Novel Projection Pursuit Indices for Feature Extraction and Classification: An Inter-comparison in a Remote Sensing Application

Charles M. Bachmann *

Naval Research Laboratory, Remote Sensing Division
Remote Sensing Hydrodynamics Branch, Code 7255
4555 Overlook Ave. SW, Washington, D. C. 20375
email: bachmann@nrl.navy.mil

May 10, 1997

Abstract

Projection Pursuit [7] [10] (PP) techniques are used to search for statistically interesting low-dimensional projections of complex, high-dimensional data. These projections reveal data structure useful for automatic classification applications. We derive a novel class of Projection Pursuit algorithms, comparing them with related PP algorithms [7] [8] [11] [2] [1]. Texture-based cloud detection in Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) imagery from the Jet Propulsion Laboratory is provided as a basis for inter-comparison.

1 Projection Pursuit: Background

In order to identify potentially meaningful data structures in high-dimensional data sets, "Projection Pursuit" [7] (PP) techniques are used to search for statistically interesting low-dimensional projections. PP is an iterative search technique that converges to extrema of a projection index or cost function,

*This work was supported in part by a grant of High Performance Computing (HPC) time from the following Department of Defense HPC Centers: Army Research Laboratory SGI Power Challenge Array, the Maui High Performance Computing Center, and the Naval Research Laboratory SGI Origin 2000.

that measures the degree of multi-modality or departure from normality of the projected data distribution. One of the first PP methods was proposed by Friedman and Tukey [7], who coined the term “Projection Pursuit.” Their cost function was the product of two functions, one that measures the spread of the projected data (a trimmed variance to ensure insensitivity to outliers), and another that measures compactness within a particular distance scale. Historically, the Friedman-Tukey cost function can be seen as an innovative step beyond Principal Component Analysis (PCA). As shown in Figure 1, using PCA to find maximal data variance of *all data samples* is not necessarily the most informative for classification. This Figure also emphasizes the fact that an orthonormal decomposition does not always reveal the most relevant information from the perspective of class separation.

Figure 1: (Left) PP vs. PCA: PCA corresponds to a special case of PP in which the Projection Index is maximal variance (power), which will not always reveal clusters in the data. More sophisticated Projection Indices can be defined to favor the discovery of multi-modal projected distributions. (Right) Three classes of data located in distinct clusters optimally separated by hyperplanes (dashed lines) perpendicular to the projection vectors; hyperplanes correspond to scalar thresholds of projected data; note that optimal PP vectors are not orthogonal.

For a unit projection vector $\hat{w}_k$, input data sample $\vec{f}_i$, and projection:

$$c_k(i) = \hat{w}_k \cdot \vec{f}_i, \quad (1)$$

the Friedman and Tukey PP algorithm is:

$$\text{Maximize : } I(c_k) = S(c_k)N(c_k) \quad (2)$$

where:

$$S(c_k) = \sqrt{\frac{\sum_{i=1}^{N_p-n} (c_k(i) - E(c_k))^2}{N_p - n}} \quad (3)$$

$$N(c_k) = \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} g(r_k(i, j)) \theta(R - r_k(i, j)) \quad (4)$$

$$\text{with } r_k(i, j) = |c_k(i) - c_k(j)|, \quad (5)$$

$$g(r_k(i, j)) : \text{Monotone Decreasing, \& } \theta : \text{a step function} \quad (6)$$

N_p is the total number of patterns, n is a small fraction of outliers removed at both extremes of the projection, $E(c_k)$ is the mean projection value, and R is a scalar cut-off outside of which pairs of points are excluded in the compactness function, $N(c_k)$. The resulting search algorithm favors the discovery of multi-modal structure in the data. In the absence of the factor, $N(c_k)$, the original PP index would reduce to $S(c_k)$ and would be equivalent to PCA. In [7], the search procedure consisted of optimizing 1D or 2D projections of the data, one at a time. A number of other PP methods were later developed. A fundamental concept in these later algorithms is to find projections that are least normal [8] [10]. Typically, the projection vectors in these methods are optimized serially with each subsequent search vector optimized on a residual after subtraction of structure from the previous projection. In this sense, they differ from unsupervised neural network learning algorithms that also implement a form of PP but do so by jointly optimizing a set of projections (see e.g. [11]). No proof exists to suggest whether serial or joint optimization is superior.

