Sparse alternatives to ridge regression: 
a random effects approach

Arief Gusnanto\textsuperscript{a}\textsuperscript{*} and Yudi Pawitan\textsuperscript{b}

\textsuperscript{a}Department of Statistics, University of Leeds, Leeds LS2 9JT, UK; \textsuperscript{b}Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-17177, Sweden

(Received 10 April 2013; accepted 27 May 2014)

In a calibration of near-infrared (NIR) instrument, we regress some chemical compositions of interest as a function of their NIR spectra. In this process, we have two immediate challenges: first, the number of variables exceeds the number of observations and, second, the multicollinearity between variables are extremely high. To deal with the challenges, prediction models that produce sparse solutions have recently been proposed. The term 'sparse' means that some model parameters are zero estimated and the other parameters are estimated naturally away from zero. In effect, a variable selection is embedded in the model to potentially achieve a better prediction. Many studies have investigated sparse solutions for latent variable models, such as partial least squares and principal component regression, and for direct regression models such as ridge regression (RR). However, in the latter, it mainly involves an $L_1$ norm penalty to the objective function such as lasso regression. In this study, we investigate new sparse alternative models for RR within a random effects model framework, where we consider Cauchy and mixture-of-normals distributions on the random effects. The results indicate that the mixture-of-normals model produces a sparse solution with good prediction and better interpretation. We illustrate the methods using NIR spectra datasets from milk and corn specimens.

Keywords: ridge regression; calibration model; sparse solution; mixture; random effects

1. Introduction

In a calibration of near-infrared (NIR) instrument, we model some chemical compositions of interest from a set of observations as a function of their NIR spectra, which are measured on several hundreds or thousands wavelengths (variables). In the calibration, we deal with the situation where the number of variables exceeds the number of observations and the variables exhibit extremely high multi-collinearity. With these characteristics of the data, we have two inherent problems in building a calibration model. Firstly, we have a problem in estimating model parameters as standard ordinary least squares (OLS) regression will fail in this situation, and secondly, some variables may not contribute towards better prediction.
To deal with the first problem, many calibration models have been proposed and shown to work well. These models can generally be categorised into two main groups. The first one is by regressing the response on linear combinations of the original variables (called latent variables or components) and the second one is directly regressing the response on the (original) variables. The first category includes partial least squares (PLS) regression and principal component regression (PCR) and the second category includes ridge regression (RR) and lasso regression. To deal with the second problem, wavelength selection or variable selection has been advocated to improve prediction (e.g. [25]). Many strategies have been proposed, including genetic algorithm (e.g. [2,20]), forward selection (e.g. [9]), self-organising map [5], and boundary subset selection [23]. See, for example, Ref. [1] for some strategies of variable selection.

Dealing with both problems separately can be sub-optimal in building a prediction model, as some selection methods are not optimal to be used with a particular calibration model. For example, forward selection method may result in a configuration of variables that does not produce better prediction than ‘best subset’ selection method in RR. To overcome this, recent studies suggested sparse alternatives to the calibration models [3,7,15,17]. The term ‘sparse’ refers to the model parameters, where some of them are zero estimated (or very close to zero), and the other parameters are estimated naturally to be away from zero. Effectively, the contribution of some variables are cancelled and the prediction model relies on the other variables that have non-zero parameter estimates. This is done in a single framework within the calibration model.

Many studies recently have investigated some sparse alternatives for PLS and PCR, mostly involving $L_1$-norm penalisation or lasso [26]. For example, Chun and Keles [4] developed a sparse PLS method by imposing sparsity in the dimension reduction step. Fu et al. [8] proposed an elastic net, a combination of ridge-type and lasso-type penalties, that imposes sparse solution within PLS. Lee et al. [16] proposed a shrinkage of singular values for principal component analysis to produce sparse loadings. Lee et al. [17] modified the NIPALS algorithm in PLS, by incorporating lasso method to produce sparse loadings.

The above studies have focused on sparse solutions for latent-variable models. For direct regression models such as ridge regression, the main sparse alternative has been the lasso regression [26] or similar models with $L_1$-norm penalty to the model objective function. Although they produce a sparse solution, the models share some concerns including interpretation of the model estimates in extremely multicollinear data.

In this paper, we propose some sparse alternatives to the ridge regression within a random effects model framework by demonstrating that a sparse solution can be achieved by modifying the distributional assumption of the random effects. The first one is a linear model where the random effects are assumed to follow a Cauchy distribution. The second one is a linear model where we assume that the random effects follow a mixture distribution of three normals. The three normal distributions correspond to those random effects that have negative, zero, and positive estimates to create a sparse solution. We compare this model to some existing models within random effects model framework such as ridge regression and lasso regression. The main results indicate that the Cauchy random effects model does not necessarily produce a sparse solution. The results also indicate that the mixture-of-normals random effects model will produce a sparse solution with interesting interpretation and a good prediction. We illustrate the application of our proposed models using NIR calibration from milk and corn datasets.

