Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Sixth Edition. Pearson Prentice Hall.

Jolliffe, I. T. (2002). *Principal Component Analysis.* Second Edition, Springer, New York.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, 1724-1735.

Li, K. -C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, K. -C., Aragon, Y., Shedden, K. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98**, 99-109.

Liu, X., Srivastava, A. and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 662-666.

Reinsel, G. C. and Velu, P. (1998). *Multivariate Reduced Rank Regression, Theory and Applications.* Lecture Notes in Statistics 136. Springer, New York.

Seber, G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *J. Amer. Statist. Assoc.* **81**, 142-149.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components. *J. Roy. Statist. Soc. Ser. B* **61**, 611-622.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Roy. Statist. Soc. Ser. B* **69**, 329-346.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Statist.* **38**, 1092-1109.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: dennis@stat.umn.edu

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.

E-mail: bing@stat.psu.edu

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.

E-mail: chiaro@stat.psu.edu

(Received April 2008; accepted June 2009)

COMMENT

Jinzhu Jia, Yuval Benjamini, Chinghway Lim, Garvesh Raskutti and Bin Yu

UC Berkeley

We thank the authors for a very interesting paper and the editors for inviting us to discuss it. Cook, Li and Chiaromonte (2010) develop the envelope model that imposes relationships between the mean parameter matrix β and covariance matrix Σ of the error vector in a linear multi-response regression model. They use the maximum likelihood estimator (MLE) under the envelope model to estimate the mean parameters. As expected, this MLE is asymptotically less variable than the usual OLS if the envelope model holds and the dimension p of the predictor is fixed while sample size n goes to infinity. The question is to what extent this superiority of envelope-MLE remains when the envelope model might not hold, which is typically the case with data.

Reducing the variability of estimates of β is critical in many modern regression settings, even more so when both the dimension p of predictors and number r of responses are large compared to sample size n (Greenland (2000)). To deal with this problem, a common strategy is to use regularization. Regularization for multi-response linear models can be achieved by constraining the parameters of the model or by pooling information from different responses to produce better estimates. Both of these aspects of regularization are found in the envelope model.

The envelope model links the linear space spanned by the parameter vectors in individual models, $\beta \in \mathbb{R}^{p\times r}$, to the covariance matrix of responses errors ($\Sigma \in \mathbb{R}^{r\times r}$), where p is the dimension of the predictor and r is the number of responses for each sample. To be precise, the envelope model assumes that the space spanned by β lies in the linear space spanned by some u eigen vectors of Σ. This link is non-trivial, and the resulting model could be computationally hard to estimate. In this discussion, we call the estimate of β under the envelope model the *envelope-MLE.* Although the authors compare their method to OLS in Cook, Li and Chiaromonte (2010), they do not compare it to standard methods used to reduce variability in estimation. One such method is ridge regression (RR) - a regularization method that uses an ℓ_2 penalty on the estimated β. Another method, Curds and Whey (CW) introduced by Breiman and Friedman (1997), exploits the multiple responses (and β structure) to improve the estimation. That is, find $B \in \mathbb{R}^{r\times r}$, such that

$$B_{i,:} = \arg\min_{b} E\|Y_i - b^T\hat{Y}_{ols}\|_2^2, \quad i = 1, \ldots, r, \tag{1}$$

where $\hat{Y}_{ols}$ is the fitted OLS responses for a given observation with predictors $X \in \mathbb{R}^p$, and $B_{i,:}$ is the ith row of matrix B.

Because it is not usually known with data whether such a link between β and Σ exists, it is crucial to evaluate the performance of different methods in cases where the link does not necessarily hold or the sample size is not large even when the link holds. We compare the performance of envelope-MLE with the algorithms Ridge and CW, both for simulated data and real data. Our experience (admittedly limited) with the envelope model suggests that

(1) the envelope model is best suited to the regime $u < p < r < n$;

(2) the envelope-MLE, as currently implemented, is computationally more intensive than Ridge and CW.

