Research Article

Smoothed Linear Modeling for Smooth Spectral Data

Douglas M. Hawkins\(^1\) and Edgard M. Maboudou-Tchao\(^2\)

\(^1\) School of Statistics, University of Minnesota, 313 Ford Hall, Minneapolis, MN 55455, USA
\(^2\) Department of Statistics, University of Central Florida, Orlando, FL 32816, USA

Correspondence should be addressed to Douglas M. Hawkins; dhawkins@umn.edu

Received 2 March 2013; Revised 4 May 2013; Accepted 19 May 2013

Copyright © 2013 D. M. Hawkins and E. M. Maboudou-Tchao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Classification and prediction problems using spectral data lead to high-dimensional data sets. Spectral data are, however, different from most other high-dimensional data sets in that information usually varies smoothly with wavelength, suggesting that fitted models should also vary smoothly with wavelength. Functional data analysis, widely used in the analysis of spectral data, meets this objective by changing perspective from the raw spectra to approximations using smooth basis functions. This paper explores linear regression and linear discriminant analysis fitted directly to the spectral data, imposing penalties on the values and roughness of the fitted coefficients, and shows by example that this can lead to better fits than existing standard methodologies.

1. Introduction

There are a number of settings in which one wishes to predict some dependent variable \(y\) from measurements of an optical spectrum. For example, Brown et al. [1] were concerned with the regression problem of predicting the composition of dough on the basis of measurements of the near infrared spectrum emitted from dough samples. Schomacker et al. [2] discussed the classification problem of deciding whether a colon polyp was benign or malignant on the basis of the optical spectrum emitted after illuminating the polyp with a laser.

Both these applications involve electromagnetic spectra, but the scope of spectral data modeling is much broader: auditory spectra and chemical chromatography data also fit the same framework as the general problem considered here.

Many fields of science now deal with high-dimensional data sets—“large \(n\)” data sets have many cases, and “large \(p\)” data sets have many measurements per case. “Small \(n\) large \(p\)” problems are particularly challenging and have been the subject of much recent research. Spectral data are generally in the “large \(p\)” class and may also have “small \(n\).” They are, however, different from “small \(n\) large \(p\)” problems in many other areas such as microarrays in that it is often expected that the regression models should be smooth: if the signal measured at 450 nm is predictive, then one would expect that measured at 449 nm and at 451 nm to be about equally predictive. This is the setting considered in this paper, where it is assumed that subject-matter knowledge motivates models in which the information varies smoothly with wavelength. This feature sets such spectral data apart from the typical statistical high-dimensional data set and leads to considering methods that fit models in which the coefficients are smooth functions of the wavelengths to which they apply.

One standard general approach is that of functional data analysis—FDA—[3], in which the spectrum is approximated by a linear combination of smooth basis functions, automatically leading to a smooth version. Replacing each spectrum by its coefficients in these basis functions then leads to approaches that perform respect smoothness. This two-step procedure, however, is not the only way to approach the problem, and we discuss old and new methods that use the measured spectra directly.

2. Notation

Let \(y\) refer to the dependent variable. This may be continuous, or may be a 0-1 indicator. Write the spectrum used for the prediction as \(x\), with elements \(x_1, \ldots, x_p\) being the energies...
measured at the \( p \) wavelengths studied. Suppose that we have \( n \) samples which we wish to use to predict a dependent \( y \). A linear predictor has the form

\[
\hat{y}_i = \sum_{j=1}^{p} x_{ij} \hat{\beta}_j, \tag{1}
\]

where the subscript \( i \) refers to a particular sample and the coefficient \( \hat{\beta}_j \) is the contribution of wavelength \( j \) to the overall model. We assume that the regression intercept has been removed from the problem by subtracting the mean from each of the variables. In settings where smoothness is expected, in the model (1) each \( \hat{\beta}_j \) should not differ dramatically from its neighbors \( \hat{\beta}_{j-1} \) and \( \hat{\beta}_{j+1} \). As pointed out by Fan and James [4], there are different ways of doing this. One is to enforce smoothness directly by imposing roughness penalties when fitting the linear model; the other is to reformulate the regression in a way that makes roughness impossible.

