Handbook of Near-Infrared Analysis

Emil W. Ciurczak, Benoît Igne, Jerome Workman, Donald A. Burns

Bayesian Regression for Chemical Calibrations of NIR Spectra

Publication details

Steven D. Brown
Published online on: 20 May 2021

How to cite: Steven D. Brown. 20 May 2021, Bayesian Regression for Chemical Calibrations of NIR Spectra from: Handbook of Near-Infrared Analysis CRC Press
Accessed on: 19 Sep 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
13 Bayesian Regression for Chemical Calibrations of NIR Spectra

Steven D. Brown*
University of Delaware
and
IRSTEA

CONTENTS

13.1 A Very Brief Introduction to Bayesian Statistics ................................................................. 236
  13.1.1 Marginal and Conditional Probabilities ................................................................. 238
  13.1.2 Bayes' Theorem ................................................................................................. 238
  13.1.3 Belief and Priors ............................................................................................... 240
  13.1.4 Establishing the Likelihood ............................................................................... 241
  13.1.5 Obtaining and Assessing the Posterior ............................................................... 241
  13.1.6 Simulation of the Posterior Distributions .......................................................... 242

13.2 Bayesian Regression .............................................................................................................. 243
  13.2.1 An Example of Simple Bayesian Regression ......................................................... 246

13.3 Bayesian Calibration of Near-Infrared Spectra ................................................................. 249
  13.3.1 Direct Variable Selection in Calibration using Bayesian Regression ................. 249
  13.3.2 Bayesian Calibration on Transformed Spectral Data ....................................... 254
    13.3.2.1 Bayesian Regression on Wavelet-Transformed Spectral Data ................ 254
    13.3.2.2 Bayesian Calibration on Principal Component and Latent Variable
      Scores of Spectra ................................................................................................. 256
  13.3.3 Nonlinear Bayesian Calibration of Near-Infrared Spectra .................................. 262
  13.3.4 Nonparametric and Empirical Bayesian Spectral Calibration Methods .......... 264

13.4 Conclusions ........................................................................................................................... 265

Acknowledgments ....................................................................................................................... 266

References ..................................................................................................................................... 266

The recent emphasis on data science has led to increasing interest in Bayesian methods for data analysis. Philosophically, the Bayesian approach to data analysis is close to that employed by practitioners in traditional measurement science, but the implementation of Bayesian methods often seems foreign to researchers who are more comfortable with conventional univariate and multivariate calibration. The flexibility offered by Bayesian methods is coupled with some mathematics that differs from the usual matrix algebra needed for the conventional regression modeling used in multivariate calibration of spectral mixtures. A Bayesian approach to regression modeling generally requires more background in statistics and computation than that needed for conventional regression modeling, and often the analyses require longer computation times, but the return gained from a Bayesian approach can be substantial, both in performance and in interpretability of the results.

* sdb@udel.edu
Because Bayesian methods are still rather obscure in chemometrics, and especially in analysis of spectral data, this chapter begins with a brief tutorial on some basic Bayesian approaches to regression. It then illustrates some ways in which Bayesian methods have been applied to problems in multivariate calibration of near-infrared spectral data and assesses the advantages and drawbacks of Bayesian methods for routine use.

13.1 A VERY BRIEF INTRODUCTION TO BAYESIAN STATISTICS

Before considering Bayesian regression, it is necessary to discuss what is meant by a “Bayesian” analysis and how that approach differs from an analysis using conventional statistical methods. These conventional methods include ordinary classical least squares regression, inverse least squares regression, principal components regression (PCR), partial least squares (PLS) regression, and other methods, linear or not, that are all based on the same basic idea of how the data behave and how to extract information from the data.

Most researchers performing multivariate calibration of spectrometric data ask “what do my data show?” There is the implicit understanding that the data carry information and that a set of unknown, true model calibration parameters summarize that information. In conventional calibration, there is an emphasis on finding estimates of those model parameters, such as regression coefficients, from the experimental data. Most commonly, a particular model form is assumed, and the calibration data, presumed to contain noise, are then fit to that form to find estimates for this set of fixed, true model parameters. From these estimates of the true parameters, a prediction of the value of the dependent variable \( y_{\text{unk}} \) in new data is estimated from the model parameters and the new predictor vector, \( x_{\text{unk}} \). The usual approach is to regard the predictor variables \( X \) as noise-free, and the dependent variable(s) \( y \) as corrupted by a source of homoscedastic, independent, additive noise. The model parameters, here the coefficients in the regression model, are always taken as unknown, noise-free constants. Probability distributions, if they are considered at all, generally are discussed in terms of means and variances of the estimated \( y_{\text{unk}} \) as reflected in the root mean squared error of many predictions (RMSEP), which ultimately connect to the uncertainty in the responses in the calibration set \( y \) and the reliability of the computation, exemplified in classical least squares regression by the accuracy of the matrix inverse \((X^TX)^{-1}\) (Mardia, et al., 1979).

This analysis has an implicit probability distribution for \( y \) that is Gaussian, centered on the average \( y \) value and with a variance specified by the noise in \( y \). This distribution is convenient for predictions of properties for data that fall in the center of the calibration set because that is where the \( y \) values most frequently occur.

Data analysts who use Bayesian methods—Bayesians for short—call this modeling approach “frequentist” because the results are based on the frequency distribution of a sample of noisy \( y \) data provided for the calibration. Under this view, the data are a sample from a larger population and the data consist of a deterministic part and a random, noisy, part. What matters in the frequentist approach to modeling is the histogram of the \( y \) values used in the calibration.

The Bayesian approach to calibration modeling takes a completely different view of the problem of finding a calibration model and using it to predict the amount unknown in new samples. In the Bayesian approach, the emphasis is again on determining the calibration model parameters and predictions of the property for new data, but as a consequence of the new information contained in the calibration data we collect. Unlike the frequentist approach described above, the Bayesian approach relies on a fundamentally different view of the data and the parameters in the model. Bayesian methods are based on probability distributions because these methods presume that the set of model parameters \( \theta \) are stochastic quantities, depending on what data are observed. The Bayesian approach also presumes that the observed data \( D \) are deterministic. This view of data analysis contrasts with the conventional (frequentist) view that the parameters are fixed, unknown constants and the data \( D \) are made of a stable, deterministic part affected by additive, stochastic error, so each observed datum is a sample drawn from some distribution with a frequency \( f \).
The parameters in a Bayesian model are taken as stochastic because these are derived from data, and they are not observed. The data—both the predictors $X$ and the response variable $y$—are taken as deterministic and without error in a Bayesian analysis because these are what we observed—they are not considered a sample from some population of possible data values, as is done in the frequentist approach. Because the distributions of stochastic variables in the model are based on our assessment of the possible value of a variable, not on the number of times a value for that variable occurs—its relative frequency—we call these belief distributions, reflecting our assessment of the value based on what we know about it. A belief distribution usually has the properties of a probability distribution.

A Bayesian approach has another major difference from the usual frequentist modeling: It relies on two models. One model incorporates any information on the desired parameters before data become available. This model is the prior distribution, often called the prior for short, and it specifies the prior belief in the stochastic parameters in the model. The prior distribution is specified by its form (e.g., Gaussian) and its parameters (usually the mean and variance). Some authors also give the prior values for the mean and variance as part of the definition, so a fully specified prior for some parameter $\beta$ might be represented as $p(\beta) = G(\mu; 0, \sigma^2)$, indicating that the prior distribution for the stochastic parameter $\beta$ is believed to be a Gaussian distribution with mean $=0$ and variance set by the value of the parameter $\sigma^2$. This prior indicates that the parameter $\beta$ is probably zero but with some uncertainty. The value for the variance of $\beta$ is an indirectly observed parameter, so it could also be set in the definition of the prior model, and then adjusted as information obtained from new data requires, or it might be specified as a hyperparameter, which can be either a preset, fixed, deterministic value that is not adjusted with information available from new data or a stochastic value that is modeled and adjusted based on observed data through another, “hierarchical” Bayesian analysis (Gelman, et al., 2013). Either approach to the prior provides a route to input information about the system being modeled before data are available, a capability that is not possible in a frequentist analysis.

Then, there is the model for new data. This is called the likelihood, and it relates the observed experimental data to the model parameters. The likelihood describes how the stochastic parameters in the model are related to the observed experimental data. This relationship is often mathematically similar to the linear algebraic model relating parameters to predictors used in the usual frequentist approach to modeling, but it, too, is a distribution in the Bayesian formalism.

The prior and likelihood distributions are then combined to give an estimate of the belief distribution for the stochastic parameters updated to consider the new information found in the data. This is the posterior distribution, often just called the posterior. It describes what is known about the stochastic model parameters after incorporation of any new information about those parameters provided in the experimental data.

In the Bayesian approach, then, the two models are each a probability distribution, and they combine to produce a posterior model probability distribution. This distribution embeds the prior information on the model system as updated by experimental observations. This posterior distribution can be queried for the value of a parameter in the model, and a calculation is known to statisticians as inference. Inference on the posterior might involve calculation of means and variances of the parameters, quantities sometimes obtained in the conventional frequentist approach, but often much more is available. The emphasis on probability distributions for all model parameters, on combining probability distributions, and on inferring values of model parameters from those distributions is what distinguishes a Bayesian method. The reliance on distributions and the need to extract parameters from those distributions make Bayesian analysis computationally demanding and difficult to learn for those with limited background in statistics.

While this introduction gives a sense of the approach and defines some of the terms common in a Bayesian analysis, those seeking more extensive background might want to start with general texts on Bayesian methods (Gelman, et al., 2013; Bolstead, 2010). There is also a very good introduction to Bayesian methods in regression in the text by Bishop (2006). For those readers comfortable with
In R software, there are an enormous number of current, public-domain software packages implementing Bayesian methods at the CRAN repository (https://cran.r-project.org/). There is also a “task view,” updated frequently, that summarizes much of the basic Bayesian material available in R and in related, special-purpose languages at https://cran.r-project.org/web/views/Bayesian.html.

### 13.1.1 Marginal and Conditional Probabilities

Bayesian methods make heavy use of marginal and conditional probabilities. A marginal probability distribution can be regarded as a part of a larger distribution, for example, a multidimensional probability distribution that varies over values for all three model parameters. Each of these parameters has its own marginal probability distribution where the probability is dependent only on values for that one parameter. For statistically independent parameters, these marginal distributions are independent of the other parameters; the product of the three marginal distributions is the full distribution, and each marginal distribution can easily be examined separately. If the parameters are not independent, getting a marginal distribution for one parameter from the full distribution requires integrating out—marginalizing—the contributions from all of the other parameters in the full distribution.

A conditional probability distribution reflects the probability distribution of some variable conditioned on the value(s) of another quantity or quantities; these conditional probabilities are indicated by a vertical bar, so, for example, \( p(y | x, \beta) \) is the probability distribution of \( y \) conditioned on specific values of \( x \) and \( \beta \). The likelihood function \( p(D | \theta) \) is a conditional probability distribution that conditions the probability of observing some value of the data \( D \) on specific values of the model parameters \( \theta \).

Given the full prior probability distribution \( p(\theta) \) for all of the model parameters \( \theta \) and the overall likelihood function, \( p(D | \theta) \), a conditional distribution describing how the data \( D \) are affected by the model parameters \( \theta \), the goal is to find the posterior conditional distribution \( p(\theta | D) \), the conditional distribution that describes how the parameters \( \theta \) are affected by the data \( D \). This distribution describes the full distribution of the set of model parameters after the information present in the data is considered. This posterior conditional distribution summarizes results of the modeling. Standard statistical inference is then needed to extract information (e.g., the mean and variance) from this distribution, or from marginalized parts of this distribution, as appropriate.

### 13.1.2 Bayes’ Theorem

Given the probabilistic models for the likelihood conditional dependence of the data on the parameters, \( p(D | \theta) \), and the prior belief \( p(\theta) \), Rev. Thomas Bayes was first to show that assuming statistical independence of the prior and the likelihood distributions and using conditional probability lead to \( p(\theta | D) \) directly. His simple approach was formalized and promoted by Laplace, but the formula is known as Bayes’ theorem. Bayes’ theorem uses the principles of conditional probability to combine the prior distribution \( p(\theta) \) and the likelihood distribution function \( p(D | \theta) \) to permit the estimation of the conditional posterior, \( p(\theta | D) \), and from it, the dependence of the parameters describing the model on the data \( D \)

\[
p(\theta | D) = p(\theta) p(D | \theta) / p(D)
\]  

where \( p(D) \), the probability distribution of the observed data, is usually unknown and often is integrated out and taken as simply a constant of normalization, so that the value of \( p(\theta | D) \) integrates to 1, as required for a probability distribution.