1. Experiments

1.1. Simulated data

Two simulation scenarios are used. The first is based on the envelope model

$$\begin{aligned} Y &= \Gamma\eta X + \epsilon, \\ \Sigma &= \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \end{aligned} \tag{1.1}$$

where $Y \in \mathbb{R}^r$, (Γ, Γ_0) are the eigenvectors of $\Sigma := \mathrm{Cov}\,(\epsilon)$, $\Omega = \Gamma^T\Sigma\Gamma \in \mathbb{R}^{u\times u}$, and $\Omega_0 = \Gamma_0^T\Sigma\Gamma_0 \in \mathbb{R}^{(r-u)\times(r-u)}$. We take $\Sigma = \Gamma^T D_r \Gamma$, where Γ are the eigenvectors of a random matrix (elements in $N(0,1)$), and D_r is a diagonal matrix with the elements $1, \ldots, r$. We simulated $\eta \in R^{u\times p}$ from $\eta_{ij} \sim N(0,1)$ and generated $\beta = \Gamma\eta$ accordingly.

In the second simulation scenario, no structural link between β and Σ is assumed. We generated Σ similar as before, and β was a random matrix, $\beta = G\eta$, where the elements of $G \in \mathbb{R}^{r\times u}$ were $N(0,1)$. Note that u in this "random model" is the rank of β, but it is not related to the structure of Σ. For both models, X was generated from $\mathcal{N}(0, V)$, where $V_{ij} = 0.5^{|i-j|}, i, j = 1, 2, \ldots, p$.

The measure used here for comparison is the overall average mean-squared prediction error. Following Breiman and Friedman (1997), the mean-squared prediction error of response i, for a particular method m, is

$$e_i^2(m) = E_X\left(\beta_{i,:}X - \hat{\beta}(m)_{i,:}X\right)^2 = (\beta_{i,:} - \hat{\beta}_{i,:}(m))V(\beta_{i,:} - \hat{\beta}_{i,:}(m))^T,$$

where V is the covariance matrix of X. We compare the average prediction error of each method normalized by the OLS average prediction error.

We considered the following four set-ups: $p > r$ or $p < r$, $u = \min(p, r)$ or $u < \min(p, r)$. When $p > r$, $p = 50$, $r = 10$; when $p < r$, $p = 10$, $r = 50$; when $u < \min(p, r)$, $u = 3$. For each of the four set-ups, we did 50 repetitions to evaluate each estimation method, and for each repetition we took sample size $n = 500$.

1.2. Data

The data is taken from Kay et al. (2008), and consists of a training set of $n = 1,750$ samples and a validation set of $n = 120$ samples. Each sample consists of $p = 64$ predictor variables and $r = 143$ responses. The data is from an experiment measuring hemodynamic response to natural image stimuli in the visual cortex of the brain using functional Magnetic Resonance Imaging (fMRI). The predictor variables measure magnitude of a spatial grating (Gabor filter) at different positions in the image (an 8 by 8 grid). The responses are measures of the fMRI response at different locations in the visual cortex. The task is to predict the fMRI responses in these locations to a new natural image stimuli – or the encoding problem in computational neuroscience.

The training set was split into an estimation set $n = 1,500$, and a set for model selection and regularization parameter optimization ($n = 250$). The best models based on the model selection set were compared on a third validation set ($n = 120$) with a better signal-to-noise ratio. Prediction accuracy was estimated using mean-square-error of the prediction to the measured responses on the validation set. These results were then normalized by the OLS mean square error for the corresponding response.

2. Results

Our results show that in most scenarios the competing methods achieved predictions as good as the envelope-MLE, and were much faster.

