Rejoinder

R. Dennis Cook[1], Bing Li[2], Francesca Chiaromonte[2], and Zhihua Su[1]

[1]University of Minnesota and [2]Pennsylvania State University

March 14, 2010

We are grateful to the discussants for their encouraging reactions. We found their comments to be stimulating, many pointing to fresh directions that suggest envelopes may indeed have a place in the future of multivariate analysis. It was not possible to respond usefully to all of the discussants' comments, and so we focused our discussion on common themes.

1 Advantages of envelopes

We begin by considering when envelopes might offer an advantage over the standard likelihood analysis when model (10) holds. A necessary condition for this is $u < \min(r, p)$ and then envelopes will perform asymptotically better than the standard analysis simply because of parsimony; that is, the envelope model is then based on fewer parameters. When $\boldsymbol{\beta}$ has full column rank, this necessary condition reduces to $u < p < r$. Jia *et al.* reported simulation results for scenarios where $u < p < r$, and where $u < r < p$ and $\boldsymbol{\beta}$ has rank u.

However, $u < \min(r, p)$ alone does not guarantee substantial gains even if $u \ll \min(r, p)$. Let $\|\mathbf{A}\|$ denote the spectral norm of the matrix $\mathbf{A}$. We have observed that the gains from envelopes is insubstantial when $\|\boldsymbol{\Sigma}_1\| \approx \|\boldsymbol{\Sigma}_2\|$. Solid gains can arise when $\|\boldsymbol{\Sigma}_1\| \gg \|\boldsymbol{\Sigma}_2\|$, while massive gains typically arise when $\|\boldsymbol{\Sigma}_1\| \ll \|\boldsymbol{\Sigma}_2\|$. The

latter observation is supported by the relative efficiency given in (34) and is what we found in the simulations of Section 7.1 and in the analysis of the wheat protein data in Section 7.2. In the first simulation scenario of Jia, *et al.* neither $\|\boldsymbol{\Sigma}_1\|$ nor $\|\boldsymbol{\Sigma}_2\|$ dominate and consequently we conjecture that the results shown their Figure 2.1 are due primarily to parsimony. We were encouraged by their simulation results overall, and anticipate that stronger relative performance of envelopes can be demonstrated when $\|\boldsymbol{\Sigma}_2\| \gg \|\boldsymbol{\Sigma}_1\|$. In contrast, Ni controlled the relative sizes of $\|\boldsymbol{\Sigma}_2\|$ (his σ_0^2) and $\|\boldsymbol{\Sigma}_1\|$ (his σ^2) and, when $\sigma_0 \gg \sigma$, he observed good relative performance for the envelope estimator in Figure 1(a). The curious dip in that figure is explained by our discussion around (34): The OLS and envelope estimators have equal asymptotic efficiency when $\sigma_0 = \sigma$, but otherwise the envelope estimator has smaller variation. Ni's Figure 1(b) will be discussed in Section 3.

During the past few months we have analyzed many data set from the literature, mostly with small to moderate values of r. In some cases envelopes demonstrated no worthwhile gains over the standard analysis and in other cases envelopes showed modest but desirable gains. However, we also observed massive gains in some analyses. For example, consider a small data set from Johnson and Wichern (2008). There are 42 measurements on air-pollution variables recorded at noon in Los Angeles on different days. Wind speed and solar radiation were taken as predictors, and the 5 responses are measurements for CO, NO, NO2, O3 and HC. With $u = 1$, which is supported by the likelihood ratio test, the ratios of the standard errors between the full model and the envelope model range from 1.80 to 176.98, $\|\widehat{\boldsymbol{\Sigma}}_1\| = 0.21$ and $\|\widehat{\boldsymbol{\Sigma}}_2\| = 31.06$, again supporting the notion that envelopes can give massive gains when $\|\boldsymbol{\Sigma}_2\| \gg \|\boldsymbol{\Sigma}_1\|$.

Our experiences from contrasting envelopes with other methods in simulations and data analysis led us to the empirical conclusion that, depending on u and the relationship between $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, envelopes typically perform about the same, better or much better than other methods in prediction and estimation. This conclusion

seems to be supported by the simulation results reported by the discussants.