Once the conditional posterior is obtained by using Equation 13.1, it is possible to make statistical inference on the posterior distribution to obtain the model parameters, including estimating their optimal values.
Thus, the Bayesian modeling approach involves four steps:

1. Establishing a full prior probability model for the model parameters, using all available information;
2. Establishing a likelihood model conditioning values of the parameters on data;
3. Establishing a full posterior probability model by conditioning the likelihood model on new data using Bayes’ theorem; and
4. Assessing the fit of the model to the data and evaluating the optimal predictions from the posterior distribution by appropriate inference on the conditional posterior.

This sequence used in a Bayesian regression is illustrated in Figure 13.1.

The literature on Bayes’ modeling agrees on these basic steps. However, there are several details in the Bayes’ theorem calculation that cause difficulty in implementing those steps.

First, a choice of a prior distribution for each parameter \( \theta \) in the model must be made. There are many ways to parameterize the model (what to use for \( \theta \)) and what to use for the prior distribution, \( p(\theta) \), for each \( \theta \) in the model. There is also the need to decide which of the \( \theta \) will be allowed to vary and which will be specified by fixed hyperparameters. The choice of priors in the Bayesian modeling is very flexible; in addition, what is used for the prior often has a significant impact on the model. Sometimes, the prior distribution \( p(\theta) \) can be taken as uniform and unbounded; this is called an “improper” prior because the integral of \( p(\theta) \) over an unbounded \( \theta \) is infinite, and \( p(\theta) \) is therefore not a “proper” probability distribution with an integral of 1. Choosing a “flat,” “data-dominated,” or “non-informative” prior model means that the analyst believes that there is no prior information available—as is implied in the usual frequentist analysis of the data. The flat prior can be used for Bayes’ regression, and often is selected because it is simplest to implement and because it should give results very close to those from a frequentist regression using maximum-likelihood estimation.

Sometimes the form of a prior is selected to ease the computation. When a closed-form “analytical” solution is desired for a Bayesian analysis, as in a text or in an example, a conjugate prior is usually selected; the equation defining this prior comes from the same mathematical family of distributions as that of the distribution used for the likelihood model. This prior simplifies the calculations used to obtain the posterior distribution. For example, a multivariate Gaussian prior can
be imposed on the parameters; this prior is conjugate to a multivariate Gaussian likelihood distribution. The value of the covariance of the multivariate Gaussian prior distribution can be set as a hyperparameter—usually taking the variances as fixed and homoscedastic—or the covariance can itself be regarded as a stochastic model parameter and allowed to adjust according to experimental data. Using either of these Gaussian priors with a Gaussian likelihood leads to a Gaussian posterior, and the computation (and the mathematics to derive and to integrate the conditional marginal posteriors) is greatly facilitated. For many situations, for example, in modeling with non-Gaussian likelihoods, using a conjugate prior is often not possible, however. Fortunately, computational methods have improved greatly so that a numerical solution can be obtained, as is necessary in many of the applications of Bayesian regression methods to spectral calibration considered here; having convenient math is not always essential.

Once the posterior distribution is obtained, some sort of calculation must be used to find an optimal value for the \( \theta \) parameter(s) in the model. To make that calculation, a decision is made on what is meant by \( L(\theta, \hat{\theta}) \) “optimal.” Optimality must be specified through choice of a loss function that determines the consequences of any difference between the true model parameter value \( \theta \) and its estimate \( \hat{\theta} \) (Berger, 2005). Much of the Bayesian literature often uses the maximum a priori (MAP) “0–1” loss function, where only an estimate made at the true value is considered useful, and all other values are considered worthless, regardless of their proximity to that true value. Choice of this loss function leads to using the mode of the posterior distribution as the optimal estimate of the parameter \( \theta \). When the prior distribution \( p(\theta) \) is uniform, the use of this loss function results in a conventional maximum-likelihood estimate (MLE) \( \hat{\theta} \). Clearly, this choice is best used with unimodal conditional posterior distributions; not all posterior distributions are, in fact, unimodal, especially in regressions.

It is also possible to base optimality on minimizing the usual sum of squared differences between the true model parameter value \( \theta \) and its estimate \( \hat{\theta} \) (the conventional “squared” \( L_2 \) loss function). This choice leads to the usual minimum-variance (least squares) estimate using the mean \( m(\theta) \) of the posterior distribution \( p(\theta|D) \), where

\[
m(\theta) = \sum_k \theta p(\theta|D) \tag{13.2}
\]

is the least squares estimate for the mean of parameter \( \theta \). This estimate does not rely on unimodality of the posterior and is commonly used when the posterior is multimodal. Thus, how the posterior distribution is interrogated matters, and different ways of interrogation used in inference often give rise to different estimates of the parameters \( \theta \).

### 13.1.3 Belief and Priors

The many choices involved in setting up and interrogating a Bayesian model permit great flexibility in the modeling. This flexibility is a significant advantage in use, but it poses a major impediment to the newcomer to Bayesian analyses. The prior model can be structured to capture any information on the model that is available before data are observed. For example, the prior may reflect the belief that an intercept should be zero in a Beer’s law model, but unlike the usual frequentist approach to modeling in which the intercept term is often dropped, the Bayesian approach puts a prior mean value of zero on this parameter, reflecting the belief that the value of the intercept parameter is usually zero, and then allows the available data to alter that value, or not, as appropriate. In a more complex multivariate model, priors may be set on all of the parameters, reflecting the belief that some terms in the regression model are more important than others in the modeling. Here, too, the data then provide information that alters these priors, reflecting the contribution of each term in the model based on the data and the loss function employed.
Because the choice of priors is subjective, reflecting the skill and prior knowledge of the researcher, any probability statement about the parameters resulting from the analysis is interpreted as the “degree of belief.” This seemingly arbitrary selection of prior belief is balanced by the fact that Bayes’ theorem provides the only consistent way to modify belief about parameters given observed data.

13.1.4 Establishing the Likelihood

The likelihood function also permits a flexibility not seen in the usual frequentist approach. Because the likelihood is a conditional probability distribution, and not a simple matrix algebraic relationship, it is possible to encode complex relationships between model parameters and data. Even empirical relationships, as discussed later, can be encoded in the likelihood. There is a cost involved with this flexibility: The full conditional distribution defining the likelihood must be defined over its region of applicability. While a frequentist relation relates data to parameters by a deterministic, algebraic equation, the Bayesian likelihood distribution is defined by the measured data, and fully defining the likelihood can sometimes require a substantial amount of data. The Bayesian model can use a deterministic likelihood structure defining what explanatory variables are included and how they are related to the response variable and to one another, as in a frequentist analysis, but the joint probability distribution for all observable and unobservable quantities in the problem must also be defined in the full probability model for the likelihood.

13.1.5 Obtaining and Assessing the Posterior

With the data defining the likelihood distribution, the appropriate posterior distribution can be computed from Equation 13.1. Getting the distribution $p(D)$ needed in Equation 13.1 is usually not an easy task, so either this quantity is conventionally omitted or it is treated as a normalization constant, and given a value so that the posterior probabilities integrate to 1 over the region of applicability.

The posterior distribution that results from Equation 13.1 is a joint probability distribution of all parameters, conditioned on the observed data. The multiple integration that is needed to estimate the marginal conditional posteriors for each of the parameters can be difficult; making this step easier is the main reason for choosing a conjugate prior. If the prior is not conjugate to the likelihood, integrating out the marginal conditional posterior distributions is challenging even when the number of parameters is small, and is almost always done numerically. However, even numerical integration may not be simple or quick when the number of parameters in the Bayesian model is large. In some cases, it is possible to simulate the posterior and then use sampling methods to estimate distributional parameters, but with “large” numbers of parameters in the model (more than about 30), even a sampling scheme as illustrated below can sometimes take substantial time, and approximate computational methods must be employed.

The numerical integrations and the significant computational burden they impose have been the largest impediment to use of Bayesian methods. Early workers in Bayesian multivariate regression used a Cray X-MP supercomputer for a Bayesian cross-validation approach to calculate contributions of only 10 predictors to a regression model (Mitchell and Beauchamp, 1988). The computational burden that these methods can impose has been substantially reduced by improvements in computation and by new approaches to extracting the marginal posteriors. These newer approaches permit the use of Bayesian regression on the large numbers of predictor variables typical in NIR spectral data, with results available in a few minutes’ time, even on a laptop.

And last, once the marginal conditional posteriors are available, inference is performed on the posterior distributions: The parameters in the model are estimated, an assessment is made of how well the model fits the data, and predictions are made from the model by adding new data $x$ to $D$ and determining $y$ from the joint posterior distribution. In almost all cases, a numerical approach is used.
because deciding on what is included in the model and deciding on the priors for those parameters and suitable values for any hyperparameters may require multiple passes of model building.

13.1.6 Simulation of the Posterior Distributions

By far, the most popular way of getting a numerical solution for a marginal conditional posterior from a Bayesian regression involves the use of a simulation. Most of these simulations are based on Markov chain Monte Carlo (MCMC) methods. An MCMC method relies on random draws (that’s the Monte Carlo part), where the next step depends only on the current status, with no memory of historical status information (the Markov chain part). The aim is to simulate the target posterior distribution by sampling, with the distribution of any draw dependent on only the last sample drawn, so that the approximate distributions improve at each step. Eventually, the simulated distribution converges on the true target posterior distribution. MCMC simulation is sometimes implemented by the Metropolis–Hastings algorithm, a kind of random walk with an acceptance/rejection rule that is used to converge on a specific target distribution. This algorithm uses a jumping distribution in which the mean value of the distribution either stays where it is currently or moves to a new value, with the probability of movement depending on the ratio of probability densities between the new location and the current one. Steps that increase the density ratio are always accepted, but some steps that decrease density are, too, to avoid any local maxima. The Gibbs sampling algorithm is an alternative to Metropolis–Hastings that makes a movement based on distributions of each variable, conditioned on the values of all other variables. More details on these numerical simulation methods go beyond the scope of this review, but can be found in texts (e.g., Robert and Casella, 2004) and in the literature.

MCMC modeling using either a Metropolis–Hastings algorithm or Gibbs sampler depends on two important practical considerations. One is the number of trials needed to remove any significant effect from initial guesses used to start the simulation—the “burn-in” period. Once the burn-in period is reached, the shape of the estimated posterior distribution stabilizes and the parameter estimates from that distribution are improved considerably. Often, the burn-in for simulation of the posterior is long, and results are sometimes discarded from about 1/2 of the Markov chain, which might be thousands or more trial calculations, so burn-in for MCMC simulation can require significant computation time. And, second, the Markov chain optimization usually shows strong serial correlation in successive guesses, regardless of the algorithm used to choose the guess at each step. Successive guess values give little new information on the shape of the posterior, so these are usually “thinned” by choosing and keeping only one result from a set number of trials to reduce effects of serial correlation. The frequency of accepted points from the Markov chain is called the thinning rate, and it involves even more computation. Because of the need for burn-in and thinning to get an accurate simulation of the marginal posterior distributions, MCMC estimation of the posterior can be quite computationally demanding; this computation limits what can be done in a Bayesian analysis of highly multivariate data. Because the posterior is very high dimensional, many of parameters in the model are not statistically independent, and a marginal posterior might be required for each of the parameters in the multivariate model.

Given the issues mentioned above, it might seem that a Bayesian analysis is quite demanding to reduce to useable computer code, but there are now several packages that can be used to implement a typical Bayesian regression easily, and with minimal coding. The software in such free packages as Bayesian inference Using Gibbs Sampling (BUGS) and its variants RBUGS and WinBUGS, in STAN, in JAGS (Just Another Gibbs Sampler), and in other packages, now makes much of the computational side of a Bayesian regression analysis very straightforward, with user-selectable priors and built-in MCMC sampling. These packages generally run as an add-on to basic R or python software and in some cases to MATLAB, too. Some, like the WinBUGS package, are stand-alone. As the BUGS site notes in their “health notice,” though, MCMC simulation is very effective, but it is not as robust as a closed-form solution, and there is no oversight against a careless application or misuse of the software, regardless of the package used.
13.2 BAYESIAN REGRESSION

The use of a prior distribution for the parameters in the model is an important main distinguishing characteristic of a Bayesian analysis. Even in the simplest case, the choice of the prior model makes a significant difference in the results from a Bayesian regression, but for the moment, the focus will be placed on the likelihood model. For example, consider the case where the data \( D = [X; y] \) are modeled with a classical least squares (CLS) regression. This regression relationship defines a likelihood model where the deterministic predictor variables \( X \) relate to the stochastic response data \( y \):

\[
y = \beta^T X + \epsilon
\]  

(13.3)

Here, the random fluctuation in the dependent variable is set by the error \( \epsilon \) for each predictor variable. Using the conventional shorthand, the distribution for the error \( \epsilon \) is taken as IID(0, \( \sigma^2 I \)) for each \( y \), so the observed responses \( y \) in the likelihood are stochastic, unlike the deterministic, observed data in \( X \). Now, Equation 13.3 defines the explicit relationship between the model parameters \( \beta \) and the observed data in \( X \) and \( y \), but this relationship is expressed as a distribution. The error term \( \epsilon \) is Gaussian-distributed with mean 0 and fixed variance \( \sigma^2 \), and since Equation 13.3 defines the relation between the observed data and the parameters, the likelihood expresses that relation as a multivariate Gaussian distribution with mean \( \beta^T X \) and variance \( \sigma^2 \), giving the likelihood distribution \( p(y|X, \beta) = G(y; \beta^T X, \sigma^2) \).