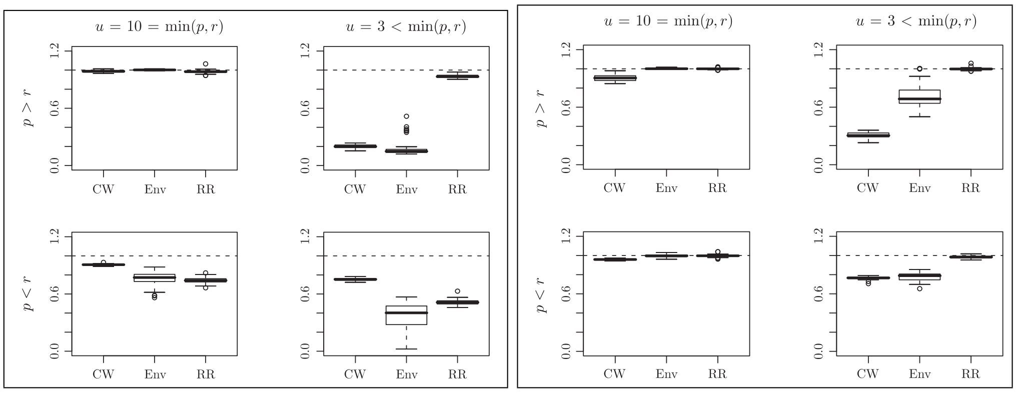
2.1. Prediction performance comparison

When the envelope model held, envelope-MLE was the most successful method when $u < p < r$ (right hand corner of (a) in Figure 2.1): it achieved lower prediction errors compared to the other methods, although it had a larger variability. In other cases, Ridge and/or CW achieved comparable results as envelope-MLE: CW when $u < r < p$, Ridge when $u = p < r$, or both when $u = r < p$.

However, this was not the case when data was not generated from an envelope model. While the envelope-MLE procedure performed better than unrestricted OLS, CW outperformed envelope-MLE (Figure 2.1 (b)), and Ridge was comparable to envelope-MLE in two cases and worse in the other two.

In the data example (Figure 2.2), all three methods gave comparable performance: the error of envelope-MLE was slightly worse than Ridge and slightly better than CW, even though envelope-MLE seems to have the largest variability. All methods performed better than OLS, ($SE \approx 0.003$, differences were significant; the results are lower than those reported in Kay et al. (2008) because of the restriction to 64 predictors).

As our results show (Figure 2.1), the envelope model is best suited to a classical regime where $u < p < r \ll n$. If $n < p$ or $n < r$, either Σ or β would be under-complete and would not be identifiable under the constraints of the model. We need $r > u$ for the regularization to be effective (otherwise the dimension of the envelope is already r and we get the OLS results).



(a) Envelope Models

(b) Random Models

Figure 2.1. The average (over responses) prediction error of each method normalized by OLS average prediction error. The distribution in each boxplot corresponds to 50 repetitions of the simulation. Left panel (a): Envelope Models. Right panel (b): Random Models. Note that u in “Random Models” (b) is not the dimension of the envelope. CW denotes the Curds and Whey, Env the Envelope model, and RR the Ridge regression with tuning parameter tuned by 5-fold cross validation. The dashed line denotes the benchmark that one method performs the same as OLS (and thus the ratio is 1).

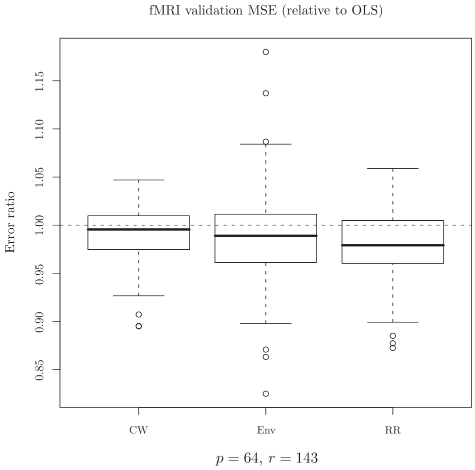


Figure 2.2. Prediction error for individual responses (normalized by OLS error) for image-fMRI data. Boxplots show distribution of $r = 143$ responses. All methods are better than OLS (Median of error ratios < 1). Ridge performs best, followed by envelope-MLE. Each boxplot corresponds to a single point in the simulations of Figure 2.1.