2. Methods

In this section, we discuss the different random effects models and how their distributional assumptions are modified to produce a sparse solution. In Section 2.1, we outline a general setting of random effects models. In Section 2.2 we describe the RR model and some current
models that produce a sparse solution. We present the Cauchy and mixture-of-normals random effects models in Sections 2.3 and 2.4, respectively.

2.1 General setting

Let $X$ be a matrix of NIR spectra of $p$ variables from $n$ observations. So the matrix $X$ is of size $n \times p$ with $n \ll p$. We assume that each variable has been centred to have zero mean. Let $y$ be an $n$-vector of response variable, and let $\beta$ be a $p$-vector of regression model parameters. We also assume that the vector $y$ is centred to have zero mean. We model the response $y$ as

$$y = X \beta + \epsilon,$$

where $\epsilon \equiv \{\epsilon_k\}, k = 1, \ldots, n$, is an $n$-vector of the error term. We assume $\epsilon$ to follow a normal distribution with mean zero and variance covariance matrix $\Sigma = \sigma^2 I_n$, where $I_n$ is the identity matrix of size $n$. Our main interest is to estimate the regression parameters $\beta$.

It is well known that the solution of OLS $\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$ is not applicable in our case because $(X^T X)$ is not invertible in the case of $n \ll p$. Even if $(X^T X)$ is invertible in the case of $n > p$, the estimates of $\hat{\beta}_{LS}$ are highly unstable because of the extreme multi-collinearity between the columns of $X$. To deal with this problem, it is natural to make a distributional assumption on $\beta$ so that we formulate the estimation problem in the context of random effects model. For example, if we assume $\beta$ to follow a normal distribution $N(0, \sigma^2 \beta)$, we have a RR model (Section 2.2.1). It is of our interest in this paper to modify this assumption on the random effects to obtain a sparse solution for $\beta$, which will be described in Sections 2.2.2 and 2.4.

Let $p(\beta)$ denote the distribution of $\beta$. In the context of random effects model, the estimation of $\beta$ is based on an extended likelihood [22] or $h$-likelihood [18]. Let $\theta$ denotes the variance(s) and other parameters. The log-likelihood of the parameters can then be written as

$$\log L(\beta, \theta) = \log p(y | \beta) + \log p(\beta).$$

The estimates of $\beta$ can be obtained by maximising $\log L(\beta, \theta)$ in Equation (2). Regardless of the distribution of the random effects $\beta$, the conditional distribution of $y$ given $\beta$ is a normal distribution with mean $E(y | \beta) = X \beta$ and variance $\Sigma$, so that the first term in the log-likelihood is given by

$$\log p(y | \beta) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X \beta)^T \Sigma^{-1} (y - X \beta).$$

In our formulation, the RR can be viewed as a random effects model where the random effects $\beta$ are assumed to follow a normal distribution (see the next section). To obtain sparse alternatives for RR, it is natural to modify the distributional assumption of the random effects and consider other distributions that have fatter tails than the normal distribution. Before we discuss about the modification, we briefly review the RR and other current models that produce sparse solutions from the random effects model framework in the next section. The proposed new sparse alternatives to the ridge regression will be presented in Sections 2.3 and 2.4.

It is interesting to note that Equation (2) can have a Bayesian interpretation. From a Bayesian perspective, $p(\beta)$, $p(y | \beta)$, and $L(\beta, \theta)$ can be considered as prior, likelihood, and posterior, respectively. We will come across this form again in the next several sections, in particular Equations (4), (7), (8), and (12).

2.2 Review of current models from random effects model point-of-view

2.2.1 Normal random effects (RR)

To view the RR from within the random effects model framework, we assume that the regression parameter $\beta$ follow a normal distribution with zero mean and variance covariance matrix
\( D \equiv \sigma^2 I_p \), i.e. \( \beta \sim N(0, \sigma^2 I_p) \). We also assume that \( \beta \) is independent of \( \varepsilon \). Combining this with the conditional log-likelihood (3), the log-likelihood of the parameters can be written as

\[
\log L(\beta, \theta) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) - \frac{1}{2} \log |D| - \frac{1}{2} \beta^T D^{-1} \beta, (4)
\]

where \( \theta \equiv \{\Sigma, D\} \).

At fixed values of \( \theta \), we take the first derivative of the log-likelihood log\( L \) with respect to \( \beta \)

\[
\frac{\partial \log L}{\partial \beta} = X \Sigma^{-1} (y - X\beta) - D^{-1} \beta.
\]

Setting this to zero and solving it for \( \beta \), we obtain the estimates of \( \beta \) as

\[
\hat{\beta}_N = (X^T \Sigma^{-1} X + D^{-1})^{-1} X^T \Sigma^{-1} y, (5)
\]

where the subscript \( N \) in \( \hat{\beta}_N \) indicates that the estimates are under the normal distribution assumption on \( \beta \). By considering that \( \Sigma \equiv \sigma^2 I_n \) and \( D \equiv \sigma^2 I_p \), the estimates for \( \beta_N \) can be rewritten into

\[
\hat{\beta}_N = (X^T X + \lambda I_p)^{-1} X^T y, (6)
\]

where \( \lambda \equiv \sigma^2 / \sigma^2_\varepsilon \). This is the RR estimates as first proposed by Hoerl and Kennard [11]. The estimation of \( \lambda \) can be done via cross-validation or through minimising Akaike’s information criterion (AIC).

