RC 20750 (02/28/97)
Computer Science/Mathematics

IBM Research Report

Computational Aspects of Exploratory Projection Pursuit

Ramesh Natarajan

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York

Computational Aspects of Exploratory Projection Pursuit

by

Ramesh Natarajan
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, N.Y. 10598.
e-mail: ramesh@watson.ibm.com

Abstract

Exploratory Projection Pursuit is a statistical technique for identifying structure in high-dimensional multivariate datasets. We discuss various computational issues associated with this technique including sphering and outlier detection, marginal density smoothing, higher-order optimization methods for index maximization, and statistical validation studies.

1 Introduction

Exploratory Projection Pursuit (PP) is a statistical technique for identifying structure (including clustering and multimodality) in high-dimensional multivariate data sets, by optimizing an index based on smoothed low-dimensional projections of the data.

PP is often used as an automatic procedure for generating low-dimensional projections of the data that provide the most interesting views for detailed examination by graphical exploratory methods (such as marginal plots, and bivariate and trivariate scatterplots, see Swayne, Cook and Buja [19]). Likewise, PP also provides a way to extend statistical density estimation techniques (which are only reliable in low dimensions due to the curse of dimensionality) to the study of high-dimensional data sets. A broad-ranging discussion of these issues can be found in Huber [9].

In this paper, we discuss various computational issues associated with a PP procedure due to Friedman [12]. Our primary interest is in tracking the PP solutions while varying the amount of smoothing in the marginal density estimates, with the idea that new insights can be obtained and new structures can be discovered in this way from the data. We also consider the use of higher-order optimization algorithms in the PP procedure, which makes it possible to obtain fully machine-precision converged results with very little extra computational effort, which simplifies the task of distinguishing identical local maxima solutions that may be obtained from different starting points.

2 Computational Details

2.1 Sphering and Outlier Removal

Let $\{\boldsymbol{a}_i\}_{i=1}^{M} \in \mathbb{R}^N$ denote the M data samples, and let $\overline{\boldsymbol{a}}$ denote their mean. Then let

$$\boldsymbol{A} = \begin{bmatrix} (\boldsymbol{a}_1 - \overline{\boldsymbol{a}})^T \\ (\boldsymbol{a}_2 - \overline{\boldsymbol{a}})^T \\ \vdots \\ (\boldsymbol{a}_M - \overline{\boldsymbol{a}})^T \end{bmatrix}, \tag{1}$$

denote the centered data matrix (if so desired, the columns of $\boldsymbol{A}$ may also be scaled in a standard way, say by the range or standard deviation of the corresponding marginal).

The SVD of $\boldsymbol{A}$ is given by (see Golub and Van Loan [7], we assume that $M > N$ and $\boldsymbol{A}$ has full column rank),

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T. \tag{2}$$

Here $\boldsymbol{U} \in I\!R^{M\times M}$ and $\boldsymbol{V} \in I\!R^{M\times N}$ are orthogonal matrices ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{N\times N}$), and the diagonal matrix $\boldsymbol{\Sigma} \in I\!R^{M\times M}$ contains the singular values in decreasing order

$$\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2, \ldots, \sigma_N, 0, \ldots, 0), \quad \text{where } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0. \tag{3}$$

We note that (2) can be equivalently written as

$$\boldsymbol{A} = \sum_{k=1}^{N} \boldsymbol{u}_k \sigma_k \boldsymbol{v}_k^T, \tag{4}$$

where $\{\boldsymbol{u}_k\}_{i=1}^N$ and $\{\boldsymbol{v}_k\}_{i=1}^N$ denote the columns of the $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices respectively. Then $\boldsymbol{v}_k$ denotes the k'th principal component direction, σ_k^2 denotes the principal variance in this direction, and $\boldsymbol{u}_k$ denotes the unit normalized projection of the data matrix $\boldsymbol{A}$ along $\boldsymbol{v}_k$.

The SVD computation yields the sphered data matrix $\boldsymbol{U}$ (in principal component coordinates), whose columns have zero mean and identity covariance matrix, and this standardization of the original data is an important preprocessing step for the PP procedure. The SVD computation can also be used to test for rank deficiency, or to identify a subspace of dominant principal components that accounts for a sufficiently large fraction of the variance in the data, so that the PP procedure can be carried out more economically in this subspace (however, we note that this dimensionality-reduction technique can sometimes lead to important structure in the data being overlooked, as is the case in one of the examples in Section 3). Finally, the SVD can used to detect and remove outliers from the data as shown below, which can only increase the sensitivity of the PP procedure to the structure in the high-density regions of the sampled data rather than in the tails.

The detection of outliers is based on comparing the Mahalanobis squared distance d_m^2 of each data point with the statistics of Chi-squared distribution with N degrees of freedom (Jolliffe [10]). This comparison tacitly assumes that the data is sampled from an underlying elliptic normal distribution (an assumption that is clearly inappropriate for any data where the PP procedure below might be expected to produce interesting results). The justification for this procedure is on the basis that PP is primarily concerned with identifying deviations from a Gaussian distribution with the same covariance structure in the high-density regions of the data sample rather than in the tails. For data samples drawn from heavy-tailed distributions, some preliminary transformation of the marginals to bring them to an approximate Gaussian form might be required prior to the use of the outlier detection and PP procedures described in this paper.