Using an invariance argument, He and Zhou reasoned that the advantages of enveloping arise from the general benefits of shrinkage rather than from any intrinsic benefits of the envelope model *per se*. We agree that shrinkage plays a role in the performance of envelopes. At the end of Section 1.1 we mentioned the related notion that enveloping can be used as means of regularization and thereby achieve a measure of "eigen sparsity" through which the estimates are shrunk. In addition, we also find value in our original motivation (Section 1.1) for the model itself as a means of characterizing regressions in which the distribution of some linear combinations $\boldsymbol{\Gamma}^T\mathbf{Y}$ of the response do not change with the predictors. Shrinkage or penalization can then be seen as ways to down weight or eliminate the linear combinations that change relatively little.

We are in agreement with He and Zhou's expression of Box's memorable statement "All models are wrong, but some are useful." They rightly point out that if $\boldsymbol{\Sigma}$ is chosen randomly with an absolutely continuous distribution, then $\boldsymbol{\beta}$ falling into any lower-dimensional envelope is a zero-probability event. Echoing Box, we can go a step further to say that essentially all useful models are zero-probability events. Take, for example, the sparse linear regression model $Y = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta}_i = 0,\ i \in A$, where A is a subset of $\{1, \ldots, p\}$. Since any proper subspace of $\mathbb{R}^p$ has Lebesgue measure 0, this model is also a zero-probability event in terms of a probability of $\boldsymbol{\beta}$ dominated by Lebesgue measure. Another example is the nonparametric variable selection model $Y \perp\!\!\!\perp \mathbf{X}|X_i, i \in A$, which again is a zero-probability event in the same sense, for the same reason. Just which specific zero-probability event we should pay attention to is a piece of transcendental intuition with which, we hope, nature can strike a cord. We argue that assuming $\boldsymbol{\beta}$ falling into a lower dimensional envelope is at least as reasonable as assuming some components of $\boldsymbol{\beta}$ are 0, because the latter is, in fact, a special case of the former when the envelope is known to be spanned by $\{e_i : i \in A\}$, where e_i is a vector whose ith component is 1 and the rest of its

components are 0. Enveloping, shrinkage and penalization are then manifestations of this basic philosophy.

2 Partial least squares

Helland pointed out a very interesting and important connection with partial least squares. The model described by Helland is actually the envelope model where the envelope is that of the covariance matrix of the predictor $\mathbf{X}$, say $\mathbf{\Sigma}_{\mathbf{X}} = \text{cov}(\mathbf{X})$. This is the model we briefly discussed in Section 8.4 but did not fully explore. The envelope model that we developed is based on the conditional covariance matrix of $\mathbf{Y}$ given $\mathbf{X}$, say $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \text{var}(\mathbf{Y}|\mathbf{X})$, where $\mathbf{\Sigma}$ is the notation used in the main article.

The connection between partial least squares and envelopes can be summed up intuitively as follows. The conditional variance $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ has r eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$; the predictor variance $\mathbf{\Sigma}_{\mathbf{X}}$ has p eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_p$; we are interested in the conditional mean $\text{E}(\mathbf{Y}|\mathbf{X})$, which is the focus of regression. Not all of the $\mathbf{v}$'s or $\mathbf{w}$'s "participate in" (so to speak) the regression. If all of the $\mathbf{v}$'s, but only some of the $\mathbf{w}$'s participate in the regression, then the model is related to the partial least squares in the way described by Helland. If all of the $\mathbf{w}$'s but only some of the $\mathbf{v}$'s participate in the regression, then the model is the main envelope model in our paper, and it is not directly related to partial least squares. As Helland pointed out in the first case there is an explicit solution: it is essentially the projection of the least-squares estimate onto the envelope. But in the second case there is *no* explicit solution, and numerical maximization over Grassmann manifold or some other iterative algorithm is necessary.

There is also the third case: not all of the $\mathbf{v}$'s and not all of the $\mathbf{w}$'s participate in the regression. This would induce further model reduction, and is a promising field to explore. We discussed this possibility of simultaneous envelopes in Section 8.5.

Hung and Huang's Proposition 1 is a nice addition to the tools for studying en-

veloping. They used it as a foundation for combining PLS and envelopes in a novel algorithm for prediction, as illustrated in their classification example with SVM. The idea of applying PLS for initial gross reduction and then refining it by using envelopes is intriguing, particularly for regressions where $n \ll p+r$. Here we would like to make two points. First, Chung and Keleş (2010) recently proved that the PLS estimator of the coefficient vector in the univariate linear regression of Y on $\mathbf{X}$ is consistent when $p/n \to 0$, but inconsistent otherwise. As a consequence, we hesitate to use PLS when $n \ll p + r$, although this requires further study in view of the results shown in Hung and Huang's Figure 1(b).

