Supplementary material

for the manuscript

Sparse Alternatives to Ridge Regression: A Random Effects Approach

by Arief Gusnanto and Yudi Pawitan

1 Adaptive-lasso and elastic-net models

Well-known sparse alternatives to the ridge regression can generally be considered as a variation or "hybrid" of the (standard) lasso regression. Two of them are the adaptive lasso regression (Zou, 2006) and elastic-net regression (Zou and Hastie, 2005), which can be considered as random effects models.

To view the adaptive-lasso model from random effects model point-of-view, let us assume the individual random effect β_i , $i = 1, 2, \dots, p$, to follow a Laplace distribution with individual scale parameter

$$p(\beta_i) = \frac{1}{2\sigma_{vi}} \exp \left\{ -\frac{|\beta_i|}{\sigma_{vi}} \right\},$$

where $\sigma_{vi} = \sigma_v/w_i$. Based on this, the likelihood of the parameters becomes

$$\begin{aligned} \log L(\beta, \theta) = & -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \\ & - p \log \{ 2\sigma_v/w_i \} - \frac{1}{\sigma_v} \sum_i w_i |\beta_i|, \end{aligned} \quad (1)$$

where $\theta \equiv \{\Sigma, \sigma_v\}$, and $\Sigma = \sigma^2 I_p$ as defined in the main manuscript. The individual weights for each variable, w_i 's, are defined as $\frac{1}{|\hat{\beta}_i^{\text{LS}}|^2}$ following the guidelines in Zou (2006), where $\hat{\beta}_i^{\text{LS}}$ is the least-squares parameter estimate of the i -th variable.

The elastic-net model is a "hybrid" between the normal random effects model (ridge regression) and the Laplace random effects model (lasso regression). Specifically, we assume that the random effects β follow a mixture of normal distribution and Laplace distribution (iid), with equal proportion (for simplicity). This is written as

$$p(\beta_i) = 0.5 Lp(0, \sigma_v) + 0.5 N(0, \sigma_\beta^2),$$

where $Lp(\cdot)$ and $N(\cdot)$ indicate the Laplace and normal density functions, respectively. The joint likelihood of the parameters becomes

$$\begin{aligned} \log L(\beta, \theta) = & -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \\ & - \frac{0.5}{2} \log |D| - \frac{0.5}{2} \beta^T D^{-1} \beta \\ & - 0.5p \log \{ 2\sigma_v \} - \frac{1}{2\sigma_v} \sum_i |\beta_i|, \end{aligned} \quad (2)$$

where $\theta \equiv \{\sigma^2, \sigma_\beta^2, \sigma_v\}$, and $D = \sigma_\beta^2 I_p$ as defined in the main manuscript.

2 Comparisons of random effects estimates

In this section we discuss the comparison between random effects estimates, using the milk dataset. To compare the different estimates, Figure 1 shows pairwise comparisons between the different estimates (not all shown). When we compare the estimates based on the Cauchy and Laplace random effects to the estimates based on the normal random effects (ridge regression), we can infer that $\hat{\beta}_C$ and $\hat{\beta}_L$ have wider range and only a small number of them are estimated away from zero. The away-from-zero estimates in $\hat{\beta}_C$ are not necessarily the ones that are highly estimated in $|\hat{\beta}_N|$ under Normal random effects model.

Figure 1: *Top left panel: The random effects estimates under Cauchy random effects $\hat{\beta}_C$ around zero. The other panels: pairwise comparisons between random effects estimates in this study (not all shown). The dotted lines are identity lines.*

In $\hat{\beta}_L$, the non-zero estimates are expected to be the ones that are also highly estimated under normal random effects, when the variables are orthonormal as previously seen in Figure 2 in the main manuscript. Due to the dependencies between variables in our data, we see that this is not the case in Figure 1. Nevertheless, a signature of this relationship is present where we see the variables that are non-zero estimates are the ones that have high magnitude under Normal random effects. An implication of these results is that, when we compare $\hat{\beta}_L$ to $\hat{\beta}_C$, there are only three non-zero estimates in agreement between the two (Figure 1 bottom left panel).

In the comparison between the mixture-of-normals and ridge-regression normal random-effects estimates, we can infer that $\hat{\beta}_M$ are classed into three groups with negative, zero, and positive means. An important aspect in the estimates $\hat{\beta}_M$ is that the range of the estimates is comparable to the range

of estimates under Normal random effects. We can also infer that, except for the zero estimates, the signs between both estimates are consistent.