One thing to note is that the RR estimates in Equation (6) are not sparse, i.e. there is no clear separation between zero and non-zero estimates. To put forward sparse alternatives for the RR, we assume \( \beta \) to follow some distributions which put a density mass around zero and have fatter tails (for non-zero estimates). A natural choice would be to consider a Laplace distribution, which is known to have fatter tails than the normal distribution (Figure 1). The result of this assumption would be the well-known Lasso regression as described in the next section.

2.2.2 **Laplace random effects**

A distribution that we commonly assume on the random effects \( \beta \) to produce a sparse solution is the Laplace distribution (Figure 1, right panel, dotted line). Specifically, we assume \( \beta \) to follow
a Laplace distribution with location 0 and scale $\sigma_v$, or

$$p(\beta) = \frac{1}{2\sigma_v} \exp \left\{ -\frac{|\beta|}{\sigma_v} \right\},$$

where $|\cdot|$ denotes the modulus or absolute value. Combining this with the conditional log-likelihood (3), the log-likelihood of the parameters can be written as

$$\log L(\beta, \theta) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)^T\Sigma^{-1}(y - X\beta) - p \log(2\sigma_v) - \frac{1}{\sigma_v} \sum_i |\beta_i|, \quad (7)$$

where $\theta \equiv \{\Sigma, \sigma_v\}$.

There is one thing to note from the estimation of $\beta$ from the log-likelihood in Equation (7). The first derivative of the last term in the log-likelihood, $\log p(\beta)$, does not contain $\beta$. In fact, the first derivative is $-1/\sigma_v$ for positive $\beta$ and $1/\sigma_v$ for negative $\beta$. There are some alternative ways to estimate $\beta$ in this case (e.g. [28]). In this study, we resort to the algorithm discussed in the original paper of Tibshirani [26] and Hastie et al. [10] which is implemented in the R package lars. The package also implements the estimation of $\sigma_v$ via cross-validation.

### 2.3 Cauchy random effects

The first alternative model that we consider is a random effects model where the random effects are assumed to follow a Cauchy distribution. This distribution is illustrated in Figure 1 (left), in comparison to the normal distribution. The Cauchy distribution has fatter tails than the normal distribution, which makes it an intuitive first alternative towards obtaining a sparse solution.

Specifically, we consider the random effects $\beta$ to follow a Cauchy distribution with location 0 and a scale $\sigma_s$, or

$$p(\beta) = \frac{1}{\pi \sigma_s} \left\{ \frac{\sigma_s^2}{\beta^2 + \sigma_s^2} \right\},$$

where $\beta$ is assumed to be independent to $\varepsilon$. Combining $p(\beta)$ with the conditional distribution of $y$ in Equation (3), the log-likelihood of the parameters is given by

$$\log L(\beta, \theta) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)^T\Sigma^{-1}(y - X\beta) - p \log(\pi \sigma_s) - \sum_i \log \left( 1 + \frac{\beta_i^2}{\sigma_s^2} \right), \quad (8)$$

where $\theta \equiv \{\Sigma, \sigma_s^2\}$.

To maximise $\log L(\beta, \theta)$ in Equation (8), an additional step is needed to approximate the log-likelihood by a quadratic form [22, pp. 464–466]. However, we can derive a more stable estimation as follows. First, note that the log likelihood of the random effects $\beta_i$ is given by $\ell(\beta_i) = -\log(1 + \beta_i^2 / \sigma_s^2)$. The first derivative of the log-likelihood is given by

$$\ell'(\beta_i) = -\frac{2\beta_i}{\beta_i^2 + \sigma_s^2}$$

or

$$\ell'(\beta) = -D_s^{-1}\beta, \quad (9)$$

where $D_s^{-1} \equiv \text{diag}(-2/(\beta_i^2 + \sigma_s^2))$. 

Secondly, we combine the expression for $\ell'(\beta)$ in Equation (9) with the derivation of log-likelihood $\log L$ in Equation (8) at fixed $\theta$ to obtain

$$
\frac{\partial \log L}{\partial \beta} = X \Sigma^{-1}(y - X \beta) - D^{-1}s\beta.
$$

At fixed $\theta$, we set this equation to zero and solve it for $\beta$ to obtain the estimates of $\beta$ as

$$
\hat{\beta}_C = \left( X^T \Sigma^{-1} X + D^{-1}s \right)^{-1} X^T \Sigma^{-1} y,
$$

(10)

where the subscript $C$ in $\hat{\beta}_C$ denotes the Cauchy distribution assumption on the random effects $\beta$. With a starting value $\beta_0^C$, the estimation of $\beta_C$ is done alternately at fixed value of $\theta$ by first computing $D^{-1}s$ and then calculating $\hat{\beta}_C$ as in Equation (10). The estimation of error variance $\sigma^2$ is done robustly by

$$
\hat{\sigma}^2 = \text{median}\{(y_k - x_k \hat{\beta}_C)^2\}
$$

and the estimation of $\sigma^2_s$ is done through cross-validation.