Functional data analysis (FDA) follows this second path. It sidesteps the representation (1) by first approximating the spectrum by a linear combination of basis functions

\[
x_{ij} \approx \sum_{k=1}^{K} \tilde{\alpha}_{sk} g_k(j), \tag{2}
\]

where the set \( g \) is a basis of smooth functions of wavelength \( j \). The spectrum \( x \) is then replaced by the \( K \) coefficients \( \tilde{\alpha}_{sk} \), which can be used as predictors, giving the model

\[
\hat{y}_i = \sum_{k=1}^{K} \tilde{\alpha}_{sk} \tilde{y}_k \tag{3}
\]

with fitted coefficients \( \tilde{y}_k \). Common choices for the basis functions \( g \) include Fourier series, B splines, and wavelets. The implicit smoothing operation in going from the initial spectral energies to their basis coefficients leads to an automatic smoothness in the dependence of the fitted model on the individual wavelengths.

The FDA approach leads to two levels of fitting and associated approximation—of \( x \) by the basis representation, and of \( y \) by the linear combination of basis coefficients. It is most successful when the shape of the individual spectra can be captured using only a few basis functions, leading to a major reduction in dimensionality. It is less successful if we are not able to get an accurate parsimonious representation in basis functions.

The other more direct approach is to fit the model (1) directly, but in a way that penalizes roughness in the coefficient series. This is the approach that will be followed here.

3. Constrained Fitting

Suppose that the calibration data set contains \( n \) cases, with \( y \) values \( y_i, i = 1, 2, \ldots, n \), and matching spectra \( x_{ij}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, p \), and write these in matrix form as the \( n \)-vector \( Y \) and \( n \times p \) matrix \( X \). A constrained least squares fit minimizes the objective function

\[
(Y - X\beta)^T (Y - X\beta) + g(\beta), \tag{4}
\]

where \( g \) is the functional form of the penalty on the coefficient vector \( \beta \). Conventional choices of \( g \) include

- The lasso: \( g(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| \) [5],
- Ridge: \( g(\beta) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 \) [6],

where \( \lambda_1 \) is a user-selected penalty constant. While these penalties penalize the sort of large and erratic coefficients that tend to emerge in unconstrained fitting, they do not address the smoothness issue. A penalty that was developed for smooth spectral data is

\[
\text{The fused lasso: } g(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} (|\beta_j - \beta_{j-1}|) \tag{7}
\]

which attempts to prevent not only wild coefficients but also coefficients that differ substantially from their neighbors.

Each of these constrained fitting criteria involves a penalty constant \( \lambda_1 \), and the fused lasso has a second penalty constant \( \lambda_2 \). The values of these penalty coefficients need to be fixed, which is commonly done using cross-validation, as outlined below.

The ridge and lasso penalties are both well suited to the generic data sets of the large-\( p \) type and to highly correlated predictors. They do not impose smoothness per se; they merely reduce the tendency for individual coefficients to be numerically large, regardless of the values of their neighbors. They are however quite different in their effects. If predictors are very similar, the ridge penalty tends to smooth their coefficients toward a common value. The lasso does not—as the penalty \( \lambda_1 \) increases, the lasso increasingly drops predictors from the regression altogether. Thus, given a set of highly correlated predictors, ridge tends to give them all roughly equal coefficients, whereas the lasso tends to select a single one to represent the whole set and to drop all its correlates.

In many regression settings, this feature selection is desirable. Sometimes feature selection helps with understanding the mechanisms connecting the predictors with the dependent variable. In drug discovery, for example, identifying the molecular features that lead to biological activity helps the user to design new and better compounds. And even when feature selection is not of direct interest, dropping unneeded predictors means that they do not have to be measured before future unknowns can be predicted, and this can save measurement cost and effort.

But this argument has little force with spectral data where the entire spectrum is measured whether or not all its individual wavelengths are used. In fact, a good case may be made that feature selection is undesirable, and that one should include all of the wavelengths in the predictor. Feature selection that leads to using only a handful of wavelengths makes the fitted model sensitive to the measurements at those wavelengths and so susceptible to malfunction, drifts,
and calibration fluctuations in the individual sensors in that instrument. It also complicates the use of the fitted model on other instruments of the same type as it means that the calibration of the "sentinel" wavelengths used must be highly accurate.