One interesting result from the comparison of $\hat{\beta}_M$ and $\hat{\beta}_L$ is in the variables that have non-zero estimates. The non-zero estimates in $\hat{\beta}_L$ are also non-zero estimated in $\hat{\beta}_M$. Specifically, for the non-zero estimates, the selected variables under Laplace random effects model are a subset of those under mixture-of-normals random effects model.

3 Standard errors: Bootstrap

In the models described in the main manuscript, we can generally approximate the standard errors of the random effects estimates. For example, in the case of Normal and Cauchy random effects, the second derivation of log likelihood will lead us to arrive at the expression for standard errors of $\hat{\beta}_N$ and $\hat{\beta}_C$. Furthermore, the approximate random effects of $\hat{\beta}_M$ can be computed from within the EM algorithm involved. However, the estimation of standard error of $\hat{\beta}_L$ can only be done via bootstrap as the solution for $\hat{\beta}_L$ is non-linear. So, to compare the standard errors for the different random effects estimates, we employ the bootstrap method in this study. We illustrate the estimation of the standard errors using the milk dataset.

Specifically, in each random effects model, the steps we take are:

1. We perform model fit in the training set and estimate the random effects estimates $\hat{\beta}$. We obtain the model fit

$$\hat{y} = X\hat{\beta}$$

and define the vector of residuals

$$\varepsilon_k = y_k - \hat{y}_k.$$

2. We select bootstrap samples from the residuals $\varepsilon^* \equiv (\varepsilon_1^*, \varepsilon_2^*, \dots, \varepsilon_{n_t}^*)^T$. From ε^* , we calculate bootstrapped response values $y^* \equiv (y_1^*, y_2^*, \dots, y_{n_t}^*)^T$, where $y^* = \hat{y} + \varepsilon^*$.
3. We regress the bootstrapped response y^* on *fixed* X to obtain bootstrapped random effects estimates $\hat{\beta}^*$. The steps 2-3 are repeated many number of times.
4. The standard deviation of the bootstrapped estimate for *each* variable is taken to be the standard deviation of the bootstrapped estimate $\hat{\beta}^*$ for the variable.

Figure 2 presents the bootstrap 95% confidence interval on $\hat{\beta}$ for each wavelength under different random effects models. The interval is calculated via $\hat{\beta}_i \pm 1.96se^*(\hat{\beta}_i^*)$ for each i . The figure indicates that the width of the individual intervals under normal and mixture-of-normals random effects models are mostly comparable. The individual intervals for the Laplace random effects model (Lasso) are relatively much wider than the normal model, even for the random effects that are zero estimated. The Cauchy random effects model also demonstrated to have relatively wider intervals compared to the normal model, especially for the random effects that are estimated away from zero. However, interestingly, the intervals for those random effects that are estimated close to zero have relatively tighter intervals compared to the Laplace random effects model.

Figure 2: The vertical lines around the random effects estimates are the bootstrap 95% confidence intervals for each random effect, $\beta_i \pm 1.96se^*(\hat{\beta}_i^*)$, and not percentile intervals.

4 Random effects estimates and prediction for the corn dataset

In this section, we present the random effects estimates and prediction for the corn dataset, for each response variable.

4.1 Spectra from mp5 spectrometer

We present the results from the corn dataset with mp5 spectrometer here for the different response variables. The results where the moisture content serves as the response variable are presented in the main manuscript.

4.1.1 Moisture as response variable

In the corn dataset, with spectra from mp5 spectrometer and moisture as the response variable, the model parameter estimates based on the different distributions of the random effects are presented in Figure 3. The figure indicates that, the estimates from both the normal and Cauchy random effects model do not produce a sparse solution, while those based on the Laplace random effects model produce a sparse solution.

In Figure 3, the estimates $\hat{\beta}_M$ under the mixture-of-normals random effects produce a sparse solution. The estimates identify only two mixture components, correspond to zero estimates and non-zero estimates. From the estimates, approximately 88.7% of the random effects are zero estimated. So approximately only 11% of the estimates are non-zeros.

Figure 3: The model estimates based on the Normal ($\hat{\beta}_N$), Cauchy ($\hat{\beta}_C$), Laplace ($\hat{\beta}_L$), and mixture-of-normals ($\hat{\beta}_M$) random effects across wavelengths in the corn dataset, with spectra from mp5 spectrometer and moisture as the response variable. The random effects estimates for the other response variables in this dataset are presented in the supplementary material.