### 2.4 Mixture-of-normals random effects

Previous approaches to obtain a sparse solution relied on the assumption that the model parameters follow a distribution that put some mass at zero and has fatter tails (for example, Laplace distribution). With this assumption, some parameters receive heavy shrinkage towards zero in the estimation, while allowing the other parameters to be estimated naturally away from zero.

In the RR, assuming a (single) normal distribution with zero mean produces random effects estimates that are shrunk towards the zero mean. To obtain a sparse solution, we need to support some parameters to be naturally estimated away from zero by constructing a random effects distribution that contains non-zero means.

Hence, a natural immediate extension to obtain a sparse solution is to assume that the parameters $\beta$ follow a mixture of three normal distributions. The first and third components of the mixture are assumed to have negative and positive means, respectively, and the second component to have zero mean. The second component with zero mean is the one that will put shrinkage on some estimates towards zero and give a sparse solution.

Specifically, we assume that the parameters $\beta \equiv \{\beta_i\}$ follow a mixture distribution

$$
p(\beta_i) = \sum_{j=1}^{g} \pi_j N(\beta_i; \mu_j, \sigma_{\beta_j}^2),
$$

(11)

where $g = 3$ is the number of components in the mixture distribution, $\pi_j$ is the mixing proportion with $0 \leq \pi_j \leq 1$ for all $j$ and $\sum_j \pi_j = 1$, $\mu_j$ is the $j$th mean of normal component, constrained to have $\mu_1 < 0$, $\mu_2 = 0$, $\mu_3 > 0$, and $\sigma_{\beta_j}^2$ is the $j$th variance of normal component in the mixture distribution.

To estimate the parameters, we first rewrite the model (1) as

$$
y_k = x_k^T \beta + \varepsilon_k,
$$

for $k = 1, \ldots, n$. Defining $\theta \equiv \{\mu_j, \sigma_{\beta_j}^2, \pi_j, \Sigma = \sigma^2 I_n; j = 1, 2, 3\}$, and combining the expressions in Equations (3) and (11), the contribution of observation $k$ to the extended likelihood of $\beta$ in
Equation (2) can be written as
\[
\log L_k(\beta, \theta) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y_k - x_k^T \beta) + \sum_{i=1}^{p} \log \left( \sum_{j=1}^{3} \pi_j p_j(\beta_i) \right),
\]  
(12)

where \( p_j(\cdot) \) is a normal density function with mean \( \mu_j \) and variance \( \sigma_j^2 \).

At fixed \( \theta \), taking the derivative of \( \log L_k(\beta, \theta) \) with respect to \( \beta \) gives us
\[
\frac{\partial}{\partial \beta} \log L_k(\beta, \theta) = \frac{1}{\sigma^2} x_k^T (y_k - x_k^T \beta) + \left\{ \frac{\sum_{j=1}^{3} \pi_j (\partial p_j(\beta_i) / \partial \beta_i)}{\sum_{j=1}^{3} \pi_j p_j(\beta_i)} \right\} \text{for } i = 1, \ldots, p
\]

\[
= \frac{1}{\sigma^2} x_k^T (y_k - x_k^T \beta) + \left\{ \frac{\sum_{j=1}^{3} \pi_j p_j(\beta_i) (- (\beta_i - \mu_j) / \sigma_j^2)}{\sum_{j=1}^{3} \pi_j p_j(\beta_i)} \right\} \text{for } i = 1, \ldots, p.
\]  
(13)

We define
\[
p_{ij} = \frac{\pi_j p_j(\beta_i)}{\sum_{l=1}^{3} \pi_l p_l(\beta_i)},
\]  
(14)

and \( P_j \) as a diagonal matrix of size \( p \times p \) with diagonal entries \( p_{ij} \) for \( i = 1, \ldots, p \). The quantity \( p_{ij} \) is the probability of \( \beta_i \) to belong to mixture component \( j \).

Expression (13) can be written as
\[
\frac{\partial}{\partial \beta} \log L_k(\beta, \theta) = -\left( \frac{1}{\sigma^2} x_k x_k^T + \text{diag} \left\{ \sum_{j=1}^{3} p_{ij} \frac{1}{\sigma_j^2} \right\} \right) \beta
\]
\[
+ \left( \frac{1}{\sigma^2} x_k y_k + \text{vector} \left\{ \sum_{j=1}^{3} p_{ij} \frac{\mu_j}{\sigma_j^2} \right\} \right).
\]  
(15)

Combining for \( k = 1, \ldots, n \), we arrive at
\[
\frac{\partial}{\partial \beta} \log L(\beta, \theta) = -\left( X^T \Sigma^{-1} X + \sum_{j=1}^{g} P_j D_j^{-1} \right) \beta + \left( X^T \Sigma^{-1} y + \sum_{j=1}^{g} P_j D_j^{-1} \mu_j \right),
\]  
(16)

with \( \Sigma = \sigma^2 I_n \) and \( D_j = \sigma^2 I_p \).