In this model, there is no error in the deterministic predictor variable \( X \), but in the Bayesian view, the model parameter \( \beta \) is stochastic because it is not an observed quantity. The Bayesian model must specify its prior probability distribution, \( p(\beta) \).

In this Bayesian model, the variance \( \sigma^2 \) is a hyperparameter; it is a deterministic parameter with a value set outside of the Bayesian modeling, so, while the value of \( \sigma^2 \) affects the performance of the model, as a hyperparameter, it is not assigned a prior and is not adjusted by the experimental data.

It is convenient to use a zero-mean Gaussian prior \( p(\beta) \) for the regression parameter \( \beta \) because this prior distribution is conjugate to the Gaussian likelihood distribution, so a zero-mean distribution with fixed variance is selected:

\[
p(\beta|\alpha) = G(\beta; 0, \alpha^{-1} I)
\]  

(13.4)

conditioned on the hyperparameter \( \alpha \), a measure of the spread of the multivariate Gaussian distribution defined in Equation 13.4 above. Note that the quantity \( \alpha^{-1} I \) indicates that the covariance term is diagonal in this model, so the prior defined in Equation 13.4 assumes statistical independence of the parameters in the \( \beta \) vector.

Then, after experimental data are used to populate the likelihood distribution \( p(y|X, \beta) \), from application of Bayes’ theorem, the un-normalized, conditional posterior in Equation 13.5 is obtained directly:

\[
p(\beta|X, y) \sim p(y|X, \beta) p(\beta|\alpha) = G(\beta; m, S)
\]  

(13.5)

where the posterior mean is

\[
m = \alpha^{-2} SX^T y
\]  

(13.6)

and the posterior variance is

\[
S = (\alpha I + \sigma^{-2} X^T X)^{-1}
\]  

(13.7)
The proportionality ~ in Equation 13.5 arises because division by \( p(X) \) has been omitted from the product of likelihood and prior, so the conditional posterior is un-normalized, and the area under the distribution defined in Equation 13.5 will not be 1.

The solutions to Equation 13.5 differ, depending on what value of the hyperparameter \( \alpha \) is used in the conditional prior distribution for the regression parameters \( \beta \). If \( \alpha = 0 \), the prior \( p(\beta|\alpha) \) is a uniform distribution (often called a “flat” distribution). If, however, either an inverse gamma distribution conjugate prior is used for the prior distribution of variance \( p(\sigma^2) \) or the variance hyperparameter is taken as fixed value \( \alpha \), the conditional posterior solution that results from using a squared loss function is the familiar maximum-likelihood estimate:

\[
\beta = (X^T X)^{-1} X^T y
\]

(13.8)

the same result as obtained from the usual frequentist solution to Equation 13.3. This agreement occurs because both maximum-likelihood regression (and its special case of least squares regression) and the Bayesian regression model developed here follow the likelihood principle. Since these models both share the same likelihood and prior, the inferences from them must be identical. By providing a flat prior distribution to the Bayesian regression, no information is available on \( \beta \) before the data are obtained, and it is reasonable that the “usual” (frequentist) approach—which also admits no prior information—and the Bayesian approach with no useful prior information should agree.

However, if \( \alpha > 0 \), there is information about the model parameters in the prior. In this case, the prior distribution of \( \beta, p(\beta|\alpha) \), is \( G(0,1/\alpha) \), a normal distribution with mean vector 0 and variance \( 1/\alpha \), and now the MLE estimate that results from the Gaussian posterior is

\[
\beta = \sigma^{-2}(\alpha I + \sigma^{-2}X^T X)^{-1} X^T y
\]

(13.9)

Equation 13.9 has the form of a conventional ridge regression with ridge parameter \( \alpha \sigma^2 \), a function of the variance \( 1/\alpha \) of the prior distribution \( p(\beta|\alpha) \) and the variance \( \sigma^2 \) of the stochastic parameters in the model for \( y \). The ridge solution in Equation 13.9 is a sparse solution to the regression given in Equation 13.3. This solution expects that most of the terms in the regression parameter \( \beta \) are nearly zero because the information provided in the Gaussian prior indicates that these coefficients are, on average, zero. In this case, a relationship results that is different from the usual frequentist model, reflecting the differing contributions of the prior, the prior belief in how close each of the parameters is to zero (as measured by the hyperparameter \( \alpha \)), the hyperparameter describing variance in the \( y \) data (\( \sigma^2 \)), and the conditional posterior probability that results from these choices.

The posterior has the form of a ridge regression in this case, but unlike conventional ridge regression, the ridge parameter is not determined by a cross-validation procedure, where multiple regression models are created over portions of the data and the root mean squared error of prediction for the withheld portion of the data is determined as a function of the ridge parameter. There are two common ways to get appropriate estimates for hyperparameters in a Bayesian regression. One is to maximize the marginal likelihood distribution as a function of the hyperparameter over the entire dataset, a method known in the literature as evidence approximation (Bishop, 2006). It is also possible to form another, higher-level Bayesian model where the hyperparameters from the lower-level model are the stochastic parameter components of a higher-level model, giving what is called a hierarchical model. The hyperparameters are then modeled with a prior and likelihood, and estimated from their posterior, just as in any other Bayesian analysis (Gelman, et al., 2013).

Prediction of the \( y \) value given a new, observed \( x \) done using any of these Bayesian models is also somewhat different than that done with a frequentist model. Once the posterior distribution of \( \beta \) is determined, a prediction of a new \( x_{\text{unk}} \) to yield \( y_{\text{pred}} \) is accomplished by integration over the posterior as shown in Equation 13.10:

\[
p(y_{\text{unk}}|y,\alpha,\sigma^2) = \int p(y_{\text{unk}}|\beta,\sigma^2)p(\beta|y,\alpha,\sigma^2) d\beta = G(y_{\text{unk}};m^T X_{\text{unk}},\sigma^2(x_{\text{unk}}))
\]

(13.10)
where

\[
\sigma^2(x_{\text{unk}}) = \sigma^2 + x_{\text{unk}}^T S x_{\text{unk}}
\]

(13.11)

Because the prior is conjugate to the likelihood distribution, the conditional marginal posterior is another Gaussian distribution, and its mean provides the maximum-likelihood estimate for the unknown \( y \) value. A benefit not seen in a frequentist solution is that the uncertainty in a single prediction—obtained from inference on the variance of the posterior after updating for the new data in \( x_{\text{unk}} \)—is readily available, too. Not surprisingly, the uncertainty in the predicted property \( y_{\text{unk}} \) depends on the uncertainty of the calibration data and the uncertainty in the estimates of the model parameters, but also on the new data \( x_{\text{unk}} \).

From the Bayesian perspective, the solutions obtained by using classical least squares and ridge regression reflect the different prior distributions chosen and our belief as to the correct values of the regression parameters. The “flat” prior distribution indicates both a lack of prior information and an openness to any value for the regression coefficients in \( \beta \). Proponents of Bayesian approaches consider the flexibility that results from this approach quite useful, but it is worth noting that, assuming Gaussian priors for each parameter, two hyperparameters (mean and variance) must be specified in these prior distributions for each of the terms in \( \beta \), reflecting prior knowledge or a best guess of each value and of its uncertainty.

When the multivariate Gaussian distribution is used as a prior, the choice of its mean location reflects an opinion as to the likely value of each term in \( \beta \), and its variances reflect the strength of that prior information—the degree of “belief” in each value. When the mean of the multivariate Gaussian prior is not zero, there is a bias introduced in the regression in the strict “frequentist” sense. In the Bayesian analysis, some of the possible values for model parameters are regarded as more likely than others, not in the sense of their frequency of occurrence, but in the sense that these values are more consistent with what is known about this system—there is a higher degree of belief in these values. This is similar to the dropping of the intercept term in fitting Beer’s law models to spectral responses as a function of concentration.

The prior belief provided in a Bayesian analysis interacts with the new data in a sensible way, too. Limited new data generally have a small effect on the posterior distribution unless a flat prior is used; the information provided in the prior dominates in the posterior because any information in the small amount of new data, even if different from that provided in the prior, is not sufficient to substantially change the information about the model provided as part of the prior. Bayesian methods are not as useful when applied to small datasets because they are so sensitive to the specifics of the prior in these cases. The practitioner should keep this aspect of Bayesian methods in mind. A poor choice of prior, including providing a flat prior in connection with having limited data, is not a good strategy to obtain a useful Bayesian analysis.

However, providing a prior that is highly consistent with information in the new data also results in a very small change in the posterior, even with a large amount of new data. Change in the posterior distribution from that of the prior distribution occurs only when there is a large enough difference between information provided in the prior and information extracted from the incoming data. The degree of belief provided in the variance hyperparameters that are part of the prior determines what difference is “large enough” to cause the posterior distribution to differ from that of the prior.

So, while consistency between the prior and posterior distributions can be a sign of either consistency of prior and new information or woefully insufficient data, when a large amount of new data is provided, and when the information in the new data is not consistent with that in the prior, the information in these new data dominates in the posterior, and the model is updated by the new information. In this case, the specific prior that is used may not matter! With enough new data, the prior is forgotten in the analysis. This updating of the model by extensive new information is now commonly used in chemical process modeling and in other applications of Bayesian multivariate time-series modeling (Nounou, et al., 2002a).
13.2.1 An Example of Simple Bayesian Regression

Unlike a conventional frequentist regression, a Bayesian regression does not give a set of fitted parameters as output, but it gives a posterior distribution. Getting the information needed from that distribution will require some work.

Two examples illustrate the steps needed.

First, consider a simple two-parameter linear regression using the model in Equation 13.3. Estimates of the model parameters $b_0$ and $b_1$ are sought, given some data $(x_i, y_i)$ shown in Figure 13.2.

We presume that the dependent variable $y$ is stochastic and is normally distributed about a mean $y$ value where

$$\bar{y} = b_0 + b_1 \bar{x}$$

with variance $\phi$, so that the conditional distribution

$$p(y_i | \beta_0, \beta_1, \phi) \sim N(\beta_0 + \beta_1 x_i, \phi)$$

defines the likelihood model.

The prior model includes our belief on the distributions of each parameter in the model. Again, this model is specified by distributions:

$$\beta_0 \sim N(m_0, t_0)$$

$$\beta_1 \sim N(m_1, t_1)$$

$$\Phi \sim IG(shape = a, rate = \gamma)$$

Here, we assume that each of the beta values is independently and normally distributed, each with a pre-specified, but possibly different mean $m$ and variance $t$, and that the uncertainty in $y$ is inverse gamma (IG)-distributed with mean $a$ and variance $\gamma$. The distributions and their set of values $(m_0, m_1, a, t_0, t_1, \gamma)$ that determine this prior model are the hyperparameters for this analysis. All are adjustable, but these are fixed at specific, preset values for the Bayesian analysis that follows. Hyperparameters are taken as deterministic in a basic Bayesian analysis.

This set of models has the non-normalized (to within a constant) posterior

$$p(\beta_0, \beta_1, \phi | y) \sim \prod_{i=1}^{n} p(y_i | \beta_0, \beta_1, \phi) p(\beta_0 | m_0, t_0) p(\beta_1 | m_1, t_1) p(\phi | a, \gamma)$$

(13.14)

FIGURE 13.2 Noisy data for the example Bayesian regression.
because Equation 13.14 does not include division by $p(x)$. It happens that, because the priors are Gaussian and inverse gamma, and the likelihood is also Gaussian, the priors are conjugate to the likelihood, and the joint posterior distribution in this special case is closed form, the parameters are all independently distributed and can be evaluated analytically. Here, though, the posterior is determined numerically. The posterior for $\varphi$ is statistically independent of those for $\beta$, so it can be considered separately, without the need to marginalize it from the joint posterior of the other parameters. This posterior is another inverse gamma distribution, with

$$
\text{shape} = a + n/2
$$

and

$$
\text{rate} = \frac{1}{2\varphi} \sum_{i=1}^{n} \left( y - (\beta_0 + \beta_1 x_i) \right)^2 + \gamma
$$

so just the joint posterior for $\beta$ is considered here. The Gibbs sampler is used to find the marginal conditional posterior distributions of the $\beta$ terms.

First, the un-normalized, marginal, conditional posterior is numerically evaluated for each of the $\beta$ parameters in this model at specific values of $\beta_0$ and $\beta_1$. Then, a grid is set up of probability density over these possible values. A kernel smoother is sometimes used to increase the resolution of entries on the grid. Then, the Gibbs sampler is used to sample from the conditional posterior at each grid point. Approximation of the joint posterior by sampling from a pre-calculated grid is a time-consuming way to avoid the algebra associated with a closed-form solution but is very useful when the solution is not closed form. The sampling is not required here, but it is usually necessary in larger problems because the conditional posteriors cannot generally be obtained in closed form.