Second, in our experience the relative performance of PLS and envelopes again depends on the relationship between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. To illustrate, we set $r = 1$ and $p = 7$ and generated $(Y, \mathbf{X})$ as multivariate normal data, with the ultimate goal of predicting the univariate response from u linear combinations of the 7 predictors. We reduced the dimension of $\mathbf{X}$ by using an envelope for the inverse regression of $\mathbf{X}$ on Y, essentially treating $\mathbf{X}$ as the response, and then predicted using the linear regression of Y on $\widehat{\mathbf{\Gamma}}^T\mathbf{X}$, the u linear combinations of $\mathbf{X}$ arising from the estimated envelope. The performance of envelopes relative to PLS in this setting is controlled by u and by the relationship between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, which now refer to the inverse regression of $\mathbf{X}$ on Y. In the simulation scenario reported here, the true $u = 2$, $n = 60$ and $\mathbf{\Sigma}$ was constructed to have eigenvalues about 0.05, 1.6, 3, 28, 80, 84 and 584. The results are shown in Figure 1, where the horizontal axis the dimension of the envelope, which corresponds to number of components in PLS, and the vertical axis is the squared prediction error determined by 5 fold cross validation. In the top panel of Figure 1, $\mathbf{\Sigma}_1$ has eigenvalues 84 and 584 and envelopes do significantly better than PLS, but in the bottom panel $\mathbf{\Sigma}_1$ has eigenvalues 80 and 84, and the performance of the two estimators is essentially the same.

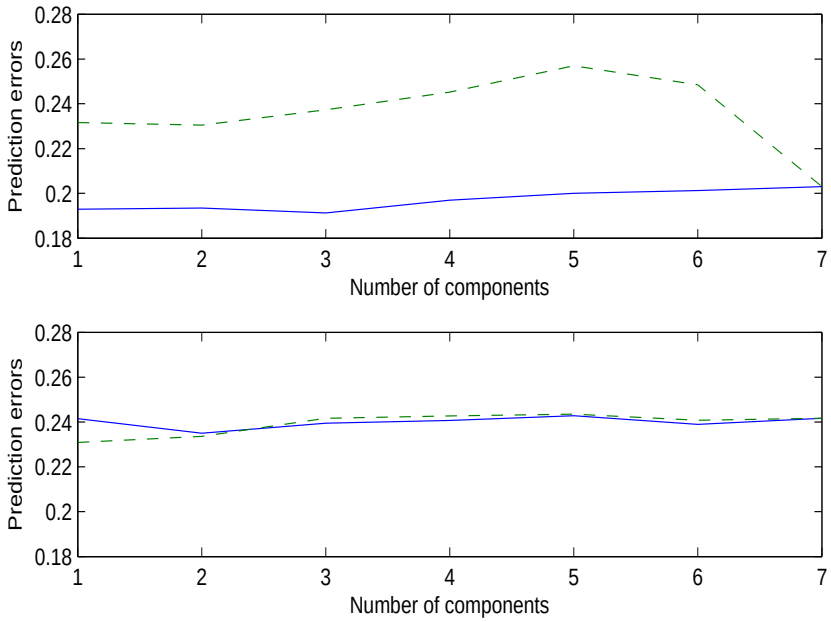


Figure 1: Simulation results on prediction errors of PLS and the envelope estimator. The solid line marks the envelope prediction error and the dashed marks the prediction error of PLS.

3 Computing

Many of the discussants raised the issue of computing, highlighting the fact that Grassmann optimization can be quite slow then r is large. This arises in part because the algebraic dimension of a Grassmann manifold is $u(r-u)$. If $u = 50$ and $r = 100$ then the optimization is taking place essentially in $\mathbb{R}^{2500}$. Our current code is useful for r up to 100 with modest values of u but is still annoyingly slow in larger problems although we are working on faster versions. Local optima can also be an issue, mostly when the signal is weak.