2 A Novel Class of Projection Pursuit Indices

In revisiting the original PP Index designed by Friedman and Tukey, one can see a limitation: the factor measuring the degree of data spread, $S(c_k)$, does not directly focus on the *spread between clusters*, but rather measures the *spread of the whole data set*. One approach to circumventing this difficulty is to define an entirely new projection index that focuses directly on the degree of departure from normality (see for e.g. [8]). Alternatively, our approach is to replace $S(c_k)$ with a function, $D(\tilde{c}_k)$, that directly measures the spread outside a clustering/nearness scale α_k . The symbol $\tilde{c}_k$ refers to the nonlinear data projection that replace the linear projections c_k defined in Equation 1:

$$\tilde{c}_i^{(n)} = \sigma(\sum_j L_{ij}^{(n)} c_j^{(n)}), \quad (7)$$

$$\text{where } \sigma(x) = a \tanh(a\lambda x), \text{ } (a, \lambda \text{ constant}), \quad (8)$$

$$c_j^{(n)} = \vec{w}_j^{(n-1)} \cdot \vec{c}^{(n-1)} + b_j^{(n)}, \quad (9)$$

and where $\vec{w}_j^{(n-1)}$ is the j th modifiable projection vector, that weights the inputs $\vec{c}^{(n-1)}$ from layer $n-1$, and $b_j^{(n)}$ is the bias. The coupling/constraint matrix $L_{ij}^{(n)}$ is either fixed or modifiable (see the next Section). The nonlinear projections are no longer constrained to be on the unit sphere, and, importantly, are expressed as saturating nonlinearities that remove sensitivity to extreme outliers in the data. In fact, this latter property allows us to retain data points originally ignored by the Friedman-Tukey Index (see (3) and (4)).

We optimize projections jointly as in PP neural network approaches. Our new PP Index is:

$$\Xi = \frac{1}{K} \sum_{k=1}^K \Xi_k = \frac{1}{K} \sum_{k=1}^K N^{cont.}(\tilde{c}_k) D^{cont.}(\tilde{c}_k) \quad (10)$$

where $N^{cont.}$ and $D^{cont.}$ are given by:¹

$$N^{cont.}(\tilde{c}_k) = E_{pairs,(\mu,\nu)}[g(\tilde{r}_k(\mu,\nu))] \quad (11)$$

$$D^{cont.}(\tilde{c}_k) = E_{pairs,(\mu,\nu)}[\tilde{r}_k^2(\mu,\nu)(1 - g(\tilde{r}_k(\mu,\nu)))] \quad (12)$$

$$\text{with } \tilde{r}_k^2(l,m) = (\tilde{c}_k(l) - \tilde{c}_k(m))^2, \text{ \& } g(\tilde{r}_k(\mu,\nu)) = e^{-\left(\frac{\tilde{r}_k(\mu,\nu)}{\alpha_k}\right)^2} \quad (13)$$

We have chosen a particular form for the nearness function $g(\tilde{r}_k(\mu,\nu))$ in order to derive specific expressions for a particular case of Equation 10. In our novel PP algorithm, each nearness function $g(\tilde{r}_k(\mu,\nu))$ has a clustering scale factor α_k associated with it. Each α_k is obtained by multiplying an initial estimate of the standard deviation of the projected data by a random number drawn from a user-determined range. The α_k can be modifiable, although for the results in this paper, they were static. Selecting a range of α_k is useful because clusters and other structure may be visible on more than one scale in the data depending on how the high-dimensional data is viewed, i.e. depending on the orientation of the PP search vector in the high-dimensional data space.

In $D^{cont.}(\tilde{c}_k)$, we use the squared distance weighted by the factor $(1 - g(\tilde{r}_k(\mu,\nu)))$, since this product will directly measure inter-point spread of projected data, weighting more heavily those distances outside the nearness/clustering scale α_k . Thus, it will be maximal for well-separated clusters existing within the scale α_k . If the projected distribution is multi-modal, $D^{cont.}(\tilde{c}_k)$ will measure the spread of the modes better than $S(c_k)$, which

¹Here $E_{pairs,(\mu,\nu)}$ means expected value over sample pairs.

simply measures overall data spread by calculating the variance about the mean of the projected distribution.