A regression that incorporates all wavelengths, by contrast, is more robust to minor differences from one instrument to another and to a few bad sensors, drifts, and calibration issues in the same instrument over time.

Like the lasso, the fused lasso also leads to feature selection, but of a different sort. The first term in its penalty leads to a tendency, like that of the lasso, to set some of the fitted coefficients to zero. The second leads to a tendency to give adjacent coefficients the same value. The resulting coefficient vectors then tend to consist of sequences of zeroes interspersed with sequences of identical coefficients.

4. Other Penalties

The successive true coefficients \( \beta_i \) are likely to vary smoothly, suggesting that they are estimated using penalty functions that encourage smoothness, but without favoring feature selection. Two new penalties that have these properties are.

**Local constancy**

\[
g(\hat{\beta}) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2,
\]

(5a)

**Local linearity**

\[
g(\hat{\beta}) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=2}^{p-1} (\beta_{j-1} - 2\beta_j + \beta_{j+1})^2.
\]

(5b)

Each of these penalty functions includes a first ridging term whose effect is to shrink the fitted coefficients towards zero, and a second term that penalizes roughness. The first penalty is an \( L2 \) (sum of squares) counterpart to the \( L1 \) (sum of absolute values) penalty used in the fused lasso. However, unlike its \( L1 \) counterpart, it will tend to avoid both feature selection and runs of identical coefficients and instead give models in which the coefficients are nonzero, of modest size, and vary smoothly with wavelength.

While the first penalty is an adaptation of the fused lasso, the second penalty does not have any counterpart in the methodologies mentioned so far. Its second term penalizes roughness, as defined by a departure from local linearity. It is a digitized implementation of the integrated squared second difference penalty widely used in functional data analysis [8, 9]. It does not reward local constancy but allows even large differences from one coefficient to the next, provided that locally the successive coefficients vary approximately linearly with wavelength.

To implement the methods, we need suitable values of the penalties \( \lambda_1 \) and \( \lambda_2 \). These are commonly found by cross-validation. In leave-one-out cross-validation, trial values of

\[\lambda_1 \text{ and } \lambda_2 \text{ are selected, and } \hat{\beta} \text{ is estimated by minimizing } (4). \]

As this is a quadratic form, this is an easy minimization with an explicit matrix solution amenable to general-purpose matrix software. Then, each spectrum of the data set and its associated \( y \) is removed in turn; the model fitted to the remaining \( n - 1 \) cases, and used to predict the held-out case leading to a predicted residual—the difference between the actual \( y \) and that predicted using the remaining \( n - 1 \) cases. The sum of squares of these predicted residuals (termed PRESS) measures the predictive quality of the coefficients found using that pair of \( \lambda \) values. This calculation is repeated, varying \( \lambda_1 \) and \( \lambda_2 \) and picking the pair giving the minimum PRESS. The model is then fitted to all the data using this pair of \( \lambda \) values to get a fitted coefficient vector that we will call \( \hat{\beta}_{raw} \).

There is one final step. As with the elastic net [10], the coefficient vectors estimated by these new approaches are overshrunk, and it is advisable to bring them back to the correct scale by scaling them back up. We obtain this scale factor by fitting the no-intercept regression of \( Y \) on the naïve predictions \( \hat{Y} = X\hat{\beta}_{raw} \) and multiplying \( \hat{\beta}_{raw} \) by the slope of this fine-tuning regression.

5. Application to a Regression Application

Brown et al. [1] discussed the prediction of four constituents of biscuit dough using near-infrared (NIR) spectra measured in \( 2 \) nm increments from wavelengths 1100 nm to 2498 nm, thus containing 700 measurements per sample. They had 40 calibration samples and 40 validation samples measured at a different time, but they excluded one sample from each of these two data sets, leaving 39 samples in each.

The problem was to use these spectra to predict four constituents of the dough—fat, sugar, flour, and water. As seen in Table 1, these constituents have some high correlations in the 39 calibration samples.