Two packages – LDR and GrassmannOpt – are available for optimization over Grassmann manifolds. LDR is written in Matlab and is available at

http://sites.google.com/sites/lilianaforzani/ldr-package.

This package implements many methods for sufficient dimension reduction, and in-

cludes routines for envelope models which require analytic first derivatives and use numerical second derivatives. As a consequence, starting at a root-n consistent estimator will result in a final estimator that is asymptotically equivalent to the MLE, even if local optima are present (see, for example, Small *et al.* (2000)). Nevertheless, like most programs for nonlinear optimization, there is no guarantee that it will always reach the global maximum. This might not be worrisome in the analysis of a single data set where it is possible to study the objective function. However, in simulation studies local optima can bias the results and be quite annoying. GrassmannOpt is written in R and is available at

http://CRAN.R-projects.org/package=GrassmannOptim.

While there are as yet no special routines for envelope models, one advantage of GrassmannOpt is that it contains an option for simulated annealing, which can avoid local optima at the expense of computing time. Computing can take a long time in both packages if r is large.

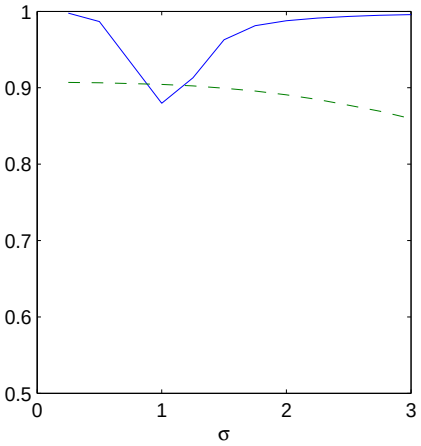
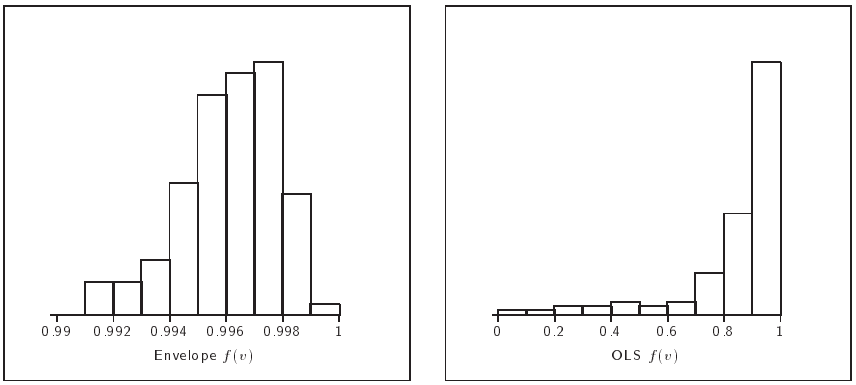


Figure 2: Rerun of Ni's Figure 1(b) using $\mathbf{\Gamma}$ as the starting value. The vertical axis is $h(v)$ and the horizontal axis is σ. The solid line marks envelope MLE and the dashed line is OLS.

We were initially a bit perplexed by the results shown in Ni's Figure 1(b). His simulation model is covered by the relative efficiency given in (34), which shows that

enveloping is asymptotically superior to OLS when $\sigma_0 \neq \sigma_1$. Thus there seems to be some disagreement between (34) and Ni's Figure 1(b). To see if Ni's routine might have gotten trapped by local optima, we reran his simulation scenario with our code using the true $\mathbf{\Gamma}$ as the starting value. The result shown in Figure 2 agrees qualitatively with the relative efficiency in (34). It seems then that algorithms using random starts might be prone to getting trapped by local maxima. To investigate this possibility, we ran our code for Ni's simulation with $\sigma = 3$ and $\sigma_0 = 1$, the right most point in Ni's Figure 1(b). The results are shown in Figure 3. The means of Ni's $h(v)$, 0.996 for envelopes and 0.840 for OLS, correspond reasonably to those at the right most plotted point of Figure 2. And the quality of the relative variations of the two estimators is predicted by (34). We expect that our implementation of Grassmann optimization worked well here because it includes an initial search over potential starting values for $\mathbf{\Gamma}$, including the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$. For instance, when $\sigma = 3$ and $\sigma_0 = 1$ in Ni's simulation model, $\mathbf{\Sigma}_{\mathbf{Y}} = 9.5\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Gamma}_0^T$, and the first eigenvector of $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ provides a root-n consistent starting value.



a. Envelopes b. OLS

Figure 3: Histograms of $f(v)$ from 100 runs with $\sigma = 3$ and $\sigma_0 = 1$.