Therefore, we have for the m'th data sample,

$$d_m^2 = (\boldsymbol{a}_m - \overline{\boldsymbol{a}})^T \boldsymbol{S}^{-1} (\boldsymbol{a}_m - \overline{\boldsymbol{a}}), \tag{5}$$

where $\boldsymbol{S}$ is the sample covariance matrix, and noting that

$$\boldsymbol{S} = \frac{1}{(M-1)} \boldsymbol{A}^T \boldsymbol{A}, \tag{6}$$

we easily obtain

$$d_m^2 = (M-1) \sum_{k=1}^{N} (\boldsymbol{u}_k)_m^2, \tag{7}$$

where $(\boldsymbol{u}_k)_m$ denotes the m'th component of the vector $\boldsymbol{u}_k$ in (4). Then, all data samples for which d_m^2 lies outside the p'th percentage point of the corresponding Chi-squared distribution with N degrees of freedom are deleted from the data set. The value of p depends on the sample size, and we typically take p in the range 0.1 to 0.001.

After deleting the outliers obtained in this way, the trimmed original data matrix is recentered, the SVD is recomputed, and the test for outliers is repeated. This cycle is repeated iteratively till eventually it stabilizes with no further outliers being obtained. More elaborate procedures can be used for robust mean, covariance matrix, and principal components estimation, based on this outlier identification method (see for example, Campbell [2] where weights are applied to each data sample) but these seem difficult to combine with the subsequent PP procedure described below.

2.2 Evaluation of the Friedman Index

Let $\boldsymbol{U}_P = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_P] \in \mathbb{R}^{M \times P}$ be the sphered data matrix in the subspace of the leading P principal components, with $\{\boldsymbol{u}_i\}_{i=1}^{P} \in \mathbb{R}^M$ being the projections along the individual principal components (with M now denoting the number of datapoints surviving the outlier removal procedure, and the entries of $\boldsymbol{U}$ from the SVD in (2) are rescaled by $\sqrt{M-1}$ so that the marginals have unit variance rather than unit norm). For $\boldsymbol{\alpha} \in \mathbb{R}^P$, let $\hat{\boldsymbol{x}} = \boldsymbol{U}_P \boldsymbol{\alpha} \in \mathbb{R}^M$ denote the projection of $\boldsymbol{U}_P$ in the direction of $\boldsymbol{\alpha}$. The resulting projected sample data are the components $\{\hat{x}_l\}_{l=1}^{M}$ of $\hat{\boldsymbol{x}}$, and this marginal will have sample mean zero and variance $\boldsymbol{\alpha}^T \boldsymbol{\alpha}$. The PP optimization procedure below is simplified by leaving norm of $\boldsymbol{\alpha}$ unconstrained, and instead working with the normalized data $x_l = (\boldsymbol{\alpha}^T \boldsymbol{\alpha})^{-1/2} \hat{x}_l$.

Friedman's PP index [12] is designed to seek out those directions $\boldsymbol{\alpha}$ in which the marginal distribution based on the projected sample data $\{x_l\}_{l=1}^{M}$, has the maximum departures from the univariate Gaussian distribution. Therefore, let x denote the continuous variable in the direction $\boldsymbol{\alpha}$, and let $\Phi(x)$ denote the cumulative density function for a univariate unit Gaussian

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{8}$$

The transformation

$$p(x) = 2\Phi'(x)\phi(r), \tag{9}$$

where

$$r = 2\Phi(x) - 1 = \begin{cases} \operatorname{erf}\left(x/\sqrt{2}\right), & x \geq 0; \\ \operatorname{erfc}\left(-x/\sqrt{2}\right) - 1, & x < 0, \end{cases} \tag{10}$$

from the density $p(x)$ in $[-\infty, \infty]$ to the density $\phi(r)$ in $[-1, +1]$ maps a unit Gaussian in x to a uniform distribution in r (this transformation maps the data samples x_l to r_l).

The required PP index for measuring departures from the Gaussian is then given by

$$\mathcal{I}(\boldsymbol{\alpha}) = \int_{-1}^{+1} (\phi - \frac{1}{2})^2 dr = \int_{-1}^{+1} \phi^2 dr - \frac{1}{2}. \tag{11}$$

The density function $\phi(r)$ is obtained using an orthogonal series estimator $\hat{\phi}(r)$, where

$$\hat{\phi}(r) = \frac{1}{2} + \sum_{k=1}^{\infty} a_k p_k(r), \tag{12}$$

where $p_k(r)$ is the Legendre polynomial of order r. In terms of this estimate, the PP Index is given by

$$\hat{\mathcal{I}}(\boldsymbol{\alpha}) = \sum_{k=1}^{\infty} \frac{2}{2k+1} a_k^2. \tag{13}$$

Unbiased regression estimates for the coefficients $\{a_k\}_{k=1}^{\infty}$ are obtained by minimizing the least-squares error

$$\int_{-1}^{1} (\phi(r) - \hat{\phi}(r))^2 dr, \tag{14}$$

for which the corresponding solution is easily obtained as

$$\frac{2}{2k+1} a_k = \int_{-1}^{1} \phi(r) p_k(r) dr = E(p_k(r)), \tag{15}$$

where using the empirical expectation computed from the sample data, we have

$$E(p_k(r)) = \frac{1}{M}\sum_{l=1}^{M} p_k(r_l). \tag{16}$$

The required calculations can be performed very rapidly using the recurrence relationship

$$\begin{aligned} p_0(r) &= 1, \quad p_1(r) = r, \\ (k+1)p_{k+1}(r) &= (2k+1)rp_k(r) - kp_{k-1}(r) \text{ for } k \geq 1. \end{aligned} \tag{17}$$

The result (15) can be equivalently written in matrix notation as

$$\boldsymbol{a} = \boldsymbol{D}^{-1}\boldsymbol{e}, \tag{18}$$

where $\boldsymbol{a}_k = a_k$, $\boldsymbol{e}_k = E(p_k(r))$, and $\boldsymbol{D}$ is the diagonal matrix with entries $\boldsymbol{D}_{kk} = 2/(2k+1)$.