Equating the above expression to zero yields the estimate
\[
\hat{\beta}_M = \left( X^T \Sigma^{-1} X + \sum_{j=1}^{3} P_j D_j^{-1} \right)^{-1} X^T \Sigma^{-1} y + \left( X^T \Sigma^{-1} X + \sum_{j=1}^{3} P_j D_j^{-1} \right)^{-1} \left( \sum_{j=1}^{3} P_j D_j^{-1} \mu_j \right).
\]  
(17)

We can infer that, as a result of assuming a mixture distribution on the distribution of random effects \( \beta \), the expression for \( \hat{\beta}_M \) in Equation (17) contains an additional term that would shift a group of parameter estimates to their corresponding \( \mu_j \), depending on their probability to belong to the mixture component \( j \). For example, if \( \beta_i \) has a higher probability to belong to the group \( j = 3 \) with a positive mean, the estimate will be shrunk towards \( \mu_3 \), and not zero. As a result, we have a non-linear shrinkage. The expression for \( \hat{\beta}_M \) in Equation (17) reduces to the ridge regression estimate \( \hat{\beta}_N \) in Equation (5), if \( \mu_j = 0 \) and \( \sigma_j^2 = \sigma^2 \) for all \( j = 1, 2, 3 \).
The estimation of $\theta$ is done jointly with the estimation of $\beta_M$ in an iterative manner. We use non-standard expectation–maximisation algorithm, where the estimation of random effects is done in the maximisation step instead of the expectation step. The advantage is that the quantities in $\hat{\beta}_M$ and $\theta \equiv \{\mu_j, \sigma_{\beta_j}^2, \pi_j, \Sigma = \sigma^2 I_n, j = 1, 2, 3\}$, are in closed form expression (tractable).

To obtain a sparse solution, $\sigma_{\beta_2}^2$ is constrained at a small value, typically in the order of $10^{-4}$–$10^{-5}$ of the estimates of $\sigma_{\beta_1}^2$ or $\sigma_{\beta_3}^2$, and we constrain $\sigma_{\beta_1}^2 = \sigma_{\beta_3}^2$. Referring to Figure 1 (right), this means that the middle component would be a ‘spike’. The variances associated with non-zero means ($\sigma_{\beta_1}^2$ and $\sigma_{\beta_3}^2$) are estimated via cross-validation.

3. Datasets and cross-validation

3.1 Datasets

In our study, we illustrate the above methods using two datasets. The first one is a dataset from a calibration of NIR instruments with spectra from milk batches as previously described in [9,14]. Seventy milk batches were drawn from a production line in milk processing factory, and their fat concentrations were measured. On each milk batch, NIR spectrum was scanned between 829 and 1145 nm with wavelength resolution of 2.45 nm, creating a spectra matrix of 70 rows (observations) and 130 variables (wavelengths). Figure 2 (left) shows the spectra of the 70 milk batches after being centred on their respective variable means.

The second dataset is from corn specimen described in Eigenvector.com website [13]. The dataset consists of NIR spectra from 80 corn specimens measured on two spectrometers (mp5 and mp6). The wavelength range is 1100–2498 nm at 2 nm intervals (creating 700 variables). The moisture, oil, protein and starch values for each of the corn specimen were measured. So,
we have four calibration models for each spectrometer, with a total eight calibration models in this dataset. Figure 2 top-right and bottom panels show the spectra of the 80 corn specimens that have been centred to have zero column means. The results of analysis on the corn dataset are presented in the supplementary material.

3.2 Cross-validation

In the cross-validation, we split \( n \) observations into two subsets. The first one is a training set with \( n_t \) observations, in which we estimate the regression parameters. The second one is a validation set with \( n_v \) observations, where we estimate the predicted fat concentration using the regression parameters obtained in the training set. In each regression method, we calculate the root mean square errors of prediction (RMSEP)

\[
\text{RMSEP}(\gamma) = \left\{ \frac{1}{n_v} \sum_{k=1}^{n_v} (y_k - \hat{y}_k)^2 \right\}^{0.5},
\]

where \( \gamma \) represents a relevant tuning parameter, the summation is done over the \( n_v \) observations in the validation set, and \( \hat{y}_k \) is the predicted fat concentration using the parameter estimates from the estimation set. The use of cross-validation to estimate a tuning parameter is appropriate in our NIR calibration problem. It can be shown that an estimate of tuning parameter based on the cross validation also maximises its profile likelihood. In a general case where the number of variables increases at a higher rate than the number of samples, careful consideration is needed to the choice of method to estimate tuning parameter (see, for example, [6,27]).