There are a few other issues that arise in using a grid. One is that the span must be selected to include all “reasonable” values that can be expected for the parameters. Deciding on the appropriate span of parameter values may take some trial and error. Second, because an un-normalized posterior is evaluated to get conditional marginal posteriors, it is possible to encounter numbers that exceed the software’s/computer’s capability to represent them. To avoid problems with underflow or overflow, the logarithm of the conditional probability is normally evaluated from the grid and then converted later.

For this simple regression, done on 100 samples of synthetic data, the Gibbs sampler, using only 1000 draws from the gridded conditional posterior and a burn-in period of 100 samples with no thinning used, provides very good estimates that come close to the true values for $\beta_0$, $\beta_1$, and $\varphi$.

Figure 13.3 shows the MCMC traces for the Gibbs sampling used to estimate the conditional posterior in this regression. The sampling has stabilized and the fluctuations of the distribution means, corresponding to MLE estimates of $\beta_0$ and $\beta_1$, are generally about the true values (indicated in the figure by the horizontal lines). The estimates for the variance $\varphi$ also are stable, though there is more fluctuation. These estimates are all positive because variance estimates cannot be negative.

The estimates of the model parameters are shown in Figure 13.4. The multivariate normal relationship of the parameters is apparent. Inference here using the mean (a maximum-likelihood estimate) gives good agreement for the slope but only fair agreement for the intercept, a consequence of the considerable additive noise in the data. The Bayesian regression gives $b_0 = 4.29 \pm 0.46$, $b_1 = 4.91 \pm 0.62$. A conventional least squares regression gives $b_0 = 4.11 \pm 0.26$ and $b_1 = 5.21 \pm 0.29$.

Even this small analysis takes some time because of the creation of the likelihood grid and the sampling of the marginal conditional posteriors. This Bayesian analysis takes about 2 minutes to complete on a recent Intel processor, while the conventional frequentist version of this regression is essentially instantaneous.
FIGURE 13.3  Markov chain Monte Carlo sequences from the Gibbs sampling for the simple Bayes regression. Only samples after the 100-point burn-in period are shown here.

FIGURE 13.4  Posterior joint probability for the regression parameter estimates from the simple Bayesian regression. The lines show iso-probability levels.
As the regression task becomes more complex, the computation time increases. A 5-parameter regression model \( y_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \phi \) is fit to the data; here, the model parameters are not assumed to be independently distributed. Using a flat, improper prior for the beta terms and an inverse-gamma-distributed variance term for the dependent variable \( y \) to model the observed data with this 5-parameter model takes about 7 minutes to run, fitting to only 50 observations, using 500,000 samples in the simulation (done here as block Gibbs sampling, where a vector draw is performed to both speed the calculation and to account for any interactions between the four model parameters), with a 100,000-sample burn-in and a trim rate of 1 in 50. Table 13.1 shows that the longer simulations and high trim rate pay off in the estimates. Other than the variance of the noise term, all parameter estimates are very close to the true values.

### 13.3 BAYESIAN CALIBRATION OF NEAR-INFRARED SPECTRA

These examples are very simple, and it is unlikely that Bayesian regression would be used to solve such simple modeling problems because the benefits will be minor and the effort more significant than that needed for the usual frequentist approach. When more complex models are fitted to large amounts of data, a Bayesian approach can show significant benefit over a frequentist model, however. Bayesian regression methods have become more common as the computational resources needed to implement them have become available. To date, however, much of the published work using Bayesian regression has been carried out by statisticians. A significant fraction of their published material has considered Bayesian regression on near-infrared (NIR) spectrometric data.

The complexity of the modeling and the computational demands make a Bayesian regression on near-infrared spectra a challenge. The benefits of performing the regression arise in three main areas: (1) selecting wavelength variables important to the modeling; (2) obtaining one-sample estimates of error in a result; and (3) modeling nonlinear and nonparametric relationships in spectral data. Each of these is discussed below.

#### 13.3.1 DIRECT VARIABLE SELECTION IN CALIBRATION USING BAYESIAN REGRESSION

One common motivation for Bayesian regression is a desire to reduce variables in some systematic way or to discover those variables most important to a model when there are many variables. That information is part of the posterior.

Selecting variables can be done by putting a prior distribution on each regression coefficient \( \beta \). Using Bayesian methods to identify important variables in multivariate data by using a prior distribution was first reported by Mitchell and Beauchamp (1988). They selected variables using a non-informative prior on the variance in \( y \), and a “spike-and-slab” prior distribution for each of the regression parameters \( \beta_j \). A spike-and-slab distribution is a proper uniform distribution between two limits \(-f_j \) and \(+f_j\)—the slab—but with extra density concentrated at 0—the spike—if the variable

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Estimator Mean</th>
<th>Estimator Variance</th>
<th>Bias</th>
<th>Percent Bias (of Truth)</th>
<th>Coverage of 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_0 )</td>
<td>1000</td>
<td>1000.55</td>
<td>26,358</td>
<td>0.55</td>
<td>0.06</td>
<td>0.95</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>50</td>
<td>50.53</td>
<td>348</td>
<td>0.53</td>
<td>1.06</td>
<td>0.95</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>-50</td>
<td>-50.05</td>
<td>2.06</td>
<td>-0.05</td>
<td>0.10</td>
<td>0.94</td>
</tr>
<tr>
<td>( b_3 )</td>
<td>10</td>
<td>9.99</td>
<td>2.64</td>
<td>-0.01</td>
<td>-0.11</td>
<td>0.96</td>
</tr>
<tr>
<td>Phi</td>
<td>10,000</td>
<td>10,401.96</td>
<td>40,97,173</td>
<td>401.96</td>
<td>4.02</td>
<td>0.96</td>
</tr>
</tbody>
</table>
is vulnerable to deletion. The parameter $\gamma_j$ relates the probability density at the “spike” to the height of the “slab” of uniform probability density. To exempt the regression coefficient corresponding to variable $j$ from removal, the prior for the parameter is set so that $\gamma_j = 0$, ensuring no prior density for the $j$th variable at the value $\beta_j = 0$. The spike-and-slab prior defined sets of possible variables in a number of submodels. To find suitable values for $\gamma$, they used a Bayesian cross-validation, in which a predictive error is calculated in a leave-one-out sense, but unlike conventional, frequentist cross-validation, the prediction of a given sample $x$ is not determined by estimating the response and then comparing it with the known value; instead, the posterior conditional distribution over the whole set is used in connection with the known response to estimate the error associated with estimating the response for each withheld sample directly, just as in a prediction of new data. Variable selection was investigated for synthetic data made with known models with and without predictor collinearity in a non-spectral dataset with ten related variables, and on a real, energy-conservation dataset. The posterior resulting from the spike-and-slab prior identified sets of submodels with different retained variables. Studies of synthetic data using this spike-and-slab prior showed that the Bayesian approach easily identified the terms in the data associated with larger regression coefficients, but variables with smaller coefficients and with zero coefficients in the true model often were misidentified as present, not zeroed out, with high posterior probability in the submodels. When the variables were made collinear in the synthetic datasets, the posterior density of several possible submodel choices was close, and choosing a single “best” submodel proved challenging, though competitive submodels with similar posterior probabilities tended to identify many, but not all, of the same predictor variables.

Even though this study did not attempt to examine spectra, because it is one of the first attempts to use the prior, as adjusted by experimental data, to decide on retention of variables in a regression model, it is important to the use of Bayesian methods in modeling spectral data. The datasets were small in size because of the computational burden imposed by the Bayesian regression and the cross-validation scheme they employed. Indeed, all of their work was done on a supercomputer, though it should be noted that in terms of computational capability, the processing power of the 1987-era Cray X-MP they used is slightly inferior to that of the 2012 Apple iPad 2 tablet (Larabel, 2012) on the LINPACK library, a conventional test!

It is also of interest to compare the model averaging done in this Bayesian approach with another common form of ensemble model averaging that does not use a Bayesian approach: stacking (Breiman, 1996; Tan and Brown, 2003). In stacking, weights are not determined from posterior probabilities of submodels, but they are determined by the errors from conventional cross-validation. Clarke has shown that when the true model is present in the set of submodels subject to Bayesian averaging, the results of the Bayesian regression are never worse than those from stacking, and may be significantly better. However, when terms in the Bayes’ submodel set are outlying to the behavior of the data because of bias or noise effects in the data, stacking performs better; overall, stacking seems more robust than Bayesian model averaging (Clarke, 2003).

The first work aimed at using Bayesian regression to perform modeling and variable identification on real near-infrared spectral data examined two sets: One set of NIR data was collected on sucrose, glucose, and fructose in aqueous solution, a set originally created to test modeling by frequentist regressions using inverse and classical least squares regression. This early work by Brown (1982) considered conventional multivariate calibration of the NIR spectrometric responses to the amounts of the sugars in aqueous solution. The calibration set consisted of 125 transmission spectra collected on designed mixtures of the 3 sugar components and measured over the range 1100–2498 nm in increments of 2 nm. Another 21 samples were provided as a prediction set. This set was constructed to pose a challenge in multivariate prediction; many of the test samples had levels of sugar that exceeded the range used in calibration. The sugar data are available from Vannucci (http://marina.blogs.rice.edu/software/).

The second NIR dataset was first studied by Osborne et al. (1993). This dataset is also available online (http://marina.blogs.rice.edu/software/). NIR spectrometry was used to measure the
composition of formed, but unbaked, pieces of cookie dough from their reflectance spectra. This set is also multivariate in response, with four constituents (fat, sucrose, dry flour, and water) subjected to calibration by varying a standard cookie recipe over a wide range. A total of 39 cookie dough samples were available in the calibration set, and another 39 samples were used for the validation set. All were measured by NIR reflectance from 1100 to 2498 nm with 2 nm resolution. The earliest work on this set made use of stepwise multiple linear regression to select wavelengths for a least squares calibration (Osborne, et al., 1993). This study treated the four target responses separately, using four stepwise, single-property classical regressions and finding different variable sets for each of the calibrated components.

Many of the published studies reporting Bayesian regression have examined one of the two datasets described above. These two sets have many of the issues typically seen in near-infrared spectra, and because they are spectral sets with far fewer samples than variables, they offer a challenge to Bayesian modeling. However, they also offer opportunities well suited to Bayesian analysis and highlight why a Bayesian approach has advantages in modeling near-infrared data. The most immediate problem with most NIR spectral datasets investigated by calibrations based on Bayesian regression lies in the large number of variables in the regression model. Both the sugar spectral data and the cookie dough spectral data have 700 variables, far too many to use in a Bayesian analysis of the sort illustrated above because of limitations in computer memory and because of the computational time needed to complete a full analysis. In addition, many of the spectral variables do not contribute independent information to the modeling. Near-infrared spectra commonly are relatively featureless and have strong correlations among the spectral variables. A third, related issue is that the number of samples collected for an analysis is less, sometimes far less, than the number of spectral variables.

For a Bayesian regression analysis of data with this many variables, variable reduction has to be performed. Thus, one focus of the many reports on Bayesian methods in analysis of NIR data has been on methods to reduce the size of the data by eliminating "unimportant" spectral variables from the data.

Brown revisited the sugar set with Fearn and Vannucci (Brown, et al., 1998) using a Bayesian approach. They began by arbitrarily reducing the number of variables from the original 700 variables available from the spectrometer to a set of 160 equally spaced, new variables located by linear interpolation of the original data. In addition to the usual priors on the regression parameters in the model, they added an additional mixture prior indicating whether or not an interpolated wavelength channel in the model was useful in modeling. The prior was implemented as a Bernoulli distribution on each of the 160 interpolated variable channels, so only a value of 0 or 1 was possible for each. In this way, the variance term in the prior model could be used to exclude a variable channel from consideration by setting both the mean and variance of the channel to 0 in the prior. Even with the reduction in variables from the original 700 wavelengths to 160 interpolated variable channels, the Bayesian regression analysis with conjugate priors done in closed form took far too long to compute on computers available then. Their estimate was that a closed-form solution to estimate conditional posteriors could be run with at most 20 spectral variable channels on the computational equipment then available. Even a version that used a Gray-code scheme to calculate the marginal conditional probabilities in an effort to speed calculations was not computationally feasible with more than about 25 spectral variables (Brown et al. 2002). On the other hand, the Gibbs sampling approach outlined above could be used with the entire reduced set of 160 variable channels. A prior that initially selected 20 channels from these data with wide latitude on admitting more (or eliminating) variable channels was implemented. Five modeling runs were performed, each with a different starting point for the analysis: all selected wavelength channels initialized in the prior as 1 in the Bernoulli distribution; half of the selected wavelength channels initialized as 1; a random set of 20 of the wavelength channels initialized as 1; all but 20 channels initialized as 1; and the first 20 wavelength channels initialized as 1. In all runs except the last, the set of 160 possible wavelength channels were randomized prior to the run in an attempt to break up the strong serial correlations present in the NIR spectra.
Within about 200 samples, the MCMC traces settled to select about 10 channels of the 160 possible. Fairly consistent sets of wavelength channels were selected over these five runs, despite the widely differing priors used to start the analysis, an indication of the information present in even these greatly reduced spectra. Of the 160 possible channels, a total 23 wavelength channels were selected from the 5 runs, where each run independently selected 10. The average of these five Bayesian models accounted for 55% of the probability examined and gave mean squared errors of 0.116, 0.361, and 0.351 for the prediction of sucrose, glucose, and fructose, respectively. Using only the 10 wavelength channels selected most often, and no model averaging, accounted for 20% of the visited probability and gave somewhat worse prediction mean squared errors of 0.210, 0.446, and 0.510. Considering the nature of the prediction set, these are fairly good results, especially given that the wavelength channels selected are not the ones that maximize the usual cross-validated error, since no cross-validation at all was performed, and all predictions were generated from integration of the marginal posterior, as discussed above.