Table 1: Correlations among dough constituents.

|          | Fat   | Sugar | Flour |
|----------|-------|-------|-------|
| Sugar    | −0.05 |       |       |
| Flour    | −0.27 | −0.94 |       |
| Water    | −0.63 | −0.73 | 0.87  |

Figures 1–4 give an idea of how the spectral shapes change with the constituents. Each shows the spectrum of the sample with the highest and that with the lowest concentration of that constituent. The shapes are smooth but quite complex. The sample with the lowest fat content has the highest energies across the spectrum, as does that with the lowest sugar content. The opposite is true for flour and water. In view of the high correlation between the flour and water concentrations of the samples, we might expect Figures 3 and 4 to be very similar, as indeed they are. Interchanging the high and low samples, the same is broadly true of the sugar and flour spectra.

Beyond the observation that the spectra do clearly differ between the high and the low end member of each dependent,
Table 2: Comparison of criteria in NIR data set.

| Method      | Fat  | Sugar | Flour | Water |
|-------------|------|-------|-------|-------|
| Wavelets    | 6.3  | 44.9  | 34.8  | 5.0   |
| Ridge regression | 68.6 | 87.4  | 62.8  | 41.7  |
| Fused lasso | 5.1  | 6.4   | 46.6  | 19.6  |
| Locally constant | 0.9  | 3.5   | 13.2  | 6.5   |
| Locally linear | 0.9  | 3.4   | 13.2  | 8.1   |

Figure 1: Spectra of samples with highest and lowest fat content.

Figure 2: Spectra of samples with highest and lowest sugar content.

Figure 3: Spectra of samples with highest and lowest flour content.

Figure 4: Spectra of extreme samples' spectra.

It is not easy to see quite how the spectra can be used to predict the four constituents. Brown et al. [1] used four traditional methods (details of which can be found in their paper), together with a new FDA approach using wavelets as basis functions. The main thrust of their discussion is that their wavelet basis is fully competitive with the other approaches and is best by a wide margin for predicting water content.

We analyzed the data set using ridge regression, the fused lasso, and the locally constant and the locally linear criteria. The penalty costs $\lambda_1$ and $\lambda_2$ were fitted separately for each dependent using leave-one-out cross-validation.

The models were fitted to the 39 calibration spectra. The predictive quality of the fits was assessed by using the fitted models to predict the 39 samples obtained in the separate study. Calculating the mean squared errors of prediction of the validation data gave the results shown in Table 2. These show the percentage of variance unexplained.

Comparing the five approaches,

(i) ridge regression performed poorly on all four dependents.

(ii) The locally constant method is a counterpart to the fused lasso. Comparing these two rows shows that it outperformed the fused lasso by a wide margin on all four dependents.

(iii) The locally constant and locally linear methods had nearly identical performance for three of the dependents, with the locally constant somewhat better for predicting the water content.

(iv) Both methods far outperformed the wavelet-based FDA approach of Brown et al., except for predicting water content, where the wavelet method was somewhat better.

In the light of Figures 1–4, it is interesting to see the behavior of the coefficients fitted using the new methods. Figures 5, 6, 7, and 8 show the coefficients of the locally linear fit to each of the four constituents.

The shapes of these curves do not correspond all that closely to what one might have expected from Figures 1–4. The coefficients of the fat regression, for example, pay most attention to the energies in the range from 1200 to 1300 nm,
with negative coefficients coming from the upper portion of this range. Figure 1, however, suggests that the greatest separation is at high wavelengths.

As flour and sugar are strongly negatively correlated, one might expect their coefficient graphs to be mirror images. And this is in fact substantially true, with both looking most closely at the 1800–2100 nm range. Figures 2 and 3, however, would not lead one to suspect the importance of this range of wavelengths or of the fact that its lower and upper portions weight in opposite directions.

The water coefficients weight most heavily towards the two ends of the spectrum, even though visually, Figure 4 shows the most separation in the middle. The relatively high correlation between water and flour does not translate into much correspondence between their coefficient vectors.

In summary, the smoothed coefficients do not correspond to intuitively obvious features that might be seen in the extremal spectra.
6. Application to Colon Cancer Diagnosis

Two-group classification problems can be analyzed by introducing a binary dependent variable to indicate group membership. Applying conventional ordinary least squares (OLS) regression to the resulting data set then gives Fisher’s linear discriminant analysis. However this OLS formulation cannot be used if $n < p$ and tends to perform poorly if $p$ is large. Regularizing the covariance matrix improves performance (see e.g., [11, 12]), and ridge regression provides a convenient way to regularize the system of linear equations defining the coefficients and get better behaved coefficients.