Minimization of our PP Index in (10) by gradient descent leads to the following modification equation for the i th projection vector $\bar{w}_i^{(n-1)}$:

$$\begin{aligned}\Delta w_{ij}^{(n-1)} &= -\eta(t) \frac{\partial \Xi^{(n)}}{\partial w_{ij}^{(n-1)}} \\ &= -\eta(t) \frac{1}{K} \sum_{k=1}^K [N^{cont.}(\bar{c}_k^{(n)}) \frac{\partial D^{cont.}(\bar{c}_k^{(n)})}{\partial w_{ij}^{(n-1)}} + \\ &\quad \frac{\partial N^{cont.}(\bar{c}_k^{(n)})}{\partial w_{ij}^{(n-1)}} D^{cont.}(\bar{c}_k^{(n)})] \end{aligned} \quad (14)$$

where:

$$\frac{\partial N^{cont.}(\bar{c}_k^{(n)})}{\partial w_{ij}^{(n-1)}} = -E_{pairs,(\mu,\nu)} \left[\frac{\psi_{kij}}{\alpha_k^2} g(\bar{r}_k(\mu, \nu)) \right] \quad (15)$$

$$\begin{aligned} \frac{\partial D^{cont.}(\bar{c}_k^{(n)})}{\partial w_{ij}^{(n-1)}} &= E_{pairs,(\mu,\nu)} \left[\psi_{kij}(\mu, \nu) \left(\left(\frac{\bar{r}_k(\mu, \nu)}{\alpha_k} \right)^2 g(\bar{r}_k(\mu, \nu)) + \right. \right. \\ &\quad \left. \left. (1 - g(\bar{r}_k(\mu, \nu))) \right) \right] \end{aligned} \quad (16)$$

$$\begin{aligned} \psi_{kij}(\mu, \nu) &= \frac{1}{2} \frac{\partial \bar{r}_k^2(\mu, \nu)}{\partial w_{ij}^{(n-1)}} \\ &= (\bar{c}_k^{(n)}(\mu) - \bar{c}_k^{(n)}(\nu)) \cdot \\ &\quad ((\bar{c}_k^{(n)}(\mu))' \bar{c}_j^{(n-1)}(\mu) - (\bar{c}_k^{(n)}(\nu))' \bar{c}_j^{(n-1)}(\nu)) L_{ki} \end{aligned} \quad (17)$$

$$(\bar{c}_i^{(n)})' = \lambda(a - \bar{c}_i^{(n)})(a + \bar{c}_i^{(n)}). \quad (18)$$

2.1 Implementation Details

Because of the pairwise calculations inherent in the quantities, D , N , $\frac{\partial D}{\partial w_{ij}}$, and $\frac{\partial D}{\partial w_{ij}}$, memory and computational requirements in a storage intensive version of the algorithm would be potentially quite severe, growing quadratically with the number of sample estimates used at each update. This potential problem is circumvented by an on-line implementation that uses stochastic gradient descent, with estimates of D , N , $\frac{\partial D}{\partial w_{ij}}$, and $\frac{\partial D}{\partial w_{ij}}$ computed by a running average. For example:

$$N^{cont.}(\bar{c}_k)(t) = \epsilon N^{cont.}(\bar{c}_k)(t-1) + (1 - \epsilon) g(\bar{r}_k(\bar{c}_k(t), \bar{c}_k(t-1))) \quad (19)$$

By choosing a small and decreasing learning rate, $\eta(t)$, and an appropriate value of ϵ , the distribution of sample pairs can be estimated sufficiently to ensure gradient descent. Typically, we let the learning rate decrease as $\eta(t) = \eta_0 / (1 + \ln(\frac{t}{\tau_\eta}))$. A theoretical justification for choosing a logarithmically decreasing learning rate can be made using arguments from simulated annealing [9].

Another detail of the implementation is that the coupling matrix $L_{ij}^{(n)}$ may be fixed or modifiable. In the latter case, $L_{ij}^{(n)}$ is modified by gradient ascent to maximize the relative entropy of the projections, $\tilde{c}$:²

$$\text{let } L_{ij}^{(n)}(0) = \begin{pmatrix} 1 - \mu, \text{ for } i = j \\ -\mu, \text{ for } i \neq j \end{pmatrix} \quad (20)$$

$$\text{and } \zeta = -\frac{1}{N^2} \sum_{k,l} \tilde{q}_k \ln(\frac{\tilde{q}_l}{\tilde{q}_k}), \text{ with } \tilde{q}_k = \frac{1}{2a} (\tilde{c}_k^{(n)} + a), \quad (21)$$

$$\Delta L_{ij}^{(n)} = \eta_L \frac{\partial \zeta}{\partial L_{ij}^{(n)}} \quad (22)$$

$$= \eta_L \frac{1}{N} c_j^{(n)} [(\tilde{c}^{(n)})'_i (1 + \ln(a + \tilde{c}_i^{(n)}) - E_p(\ln(a + \tilde{c}^{(n)}))) - (a - \tilde{c}_i^{(n)}) E_p(a + \tilde{c}_i^{(n)})]. \quad (23)$$