Hung and Huang's proposal of using PLS for gross reduction followed by enveloping is appealing from a computational view. With that method they were able to analyze a classification problem with $r = 15109$, $p = 5$ and $u = 4$, which we found to be impressive since it is not possible to handle problems of that size with our

current implementation of Grassmann optimization. We expect that the numerical instability they noticed in their Figure 3 for larger component numbers was caused by convergence to local optima and does not reflect an intrinsic property of envelopes. Optimization issues like local minima may also be the cause of the extra variation of envelope predictions that was observed by Jia, *et al.*

He and Zou raised the possibility of using canonical correlations for dimension reduction, in part because they are simpler to compute and only standard software is needed. This raises the more general possibility of using canonical correlations for gross reduction followed by enveloping, in much the same way that Hung and Huang used PLS and envelopes. Nevertheless, one possible criticism of the C-estimator is that it fails to take into account the asymmetric nature of regression, and this is related to our rejoinder to Helland in Section 2. In regression we are often interested in estimating $\mathrm{E}(\mathbf{Y}|\mathbf{X})$. This implies that reducing the dimension of $\mathbf{Y}$ is of a different nature reducing the dimension of $\mathbf{X}$. Using the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ to determine the envelope for projecting $\boldsymbol{\beta}$ seems to be treating $\mathbf{X}$ and $\mathbf{Y}$ symmetrically. However, it may well be the case that the C-estimator is more appropriate in a different context where $\mathbf{X}$ and $\mathbf{Y}$ play symmetric roles.

Ni proposed another interesting algorithm based on one at a time minimization over basis vectors. This algorithm also merits further study, but its value may depend heavily on having good starting values to avoid local minima.

4 Extensions and combinations

The discussants mentioned several thought-provoking ways in which enveloping might be extended or combined with other methods. Wen's result on Fisher consistency of the envelope MLE under a misspecified link function is intriguing because it suggests that there may well be a useful link-free version of enveloping methodology. Along similar lines, Helland hinted that it may be possible to adapt enveloping for

application with generalized linear models.

He and Zhou expressed the view that "More efficient estimation is often achieved without reliance on any formal dimension reduction method," and then went on to illustrate their point by using a penalized full model log likelihood to demonstrate that penalization can result in improvements beyond those for enveloping alone. The differences between shrinkage and enveloping are certainly worth exploring, but we emphasize that this is not an either-or situation: there is nothing in principal that would prevent us from penalizing an envelope log likelihood and thereby combining the benefits of both approaches.

Indeed, using a penalized envelope log likelihood is exactly one of the approaches proposed by Yu and Zhu. We think that this is a promising direction to pursue. Consider, for example, the penalty function $\rho(\boldsymbol{\Gamma}) = \lambda \sum_{i=1}^{r} (\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)_{ii}^{1/2}$, which is like the penalty suggested by Yu and Zhu, except only the diagonal terms are used. For any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u \times u}$, $\rho(\boldsymbol{\Gamma}) = \rho(\boldsymbol{\Gamma}\mathbf{O})$ and consequently ρ depends only on span($\boldsymbol{\Gamma}$). In effect, ρ penalizes the rows of $\boldsymbol{\Gamma}$ and this is exactly what is needed to tell which responses are independent of changes in $\mathbf{X}$. Chen, Zou and Cook (2010) used ρ in combination with standard methods like SIR and SAVE to produce sparse estimates of the central subspace. They showed that their penalized subspace estimator CISE has the oracle property and that it dominates various other methods, including methods that penalize individual elements of $\boldsymbol{\Gamma}$. This is a specific instance of the synergy between dimension reduction and penalization. X. Chen (personal communication) has also conducted some small simulations to explore the potential advantages of using ρ in combination with enveloping based on minimizing

$$\log\det(\boldsymbol{\Gamma}^T \widehat{\boldsymbol{\Sigma}}_{\text{res}} \boldsymbol{\Gamma}) + \log\det(\boldsymbol{\Gamma}_0^T \widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}} \boldsymbol{\Gamma}_0) + \rho(\boldsymbol{\Gamma})$$

over the Grassmann manifold $\mathbb{G}^{r \times u}$. Here also, the results support the notion of a synergy between dimension reduction and penalization.