An extreme example of regularization is the naïve Bayes (NB) estimator, in which all off-diagonal elements of the cross-product matrix $X'X$ are set to zero. NB has both a good empirical record [13] and some theoretical justification [14] for high-dimensional classification problems, making it a good candidate for consideration.

A second example illustrates this classification setting with a problem arising in colonoscopy. Polyps are often found during colonoscopy. Some are adenomas—premalignant growths—and some are benign. The problem is to distinguish the two so that adenomas can be excised.

When illuminated with ultraviolet light, normal and adenomatous tissue fluoresce, but at different wavelengths. Figure 9 shows the laser-induced fluorescence (LIF) spectra of a normal and an adenomatous polyp measured at 1 nm spacing from 375 to 550 nm. As the figure shows, adenomatous tissue tends to emit lower energy at wavelengths below 430 nm and higher at wavelengths above 430 nm. Schomacker et al. [2] provided a biological rationale for this.

A colonoscopy data set comprising a total of 318 polyps, 119 of which were adenomatous and which provided a total of 1409 LIF spectra, was used as a test bench. Several of the methods described earlier were applied, along with an FDA approach implemented using Fourier series as basis functions.

![LIF spectra, normal and adenomatous](image)

**Figure 9**: Examples of normal and adenomatous tissue spectra.

| Method      | LDA | NB  | FDA  | Fused lasso | Constant | Linear |
|-------------|-----|-----|------|-------------|----------|--------|
| AUC         | 0.6293 | 0.7820 | 0.7921 | 0.8005 | 0.8008 | 0.8010 |

To illustrate the performance of the methods with a calibration sample of modest size, the available data were split into a calibration sample of 268 spectra and a validation sample of 1141 spectra.

The normalized spectral energies were used as predictors. The dependent variable $y$ was coded as 1 for adenomas and 0 for normal tissue.

Six estimators were studied:

- FDA, with Fourier series as basis functions,
- full-sample linear regression, (i.e., conventional linear discriminant analysis),
- the naïve Bayes estimator,
- the fused lasso,
- the locally constant penalty model,
- the locally linear penalty model.

The penalty constants were found by leave-one-out cross-validation using the calibration data set and the final models applied to the validation data set. The fits were then evaluated by their area under the ROC curve of the validation data set. These areas were as follows.

(i) Not unexpectedly (given the high dimensions), conventional LDA has by far the lowest area under the ROC curve, not being much better than the 0.5 given by blind guessing.

(ii) The naïve Bayes method is considerably better, in line with the Dudoit et al. [13] observation that NB often performs quite well in high-dimensional settings.

(iii) The remaining three methods in order of fit are the FDA approach, fused lasso, and the locally constant and the locally linear methods; see Table 3.

Though the performance difference of the three smoothing approaches is small, they reach their similar performance in strikingly different ways. Figure 10 shows the coefficients of the fused lasso fits. Of the 176 coefficients, 111 are zero, so only about one third of the sensors are used in calculating the fused lasso score. The nonzero coefficients range in value from $-0.69$ to $0.44$.

Figure 11 shows the coefficients of the locally constant and the locally linear fit coefficients. The locally linear coefficients match what one would expect from Figure 9: they follow the pattern of differences between the mean spectra of adenomatous and normal tissue. However, neither the locally constant nor the fused lasso coefficients have a shape that follows this mean profile.

Another major difference is perhaps less apparent from the plots. The coefficients of the fused lasso are some two orders of magnitude bigger than those of the two L2 fits even though all three fits are scaled to have the same variability as
the calibration data. The coefficients from the locally constant approach range from $-2 \times 10^{-3}$ to $4 \times 10^{-3}$. Those from the locally linear approach, ranging from $-5 \times 10^{-4}$ to $8 \times 10^{-4}$, are even smaller.

This means, the fused lasso fit would be much more sensitive than the two L2 methods to any failures or calibration drift in the sensors. Thus, even though the three penalized fits produce essentially the same performance on the validation data set and are scaled to give the same variance of fitted values, the locally linear fit is the most attractive in that it is by far the least sensitive of the three to the vagaries of the individual sensors.