If the scale of the initial coupling matrix, set by μ , is sufficiently larger than the scale used to randomly initialize the weights w_{ij} , and the learning rates for L and $\vec{w}$ are chosen appropriately, then the change in the effective projection vectors $\delta \vec{w}_{eff} = \delta(L \cdot \vec{w}) = L \delta \vec{w} + (\delta L) \vec{w} = L \eta_{\vec{w}} \nabla_{\vec{w}} \Xi + \eta_L (\nabla_L \zeta) \vec{w}$, will be dominated by the PP gradient term with adjustments from the second term dependent on $\nabla_L \zeta$. Our experiments to date suggest that using the modifiable coupling matrix rather than the fixed constraint may accelerate the maximization of the PP Index Ξ . At present we use both forms in experiments, although results reported here were obtained with the modifiable constraint matrix.

2.2 Results for a Remote Sensing Application

To investigate the potential usefulness of PP techniques for the extraction of textural features in future Multi-Angle Imaging Spectro-Radiometer (MISR) [6] data, we have analyzed 17 high resolution images from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [13] operated by the Jet Propulsion Laboratory. We use the four (out of 224) AVIRIS channels centered at the MISR wavelengths of 443, 555, 670, and 865 nm [6]. An example of

² $E_p()$ refers to the expected value across projections in the network.

Figure 2: Two-dimensional histogram of data projected onto a pair of projections in our new PP network after training; each axis is the degree of overlap with one of the selected projections. Inputs were gray-level difference vector (GLDV) histograms [5] [15] derived from AVIRIS band 5; input windows were 12x12 pixels.

multi-modal structure found in the AVIRIS data by our new PP algorithm is shown in Figure 2. A typical end result for one of the novel AVIRIS test images using related PP techniques [1] [2] from our earlier research along with standard texture features [14] is illustrated in Figure 3. As Figure 3 demonstrates, the majority of incorrect identifications are at the edges of the clouds where the features are a mixture of cloudy and cloud-free imagery. In Figure 4, results obtained with our new PP algorithm combined with cross-entropy based backward propagation (BPCE) [12] as a back-end classifier are compared against these other approaches. For comparable architecture sizes, the new PP algorithm with BPCE appears to be better than the other PP algorithms individually combined with BPCE. Also, the new PP algorithm with BPCE is statistically closest to the “all” cases that achieved the best performance by combining features derived by all of the other PP algorithms along with standard statistical features [14]. Mean performance on the novel held-out image for the best “all” case is $(93.5 \pm 6.8) \%$ cloud pixels detected with false alarm rate of $(10.6 \pm 10.0) \%$. The size of the errors bars in Figure 4 is somewhat large due to several factors: the limited number of trials (eight - ten per paradigm), the fact that there is a high degree of variability within even this limited database, and the complexity of the search space, that is full of local minima. Also, this database is significantly smaller than is necessary to completely characterize class variability and adequately represent the cloud detection problem domain in the training set. A larger statistical sample of experiments and further tuning of the individual PP algorithms is needed to obtain a better estimate of the relative merits of the approaches,

Figure 3: Cloud detection in a novel AVIRIS test scene. (Upper left) Near infra-red band (865nm) for an AVIRIS scene in the test set; band is one of 4 chosen to correspond spectrally to future MISR bands; colormap is black (low) to white (high). (Upper right) Human interpretation of cloud pixel location, white = cloud, black = no cloud. (Lower left) Cloud detection result from an ensemble network using extracted Wavelet Projection Pursuit (WPP) features [2], BCM-PP [4] [11] [1] features from GLDV histograms, BCM-PP features from simply normalized pixel intensities [1], and standard standard statistical moments [14] from GLDV; features in the ensemble model were extracted from all four spectral channels; white = cloud, black = no cloud. (Lower right) difference mask: black = no error, gray = false-alarm, white = false negative; most errors are on cloud/no-cloud boundaries. Cloud detection rate for this novel test image was 94.0 % with false alarm rate 4.8 % .

but these first results are encouraging and suggest that features from the new PP algorithm might be more robust than those found by the other PP algorithms and should be added to the “all” case to look for further improvement in future experiments.

References

- [1] C. M. Bachmann, E. E. Clothiaux, J. W. Moore, K. J. Andreano, “An Ensemble Approach to Automatic Cloud Identification in AVIRIS Imagery Using BCM Projection Pursuit,” in *Neural Networks for Signal Processing IV: Proceedings of the 1994 IEEE Workshop*, Sept 6-8, 1994, Ermioni, Greece, pp. 394-403.