The series expansion for the terms in the estimator $\hat{\phi}(r)$ must be truncated to avoid overfitting (since the unbiased sample estimates of the higher-order coefficients will have a very large variance unless the sample size is also very large). Furthermore, the primary interest is often in finding directions in which the projected data have significant low-frequency departures from the Gaussian distribution, and these are primarily determined by the leading-order terms in (12) and (13) respectively (although the higher-order terms can sharpen the results). Also, the inclusion of the higher-order terms in the PP index (13) leads to proliferation of local maxima solutions, which mostly have only high-frequency departures from normality due to artifacts of sampling fluctuations. This proliferation of local maxima solutions also makes it difficult for the optimization procedure to identify the more robust and substantial local maxima solutions.

The approach taken by Friedman [12] (see also Sun [17]) is to truncate the series expansion to K terms, with K taking small values typically in the range 6–8. However, when the sample size is large enough, it is often useful to vary K starting from small values where the optimization procedure is relatively successful in identifying all the substantial local maxima solutions. The accuracy of these local maxima solutions can be subsequently refined by increasing K. However, the discrete nature of the K parameter is quite inconvenient for implementing this continuation procedure, and the same effect is obtained more conveniently by estimating $\hat{\phi}(r)$ from the following regularized least-squares problem,

$$\mathcal{L}(\boldsymbol{a}, \lambda) = \int_{-1}^{1} (\phi(r) - \hat{\phi}(r))^2 dr + \lambda \int_{-1}^{1} (\hat{\phi}''(r))^2 dr, \tag{19}$$

with $\lambda > 0$, where the second term is chosen to penalize the curvature semi-norm of the solution. The solution of (19) is obtained from

$$(\boldsymbol{D} + \lambda \boldsymbol{A})\boldsymbol{a} = \boldsymbol{e}, \tag{20}$$

The matrix entries $\boldsymbol{A}_{kl}$ can be obtained using standard manipulations, and these entries are zero for $k = 1$ or $l = 1$, or if $k + l$ is odd. The only nonzero entries occur for $k + l$ is even with $k \neq 1, l \neq 1$, when

$$\boldsymbol{A}_{kl} = \sum_{n=0}^{\min(k-2,l-2)} \frac{2}{2n+1}(n + \frac{1}{2})^2[k(k+1) - n(n+1)][l(l+1) - n(n+1)]. \tag{21}$$

The solution to (19) is obtained by truncating to K terms, and obtaining the $\{a_k\}_{k=1}^{K}$ from a Cholesky factorization.

The inclusion of the penalty term in (19) ensures that for large enough values of k, the coefficients a_k in the expansion (12) will be damped to zero. Therefore, in contrast to the earlier procedure where the choice of K was critical to the smoothing, in the present case K needs to only be chosen large enough so that for a given λ, all a_k with $k > K$ are negligibly small (typically we have taken K in the range 20–30, depending on λ). The advantage is that the parameter λ can now be varied continuously in order to unfold the multivariate structure at different levels of smoothing in the marginal density estimation. Note that, unlike traditional density estimation, no attempt is made here to find an "optimal" value of λ (such as by cross-validation or other procedures discussed for example in Silverman [16]).

2.3 Optimization Procedure for Projection Pursuit

The optimization problem

$$\arg\max_{\boldsymbol{\alpha}} \hat{\mathcal{I}}(\boldsymbol{\alpha}), \tag{22}$$

for the PP Index in (13) is carried out using the unconstrained optimization package of Chow, Eskow and Schnabel [4], which uses analytical gradient and Hessian information, and implements the second-order Gauss-Newton method with a line-search procedure to ensure convergence from arbitrary starting points (we have also used with equally successful results, an optional fourth-order tensor method in this package in which the higher-order terms are estimated by interpolating function and gradient evaluations from previous iterates). The second-order convergence rate of the Gauss-Newton procedure enables machine-precision converged results to be obtained with very little extra effort, and this simplifies the task of distinguishing identical local maxima that may be obtained from different starting points.

The components of the required analytical gradient and Hessian of $\hat{\mathcal{I}}(\boldsymbol{a})$ are given by

$$\begin{aligned}\frac{\partial \hat{\mathcal{I}}}{\partial \alpha_i} &= \sum_{k=1}^{K} \frac{4}{2k+1} a_k \frac{\partial a_k}{\partial \alpha_i}, &(23)\\ \frac{\partial^2 \hat{\mathcal{I}}}{\partial \alpha_i \partial \alpha_j} &= \sum_{k=1}^{K} \frac{4}{2k+1} \left(a_k \frac{\partial^2 a_k}{\partial \alpha_i \partial \alpha_j} + \frac{\partial a_k}{\partial \alpha_i} \frac{\partial a_k}{\partial \alpha_j} \right), &(24)\end{aligned}$$