We framed our development in the context of the multivariate normal linear model,

but the underlying idea and formal definition of an envelope are based only on moments and do not require normality. Consequently, we are free to pursue envelope estimation in ways that rely less on an underlying distribution. Thinking along these lines we pose the following approach: For each fixed $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, let $\mathbf{v}_1(\boldsymbol{\beta}), \ldots, \mathbf{v}_r(\boldsymbol{\beta})$ be the eigenvectors of

$$\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i)(\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i)^T.$$

We want $\boldsymbol{\beta}$ to be such that

1. $(\boldsymbol{\beta}\mathbf{X}_1, \ldots, \boldsymbol{\beta}\mathbf{X}_n)$ is as close to $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ as possible, and
2. it is orthogonal to as many as eigenvectors $\mathbf{v}_\ell(\boldsymbol{\beta})$ of $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ as possible, so that the remaining eigenvectors effectively form an envelope.

It seems reasonable then to minimize the following objective function

$$\sum_{i=1}^{n}\|\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i\|^2 + \sum_{\ell=1}^{r}\lambda_\ell\sqrt{\mathbf{v}_\ell^T(\boldsymbol{\beta})\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}\mathbf{v}_\ell(\boldsymbol{\beta})}.$$

Intuitive, minimizing this function would bring $\boldsymbol{\beta}\mathbf{X}$ close $\mathbf{Y}$ and at the same time force $\boldsymbol{\beta}$ to be orthogonal to a subset of the eigenvectors, depending on the tuning parameters λ_ℓ.

One can also consider the sparsity of $\mathbf{X}$ and the eigen-sparsity of $\mathbf{Y}$ together — for example by minimizing the function

$$\sum_{i=1}^{n}\|\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i\|^2 + \sum_{\ell=1}^{r}\lambda_\ell\sqrt{\mathbf{v}_\ell^T(\boldsymbol{\beta})\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}\mathbf{v}_\ell(\boldsymbol{\beta})} + \sum_{k=1}^{r}\sum_{\ell=1}^{p}\tau_{k\ell}|\beta_{k\ell}|.$$

5 Discrimination

Hung and Huang and Dong and Zhu considered the value of enveloping in the context of discrimination, although from different perspectives. Hung and Huang demon-

strated good performance of PLS and enveloping relative to SVM.

Dong and Zhu addressed directly a conjecture we made in Section 8.2: When $\|\boldsymbol{\Omega}_0\| \ll \|\boldsymbol{\Omega}\|$ (equivalently $\|\boldsymbol{\Sigma}_2\| \ll \|\boldsymbol{\Sigma}_1\|$) the misclassification rates based on enveloping will be substantially less than those based on Fisher's linear discriminant. The required relation here $\|\boldsymbol{\Sigma}_2\| \ll \|\boldsymbol{\Sigma}_1\|$ is the reverse of the most desirable relation $\|\boldsymbol{\Sigma}_2\| \gg \|\boldsymbol{\Sigma}_1\|$ that we stated in Section 1 of this rejoinder. Intuitively, that difference arises because from a predictive view the roles of $\mathbf{Y}$ and $\mathbf{X}$ are reversed: In Section 1 we were concerned with predicting $\mathbf{Y}$ from $\mathbf{X}$ while discriminant analysis deals with predicting $\mathbf{X}$ from $\mathbf{Y}$. We were pleased to see that Dong and Zhu confirmed our conjecture with up to 30 percent gains in prediction error for enveloping. However, similar to the circumstances surrounding Ni's Figure 1(b), we did not anticipate that enveloping would be inferior to Fisher's linear discriminant when $\|\boldsymbol{\Sigma}_2\| \gg \|\boldsymbol{\Sigma}_1\|$; that is, when $\sigma_0 \gg \sigma$. Our intuition suggests that the results shown in Dong and Zhu's Table 1 for $\sigma_0 = 9$ are again due to local maxima and starting values. To check on this possibility we ran two instances from their Table 1, both with $n = 50$ in scenario (i), but using 2000 replications. In the first, we set $\sigma_0 = 1$ and obtained the misclassification rates (Full, Envelope) $= (8.707, 5.877)$, which agrees well with their results. In the second instance, we set $\sigma_0 = 9$ and obtained (Full, Envelope) $= (8.862, 9.089)$, which shows a much closer agreement between the methods. We do not know why our results differ, but suspect the reason rests again operationally with starting values. In any case, we would like to emphasize that this also is not an either-or situation. Fisher's linear discriminant arises as a special case of enveloping when $u = r$. Consequently, in practical problems we might use cross-validation to choose u, perhaps arriving at Fisher's discriminant when $\|\boldsymbol{\Sigma}_2\| \gg \|\boldsymbol{\Sigma}_1\|$, but typically using proper envelope classification $u < r$ with improved performance when $\|\boldsymbol{\Sigma}_2\| \ll \|\boldsymbol{\Sigma}_1\|$. Finally, revisiting a theme we expressed in Section 4, penalization might be combined with envelope discrimination to improve classification rates even further.