Figure 4: An inter-comparison of cloud detection results over multiple trials: $F_{merit} = \sqrt{\epsilon_{false\ alarm}^2 + \epsilon_{false\ negative}^2}$ for the novel held out image vs. the cross validation set; $\epsilon_{false\ alarm}$ is the false alarm rate, $\epsilon_{false\ negative} = 1 - \rho_{detect}$ is the false negative rate, and ρ_{detect} is the rate of detection. F_{merit} (unbiased error) is the distance of network performance from an ideal detector on a receiver operating curve (ROC). Results for the novel PP algorithm combined with BPCE as back-end classifier achieve lower error than the other PP algorithms with BPCE; the new algorithm is statistically closest to the ensemble “all” cases that combine features from all of the other PP algorithms and achieve the best performance. “All” paradigms correspond to BPCE networks receiving input features from Wavelet Projection Pursuit (WPP)[2] projections of albedos, BCM-PP [4] [11] [1] projections of albedos, BCM-PP applied to GLDV histograms [1], and standard statistical features [14] from GLDV. “All” cases showed statistically significant improvement over any other combination of features with average held-out image performance for the best: $(93.5 \pm 6.8) \%$ cloud pixels detected with false alarm rate $(10.6 \pm 10.0) \%$.

- [2] C. M. Bachmann, E. E. Clothiaux, D. Q. Luong, "Wavelet Projection Pursuit for Feature Extraction and Cloud Detection in AVIRIS and AVHRR Imagery," in *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS96)*, Lincoln, Nebraska, May 27-31, 1996, Vol. I, pp. 356-359.
- [3] C. M. Bachmann, E. E. Clothiaux, D. Q. Luong, "Texture Feature Extraction by Wavelet Projection Pursuit for Cloud Detection in AVIRIS and AVHRR Imagery," in *Neural Networks for Signal Processing VI: Proceedings of the 1996 IEEE Workshop*, Kyoto, Japan, Sept. 4-6, 1996, pp. 351-360.
- [4] E. L. Bienenstock, L. N. Cooper, P. W. Munro, "Theory for the Development of Neuron Selectivity: Orientation Specificity and Binocular Interaction in Visual Cortex," *J. Neuroscience*, Vol. 2, No. 1, pp. 32-48, 1982.
- [5] D. W. Chen, S.K. Sengupta, and R.M. Welch, "Cloud Field Classification Based upon High Spatial Resolution Textural Features: 2. Simplified Vector Approaches," *J. Geophys. Res.*, 94, No. D12, 14749-1
- [6] D. J. Diner, C. J. Bruegge, J. V. Martonchik, G. W. Bothwell, E. D. Danielson, E. L. Floyd, V. G. Ford, L. E. Hovland, K. L. Jones, M. L. White, "A Multiangle Imaging SpectroRadiometer for Terrestrial Remote Sensing from the Earth Observing System," *Int. J. of Imaging Systems and Technol.*, Vol. 3, 92-107, 1991.
- [7] J. H. Friedman, J. W. Tukey, "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Trans. Computers*, Vol. c-23, No. 9, pp. 881 - 890, 1974.
- [8] J. H. Friedman, "Exploratory Projection Pursuit," *J. of the Amer. Stat. Assoc.*, March, 1987, Vol. 82, No. 397, Theory and Methods, pp. 249-266.
- [9] T. M. Heskes, E. T. P. Slijpen, B. Kappen, "Cooling Schedules for Learning in Neural Networks," *Phys. Rev. E*, Vol. 47, No. 6, pp. 4457-4464.
- [10] P. J. Huber, "Projection Pursuit," *Annals of Statistics*, Vol. 13, No.2, 435-475, 1985.
- [11] N. Intrator, L. N. Cooper, "Objective Function Formulation of the BCM Theory of Visual Cortical Plasticity: Statistical Connections, Stability Conditions," *Neural Networks*, Vol. 5, pp. 3-17, 1992.
- [12] M. D. Richard, R. P. Lippman, "Neural Network Classifiers Estimate Bayesian a posteriori Probabilities," *Neural Computation*, 3, 461-483, 1991.
- [13] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, W. M. Porter, *The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)*, *Remote Sens. Environ.*, 44: 127-143 (1993).
- [14] R. M. Welch, S. K. Sengupta, A. K. Goroch, P. Rabindra, N. Rangaraj, M. S. Navar, "Polar Cloud and Surface Classification Using AVHRR Imagery: An Intercomparison of Methods," *J. Appl. Meteorology*, Vol. 31, pp. 405-420, May, 1992.
- [15] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative Study of Texture Measures for Terrain Classification," *IEEE Trans. Syst., Man, Cybern.*, SMC-6, 269-285, 1976.