Chen and Martin (Chen, et al., 2007c; Chen and Martin, 2009) also used a Bayesian regression to select variables from NIR. Like Brown et al. (1998), they used a Gaussian prior on the regression parameters in $\beta$, but they considered the multivariate case with M multiple, statistically independent response variables in $Y$. While the covariance of the $Y$ is diagonal, the analysis does not reduce to a set of separate Bayesian regression models developed independently on univariate responses because the prior distribution $p(\beta|\alpha^{-1})$ for each set of parameters is based on the same hyperparameter $\alpha^{-1}$, reflecting Gaussian distributions with the same prior mean but with different variances for the regression parameters in the M response models. They took the four responses as independent and assumed independence of the regression coefficients for each response, so

$$p(\beta|\alpha) = \prod_{i=1}^{M} G(\beta_i; 0, \alpha_i^{-1})$$  (13.17)

from which $S$, the vector of posterior variances of $\beta$, is obtained

$$S = (A + \sigma^{-2}X^TX)^{-1}$$  (13.18)

where $A = \text{diag}(\alpha_1, \alpha_2, \alpha_3,...,\alpha_M)$. Evidence approximation (Bishop, 2006) is then used to find maximum-likelihood estimates for $\alpha$ and $\sigma$. The hyperparameter $\alpha_i$ is used to decide on retention of variables: When the value of $\alpha_i$ falls below a threshold, the corresponding $\beta_i$ can only be close to zero, and the $i$th predictor has little impact on the regression, effectively removing it from the regression model. The authors claim that the size of the threshold is not important to the predictive performance of the model and to the variables that are ultimately excluded, so long as the threshold is set high enough. This approach to identifying variables in a Bayesian regression is known as automatic relevance determination (ARD) (MacKay, 1992; Neal, 1996), and it is increasingly used in imaging NIR spectrometry and NIR tomography (Miyamoto, et al., 2015).

The variable selection was used in developing calibrations in two datasets based on NIR spectra. The first set studied was the wheat kernel dataset (Pedersen, et al., 2002), available from http://www.models.life.ku.dk/wheat_kernels as a MATLAB “mat” file. This set consists of a training set of 415 samples of wheat measured by reflectance over 100 wavelengths from 850 to 1050 nm. A test set of 108 additional samples is also included. The calibration related the percentage of protein to the NIR spectra. This set is challenging to calibrate because of the large amount of variation in the spectral data caused by the use of different varieties of wheat grown at different locations, and the delay in measurement of the test samples relative to the calibration samples.

The second study reported results for calibration of the four response variables moisture, oil, protein, and starch in the extensively studied Cargill cornmeal dataset, available as a MATLAB “mat” file from Eigenvector Technologies (http://www.eigenvector.com/data/Corn/). This heavily studied
dataset consists of near-infrared spectra of 80 cornmeal samples, each measured in reflectance on 3 instruments; one instrument (labeled m5) was a NIRSystems 6500 spectrometer and the other two (labeled mp5 and mp6) were different NIRSystems 5500 spectrometers. All spectra were originally measured over the wavelength range 1100–2498 nm at 2 nm resolution.

Both studies compared the predictive performance of Bayesian regression with that of conventional partial least squares (PLS) regression. Full Bayesian regression, using all variables, and regression with the adaptive selection of variables discussed above were compared with PLS regression using the wheat spectra. The root mean squared error of prediction (RMSEP) from PLS regression was 0.70%. Bayesian regression using the Gaussian prior, but without variable selection, gave an RMSEP of 0.62% on the same data. When the adaptive variable selection discussed above was employed, the Bayesian regression gave an RMSEP of 0.55%, using only 16 of the predictors. The sparse set of 16 predictors used in the adaptive model generally agreed with those wavelength regions having larger posterior means for the regression parameters found without adaptive variable selection.

Both separate and multivariate Bayesian regression models were developed on the data for the three NIR instruments in the cornmeal dataset. The results from the four univariate response Bayesian regressions were compared to those from univariate PLS-1 regression models. Results from the multivariate Bayesian regression were compared to those from a multivariate PLS-2 regression model. The PLS models generally underperformed the Bayesian models in all cases. Table 13.2 summarizes their results. The m5 dataset is of notably higher quality, and many regression methods do well on predictions of properties using these spectral data, so it was not surprising that on this set, the univariate Bayesian regression for moisture using adaptive variable selection had a very low RMSEP, about 10% of that reported from PLS-1 regression. It was also noteworthy that moisture was so well predicted by a very sparse model with only two spectral variables, at about 1900 and 2100 nm. The PLS-2 model for the cornmeal produced mixed results: Predictions of protein were much superior to those from a PLS-1 model built on only protein, but predictions were worse for the other responses than those obtained from the corresponding univariate PLS-1 models. The multivariate Bayesian regression with adaptive variable selection also underperformed the four

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Moisture</th>
<th>Oil</th>
<th>Protein</th>
<th>Starch</th>
</tr>
</thead>
<tbody>
<tr>
<td>m5</td>
<td>PLS-1</td>
<td>0.036</td>
<td>0.154</td>
<td>0.317</td>
<td>0.261</td>
</tr>
<tr>
<td>m5</td>
<td>Bayes</td>
<td>0.018</td>
<td>0.140</td>
<td>0.127</td>
<td>0.230</td>
</tr>
<tr>
<td>m5</td>
<td>Bayes-VS</td>
<td>0.004(2)</td>
<td>0.101(23)</td>
<td>0.072(27)</td>
<td>0.211(21)</td>
</tr>
<tr>
<td>mp5</td>
<td>PLS-1</td>
<td>0.184</td>
<td>0.207</td>
<td>0.365</td>
<td>0.698</td>
</tr>
<tr>
<td>mp5</td>
<td>Bayes</td>
<td>0.190</td>
<td>0.172</td>
<td>0.297</td>
<td>0.556</td>
</tr>
<tr>
<td>mp5</td>
<td>Bayes-VS</td>
<td>0.164(20)</td>
<td>0.153(32)</td>
<td>0.237(20)</td>
<td>0.531(22)</td>
</tr>
<tr>
<td>mp6</td>
<td>PLS-1</td>
<td>0.280</td>
<td>0.196</td>
<td>0.262</td>
<td>0.836</td>
</tr>
<tr>
<td>mp6</td>
<td>Bayes</td>
<td>0.211</td>
<td>0.192</td>
<td>0.218</td>
<td>0.559</td>
</tr>
<tr>
<td>mp6</td>
<td>Bayes-VS</td>
<td>0.194(18)</td>
<td>0.180(17)</td>
<td>0.198(30)</td>
<td>0.508(23)</td>
</tr>
<tr>
<td>m5</td>
<td>PLS-2</td>
<td>0.046</td>
<td>0.206</td>
<td>0.144</td>
<td>0.408</td>
</tr>
<tr>
<td>m5</td>
<td>mBayes</td>
<td>0.038</td>
<td>0.142</td>
<td>0.143</td>
<td>0.330</td>
</tr>
<tr>
<td>m5</td>
<td>mBayes-VS (10)</td>
<td>0.005</td>
<td>0.127</td>
<td>0.130</td>
<td>0.311</td>
</tr>
</tbody>
</table>

The number in parentheses indicates the number of wavelengths used in the sparse Bayesian model. mBayes is the Bayes regression using multivariate response variables. VS indicates variable selection by adaptive modeling. Data adapted from Chen, et al. (2007c).
univariate adaptive Bayesian regressions, no doubt because the number of wavelengths chosen for inclusion in the multivariate regression was a clear compromise: larger by far than those needed by a univariate Bayesian model for moisture, but smaller by far than univariate Bayesian models for the remaining three response variables. The mp5 and mp6 data again showed that the predictions from univariate Bayesian regressions were mostly superior to those from PLS-1 and that the adaptive variable selection further lowered prediction errors, though not to the levels seen from the analysis of the m5 data.

The cookie dough dataset has also been examined using variable selection based on a spike-and-slab approach (Hérnandez-Lobato, et al., 2015) in conjunction with an expectation propagation (Minka, 2001) to approximate the posterior distribution by a simpler parametric distribution, chosen to permit fast calculation. This combination gives RMSEP errors for the dry flour and moisture responses in the cookie dough data that were comparable to those found by other means, but speeds calculations by a factor of about 2000x or more.

The studies cited above demonstrate that some form of Bayesian regression using a prior on the variables included in the regression model is clearly an effective way to discover a set of variables that are highly informative about the regression relationship. While the set of informative variables has often been generated as a kind of by-product in the Bayesian analysis of NIR spectra where the main goal is reduced model size to reduce computational demands, the relative success of Bayesian variable selection has not yet been systematically compared to variable selection achieved by using other sparse modeling methods. Nott and Leng (2010) focused on studying the success of a Bayesian regression at identifying variables for inclusion in a regression model. As part of this work, they considered the situation where there were many more variables than samples, and, as in the studies on NIR spectra cited above, they found that the posterior distribution has reliable, though not always definitive, information on the active and inactive variables in the true model. Using synthetic data, they employed that a Bayesian version of the lasso, a sparse regression where, instead of the usual $L_2$ loss function based on the sum of squared differences, the $L_1$ loss function based on the sum of absolute differences is minimized (Tibshirani, 1996). This change in the loss function better concentrates the posterior probability on those variables present in the true model underlying the data. Lasso methods have been reported in Bayesian regressions on NIR spectra, and a lasso regression using empirically selected variables from the 700 available in the sugar dataset performed about as well as the much more computationally demanding method of Brown discussed above, but one using generalized cross-validation selected too many predictors and failed to predict the test data well (Yuan and Lin, 2005).

### 13.3.2 Bayesian Calibration on Transformed Spectral Data

Instead of performing interpolation to downsample spectra to data with far fewer channels, reducing the set of variables in a typical set of NIR spectra can also be done by use of a linear transform on the spectral data. Several linear transforms are often employed to reduce data variables.

#### 13.3.2.1 Bayesian Regression on Wavelet-Transformed Spectral Data

One linear transform is the discrete wavelet transform (DWT). Like its relative the discrete (fast) Fourier transform, the discrete wavelet transform converts data from the time domain to the frequency (time$^{-1}$) domain, but because the wavelet transform is performed with a set of orthogonal basis functions of differing frequency ranges, called scales in the wavelet terminology, the wavelet transform converts a conventional time-domain spectrum into a series of wavelet coefficients for each scale. A wavelet transform is a dyadic decomposition of the original signal, meaning that each new scale representation of the input data includes progressively 1/2 fewer wavelet coefficients. Thus, the DWT of an n-point NIR spectrum is represented by a set of l scales, with the highest scale frequencies in the spectral data contained in the first scale, expressed by $n/2$ wavelet coefficients. A lower set of scale frequencies is expressed by the second scale, using $n/4$ wavelet coefficients, a
still lower set of scale frequencies is expressed in the third scale, by \( n/8 \) wavelet coefficients, and so on down to the last scale, which describes the lowest scale frequency with 1 wavelet coefficient. The scales generated by the DWT depend on the wavelet basis set used to define the wavelet filter function. The wavelet basis is defined by the so-called “mother” wavelet, a filter function that generates the orthonormal set of wavelet scale basis functions that transform the data. Popular wavelet bases for transforming NIR spectral data include the Daubechies 4, Symlet 4, and Symlet 8 wavelet filters; here, the numbers refer to the number of “vanishing moments” in the wavelet filter function, a value that represents the highest order polynomial that can be described exactly by the wavelet filter. The results of the wavelet transform filtering are a set of coefficients that describe the behavior of the spectral data in different frequency ranges, from the highest frequencies spanned by the first scale, down to DC (0) frequency in the last scale.