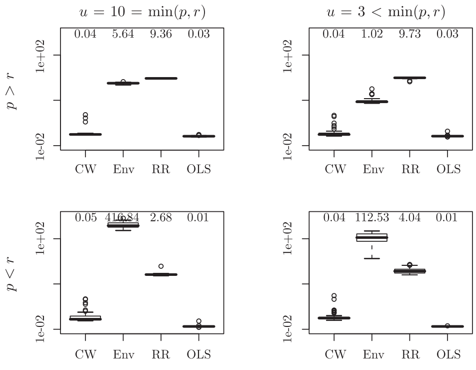


Figure 2.3. Run Time (in Seconds) for each iteration. The times are shown in log scale. CW denotes the Curds and Whey, Env the Envelope model, RR the Ridge regression with tuning parameter tuned by 5-fold cross validation, OLS the ordinary least square method.

2.2. Runtime comparison

The envelope-MLE method was always more computationally intensive than all other methods used in the regime $p < r < n$ (see Figure 2.3). Since this is the regime where the envelope model is most useful, the result highlights the need to improve efficiency of the algorithm. In fact, for the fMRI data, using the original implementation of envelope-MLE, the algorithm could not run in reasonable time. Instead we used a parameter tuning set to tune u (the dimension of $\mathcal{E}_{\Sigma}(\mathcal{B})$) to improve computation speed. It took the Envelope model estimation more than 300 seconds to run in the optimal setting $u = 20$, and up to 2,500 seconds for an envelope dimension of $u = 100$, while the other methods are very efficient (less than 0.1 seconds for each).

The model selection procedure for the envelope model was fairly stable. The error in the training set reduces as u increases (when $u = 143$ the model is equivalent to OLS which minimizes the training set error). However, in both the parameter tuning set and validation set the prediction errors were minimized by the same value, $u = 20$ (see Figure 2.5).

3. Conclusion

To summarize, the envelope model provides a novel way of regularization for multi-response linear regression problems. Our simulation and data results suggest that the envelope model works best in the classical domain when $u < p < r < n$ and the envelope model holds. More experience is needed for us to better understand the envelope model relative to other regularization methods such as Ridge regression and CW, especially when $\min(r, p) \gg n$. This is feasible only when faster codes become available for fitting the envelope model.

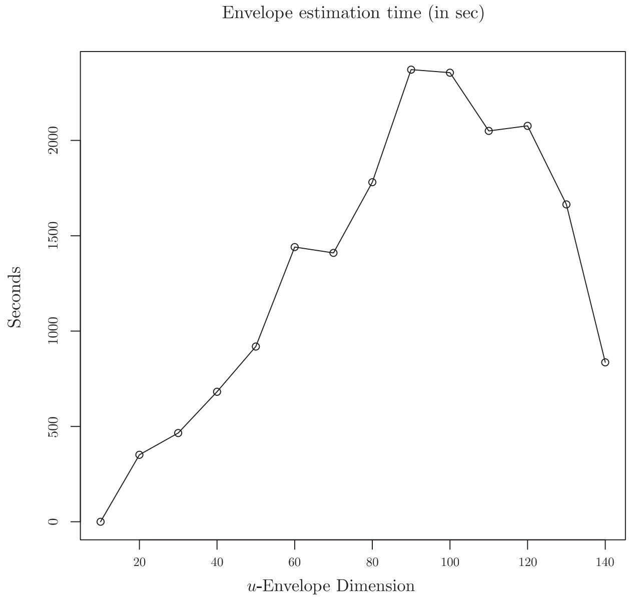


Figure 2.4. Run time for different u on the fMRI data. The time required for the envelope-MLE estimation (300-2,400 seconds) restricted testing larger parameter and responses sets.

Acknowledgement

This research is partially supported by NSF grants DMS-0605165, DMS-0907632, and SES-0835531. Yuval Benjamini is partially supported by an NSF VIGRE Graduate Fellowship. In addition, Bin Yu is partially supported by a grant from MSRA.