6 Other issues

6.1 Second order bias

Yu and Zhu raised the possibility that there might be a worrisome second order bias in the envelope estimator of $\boldsymbol{\beta}$ and pointed to a couple of ways in which that bias can be mitigated. However, we wonder if their numerical results in Table 1 show that the gain could be worth the effort. The squared bias norms given in their Table 1 relate to $\boldsymbol{\beta} = (\sqrt{10}, \ldots, \sqrt{10})^T$. Assuming that each element of $\sum_{i=1}^{200} \boldsymbol{\beta}_{\text{em}}^i - \boldsymbol{\beta}$ is the same order of magnitude, the element-wise bias they report in the worst case ($\sigma_0 = 2$, $n = 20$) is only about one percent of the common element ($\sqrt{10}$) of $\boldsymbol{\beta}$. A one percent bias will often be swamped by variation and may be unimportant for the scientific substance of the analysis.

6.2 Heterscedasticity

Lue and Su bring up the important issue of heteroscedasticity. Their simulation results illuminate the following point. In any linear regression model, the linear coefficient cannot fully recover a function of $\mathbf{X}$ in the variance of the error, unless that function depends only on the effective predictor. More specifically, consider the model

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \mathbf{F}(\boldsymbol{\gamma}^T\mathbf{X})\boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, $\boldsymbol{\gamma} \in \mathbb{R}^{p\times s}$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{F} : \mathbb{R}^p \to \mathbb{R}^{r\times p}$, and $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{Y}$. In this model, unless $\text{span}(\boldsymbol{\gamma}) \subseteq \text{span}(\boldsymbol{\beta}^T)$, no consistent estimator of $\boldsymbol{\beta}$ can fully recover $\text{span}(\boldsymbol{\gamma})$. Lue and Su's model (1) is a special case of (1) with

$$\boldsymbol{\beta} = (\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2, \mathbf{0}, \mathbf{0}, \mathbf{0})^T, \quad \boldsymbol{\gamma} = \mathbf{e}_3.$$

Hence span($\boldsymbol{\gamma}$) $\perp$ span($\boldsymbol{\beta}$), which falls in the described scenario. At the same time, mrSIR is capable of recovering the directions in the variance. We suspect that the same trend displayed in Table 1 in Lue and Su's comments would uphold even if the true $\boldsymbol{\beta}$ is used in place of the MLE under envelope model. In this case (span($\boldsymbol{\gamma}$) $\perp$ span($\boldsymbol{\beta}$)) the information about $\boldsymbol{\gamma}$ can only be found in the residuals.

Another interesting point related to Professor Lue and Su's comments is how to construct an envelope model when the conditional variance var($\mathbf{Y}|\mathbf{X}$) depends on $\mathbf{X}$. In this case a reducing subspace $\mathcal{S}$ of var($\mathbf{Y}|\mathbf{X}$) also depends on $\mathbf{X}$. At present we do not yet have an answer to this question.

References

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. Under revision.

Chung, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society B*, **72**, 3–25.

Johnson, R. A. and Wichern W. A. (2008). *Applied Multivariate Statistical Analysis* New Jersey: Prentice Hall.

Small, C. G., Wang, J., and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statistical Science* **15**, 313–332.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
Emails: dennis@stat.umn.edu and zhihua.sophia@gmail.com.

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.
Emails: bing@stat.psu.edu and chiaro@stat.psu.edu.