Using the full set of scales, the spectral data can be reconstructed perfectly from their wavelet representation. However, because the higher frequencies often describe spectral noise in the NIR spectral data, these can sometimes be omitted without much loss of true signal. It also may be that the lower frequencies describe baseline effects in the NIR data, and these also may be omitted. A 700-variable NIR spectrum, padded to a total of 1024 variables to make the length of the signal a power of 2 and to avoid signal “leakage” issues at the ends of the spectrum from application of the wavelet filtering, and then expressed by a total of 1024 wavelet coefficients in all, might be well represented by only 100 or fewer wavelet coefficients. Note that these wavelet coefficients are not spectral wavelengths, nor are they readily converted to represent discrete wavelengths (note, though, that conversion is possible using a particular type of wavelet reconstruction; Tan and Brown, 2003), so wavelet representation of the data is not like a conventional wavelength selection scheme. Readers seeking understanding of the wavelet transform might want to start with two excellent tutorials (Graps, 1995; and Polikar, 2006) before tackling textbooks (e.g., Mallat, 1999) or the literature concerned with wavelet signal analysis.

The wavelet transform can be applied to selection of portions of NIR spectra included in a calibration model in a manner analogous to that used to select channels. In this case, the usual regression model is used as the likelihood, and the prior distributions of the parameters for the intercept, the regression coefficients, and the covariance matrix are specified, with an additional normal prior indicating whether or not the wavelet coefficient is used in the Bayesian model. In this way, the posterior incorporates the probability that a wavelet coefficient is selected (or not) and has the benefit of highly weighting subsets of the spectra that account for large variation in the response variable while, because of the prior, penalizing choices with many wavelet coefficients. An 8-scale Daubechies 4 wavelet decomposition was used to reduce NIR spectral data after some data at the ends of the spectra were removed as non-informative because of noise effects (Brown, et al., 2001). Of the 10 terms in the best 8-scale predictive model, 37% of the wavelet coefficients in the mid-range scale (scale 4) were included, along with a few terms from higher frequencies (6%, 6%, 1%, and 2% of scales 5–8, respectively), but no wavelet coefficients from the low-frequency scales (1–3) appeared in the posterior model, even though these would be important in any reconstruction of spectra from the coefficients. Similar findings are reported for other calibrations of near-infrared data based on wavelets (Tan and Brown, 2003). Inverse wavelet transformation of the coefficients in the model showed the spectral information in the posterior to be concentrated in two main bands: one at 1400–1450 nm, and the second at 1650–1730 nm, a region known to be active in spectra of fats and oils.

It is also possible to build an adaptive Bayesian model using families of submodels based on different wavelet decompositions, with different mother wavelets, and different numbers of decomposition levels, with basis functions employing different numbers of vanishing moments, and set Gaussian priors on the weights for each submodel. The posterior weights of these submodels are adapted in an MCMC search, as outlined above. This adaptive search produced a regression model where different wavelet bases contributed over different portions of the spectrum. The Bayesian wavelet model was compared to an approach based on stacking of the sets of wavelet submodels, with the stacking weights set by a bootstrap procedure (Breiman, 1996). These two modeling
methods were applied to the biscuit dough dataset discussed above (Donald, et al., 2011). All forms of the Bayesian regression produced results substantially better than those from conventional step-wise multiple linear regression, principal components regression (PCR), or partial least squares (PLS) regression applied to the spectra of dough, as shown in Table 13.2, but the stacked ensemble of Bayesian models generally showed superior predictive performance, in agreement with theory (Clarke, 2003). The stacked models consistently produced superior errors of prediction, but resampled stacking with different wavelet types performed worse than when only a single wavelet type was used, possibly because of the very large number of models used in the ensembles used to model, according to the authors.

### 13.3.2.2 Bayesian Calibration on Principal Component and Latent Variable Scores of Spectra

Another way to reduce the number of variables in a spectral dataset is to perform a singular value decomposition (SVD) on the spectral data to obtain the principal components of the data. If many of the singular values of the data are zero, or can be taken as zero, the number of loadings retained after the SVD is often far fewer, or can be taken as far fewer, than the number of original variables in the data. This sort of variable reduction is well established in chemometrics.

The first report of a Bayesian regression performed on the SVD loadings [32] parameterized the Bayesian model in terms of the regression coefficients and the SVD loadings, so that, if the usual SVD is written as

\[ X = U S V^T = T P^T \] (13.19)

with scores matrix \( T \) and loadings matrix \( P^T \), the principal components regression (PCR) model is then

\[ y = X \beta = T \theta \] (13.20)

where, for \( k \) “significant” loadings, the \( 1 \times k \) regression vector \( \theta \) is expressed as

\[ y = X \beta = T \theta \] (13.21)

Now, under these transforms, observing the response \( y \) provides information only on the loadings regression parameters \( \theta \), so the Bayesian model is specified in the loadings space, not in terms of the original data. This is an unusual Bayesian model because the loadings space is data-dependent; since the loadings span independent components of variance in the data, the loadings space often depends on the experimental design used to obtain the data and on the specific data \( X \) included in the model. Because of this, the principal components model changes as the number of singular values taken to be nonzero changes. This model is one that will have priors and posteriors on \( \theta \), not \( \beta \), so there is a need to establish the inverse relationship between \( \theta \) and \( \beta \), since knowing what wavelengths contribute significantly is an important part of the modeling. Using the full 700 points of the biscuit dough data and the first 16 loadings, and employing 100,000-sample MCMC estimation of the posterior distributions with a burn-in of 1000 samples and a thinning rate of 1/5 gave results very similar to those produced by the wavelet decomposition of the reduced spectra. Bayesian regressions based on wavelet and SVD data compression both showed that wavelengths near 1718 nm were especially important in the determination of the properties of the dough. The SVD analysis also found significant wavelengths at about 1520 and 2350 nm. Interestingly, the number of loadings here was not a parameter in the Bayesian model, and so the number of loadings retained in the model was not subject to setting a prior and to adjustment by information in the data, and as a consequence, the spectral data seem overfit, judging from the regression vector plotted in the paper (West, 2003).

In the SVD-based analysis reported here, an amazing 99.995% of the variance in the spectral data
was incorporated in the Bayesian model, and the number of components was set “consistent with the inherent smoothness of the predictor variable.”

The construction of a Bayesian model for a variable reduction using SVD was also considered (Nounou, et al., 2002a,b; Chen, et al., 2007a,b). This chapter provides a clear introduction to Bayesian modeling and considers the many complex issues arising from a Bayesian principal component analysis (Nounou, et al., 2002a). Their treatment illustrates the complexity of this approach; unfortunately, their results do not show significant predictive improvement over other approaches used on the same data.

The usual way to perform a principal components regression by singular value decomposition is to first decompose the spectral data in $X$, decide on an appropriate number of principal component scores to retain in the model by cross-validation or by examining the predictive performance of the various models on a validation set, and then use the model to predict the property value $y$ for new data. That approach cannot be used in a Bayesian setting because the cross-validation step is computationally expensive and because it assumes equivalence of models. A different approach is necessary.

For a principal components decomposition of $X$, Equation 13.19 applies, but here both $T$ (the scores of $X$) and $P^T$ (the $x$-loadings) are taken as stochastic, because they are estimated from $X$, which is deterministic, but with additive error. This approach differs completely from that mentioned earlier, and it highlights an issue with the great flexibility of a Bayesian approach: Deciding how to parameterize a model, what is and what is not stochastic, what the distributions of the stochastic variables are, and what are suitable hyperparameters is not always clear, nor is there a ready set of guidelines to follow.

Under these authors’ view of the decomposition, a simple principal components regression (PCR), where principal components decomposition of the spectral data $X$ is combined with a regression step, the Bayesian analysis becomes much more complex because of the need to specify all the distributions needed. Here, $T$, $P$, and the regression coefficient vector $\beta$ are all treated as stochastic, and so each must have a prior distribution specified. In addition, conditional distributions for $X$ and $y$ must be set. There is also the question of statistical independence of these distributions, because it is conceivable that some stochastic parameters depend on other parameters. So far, no one has checked to see what the consequences of a badly selected distribution are on the result of this sort of modeling, and only the effects of noise have been reported (Chen, et al., 2007a,b).

A Bayesian approach to latent variables assumes that both the predictor variables $X$ and the response variable $y$ are contaminated by independent, Gaussian noise $\epsilon_x \sim N(0,R_x)$ and $\epsilon_y \sim N(0,R_y)$, where $R_x$ is the covariance matrix of the set of predictor variables and $R_y$ is the variance of the noise in the univariate response variable. Measurement noise in the predictors is presumed independent of the measurement noise in the response variable. Since some principal component directions in the data are dropped, the observed data are regarded as containing noise and therefore stochastic, since

$$X_i = t_i p^T + \epsilon_{x_i} = \bar{x}_i + \epsilon_x$$  

and

$$y_i = \beta t_i + \epsilon_{y_i} = \bar{y}_i + \epsilon_x$$  

where the tilde in the equations above indicates a noise-free quantity.

Then, the regression can be implemented with the usual PCR model:

$$\tilde{X} = TP^T$$  

and

$$\tilde{y} = T\beta$$
The wavelength–space regression parameters $b$ normally reported for a NIR calibration are:

$$b = P\beta$$  \hspace{1cm} (13.26)

Using this set of relations as the model, the marginal likelihoods for $X$ and $y$ are:

$$p(x_i | P, t_i) - N(P_{t_i}, R_x)$$  \hspace{1cm} (13.27)

and

$$p(y_i | \beta, t_i) - N(\beta^T t_i, R_y)$$  \hspace{1cm} (13.28)

and because these are statistically independent, the full likelihood distribution is just the product of Equations 13.27 and 13.28.

Prior information must be provided for $b$ and $\tilde{X}$ because we observe the noisy $X$. These priors can be non-informative or Gaussian. The Gaussian priors are $p(b) - N(\mu_b, Q_b)$ and $p(\tilde{x}) - N(\mu_x, Q_x)$, where $\mu_x$ is the prior mean of the noisy $X$ because the noise in $X$ is assumed zero-mean. The prior for the response is just $p(y) - N(\mu_y, R_y)$, and the prior for $\beta$ is $p(\beta) - N(P^T\mu_b, P^TQ_bP)$.

Because all priors are conjugate to the likelihoods here, once all priors and likelihoods are specified, application of Bayes’ theorem leads to a closed-form conditional posterior probability expression describing Bayesian PCR:

$$p(P,T,\beta | X, y) - p(X,y | P,T,\beta)^* p(P,T,\beta) - p(X | P,T)^* p(y | T,\beta)^* p(T | P)^* p(\beta | P)$$  \hspace{1cm} (13.29)

presuming that the prior $p(P)$ is uniform and that the rank of $X$ is a fixed hyperparameter. This solution is only proportional to the full posterior because $p(X)$ is not readily calculated and is omitted from the calculation defined in Equation 13.1.

Equation 13.29 describes a constrained least squares problem that is not easily or quickly solved for large numbers of variables or observations, so MCMC methods and Gibbs sampling were employed to sample the full posterior (Chen, et al., 2007a). After extensive algebra, the uniform prior led to the full conditional posterior

$$\beta - N\left(\left(T^T T \right)^{-1} y, (T^T T)^{-1} \otimes R_y \right)$$  \hspace{1cm} (13.30)

where $\otimes$ is the Kronecker product. This is a maximum-likelihood solution to PCR, as is expected from the use of a uniform prior.

Algebra using the Gaussian prior on Equation 13.29 led to the full conditional posterior

$$p(\beta | y, P, T) - N(m,v)$$  \hspace{1cm} (13.31)

where

$$m = P^T \mu_b + P^T Q_b P T^T \left(T P^T Q_b P T^T + I \otimes R \right)^{-1} \left(y - T P^T \mu_b \right)$$  \hspace{1cm} (13.32)

and variance is

$$v = P^T Q_b P \left(I - T^T \left(T P^T Q_b P T^T + I \otimes R \right)^{-1} T P^T Q_b P \right)$$  \hspace{1cm} (13.33)

If the prior mean $P^T \mu_b = 0$ and the prior covariance $P^T Q_b P = I \otimes k$, the mean of this distribution is a ridge regression solution with ridge parameter $\lambda = R/k$, again as expected from the use of a Gaussian prior.
Gibbs sampling was applied to these conditional posteriors. Because of the orthogonality constraint on the principal component loadings, it is not possible to use the MCMC simulation to draw samples to simulate the posterior distributions. The solution was to use a closed-form solution for loadings obtained from the PCA or PLS method, then use those latent variable loadings to perform MCMC draws, and then update the Bayesian model from these to update the loadings.

This Bayesian PCR modeling approach was applied to calibrate NIR data from the wheat flour dataset (Kalivas, 1997). This dataset is no longer available at the Web site indicated in the published paper, but is included with the OHPL package available on CRAN (https://cran.r-project.org/) and can be obtained from there. The set consists of the NIR spectra of 100 wheat samples over the range 1100–2498 nm, measured in increments of 2 nm. The response variables are moisture and protein. There are 50 samples in the calibration set and 20 samples in each of the two validation sets. Here, every fifth wavelength measurement was used in this analysis to reduce the computational burden. For PCR and PLS calibrations, 10-fold cross-validation was used, in which 90% of the data were used in training and 10% for testing.