4. Results

In this section, we present the results of our analysis on the milk dataset; the results on the corn dataset are presented in the supplementary material, except the summary of prediction error which are presented in Section 4.4.

Figure 3. Left: Residual sum of squares in the validation set, as a function of \( \lambda \equiv \sigma^2 / \sigma_N^2 \) in the normal random effects (RR). The residual sum of squares in the validation set is calculated using estimates from the training set at each given \( \lambda \). Right: Random effects estimates across 130 wavelengths under normal random effect model \( \hat{\beta}_N \). The dotted lines are the top and bottom spectra in the data to indicate the variability of spectra across wavelengths.
4.1 Normal random effects model (RR)

Figure 3 shows the results of cross-validation and the estimates of normal random effects. The figure indicates that the optimal $\lambda$ is estimated at $\exp(-9)$. The estimates $\hat{\beta}_N$ (from the training set) in Figure 3 suggest that we cannot obtain a sparse solution; the estimates are all non-zero. From variable selection point of view, we do not have a simple interpretation of the estimates as all of the wavelengths are taken into account in the prediction. The estimates $\hat{\beta}_N$ follow approximately a normal distribution (not shown), reflecting our assumption on $\beta_N$.

4.2 Cauchy and Laplace random effects models

Figure 4 presents the results of cross validation and random effects estimates under the Cauchy and Laplace random effects models. For the Cauchy random effects model, the estimate of $\sigma^2_s$ is $\exp(4)$, which minimises the validation residual sum of squares. The figure shows some parameters that are estimated away from zero while the others are estimated near zero. When we ‘zoom

![Cross validation and Estimates](image)

Figure 4. Left column: Residual sum of squares in the validation set, as a function of $\sigma^2_s$ in the Cauchy random effects model and $\lambda_L \equiv 1/\sigma_v$ in the Laplace random effects model (lasso regression). Right column: Random effects estimates across 130 wavelengths under Cauchy ($\hat{\beta}_C$) and Laplace ($\hat{\beta}_L$) random effects models. The dotted lines are the top and bottom spectra in the data to indicate the variability of spectra across wavelengths.
in’ to the estimates $\hat{\beta}_C$ around zero, these estimates are not zero exactly, but only distributed tightly around zero (see Figure 1 in the supplementary material, top-left panel). This indicates that the Cauchy distribution with fatter tails allows some estimates to be away from zero, but not enough to shrink some estimates to zero. The behaviour of the estimates $\hat{\beta}_C$ around zero is similar to that of the normal random effects $\hat{\beta}_N$.

For the lasso regression, Figure 4 indicates that the estimate of $\lambda_L \equiv 1/\sigma_v$ is $\exp(-5.079)$. The algorithm to estimate the random effects was implemented in ‘steps’, hence the value of $\lambda_L$ evaluated cannot be made equally spaced. Looking at the bottom-right figure on the estimates $\hat{\beta}_L$, we have a sparse solution where many random effects are shrunk to zero, and the others are non-zero estimated. Unlike the characteristics of the estimates $\hat{\beta}_C$, some of the estimates under the lasso regression $\hat{\beta}_L$ are zero estimated. In practice, the estimates are very small (in the order of $10^{-7}$–$10^{-5}$) and reported as zero.

4.3 Mixture-of-normals random effects model

For the mixture-of-normals random effects model, the estimates of $\sigma^2_{\beta_j}$ for $j = 1, 3$ are presented in Figure 5. Based on the cross-validation, the variances $\sigma^2_{\beta_1} = \sigma^2_{\beta_3}$ are estimated at $\exp(5)$. The estimates $\hat{\beta}_M$ are presented in the right panel of the figure and exhibit a sparse solution, where some estimates are shrunk to zero.

Compared to the lasso regression solution, the solution under mixture-of-normals random effects has less number of zero estimates. The estimates of mixture proportion $\pi_j$’s are $0.471, 0.113, \text{and } 0.416$ for $j = 1, 2, 3$. So, the number of zero estimates are approximately 11% out of 130 wavelengths. The estimated means of the mixture components $\mu_j$’s are $-25.4$ and $28.7$ for $j = 1, 3$ (the middle component is constrained to have zero mean). This is reflected in the estimates $\hat{\beta}_M$ where they are separated into three groups as displayed in Figure 5.

From Figure 5, we have an interesting note. There are six regions in the wavelength that have ‘blocks’ of non-zero random effects estimates, three in each sign. It has been suggested in the context of PLS regression that selecting regions of wavelengths, rather than selecting individual

![Cross validation](image1.png)

![Estimates](image2.png)

Figure 5. Left: Cross validation of $\sigma^2_{\beta_j}$ for $j = 1, 3$ in the mixture-of-normals random effects model. Right: Random effects estimates across 130 wavelengths under mixture-of-normals random effects model ($\hat{\beta}_M$). The dotted lines are the top and bottom spectra in the data to indicate the variability of spectra across wavelengths.
wavelengths, can potentially give better prediction (e.g. [19,21,24]) or give better stability of the estimates [12]. These studies support the notion that within the NIR regions of wavelength, there are some intervals where the main information lie.