where

$$\frac{\partial a_k}{\partial \alpha_i} = \frac{2k+1}{M} \sum_{l=1}^{M} p_k'(r_l) \Phi'(x_l) \frac{\partial x_l}{\partial \alpha_i}, \tag{25}$$

$$\frac{\partial^2 a_k}{\partial \alpha_i \alpha_j} = \frac{2k+1}{M} \sum_{l=1}^{M} \left[\left(2p_k''(r_l)(\Phi'(x_l))^2 + p_k'(r_l)\Phi''(x_l) \right) \frac{\partial x_l}{\partial \alpha_i} \frac{\partial x_l}{\partial \alpha_j} - p_k'(r_l)\Phi'(x_l) \frac{\partial^2 x_l}{\partial \alpha_i \partial \alpha_j} \right], \tag{26}$$

and

$$\frac{\partial x_l}{\partial \alpha_i} = \frac{(\boldsymbol{U}_i)_l}{\sqrt{\alpha^T \alpha}} - \frac{\alpha_i x_l}{\alpha^T \alpha}, \tag{27}$$

$$\frac{\partial^2 x_l}{\partial \alpha_i \partial \alpha_j} = -\frac{\delta_{ij} x_l}{\alpha^T \alpha} - \frac{(\alpha_j (\boldsymbol{U}_i)_l + \alpha_i (\boldsymbol{U}_j)_l)}{(\alpha^T \alpha)^{3/2}} + \frac{3\alpha_i \alpha_j x_l}{(\alpha^T \alpha)^2}. \tag{28}$$

Also note that

$$\Phi'(x_l) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_l^2}, \quad \Phi''(x_l) = -x_l \Phi'(x_l). \tag{29}$$

The quantities in (25) and (26) can be rapidly evaluated using the following recurrences for the derivatives of the Legendre polynomials

$$p_1'(r) = 1, \quad p_2'(r) = 3r, \quad p_k'(r) = r p_{k-1}'(r) - k p_{k-1}(r) \text{ for } k \geq 2, \tag{30}$$

$$p_1''(r) = 0, \quad p_k''(r) = p_{k-1}''(r) + (k+1) p_{k-1}'(r) \text{ for } k \geq 2. \tag{31}$$

The correctness and consistency of the evaluation and programming of the analytical form of the function, gradient and the Hessian is verified by comparing function evaluations with converged Taylor's series estimates (as suggested in Averbukh, Figueroa and Schlick [1]).

λ	"Pure Noise" PP indices		Threshold PP index	Significant PP indices from FB Dataset
	Mean	Std. Dev.		
0.01	0.00472	0.00196	0.0093	$0.0171^{(a)}, 0.0114^{(b)}$
0.001	0.0178	0.00613	0.0321	$0.0913^{(b)}, 0.0599^{(a)}, 0.0442^{(c)}$
0.0001	0.0355	0.0139	0.0680	$0.181^{(b)}, 0.156^{(a)}, 0.0844^{(c)}$
0.00001	0.0563	0.0172	0.0965	$0.227^{(a)}, 0.226^{(b)}, 0.105^{(c)}, 0.102^{(d)}$

Table 1: Significant PP Index values computed for different values of λ for the FB dataset. The superscripts $(a), (b), (c), (d)$ identify the corresponding Index values with the same projection directions for different λ.

PP Index	Projection directions
$0.227(a)$	$(+0.813, +0.547, -0.009, +0.185, +0.081, -0.001)$
$0.226(b)$	$(+0.904, -0.157, +0.203, -0.137, -0.270, -0.160)$
$0.105(c)$	$(+0.058, -0.878, +0.220, -0.288, -0.250, -0.181)$
$0.102(d)$	$(+0.145, -0.675, +0.114, -0.587, -0.402, -0.073)$

Table 2: PP solutions (in principal component coordinates) corresponding to the case $\lambda = 0.00001$ in Table 1.

PP Index	Projection directions
$0.227(a)$	$(-0.595, +0.243, +0.010, +0.576, +0.000, +0.505)$
$0.226(b)$	$(-0.897, +0.099, +0.203, -0.092, -0.319, +0.183)$
$0.105(c)$	$(-0.331, -0.202, +0.074, -0.734, -0.300, -0.464)$
$0.102(d)$	$(-0.314, -0.194, +0.256, -0.847, -0.266, -0.095)$

Table 3: PP solutions (in original variable coordinates) corresponding to the case $\lambda = 0.00001$ in Table 1.

3 Examples

In this section, we describe the use of PP on two well-known datasets from the literature. Our description of the analysis of these datasets is not a detailed one, since our goal is to only amplify on the applicability and usage of the PP methodology.