The Bayesian SVD method (Nounou, 2002b; Chen, et al., 2007a) requires an estimate of the covariance of measurement noise over the spectral variables used. That was not available from the published data, so, based on older literature, the noise variance in the spectral data was assumed to be piecewise linearly dependent on the wavelength of the NIR spectrum, with a minimum noise value set at 1800 nm and symmetric, linear increases in the noise variance below and above 1800 nm using a literature value for the standard deviation of noise over the NIR spectra (Williams and Norris, 2001). Estimates of the uncertainty in the laboratory values for protein and moisture are also needed and were again not reported with the dataset. Estimates for these were obtained from other published work (Centner, et al., 2000).

This approach can be extended to one using PLS latent variables, and a Bayesian partial least squares regression can be developed along the same lines as the Bayesian principal components regression described above. For convenience, the Bayesian regression is called a Bayesian regression on latent variables (BRLV) below.

As shown in Table 13.3, the results of Bayesian regression on latent variables reflect the amount of information provided in the prior. The “u” refers to the uniform, non-informative prior, and the “g” refers to the Gaussian (normal) prior.

Uninformative priors led to BRLV models with very little improvement over that obtained from the usual latent variable regressions with principal components regression (PCR) and partial least squares (PLS) regression. This result is no surprise: As discussed above, a uniform prior provides

<table>
<thead>
<tr>
<th>Regression Method</th>
<th>Fat</th>
<th>Sugar</th>
<th>Flour</th>
<th>Moisture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stepwise MLR</td>
<td>0.044</td>
<td>1.188</td>
<td>0.722</td>
<td>0.221</td>
</tr>
<tr>
<td>Bayesian regression</td>
<td>0.076</td>
<td>0.566</td>
<td>0.265</td>
<td>0.176</td>
</tr>
<tr>
<td>Wavelet Bayesian regression</td>
<td>0.059</td>
<td>0.466</td>
<td>0.351</td>
<td>0.047</td>
</tr>
<tr>
<td>Wavelet Bayesian regression (best performing model)</td>
<td>0.063</td>
<td>0.449</td>
<td>0.348</td>
<td>0.050</td>
</tr>
<tr>
<td>PLS regression</td>
<td>0.151</td>
<td>0.583</td>
<td>0.375</td>
<td>0.105</td>
</tr>
<tr>
<td>PCR</td>
<td>0.160</td>
<td>0.614</td>
<td>0.388</td>
<td>0.106</td>
</tr>
<tr>
<td>Stacked wavelet regression</td>
<td>0.043</td>
<td>0.340</td>
<td>0.182</td>
<td>0.037</td>
</tr>
</tbody>
</table>

Adapted from Donald et al. (2011).

a Results from using a stacked ensemble of Bayesian wavelet regression models built using a Coiflet wavelet with 1 level of decomposition and one vanishing moment. Other forms of the Coiflet wavelet performed less well on this dataset.
no additional prior information, equivalent to that used in a frequentist regression, so BLVR modeling with a uniform prior provides the most direct comparison with PCR and PLS regression. The comparison is especially easy here because the PCR used the same loadings as the Bayesian regression in this case. Results are similar, but slightly worse, for a Bayesian regression on latent variables using the uniform prior as compared to those from conventional PLS regression for both moisture and protein. Because both approaches used the same 15 latent variables to develop the regressions used to calibrate, the main difference in the results may be a consequence of the somewhat arbitrary spectral and laboratory uncertainty estimates used in the Bayesian regression. This result shows the challenge of performing a Bayesian regression when the prior information provided is omitted from the data collection and must be guessed from other sources.

When the Gaussian prior is used, the BLVR modeling method uses fewer latent variables than conventional PLS for both calibrated responses. BLVR with the Gaussian prior predicts slightly better than PLS for moisture, but worse for protein. Overall, the difference between the results of conventional and BLVR predictive modeling is slight.

There is a third approach to Bayesian regression, the one not yet considered. This makes more extensive use of prior information, but the approach taken here is to use expert knowledge about the spectral responses to set the prior. Unlike other approaches in which the wavelengths were selected in the Bayesian analysis by allowing data to adjust prior weights, the approach taken here was to attempt to focus the model on a small set of wavelengths selected from previous experience with the calibration. This is an “informed” prior, labeled with “i” in Tables 13.4 and 13.5. In this case, the prior on wavelength participation in the regression is set with a hyperparameter vector to reflect known bands for hydroxyl, which get high weights in the prior for the regression vector $b$, and to exclude those portions of the spectrum deemed irrelevant to this calibration, which get very small weights in $b$. The wavelength regions associated with protein absorption are less well established.

---

### TABLE 13.4

**Validation Error for Moisture in Wheat Data (Kalivas, 1997)**

<table>
<thead>
<tr>
<th></th>
<th>PCR</th>
<th>PLS</th>
<th>BLVR (u)</th>
<th>BLVR (g)</th>
<th>BLVR (i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank of spectral model</td>
<td>23</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>MSE for validation set 1</td>
<td>0.0667</td>
<td>0.0689</td>
<td>0.0729</td>
<td>0.0434</td>
<td>0.0371</td>
</tr>
<tr>
<td>MSE for validation set 2</td>
<td>0.1322</td>
<td>0.1195</td>
<td>0.1231</td>
<td>0.1003</td>
<td>0.0817</td>
</tr>
<tr>
<td>Overall MSE</td>
<td>0.0995</td>
<td>0.</td>
<td>0.0980</td>
<td>0.0718</td>
<td>0.0594</td>
</tr>
</tbody>
</table>

Adapted from data provided in Chen et al. (2007).

### TABLE 13.5

**Validation Error for Protein in Wheat Data (Kalivas, 1997)**

<table>
<thead>
<tr>
<th></th>
<th>PCR</th>
<th>PLS</th>
<th>BLVR (u)</th>
<th>BLVR (g)</th>
<th>BLVR (i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank of spectral model</td>
<td>24</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>MSE for validation set 1</td>
<td>0.1850</td>
<td>0.1503</td>
<td>0.1542</td>
<td>0.1728</td>
<td>0.1621</td>
</tr>
<tr>
<td>MSE for validation set 2</td>
<td>0.1401</td>
<td>0.1350</td>
<td>0.1418</td>
<td>0.1442</td>
<td>0.1177</td>
</tr>
<tr>
<td>Overall MSE</td>
<td>0.1625</td>
<td>0.1427</td>
<td>0.1480</td>
<td>0.1585</td>
<td>0.1399</td>
</tr>
</tbody>
</table>

Adapted from data provided in Chen et al. (2007b).
Chemical Calibrations of NIR Spectra

than those for moisture, so the priors used here were informed by looking at the size of the regression coefficients that resulted from the use of the regression with the uniform prior. The larger regression coefficients were deemed relevant to the protein calibration, and smaller values were deemed irrelevant. The portions considered relevant to protein in these empirically selected wavelength regions were again weighted highly. This approach requires confidence in the regressions with uninformed priors, since it biases the model strongly, but as Table 13.3 shows, the additional information available on hydroxyl in the prior led to much improved predictions. Now, the MSEP from the Bayesian latent variable regression is much lower than that of PLS regression, and with a far more parsimonious model. Even though the information about hydroxyl absorptions is well known, there is no easy way to take advantage of this knowledge in PLS, other than to rebuild a PLS model using only these wavelengths, because of the implied flat prior in frequentist regression. When the information provided is less well known and more broadly dispersed in the spectrum, however, in the case of the regions corresponding to protein absorption, the gains are more modest, as shown in Tables 13.4 and 13.5, but even here, the Bayesian regression proves superior to PLS modeling in predictions. It would be interesting to compare PLS models developed using the same set of wavelengths with the Bayesian approach to determine if the wavelength selection is truly the reason for the improvement in these models.

Several different types of Bayesian regression were compared against PLS-1 regression for the calibration of fatty acids in milk (Ferragina, et al., 2015). The authors collected 1264 samples of milk from 85 herds and measured the spectra of the milk samples in the near-infrared and infrared, from 925 to 5011 cm$^{-1}$. Four target fatty acids in the milk were measured by gas chromatography; then, an individual model cheese was produced from each milk sample. The spectra were then related to the fatty acid levels in the milk and various properties of the cheese and the whey. Calibration of the spectra to these properties was then attempted using conventional PLS-1 regression, the WinISI modified PLS (MPLS) regression, and three Bayesian regression methods from the BGLR package available at the CRAN repository (https://cran.r-project.org/web/packages/BGLR/index.html). The three Bayesian regressions employed here were similar, and all used the same likelihood distribution, but they employed different priors. Bayesian ridge regression used a Gaussian prior, with no wavelength selection, similar to that done in the Gaussian process regression methods discussed above. Bayes A regression used a prior distribution for the regression parameters based on a Student t-distribution, with more density at zero and with heavier tails than the Gaussian distribution. This prior distribution induces some shrinkage of the regression parameters, resulting in higher posterior density in parameters with higher association with the property, and lower density in the posterior for parameters with little association. Bayes B regression used a variant of the spike-and-slab prior mentioned earlier. This prior combines a density spike at zero with a t-distribution for the parameter to enhance the shrinkage in the posterior of any regression parameters without strong correlations with the property. Here, the data must provide strong evidence that the term is clearly nonzero to overcome the prior’s high probability that the parameter is zero. This prior tends to create sparse models with only variables highly correlated with the property.

While MPLS and the three Bayesian regression methods all outperformed conventional PLS-1 on these data, the three Bayesian methods all had higher $R^2$ values and lower residual variance than MPLS for all of the many properties calibrated. Bayes A and Bayes B regressions gave the most accurate results, except for the fat property, where MPLS regression and Bayes B regression produced the same results. Bayes A and B regressions had the best lines of concordance for all properties, with intercepts nearest 0 and slopes nearest 1.0 and with lower bias than PLS or MPLS. Bayes B regression used very few wavelengths, suppressing all of the usual water bands, while the far less selective Bayes ridge regression had some posterior density at almost all wavelengths. The authors mention, but did not compare these Bayes results with those from, PLS regression combined with a genetic algorithm, but they note the high information content in the wavelengths selected in the Bayes A and B regressions.
13.3.3 **Nonlinear Bayesian Calibration of Near-Infrared Spectra**

The studies discussed so far have focused on variable reduction and on the use of the prior to enhance modeling power in a calibration of a multivariate response as found in spectra. Another area of research in Bayesian regression modeling concerns ways of constructing the likelihood. The likelihood distribution in a Bayesian regression is not constrained to be linear. While it is much easier to map the density as a Gaussian based on some predetermined relation developed from linear matrix algebra, and to set the variance from an estimate of error, how the likelihood is created is flexible.

One place where the flexibility of a Bayesian approach offers a unique advantage over the usual, frequentist approach is in nonlinear regression applied to spectral data. A conventional, frequentist nonlinear regression is not possible when the number of predictor variables is much larger than the number of samples available, the case that arises for most data based on near-infrared spectrometry. However, nonlinear regression based on a reproducing kernel Hilbert space (Cristianini and Shawe-Taylor, 2000) can be used. This approach provides automatic dimension reduction of the predictor variables. The Bayesian formulation of this approach for a dataset that has $n$ samples and $p$ predictors and a $q$-dimensional multivariate response $y$ begins with the nonlinear relationship

$$ y = f(X) + \epsilon $$

(13.34)

where $f$ refers to some unknown, nonlinear function and $\epsilon$ is random error. The prediction of multivariate $y$ can be structured as the minimization of some loss function $L$:

$$ \min \left[ \sum_{i=1}^{n} L(y_i, f(x)) + \lambda J(f) \right] $$

(13.35)

where $J(f)$ is a penalty function, $\lambda$ is the regularization parameter, and $n$ is the number of samples available for the modeling. The kernel function $K$ is a solution to this optimization where

$$ f_i(X) = \sum_{i=1}^{n} \beta_i K(X_i, X_j | \psi), r = 1, \ldots, q $$

(13.36)

where $\psi$, the kernel parameter, controls the shape of the kernel. Several possible kernels can be used, but for any kernel, the relation

$$ y_i = BK_i + \epsilon_i $$

(13.37)

describes the regression. The kernel has converted the problem from a $p$-dimensional task to an $n$-dimensional task, since the number of calibrated components $q$ in the mixture should be far less than $n$. A polynomial kernel function was selected in one paper using this approach (Chakraborty, 2012; Chakraborty, et al., 2012). Putting this relationship into a Bayesian regression leads to a complex likelihood, requiring the use of Bayesian latent variables to speed the Gibbs sampling. Proper, diffuse priors were used, each with its own hyperparameter. In this study, the regression is structured hierarchically, and the set of hyperparameters used in the regression were subject to a higher-level Bayes analysis. This hierarchical structure provided for estimates of the uncertainty in the regression parameters $\beta$ and the variance of the priors.