Our results based on the mixture-of-normals random effects model in the milk data seem to support this idea. Although we do not model explicitly the spatial information between wavelengths, the existing correlation structure between wavelengths and the mixture model assumption on the random effects have resulted in accentuating the regions of wavelengths that positively or negatively associated with the fat concentration. This result is not specific only to this dataset. Similar interpretation can be inferred in different calibration models in the corn dataset (presented in the supplementary material).

4.4 Prediction

To check whether the models are reasonable to be used in real application, we compare the predicted fat concentration in the validation set to their observed value under different models as presented in Figure 6 for the milk dataset. The predicted fat concentrations are based on the

![Figure 6. Predicted versus observed fat concentration in the validation set, under different random effects distributions in the milk dataset. The predicted fat concentrations are based on the parameter estimates from fitting in the training set. The root mean square error of prediction (RMSEP) is printed for each model. The dotted lines are identity lines. The figures for the corn datasets are presented in the supplementary material.](image-url)
Table 1. RMSEP’s of the calibration models in the milk and corn datasets, under the different distributional assumptions of random effects.

| Data     | Response variable | Normal (ridge) | Cauchy | Laplace (lasso) | Mixture of normals | Elastic net | Adaptive lasso |
|----------|-------------------|----------------|--------|-----------------|-------------------|-------------|----------------|
| Corn (mp5) | Moisture         | 0.1869         | 0.1874 | 0.1747          | 0.1758            | 0.1743      | 0.1964         |
|          | Oil              | 0.0948         | 0.0998 | 0.0943          | 0.0898            | 0.0936      | 0.0931         |
|          | Protein          | 0.1395         | 0.1454 | 0.1347          | 0.1282            | 0.1337      | 0.2630         |
|          | Starch           | 0.3432         | 0.3456 | 0.3704          | 0.3538            | 0.3432      | 0.4498         |
| Corn (mp6) | Moisture         | 0.1816         | 0.1828 | 0.1730          | 0.1664            | 0.1696      | 0.1807         |
|          | Oil              | 0.0993         | 0.0995 | 0.1005          | 0.0959            | 0.0994      | 0.0986         |
|          | Protein          | 0.1484         | 0.1481 | 0.1466          | 0.1443            | 0.1584      | 0.1854         |
|          | Starch           | 0.3860         | 0.3799 | 0.3643          | 0.3488            | 0.3682      | 0.4304         |
| Milk     | Fat              | 0.6281         | 0.6258 | 0.5775          | 0.6094            | 0.5778      | 1.2895         |

The description for elastic-net and adaptive lasso models are presented in the supplementary material.

parameter estimates obtained in the training set. In general, all of the models that we consider in this study suggest a good prediction. The figure indicates that the Laplace random effects model exhibits the lowest RMSEP, while the RR exhibits the highest RMSEP in the milk dataset.

The different RMSEP’s for the different calibration models in the corn and milk datasets are presented in Table 1. Overall, the different calibration models indicate a good prediction error. It is not categorically conclusive that one model is always superior to the other models. Some reduction of prediction error that we can observe from those datasets are marginal. The mixture-of-normals random effects model tend to have lower prediction error in majority of calibration models.

5. Discussion

In this study, we have investigated some sparse alternatives for RR in NIR calibration problem from random effects approach. This approach gives us the flexibility in modifying the distributional assumption of the parameters to obtain a sparse solution. We concentrate on the distributions that put more density mass around zero, but still allow some random effects to be estimated away from zero. In this paper, we investigate the Cauchy, and mixture-of-normals distributions for the random effects as alternatives to the normal distribution, and briefly review lasso regression as a well-known alternative.

In a strict sense, assuming Cauchy distribution for the random effects does not produce a sparse solution, as none of the variables is shrunk to zero. However, apart from the Laplace distribution, the Cauchy distribution is another immediate natural alternative to the normal distribution. It has fatter tails that we expect to allow some random effects to be estimated away from zero. Although some parameters are estimated away from zero, the model does not generally improve the prediction compared to the RR.

We found that the mixture-of-normals random effects model is a better alternative to the RR compared to the Cauchy and Laplace random effects models. In the mixture-of-normals random effects model, the estimates form a sparse solution and, with extremely high multicollinearity presents, the estimates have a more intuitive interpretation compared to those of the Cauchy or Laplace random effects model. Our results suggest that the model shows a good prediction in a cross validation.
6. Conclusion

The challenges in calibration of NIR instruments have recently been dealt with sparse regression models. We investigated a novel sparse alternative to RR within a random effects model framework where we assume the random effects to follow a mixture of three normal distributions. This assumption has resulted in a sparse solution where some estimates are zero estimated. Compared to the RR or lasso regression, the mixture-of-normals random effects model has a more intuitive interpretation with relatively better prediction.¹