References

Breiman, L. and Friedman, J. (1997). Predicting multivariate responses in multiple linear regression. *J. Roy. Statist. Soc.* **59**, 3-54.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica.* **20**, 927-960.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology* **29**, 158-167.

Kay, K., Naselaris, T., Prenger, R. and Gallant, J. (2008). Identifying natural images from human brain activity. *Nature* **452**, 352 - 355.

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: jzjia@stat.berkeley.edu

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: yuvalb@stat.berkeley.edu

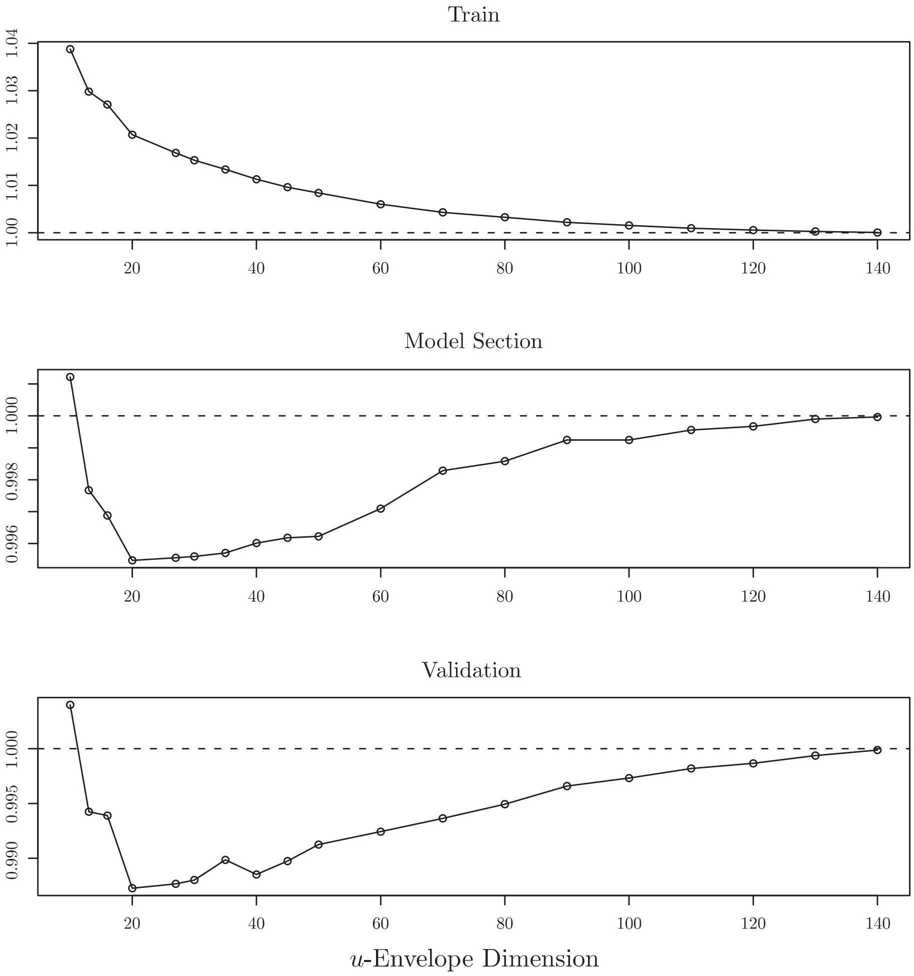


Figure 2.5. Prediction errors of envelope-MLE (normalized by OLS errors) on fMRI data for training, parameter selection, and validation sets when u varies. For $u = r$ the envelope errors are similar to OLS errors. Both parameter tuning set and validation set prediction errors were minimized at $u = 20$.

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: lim@stat.berkeley.edu

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: garveshr@stat.berkeley.edu

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: binyu@stat.berkeley.edu

(Received January 2010; accepted February 2010)