7. Conclusions

Many spectral data sets tend to follow the “first law of geography,” that nearby things tend to be more similar than far-apart things, and so analysis methods that respect this law have an immediate intuitive attraction. This motivates methods in which fitted coefficients vary slowly from one wavelength to the next. Functional data analysis approaches achieve this end by approximating the spectra by superpositions of smooth basis functions, but it is also possible to work directly with the spectra, using penalty functions to penalize roughness in the fitted models.

One existing method of implementing this idea is the fused lasso, which shrinks coefficients toward their neighbors and also does feature selection by dropping some of the measurement. But with spectral data, there is no particular virtue in feature selection. It does not make data gathering easier or cheaper, and to the contrary, using all wavelengths gives models that, because they aggregate information from all sensors, are less susceptible to problems arising in a minority of the sensors.

These two considerations lead to the idea of imposing L2 penalties on the roughness of the series of coefficients fitted to spectral data for prediction as L2 penalties achieve smoothness without doing feature selection. One possible penalty motivated by this thinking encourages local constancy, another local linearity.

These two approaches are illustrated with a regression problem using near infrared spectroscopy and with a classification problem using laser-induced fluorescence. In both examples, the locally constant and the locally linear approach give more accurate fits than those given by both the fused lasso and FDA methods involving basis expansions. The LIF example also illustrates the virtue of the L2 norm in giving large numbers of relatively small coefficients rather than fewer larger ones, as this gives a model that is more robust to instrumental drifts, sensor failures, and random errors.

Acknowledgments

Tom Fearn kindly provided the NIR data for the first example. The work leading to this paper was supported by a research grant from SpectraScience, which also provided the data for the second example. The authors are grateful to the referee for several helpful suggestions for improving the paper.

References

[1] P. J. Brown, T. Fearn, and M. Vannucci, “Bayesian wavelet regression on curves with application to a spectroscopic calibration problem,” *The Journal of the American Statistical Association*, vol. 96, no. 454, pp. 398–408, 2001.
[2] K. T. Schomacker, J. K. Frisoli, C. C. Compton et al., “Ultraviolet laser-induced fluorescence of colonic tissue: basic biology and diagnostic potential,” *Lasers in Surgery and Medicine*, vol. 12, no. 1, pp. 63–78, 1992.
[3] W. Saeyes, B. de Ketelaere, and P. Darius, “Potential applications of functional data analysis in chemometrics,” *Journal of Chemometrics*, vol. 22, no. 5, pp. 335–344, 2008.
[4] Y. Fan and G. James, “Functional additive regression,” 2012, http://www-bcf.usc.edu/~gareth/.
[5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society B*, vol. 58, pp. 267–288, 1996.
[6] A. E. Hoerl and R. W. Kennard, “Ridge regression: applications to nonorthogonal problems,” Technometrics, vol. 12, pp. 27–51, 1970.

[7] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, “Sparsity and smoothness via the fused lasso,” Journal of the Royal Statistical Society B, vol. 67, no. 1, pp. 91–108, 2005.

[8] J. O. Ramsey and B. W. Silverman, Applied Functional Data Analysis, Methods and Case Studies, Springer, New York, NY, USA, 2002.

[9] J. O. Ramsey and B. W. Silverman, Functional Data Analysis, Springer, New York, NY, USA, 2nd edition, 2005.

[10] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” Journal of the Royal Statistical Society B, vol. 67, no. 2, pp. 301–320, 2005.

[11] D. M. Witten and R. Tibshirani, “Covariance-regularized regression and classification for high dimensional problems,” Journal of the Royal Statistical Society B, vol. 71, no. 3, pp. 615–636, 2009.

[12] L. S. Chen, D. Paul, R. L. Prentice, and P. Wang, “A regularized Hotelling’s T2 test for pathway analysis in proteomic studies,” The Journal of the American Statistical Association, vol. 106, no. 496, pp. 1345–1360, 2011.

[13] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” The Journal of the American Statistical Association, vol. 97, no. 457, pp. 77–87, 2002.

[14] P. J. Bickel and E. Levina, “Some theory for Fisher’s linear discriminant function, “naive Bayes”, and some alternatives when there are many more variables than observations,” Bernoulli, vol. 10, no. 6, pp. 989–1010, 2004.
