–356.

Pasolli, L., Melgani, F., & Blanzieri, E. (2010). Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, *7*, 464–468.

Richardson, A. D., Duigan, S. P., & Berlyn, G. P. (2002). An evaluation of noninvasive methods to estimate foliar chlorophyll content. *New Phytologist*, *153*, 185–194.

Schlerf, M., Atzberger, C., Hill, J., Buddenbaum, H., Werner, W., & Schuler, G. (2010). Retrieval of chlorophyll and nitrogen in Norway spruce (*Picea abies* L. Karst.) using imaging spectroscopy. *International Journal of Applied Earth Observation and Geoinformation*, *12*, 17–26.

Serrano, L., Penuelas, J., & Ustin, S. L. (2002). Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: Decomposing biochemical from structural signals. *Remote Sensing of Environment*, *81*, 355–364.

Sims, D. A., & Gamon, J. A. (2003). Estimation of vegetation water content and photosynthetic tissue area from spectral reflectance: A comparison of indices based on liquid water and chlorophyll absorption features. *Remote Sensing of Environment*, *84*, 526–537.

Sloughter, J. M., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, *135*, 3209–3220.

Townsend, P. A., Foster, J. R., Chastain, R. A., & Currie, W. S. (2003). Application of imaging spectroscopy to mapping canopy nitrogen in the forests of the central Appalachian Mountains using Hyperion and AVIRIS. *IEEE Transactions on Geoscience and Remote Sensing*, *41*, 1347–1354.

Ustin, S. L., Roberts, D. A., Gamon, J. A., Asner, G. P., & Green, R. O. (2004). Using imaging spectroscopy to study ecosystem processes and properties. *Bioscience*, *54*, 523–534.

Wintle, B. A., McCarthy, M. A., Volinsky, C. T., & Kavanagh, R. P. (2003). The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, *17*, 1579–1590.

Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 109–130.

Zarco-Tejada, P. J., Miller, J. R., Noland, T. L., Mohammed, G. H., & Sampson, P. H. (2001). Scaling-up and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *39*, 1491–1507.

Zhang, X., & Zhao, K. (2012). Bayesian neural networks for uncertainty analysis of hydrologic modeling: A comparison of two schemes. *Water Resources Management*, 1–18.

Zhang, Y. Q., Chen, J. M., Miller, J. R., & Noland, T. L. (2008). Leaf chlorophyll content retrieval from airborne hyperspectral remote sensing imagery. *Remote Sensing of Environment*, *112*, 3234–3247.

Zhao, K. G., & Popescu, S. (2009). Lidar-based mapping of leaf area index and its use for validating GLOBCARBON satellite LAI product in a temperate forest of the southern USA. *Remote Sensing of Environment*, *113*, 1628–1645.

Zhao, K. G., Popescu, S., Meng, X. L., Pang, Y., & Agca, M. (2011). Characterizing forest canopy structure with lidar composite metrics and machine learning. *Remote Sensing of Environment*, *115*, 1978–1996.

Zhao, K. G., Popescu, S., & Zhang, X. S. (2008). Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. *Photogrammetric Engineering and Remote Sensing*, *74*, 1223–1234.