The PP solutions were computed using about 100 starting guesses for moderate λ, and increasing to about 500 starting guesses for smaller λ. The starting guesses are chosen randomly, or based on previous computations with different λ. The distinct local maxima computed in this way are ranked in terms of the magnitude of the corresponding PP indices. Typically, for moderate λ only a few distinct local maxima solutions are obtained (irrespective of how many starting guesses are used), and the number of distinct

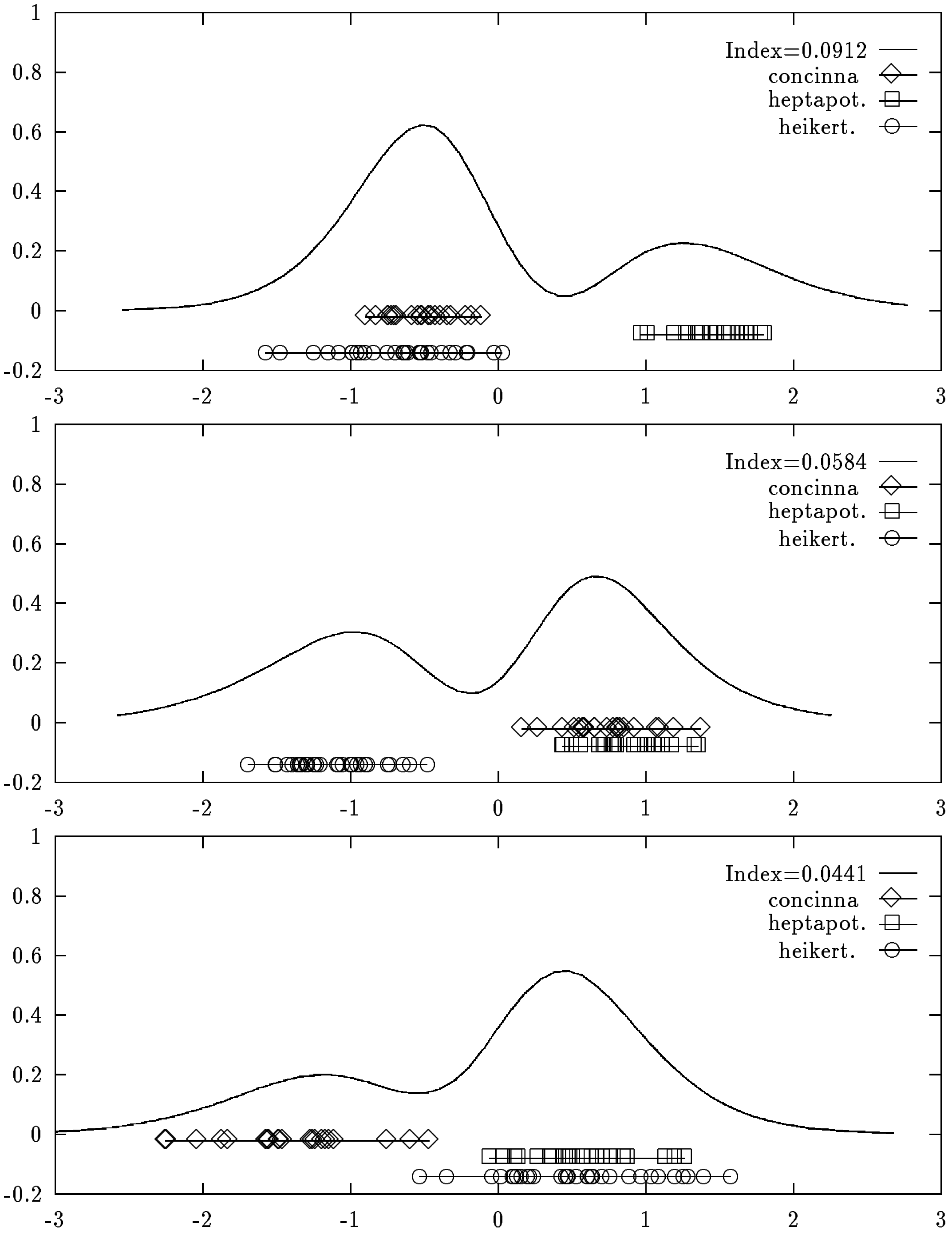

Figure 1: Projection of the labelled data samples on the PP solutions in Table 1, along with marginal density estimates for $\lambda = 0.001$

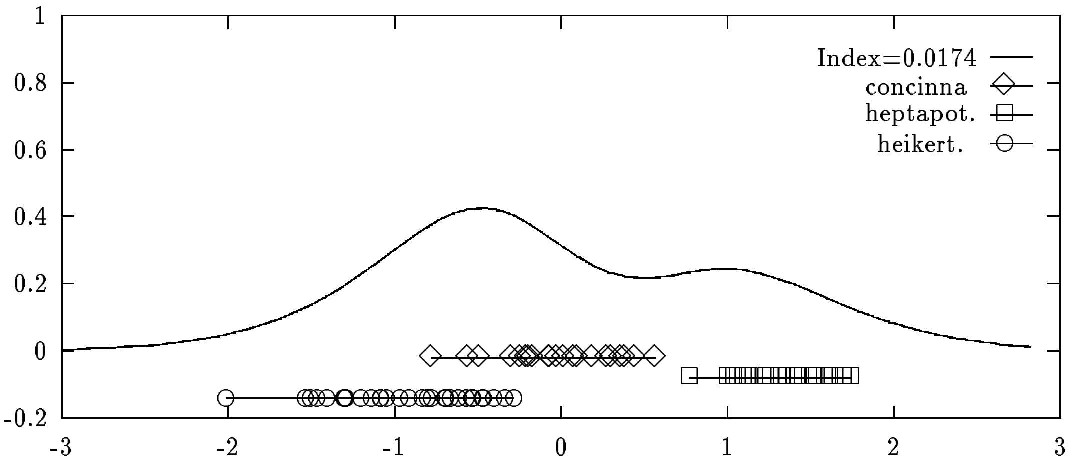

Figure 2: Projection of the labelled data samples of the FB dataset on the first principal component, along with marginal density estimates for $\lambda = 0.001$

solutions increases rapidly with decreasing λ. As a result, with small λ the substantial local maxima can even be missed unless a continuation procedure from larger values of λ is used to generate suitable starting guesses.