The resulting kernel model permitted a complex relationship between the predictor variables and the multiple responses in $y$, but unlike earlier studies using linear Bayesian regression, no variable selection was performed because of the projection of the predictors into the kernel space. Since the posterior model was not closed form, Gibbs sampling was necessary, and the MCMC fitting of the
hierarchy required two independent MCMC chains of 20000 iterations, with a burn-in of 10,000 samples and a thinning rate of 5 (Chakraborty, 2012).

This modeling approach was applied to the biscuit dough data discussed above. While this dataset has been studied by many other groups, unlike the other studies reported for this dataset, the calibration set used here was developed from randomly selecting 52 of the 78 dough samples available, with the remaining 26 samples used as the test set, and with this modeling repeated 100 times. Also, unlike most of the other studies, multivariate response modeling was employed in the kernel regressions, and all four properties were predicted from the same model; the pairwise correlations of the four response properties were fairly high. The modeling led to a posterior estimate of the kernel parameter \( \psi \) of 3 in over 50% of the runs and 2 in over 30%, strongly supporting the idea that the NIR spectral data are nonlinearly related to the properties in the cookie dough data. While results from these studies were not exactly comparable to those from other attempts made to model the biscuit dough, the RMSEP errors for all four properties reported in this study were generally lower than those from other Bayesian methods, including those from multivariate Gaussian process regression, and they were considerably lower than those from kernel PLS regression (Kuhn and Johnson, 2013) or from conventional, linear PCR. The predictive errors in the Bayesian results were also far lower than those from methods using random forest regression and support vector machines using either a polynomial or Gaussian kernel function (Rosipal and Trejo, 2001).

The NIR wheat data were also modeled using the same multivariate Bayesian regression approach. Here, 66 samples, again randomly selected, were used to train, with the remaining 34 used in the test set. The training and predictions were again repeated 100 times. Here, too, the polynomial kernel parameter \( \psi \) was found to be 3 in about 25% of the runs and 2 in over 60%, again strongly supporting the idea that the near-infrared spectral data are nonlinearly related to the wheat properties. As with the biscuit dough data, errors of prediction for moisture and protein were substantially lower than those from the support vector machine or the random forest regression. The predictive errors from the multivariate kernel Bayesian regression were also substantially lower than those from Gaussian process regression and kernel PLS regression, but here, surprisingly, the multivariate kernel Bayesian regression gave predictive errors only slightly lower than those from principal components regression (Chakraborty, 2012). The time required for these calibrations and predictions to run is not given in either of these papers, but it is noteworthy that, even though the projection to the kernel space greatly reduced the size of the model, the Bayesian regression runs were made on a computer cluster system.

Cui and Fearn (2017) also compared the performance of nonlinear Bayesian methods, as well as other linear and nonlinear regression methods, on near-infrared data. They developed a Gaussian likelihood function for the univariate property in terms of three hyperparameters: the signal and length scale of the radial basis function kernel distribution, and the regularization parameter \( \lambda \). The Gaussian process regression used in this study was implemented using the GPML package available in the MATLAB programming language (Rasmussen and Nickisch, 2010). Rather than employ hierarchical Bayesian modeling to estimate the hyperparameters in the Bayesian model, they used evidence approximation methods (Bishop, 2006) to determine appropriate values, which were then held constant. A near-infrared dataset taken on 1240 agricultural seeds measured by a hyperspectral, near-infrared instrument and calibrated to the percent nitrogen was used to compare the regression methods in this study. The spectra were measured using 239 wavelengths with a resolution of 6.26 nm over the range 993.0–2488.4 nm. The predictive error from kernel Gaussian process regression (Gramacy and Lee, 2008) was compared to that from support vector machine (SVM) regression and conventional partial least squares regression. Like the kernel Gaussian process regression, the SVM regression also used projection to a kernel space based on a radial basis function. The hyperparameters in the SVM modeling were set by cross-validation using an \( L_2 \) norm. In all cases, both of the nonlinear methods easily outperformed PLS predictive modeling on the near-infrared data. No substantial difference was found here between the support vector machine and the Gaussian process regression modeling on this dataset, as might be expected from the use
of the same kernel function and the weakly informative prior. The benefit of getting full posterior information for each prediction from the Bayesian regression, which can help with outlier detection, balanced against the difficulty of setting suitable values of hyperparameters by evidence approximation in smaller datasets with relatively few degrees of freedom.

Gaussian process regression modeling has also been used in modeling moisture content in three freeze-dried pharmaceutical compounds. A fourth compound was used as the external validation set in testing the global models for moisture (Clavaud, et al., 2017). The authors compared many linear and nonlinear regression methods on near-infrared data of 3822 samples taken on two purportedly identical spectrometers. Because these freeze-dried samples all had residual humidity levels within a very narrow range centered near 1.00%, a range of moisture levels was created in the samples by carefully controlled humidification or by drying of samples from the three compounds used for calibration, but not the samples of the test set. Spectral measurements were made in triplicate through the glass storage vial in reflectance mode on each of the two Fourier transform NIR spectrometers by averaging 32 scans made over a spectral range of 4000–10,000 cm⁻¹, giving a total of 3822 spectra of the 637 vials for the calibration. Another 168 spectra of the fourth compound were collected in the same manner. The authors tested the usual PLS-1 regression models for moisture in each of the three compounds separately as well as a global PLS-1 model, where all three pharmaceuticals were included in the calibration for moisture. They focused modeling efforts on well-known water bands near 5000 cm⁻¹ (2000 nm) and 7000 cm⁻¹ (1450 nm). In this study, support vector regression with a radial basis kernel outperformed linear Bayesian Gaussian process regression modeling. Bayesian ridge regression was implemented using Bayesian routines taken from the Anaconda python software package, a conventional, linear Bayes likelihood, ridge-regression model with a Gaussian prior. It is not at all surprising that the linear Bayesian regression method was outperformed by the nonlinear regression in this study. Plots of predicted vs. measured moisture for both calibration and validation all appear slightly curved downward for PLS-1 regression, but appear strictly linear for the SVR model, lending strong support to the claim of a nonlinear relationship between the moisture content and the spectral response (Clavaud, et al., 2017).

### 13.3.4 Nonparametric and Empirical Bayesian Spectral Calibration Methods

The linear Bayes regression with Gaussian prior was also outperformed slightly by a nonparametric prediction method. This method is based on choosing the property associated with the K-nearest calibration samples, where the unknown is predicted from a calibration set by finding the calibration sample closest to the spectrum of the unknown, or in some cases by calculating the average property value from a group of K samples that comprise the set of K-nearest neighbors in spectral space to the unknown spectrum (Clavaud, et al., 2017). Prediction from a K-nearest neighbor calibration requires no calibration relationship other than a distance, but it is of limited use without a large number of calibration samples fully spanning the space of the analysis (Pérez-Marín, et al., 2005). With limited or insufficiently distributed data, an unknown spectrum may fall in a region far from calibration samples, and the K-nearest neighbor algorithm may decline to estimate its property because the distances to its nearest neighbors are too large. When there are sufficient numbers of sufficiently well-dispersed data, a local calibration based on nearest neighbor spectra offers real advantages over conventional linear and even nonlinear approaches to modeling because the distance-based calibration permits variably nonlinear, flexible calibration relationships that are otherwise very difficult to model.

A Bayesian approach to K-nearest neighbor calibration, where the SVD scores of the NIR spectra of individual calibration samples are converted to a likelihood distribution by application of a simple Gaussian kernel function, has also been reported (Fearn, et al., 2010). This approach to distance-based calibration has the advantage of smoothing the calibration distribution in score space so that the likelihood distribution is defined over the space of the calibration. This permits calibration on smaller datasets where gaps may be present in the calibration data. The use of a likelihood
also has the advantage over conventional, local smoothing in that all of the calibration data contribute to any prediction, unlike the K-nearest neighbor approach, where only K samples participate. In the Bayesian approach, samples contributing to likelihood density nearer the unknown contribute more heavily to the prediction. And, because it is a Bayesian approach, it also makes use of prior information, as in any other Bayesian analysis.

The weak point in any kernelization of data, as is done here, is deciding on the nature of the kernel function and setting its hyperparameters. While there is no need for cross-validation in a Bayesian model, validation using a tuning set is necessary to determine the optimal tuning hyperparameters for the kernel by testing the kernel values against predictive performance (Pérez-Marín, et al., 2012). Study of the same NIR feedstuff data and same likelihood kernel as used in earlier work (Fearn, et al., 2010) but with a different tuning hyperparameter gave different results; apparently, small values for the tuning hyperparameter in the Bayesian kernel regression resulted in lower SEP values, but produced occasional outlier results, while larger tuning hyperparameter values led to more consistent results, but with larger (worse) SEP values, and a robustness–performance tradeoff resulted. Using either the K-nearest neighbor model or the Bayes kernel regression on scores of the preprocessed spectral data, predictive results on NIR spectra of the feedstuff data were consistently better than those from PLS-1 regression in 24 of the 26 reported properties analyzed, reducing the SEP values by about 50% on average, with some variation between ingredients in the feedstuffs. Here, the K-nearest neighbor method slightly outperforms the Bayes nonparametric regression in terms of SEP, but those slightly better SEP results come at a significant cost in reliability. The K-nearest neighbor calibration approach fails to predict on 10–40% of samples, depending on the calibrated ingredient, while the Bayes regression predicts on all samples, and gives realistic credible intervals—the Bayesian equivalent of a confidence interval (see Gelman, et al., 2013 for more on this subject)—from the analysis.

Fearn et al. (2010) also demonstrate, using a Bayesian formulation, how an empirical prior can be used to improve calibrations. Somewhat akin to the informed prior used in selecting wavelengths for a regression model, the prior for the property can be set to represent the distribution of property data actually used in a calibration. This approach, called empirical Bayes calibration (Lwin and Maritz, 1980; Brown, 1982), relies on the distribution of property values present in the training set being a good description of future data. Like the K-nearest neighbor calibration, large gaps in property values or a lack of representation over the full range of calibration will lead to very disappointing predictive results. In many cases where inverse regression models (such as those in PCR and PLS-1) are used to form the likelihood, a normal prior will generally produce the best results because this prior often accurately describes the data used in calibration. However, when a calibration set is not well represented by a normal distribution, as is the case in some trace analyses, using the distribution that better reflects the calibration data will lead to superior results (Fearn, et al., 2010). For example, calibration of NIR data for a property with a lognormal distribution, as occurs in measurements of organic carbon in soils (Bellon-Maurel, et al., 2010), benefits from the use of the empirical, approximately lognormal prior. Both the bias and the SEP are reduced compared to the usual PLS-1 methods. An empirical distribution extends that benefit to data where the form of the distribution of the property value is not easily determined as lognormal or some other well-known function, or where the distribution may be mixed.

### 13.4 CONCLUSIONS

As this chapter demonstrates, Bayesian methods offer many advantages in multivariate calibration of near-infrared spectra. The great flexibility of the Bayesian modeling permits easy examination of linear modeling of univariate properties and linear modeling of multivariate properties, like many chemometric methods, but also permits nonlinear modeling and even nonparametric modeling with minimal changes to the likelihood model. This flexibility is a real advantage for accurate modeling of NIR spectra where nonlinear relationships to the property may be present. The prior offers a way
to select variables and to provide information about the data and about parameters that is not easily incorporated in the conventional measurement model. Together, the two models allow an unusually capable approach to multivariate calibration of spectra.

While Bayesian regression is best done on larger datasets, the main difficulties associated with Bayesian regression have been associated with its substantial computational burden, the need to write complex software incorporating a different perspective on data analysis, and the need for practitioners to develop a background in probability and statistics that goes beyond basic linear algebra. The great advances in computational power available on small computer systems and the increasingly common multi-core, parallel computational systems make running most Bayesian regressions possible on small computer systems, even on laptop computers. Improvements in software have greatly reduced the need for user-written Bayesian regression software. Public-domain packages and public-domain Bayesian regression software are now easily obtained and used. The remaining hurdle is the training of users of this computational power and readily available software in Bayesian methods. It seems that this, too, is occurring, if gradually. The use of Bayesian regression in spectroscopic calibrations is much like the use of PLS regression was in 1987: The capability is there, some software is there, and computers can run the analyses in reasonable times, but reports are somewhat scarce and users have not yet seen a clear reason to invest in the methods. This chapter offers motivation for spectroscopists to make that investment and provides key references, links to software, and background reading to get the motivated user ready to try Bayesian regression methods.

ACKNOWLEDGMENTS

This work was supported by Grant 1506853 from the United States National Science Foundation.

Initial work on this chapter was carried out when SDB was a visitor at IRSTEA in Montpellier, France. The author acknowledges, with gratitude, the assistance of Dr. Jean-Michel Roger and Dr. Jean-Claude Boullet. He also thanks Dr. Javier Haddoux and others at IRSTEA for their hospitality.

REFERENCES