Similar to the milk data, the estimates $\hat{\beta}_M$ in Figure 3 also indicates some 'blocks' of non-zero estimates. In the lasso regression, we only have individual wavelengths being selected to have non-zero estimates $\hat{\beta}_L$. The correlation between wavelengths and the mixture-of-normals assumption have resulted in accentuating some regions of wavelengths.

The prediction of moisture in the validation set for the corn dataset is presented in Figure 4 for the different random effects models, using mp5 spectra data. Compared to the ridge regression, the mixture-of-normal random effects model has lower RMSEP, which comparable to that of the lasso regression. Only the Cauchy random effects model that has slightly higher RMSEP than the ridge regression.

Figure 4: Predicted versus observed moisture of corn in the validation set, under different random effects distributions, using mp5 spectra data. The predicted moistures 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 prediction for the other response variables is available in the supplementary material.

4.1.2 Oil as response variable

Figure 5: *The model estimates based on the different random effects distributions in the corn dataset, with spectra from mp5 spectrometer and oil content as the response variable.*

Figure 6: *Predicted versus observed oil of corn in the validation set, under different random effects distributions, using mp5 spectra data. The dotted lines are identity lines.*

4.1.3 Protein as response variable

Figure 7: *The model estimates based on the different random effects distributions in the corn dataset, with spectra from mp5 spectrometer and protein as the response variable.*

Figure 8: *Predicted versus observed protein of corn in the validation set, under different random effects distributions, using mp5 spectra data. The dotted lines are identity lines.*

4.1.4 Starch as response variable

Figure 9: *The model estimates based on the different random effects distributions in the corn dataset, with spectra from mp5 spectrometer and starch value as the response variable.*

Figure 10: *Predicted versus observed starch of corn in the validation set, under different random effects distributions, using mp5 spectra data. The dotted lines are identity lines.*

4.2 Spectra from mp6 spectrometer

We present the results from the corn dataset with mp6 spectrometer here for the different response variables. The results where the starch content serves as the response variable are presented in the main manuscript.

4.2.1 Moisture as response variable

Figure 11: *The model estimates based on the different random effects distributions in the corn dataset, with spectra from mp6 spectrometer and moisture as the response variable.*

Figure 12: *Predicted versus observed moisture of corn in the validation set, under different random effects distributions, using mp6 spectra data. The dotted lines are identity lines.*

4.2.2 Oil as response variable

Figure 13: *The model estimates based on the different random effects distributions in the corn dataset, with spectra from mp6 spectrometer and oil content as the response variable.*

Figure 14: *Predicted versus observed oil of corn in the validation set, under different random effects distributions, using mp6 spectra data. The dotted lines are identity lines.*

4.2.3 Protein as response variable

Figure 15: *The model estimates based on the different random effects distributions in the corn dataset, with spectra from mp6 spectrometer and protein as the response variable.*

Figure 16: *Predicted versus observed protein of corn in the validation set, under different random effects distributions, using mp6 spectra data. The dotted lines are identity lines.*

4.2.4 Starch as response variable

Figure 17: *The model estimates based on the Normal ($\hat{\beta}_N$), Cauchy ($\hat{\beta}_C$), Laplace ($\hat{\beta}_L$), and mixture-of-normals ($\hat{\beta}_M$) random effects across wavelengths in the corn dataset, with spectra from mp6 spectrometer and starch value as the response variable. The random effects estimates for the other response variables in this dataset are presented in the supplementary material.*

For the corn dataset using mp6 spectra and starch value as the response variable, the random effects estimates based on the different random effects models are presented in Figure 17. The figures indicates that the Cauchy random effects model has some estimates that are estimated away from zero, while the other estimates are close to zero. The estimates based on the Laplace and mixture-of-normals random effects produce a sparse solution. In the mixture-of-normals model, approximately 63.6% of the estimates are penalised to zero. We also have the indication that the non-zero estimates, unlike in the Laplace random effects model, form 'blocks' along the range of the wavelength.

The prediction of starch values in the validation set for the corn dataset is presented in Figure 18 for the different random effects models, using mp6 spectra data. The figure indicates that all of the alternative models give lower RMSEP compared to that of ridge regression. The lowest RMSEP is achieved using the mixture-of-normals random effects model.

Figure 18: *Predicted versus observed starch values of corn in the validation set, under different random effects distributions, using mp6 spectra data. The predicted starch values 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 prediction for the other response variables is available in the supplementary material.*

References

- Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B*, **67**: 301-320
- Zou, H. (2006) The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**:1418-1429