In order to generate a “null hypothesis” test for the PP results, 5-10 “pure noise” datasets of the same size and dimension were generated by sampling from a multivariate Gaussian distribution with zero mean and identity covariance matrix. For a given λ, these datasets were used to obtain at least 100-200 PP solutions and corresponding index values. These index values may be used to obtain empirical P values for the PP solutions from the real data set. Since these empirical P values attempt to estimate the tail probability of the distribution of PP indices from the pure noise data set, it is clear that a very large sample size will be required to obtain accurate estimates (making this overall procedure quite computationally intensive). One approach to reduce the computational cost is to fit a parametric or non-parametric model to the sample data, and estimate P values from this model. For example, as done below, the sample data can be used to estimate the parameters of a normal distribution when the examination of the data suggests that this is appropriate. A one-sided hypothesis test at the 1% significance level was then used to reject the null hypothesis that a computed PP index (from the real data set) was produced by a pure noise data set.

3.1 Flea-Beetle (FB) Data set

This data set is from Lubischew [15] and has been studied using PP by Jones and Sibson [11] among others. It consists of 6 scaled biometric measurements (width of first joint of first tarsus, width of second joint of first tarsus, maximal width of the aedeagus in the fore-part, front angle of aedeagus, maximal width of head between external edges of eyes, and aedeagus width from side) on 74 flea-beetles from 3 species (Ch. concinna, Ch. heptapotamica, and Ch. heikertingeri). It is of interest to see if PP on the unlabelled data can obtain good projections for separating the three species.

Table (1) shows the statistics for the "pure noise" data sets, the threshold index values from the significance test at the 1% level, and the PP solutions for the FB data set that passed the significance test. The solutions labelled (a),(b) and (c) in this table appear at fairly moderate values of λ, and the projected data along with the smoothed marginals are shown in Figure (1). The projections (a) and (b) cleanly isolate 2 of the 3 species, and these 2 projections taken together are sufficient to identify and separate all 3 species. Even though the marginal density plots in these 2 projections are oversmoothed, there is a clear bimodal structure with very little overlap in the tails of the two modes. The third projection attempts to isolate the concinna species, but there is now some overlap in the tails of the smoothed bimodal marginal (as well as in the projected labelled data).

Table (2) gives the projection directions in principal component coordinates corresponding to the solutions in Figure (1). The projection directions (a) and (b) have a large loading on the first principal component. While the marginal along this principal component in Figure (2) shows some bimodal structure, it does not provide a clear separation of the modes seen in the PP solutions in Figure (1). Table 3 shows the results as Table (2) reported in original variable coordinates. The solution (b) which separates Ch. heikertingeri from the other 2 species has a large loading on the the first variable, while the solutions (a) and (c) are non-trivial combinations of the original variables.

Table (1) shows that as λ is varied, the relative significance of different PP solutions can change, and new PP solutions can appear above the noise level. This illustrates our point that it is often desirable to use a wide range of values of λ in order to obtain new insights and identify new structure in the data.

3.2 Olive Oil (OO) Data set

This data set contains the measurements of the percentage composition of 8 different

Variable	Principal Components							
	I	II	III	IV	V	VI	VII	VIII
1	0.281	-0.637	-0.446	-0.079	0.452	-0.129	0.091	0.283
2	0.093	-0.097	-0.163	-0.512	-0.715	-0.313	-0.005	0.291
3	-0.011	-0.014	0.731	-0.427	0.366	-0.233	0.043	0.306
4	-0.842	0.173	-0.333	-0.112	0.192	-0.124	0.076	0.282
5	0.450	0.742	-0.301	-0.090	0.242	-0.089	0.071	0.276
6	0.004	-0.034	0.084	0.302	-0.087	0.246	-0.682	0.605
7	0.013	-0.007	0.140	0.653	-0.132	-0.651	0.268	0.201
8	0.011	-0.043	0.113	0.104	-0.174	0.569	0.665	0.422
% Variance	89.72	98.58	99.38	99.68	99.91	99.96	99.98	100.00

Table 4: Principal Components for OO Data set, with the last row giving the cumulative percent variance due to the leading principal components.

λ	"Pure Noise" PP indices		Threshold	Significant PP indices
	Mean	Std. Dev.	PP index	from OO Data set
10^{-4}	0.00278	0.00117	0.0055	$0.134^{(a)}, 0.130^{(b)}, 0.127, 0.124, 0.112$
10^{-5}	0.0051	0.00173	0.0092	$0.270^{(a)}, 0.232, 0.194^{(b)}, 0.177, 0.143$

Table 5: Significant PP Index values computed for different values of λ for the OO data set. The superscripts $(a), (b)$ identify the corresponding Index values with the same projection directions for different λ.

λ	Projection directions
10^{-4}	$(+0.586, -0.514, +0.200, +0.113, -0.308, +0.412, +0.202, +0.186)$
10^{-5}	$(+0.547, -0.514, +0.238, +0.134, -0.306, +0.417, +0.252, +0.176)$

Table 6: The PP solution (in principal component coordinates) corresponding to the case denoted by (a) in Table 5.