Note

1. Supplementary Content may be viewed online at http://dx.doi.org/10.1080/02664763.2014.929640

References

[1] C.M. Andersen and R. Bro, Variable selection in regression – a tutorial, J. Chemometrics 24 (2010), pp. 728–737.
[2] M. Arakawa, Y. Yamashita, and K. Funatsu, Genetic algorithm-based wavelength selection method for spectral calibration, J. Chemometrics 25 (2011), pp. 10–19.
[3] C.J.F. ter Braak, Regression by L₁ regularization of smart contrasts and sums (ROSCAS) beats PLS and elastic net in latent variable model, J. Chemometrics 23 (2009), pp. 217–228.
[4] H. Chun and S. Keles, Sparse partial least squares for simultaneous dimension reduction and variable selection, J. R. Statist. Soc. – Ser. B 72 (2010), pp. 3–25.
[5] F. Corona, S.P. Reinikainen, K. Aaljoki, A. Perkiö, E. Liitiäinen, R. Baratti, O. Simula, and A. Lendasse, Wavelength selection using the measure of topological relevance on the self-organizing map, J. Chemometrics 22 (2008), pp. 610–620.
[6] Y. Fan and C.Y. Tang, Tuning parameter selection in high dimensional penalized likelihood, J. R. Statist. Soc. Ser. B 75 (2013), pp. 531–552.
[7] P. Filzmoser, M. Gschwandtner, and V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, J. Chemometrics 26 (2012), pp. 42–51.
[8] G.H. Fu, Q.S. Xu, H.D. Li, D.S. Cao, and Y.Z. Liang, Elastic net grouping variable selection combined with partial least squares regression (EN-PLSR) for the analysis of strongly multi-collinear spectroscopic data, Appl. Spectrosc. 65 (2011), pp. 402–408.
[9] A. Gusnanto, Y. Pawitan, J. Huang, and B. Lane, Variable selection in random calibration of near infrared instruments: Ridge regression and partial least squares regression settings, J. Chemometrics 17 (2003), pp. 174–185.
[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd ed., Springer, New York, 2009.
[11] A. Hoerl and R. Kennard, Ridge regression: Applications to nonorthogonal problems, Technometrics 12 (1970), pp. 69–82.
[12] A. Höskuldsson, Variable and subset selection in PLS regression, Chemometrics Intell. Lab. Syst. 55 (2001), pp. 23–38.
[13] Eigenvector Research, Inc. NIR of Corn Samples for Standardization Benchmarking. http://www.eigenvector.com/data/Corn/index.html
[14] J. Huang, D. Brennan, L. Sattler, J. Alderman, B. Lane, and B. O’Mathuna, A comparison of calibration methods based on calibration data size and robustness, Chemometrics Intell. Lab. Syst. 62 (2002), pp. 25–35.
[15] J.H. Kalivas, Overview of two-norm (L₂) and one-norm (L₁) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, J. Chemometrics 26 (2012), pp. 218–230.
[16] D. Lee, W. Lee, Y. Lee, and Y. Pawitan, Super-sparse principal component analyses for high-throughput genomic data, BMC Bioinformatics 11 (2010), p. 296.
[17] D. Lee, W. Lee, Y. Lee, and Y. Pawitan, Sparse partial least-squares regression and its applications to high-throughput data analysis, Chemometrics Intell. Lab. Syst. 109 (2011), pp. 1–8.
[18] Y. Lee and J.A. Nelder, Hierarchical generalized linear models (with discussion), J. R. Statist. Soc. Ser. B 58 (1996), pp. 619–678.
[19] L. Munck, J. Pram Nielsen, B. Moller, S. Jacobsen, I. Sondergaard, S.B. Engelsen, L. Norgaard, and R. Bro, Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics, Anal. Chim. Acta 446 (2001), pp. 169–184.
[20] A. Niazi and R. Leardi, Genetic algorithms in chemometrics, J. Chemometrics 26 (2012), pp. 345–351.
[21] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, and S.B. Engelsen, *Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy*, Appl. Spectrosc. 54 (2000), pp. 413–419.

[22] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference using Likelihood*, Oxford University Press, New York, 2001.

[23] O.Y. Rodionova and A.L. Pomerantsev, *Subset selection strategy*, J. Chemometrics 22 (2008), pp. 674–685.

[24] M. Shariati-Rad and M. Hasani, *Selection of individual variables versus intervals of variables in PLSR*, J. Chemometrics 24 (2010), pp. 45–56.

[25] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, and G.L. Cote, *Theoretical justification of wavelength selection in PLS calibration: Development of a new algorithm*, Anal. Chem. 70 (1998), pp. 35–44.

[26] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Statist. Soc. Ser. B 58 (1996), pp. 267–288.

[27] H. Wang, B. Li, and C. Leng, *Shrinkage tuning parameter selection with a diverging number of parameters*, J. R. Statist. Soc. Ser. B 71 (2009), pp. 671–683.

[28] T.T. Wu and K. Lange, *Coordinate descent algorithms for lasso penalized regression*, Ann. Appl. Statist. 2 (2008), pp. 224–244.