λ	Projection directions
10^{-4}	$(+0.273, +0.158, -0.044, -0.704, -0.234, +0.174, -0.022, +0.564)$
10^{-5}	$(+0.246, +0.131, -0.027, -0.685, -0.264, +0.145, +0.004, +0.602)$

Table 7: PP solutions (in original variable coordinates) corresponding to the case denoted by (a) in Table (5).

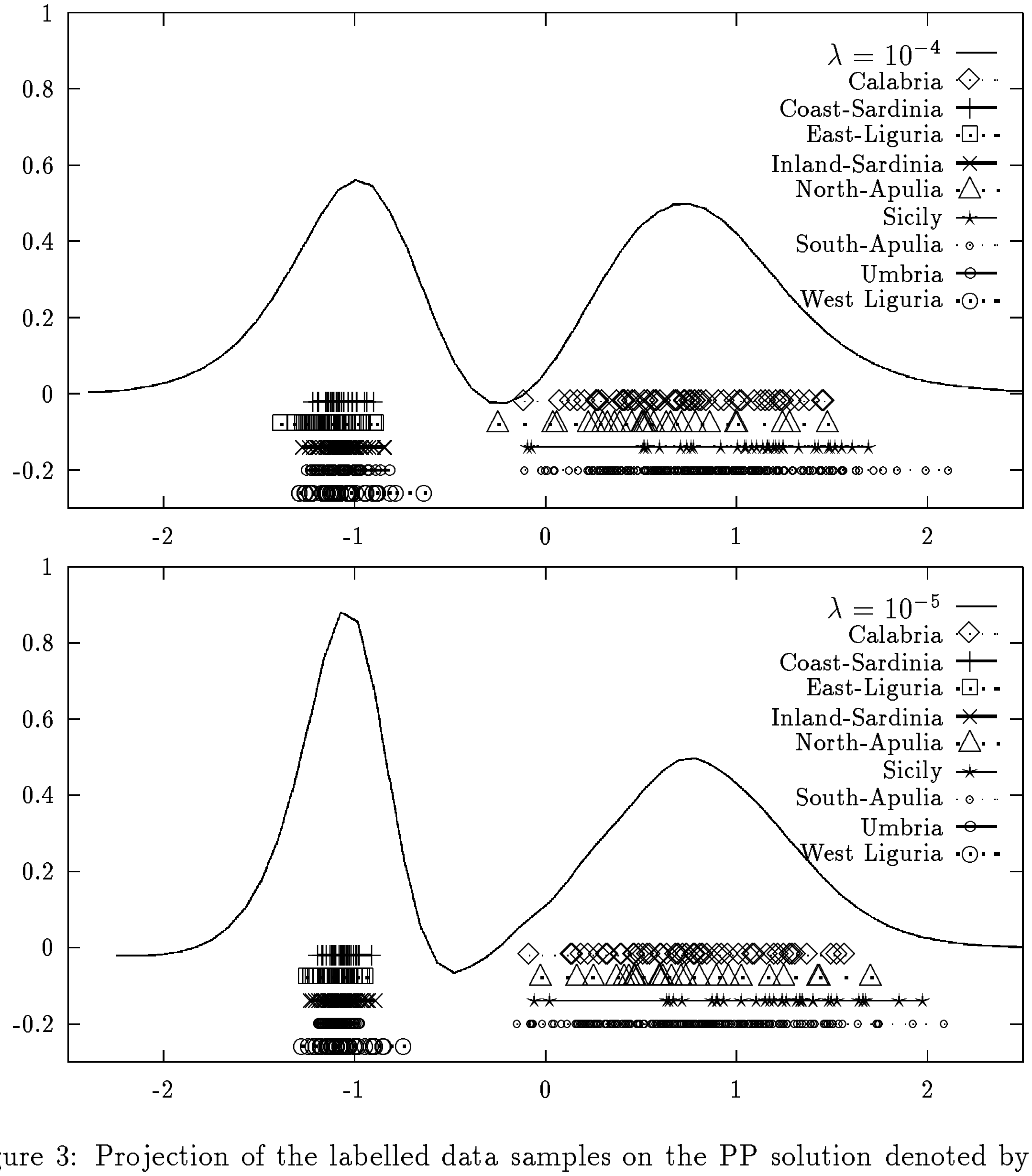

Figure 3: Projection of the labelled data samples on the PP solution denoted by (a) in Table (5) along with marginal density estimates for $\lambda = 0.0001$ and $\lambda = 0.00001$.

fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, eicosenoic) in the lipid fraction of 572 Italian olive oils from 9 different collection areas. These areas are North and South Apulia, Calabria and Sicily in Southern Italy, Coastal and Inland Sardinia, and Umbria, East and West Liguria in Northern Italy. Andreas Buja (personal communication) has referred to other previous analyes of this data using several different statistical methods in Forina et al. [6], and using PP in Glover and Hopke [13]. Again, it is of interest to see if PP can lead to the good projections for distinguishing between the labelled data for the oils from different regions and collection areas.

The iterative outlier removal scheme described in Section 2.1 with $p = 0.01$ led to 4 outliers being trimmed, and with $p = 0.1$ led to 19 outliers being trimmed (in both cases the iterative procedure stabilized after 3 iterations). Since the subsequent results with both these trimmed data sets are quite similar, we focus only on that obtained with $p = 0.01$ here. The principal components for the trimmed data set in Table (4) show the primary variation in the data is due to the oleic, linoleic and palmitic acid compositions (in that order).

Table (5) shows the statistics for the pure noise data sets, and the index values for the 5 top-ranked PP solutions that passed the significance test (these index values for the latter are well above those from the pure noise data sets).

The most distinctive PP solution is the one identified by (a) in Table (5), which isolates the oils from the 4 Southern Italian regions. This projection is given in Table (6) and (7), and the separation of the clusters in shown in Figure (3), where it can be seen that the adjustments from $\lambda = 10^{-4}$ to $\lambda = 10^{-5}$ leads to sharper result with a better separation of the 2 clusters. This projection direction is seen to have a large loading from variables 4 and 9 (oleic and eicosenoic acids respectively). It is interesting that variable 9 has a significant loading only on the last 3 principal components in in Table (4), which collectively account for less than 0.1% of the variation in the data. We have verified that in this case, a PP analysis in the subspace of the dominant 6 principal components can lead to the distinctive PP solution described above being completely missed.

The sharp separation between the 2 clusters in (a) suggests isolating the corresponding data samples and repeating the PP procedure individually might uncover further substructure in the data. This aspect, however, has not been pursued further here.

4 Summary Remarks

The two examples studied in Section (3) show that PP is able to uncover some of the structure due to the known class labels in the data. In our view, irrespective of whether the data is labelled or unlabelled, the use of PP nicely complements existing statistical methods such as principal components (what if the expected structure is not found in the plots of the leading principal components?) and discriminant analysis (what is the intrinsic structure of the data that does not require a knowledge of the class labels?).

The PP procedure is computationally intensive because of the difficulties with numerous suboptimal local maxima in the optimization procedure, and because of the extensive simulations required to establish the statistical significance of the results.

The inability of the PP procedure to find useful structure in a data set does not preclude the existence of such structure (Diaconis and Freedman [5] have shown that most low-dimensional projections of many structured multivariate data sets are approximately Gaussian). In addition, even when PP can be successful, the most useful projections may not be obtained from the starting guesses selected for the optimization procedure. In other cases, it may be necessary to look for useful PP solutions in higher-dimensional projections (say 2-D or 3-D) rather than the 1-D case considered in this paper.

Our final remark concerns the importance of interactive graphics for visualizing marginal scatterplots, histograms and smoothed density estimates. For unlabelled data especially, it is valuable to be able to look at the marginals of several PP solutions simultaneously, and be able to highlight and link the individual sample data points in each one of them, for a better understanding of the data.

References

[1] V. Averbukh, S. Figueroa and T. Schlick, *ALGORITHM 566: HESFCN - A FORTRAN Package of Hessian Subroutines for Testing Nonlinear Optimization Software*, ACM Transactions on Mathematical Software, 20 (1994), pp. 282-285.

[2] N. A. Campbell, *Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation*, Appl. Statist. (1980), pp. 231-237.

[3] D. Cook, A. Buja and J. Cabrera, *Projection Pursuit Indexes Based on Orthonormal Function Expansions* J. Computational and Graphical Statistics, 2 (1993), pp. 225-250.

[4] T. Chow, E. Eskow and R. Schnabel, *ALGORITHM 739: A Software Package for Unconstrained Optimization Using Tensor Methods*, ACM Transactions on Mathematical Software, 20 (1994), pp. 518-530.

[5] P. Diaconis and D. Freedman, *Asymptotics of Graphical Projection Pursuit*, Annals of Statistics, 12 (1984) pp.793-815.

[6] Forina, M. and Armanino, C. and Lanteri, S. and Tiscornia, E. *Classification of olive oils from their fatty acid composition*, (eds. H. Martens and H. Russwurm Jr.), Food Research and Data Analysis, Applied Science Publishers, London, (1983), pp. 189-214.

[7] G. H. Golub and C. F. van Loan, *Matrix Computations*, Second Edition, Johns Hopkins University Press, Baltimore (1989).

[8] R. Gnanadesikan, *Methods for Statistical Data Analysis*, Wiley (1977).

[9] P. J. Huber, *Projection Pursuit (With Discussion)*, Annals of Statistics, 9 (1985), pp. 435-475.

[10] I. T. Jolliffe, *Principal Components Analysis*, Springer (1986).

[11] M. C. Jones and R. Sibson, *What is projection pursuit?*, J. Royal Stat. Soc., 150 (1987), pp. 1-37.

[12] J. H. Friedman, *Exploratory projection pursuit*, J. Amer. Stat. Assoc., 820 (1987), pp. 249-266.

[13] D. M. Glover and P. K. Hopke, *Exploration of Multivariate Chemical Data by Projection Pursuit*, Chemometrics and Intelligent Laboratory Systems, Elsevier Publishers, 16 (1992), pp. 45-59.

[14] P. Hall, *On Polynomial-Based Projection Indices for Exploratory Projection Pursuit*, Annals of Statistics, 17 (1989), pp. 589-605.

[15] A. A. Lubischew, *On the Use of Discriminant Functions in Taxonomy*, Biometrics, 18 (1962) pp. 455-477.

[16] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London (1986).

[17] J. Sun, *Significance Levels in Exploratory Projection Pursuit*, Biometrika, 78 (1991), pp. 759-769.

[18] J. Sun, *Some Practical Aspects of Exploratory Projection Pursuit*, SIAM J. Sci. Comput., 14 (1993), pp. 68-80.

[19] D. F. Swayne, D. Cook, and A. Buja, *User's Manual for XGobi, a Dynamic Graphics Program for Data Analysis Implemented in the X Window System*, Bellcore Technical Memorandum (1991).