LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-220726

Principal component, varimax rotation and cost analysis of volume effects in rectal bleeding in patients treated with 3D-CRT for prostate cancer

J. D. Bauer, A. Jackson, M. Skwarchuk, M. Zelefsky

April 20, 2006

Physics in Medicine and Biology

Principal component, varimax rotation and cost analysis of volume effects in rectal bleeding in patients treated with 3D-CRT for prostate cancer

J D Bauer[1,2], Andrew Jackson[1,3], Mark Skwarchuk[1,4] , and Michael Zelefsky[1]

[1]Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, NY, NY, 10021

[2] Current Address, Lawrence Livermore National Laboratory, 7000 East Avenue, Box L-18 , Livermore CA, 94550

[3] Corresponding Author

[4] Current Address, Radiation Oncology/KRMC, 343 Sunnyview Lane Kalispell, MT 59901

Abstract

We investigate the utility of principal component analysis as a tool for obtaining dose-volume combinations related to rectal bleeding after radiotherapy for prostate cancer. A direct implementation of principal component analysis reduces the number of degrees of freedom from the patient's dose-volume histograms that are associated with bleeding. However when low variance principal components are strongly correlated to outcome, their interpretation is problematic. A Varimax rotation is employed to aid in interpretability of the low variance principal components. This procedure brings us closer to finding unique dose-volume combinations related to outcome but reintroduces correlation, requiring analysis of the overlap of information contained in such modes. Finally, we present examples of cost - benefit analyses for candidate dose-volume constraints for use in treatment planning.

PACS: 87.53.Tf

Introduction

Dose-Volume Histograms (DVHs) are widely used to investigate volume effects in the outcome of radiotherapy and play a critical role in treatment planning, where dose-volume constraints on target and normal tissues ensure safe and efficacious treatment. Unique dose-volume combinations that give rise to outcome are difficult to obtain directly from statistical analysis of the DVHs of treated patients since the dose-volume combinations present in patient populations are correlated with each other. In the context of external beam radiotherapy, these correlations arise from the fixed treatment techniques and beam geometries used. In this work we employ the well known statistical method of Principal Component Analysis (PCA) (Krzanowski 2000) as a means to investigate the information contained in dose-volume histograms. The method is applied to the analysis of rectal bleeding using the dose-volume histograms of patients receiving 3D-conformal radiation therapy for prostate cancer at Memorial Sloan-Kettering Cancer Center (Bauer et al. 2004) Recently, Dawson et al. have applied PCA to analysis of RILD in patients treated with radiotherapy for tumors in the liver (Dawson et al 2005).

We find that when PCA is applied straightforwardly, it allows us to produce models of patient outcome that contain only a small number of uncorrelated degrees of freedom. These explanatory variables for outcome consist of projections of the DVH data onto particular basis vectors produced by the PCA procedure. This direct implementation of the PCA method forfeits ease of interpretation; the modes produced still cannot be simply interpreted as representing unique dose-volume combinations correlated to outcome; although insight into the qualities of the DVHs that correlate with outcome may be ascertained. This potential shortcoming of PCA should not be surprising. PCA partitions the data by variance using linear combinations of "original" variables, in this case the volumes-exposed. The contribution of volumes-exposed in the various principal components is determined by the size of their variance and the redundancy of the information they contain, and the association of these qualities with outcome is not always straightforward.

We subsequently employ a Varimax Rotation (Harmon 1970). The Varimax Rotation is an orthogonal rotation applied to a truncated set of principal components (consisting of the smallest set whose combined variance constitutes the vast majority of the total). Its application is an attempt to obtain modes that are simple to interpret. The simplicity of a Varimax mode is characterized by non-negligible amplitude in only narrow dose regions. Despite the orthogonality of the Varimax modes, the projection of the DVH data onto them results in explanatory variables that are once again correlated (Joliffe 1995). Thus, when ease of interpretability is obtained by the Varimax Rotation technique, one must consider the relationship of the Varimax variables to each other when determining their relationship to outcome.

1. Method

We illustrate our method with data from patients treated for prostate cancer to 70.2 Gy and 75.6 Gy with 3D-Conformal Radiation Therapy (3D-CRT) discussed in previous work (Skwarchuk 2000, Jackson 2001). We consider the relationship between patient DVH and ≥ grade 2 rectal bleeding (outcome) arising from treatment (Lawton 1991). For

each dose group, DVHs of the rectal wall defined as extending from the sigmoid colon to the anal verge were examined. In the case of patients treated to 75.6 Gy the patient sample consists of 36 out of 38 patients who developed rectal bleeding before 30 months from the end of treatment and a random sample of 83 out of 192 patients who showed no bleeding at this time. For patients treated to 70.2 Gy the analysis utilizes data from13 patients who developed rectal bleeding before 30 months from the end of treatment and a random sample of 39 out of 208 patients without bleeding at this time. We remind the reader that a six-field arrangement was used to treat the patients. This arrangement consisted of one pair of opposed lateral beams and two pairs of opposed oblique beams, resulting in three regions of correlated dose to the rectal wall. The first high dose region was contained in the overlap region of all beams, the second intermediate dose region was contained in the overlap region of all oblique beams, the third low dose region was contained in only one oblique pair (Jackson 2001).

Once the DVHs have been constructed they are imported into MATLAB (The Mathworks) in which all mathematical manipulations are performed.

1.1 PCA

PCA is a technique used in multivariate data analysis to reduce the number of variables necessary to explain the dispersion in a data set. Given a data set in which the variables of interest are statistically correlated, PCA constructs a new set of uncorrelated variables via an appropriate linear combination of the original variables. Each of the new uncorrelated variables explains a unique portion of the data. Typically only a small subset of these variables that explains a majority of the variance in the data are retained. The rest account for arbitrarily small variations in the data and are discarded.

The principal component analysis of the patient data proceeds by forming the mean-adjusted variance-covariance matrix. The mean-adjusted volume for each dose bin is obtained by subtracting the dose-bin mean over all patients from each individual patient's DVH as indicated below. The covariance matrix **C** consists of correlations between mean-adjusted volumes exposed to a particular dose or greater.

$$C_{\alpha\beta} = \left(\frac{N_{pt}}{N_{pt} - 1} \right) \frac{\sum_{i \in patients} w_i \left(V_{i\alpha} - \overline{V_\alpha} \right) \left(V_{i\beta} - \overline{V_\beta} \right)}{\sum_{i \in patients} w_i} \qquad (1)$$

$$\overline{V_\alpha} = \frac{\sum_{i \in patients} w_i V_{i\alpha}}{\sum_{i \in patients} w_i} \qquad (2)$$

In the above equations, $V_{i\alpha}$ is the % volume exposed to dose ≥ α for patient i, w_i represents the weight of patient i due to the sampling of the patient population. N_{pt} denotes the actual number of patients used and $\overline{V_\alpha}$ is the mean-volume exposed to dose ≥ α. All patients, bleeding and non-bleeding, are included in the above summations. The prefactor, $\frac{N_{pt}}{N_{pt}-1}$, ensures the appropriate divisor for the correlation matrix when all weights are equal. For the purposes of PCA this prefactor is irrelevant as it serves only as an overall scale factor on the variance-covariance matrix. The diagonal elements of **C** represent the variance of the DVH data while the off-diagonal elements represent the covariance between volumes. The average volumes $\overline{V_\alpha}$ for the 75.6 Gy and 70.2 Gy patient data are shown in [1(a)-(b)]

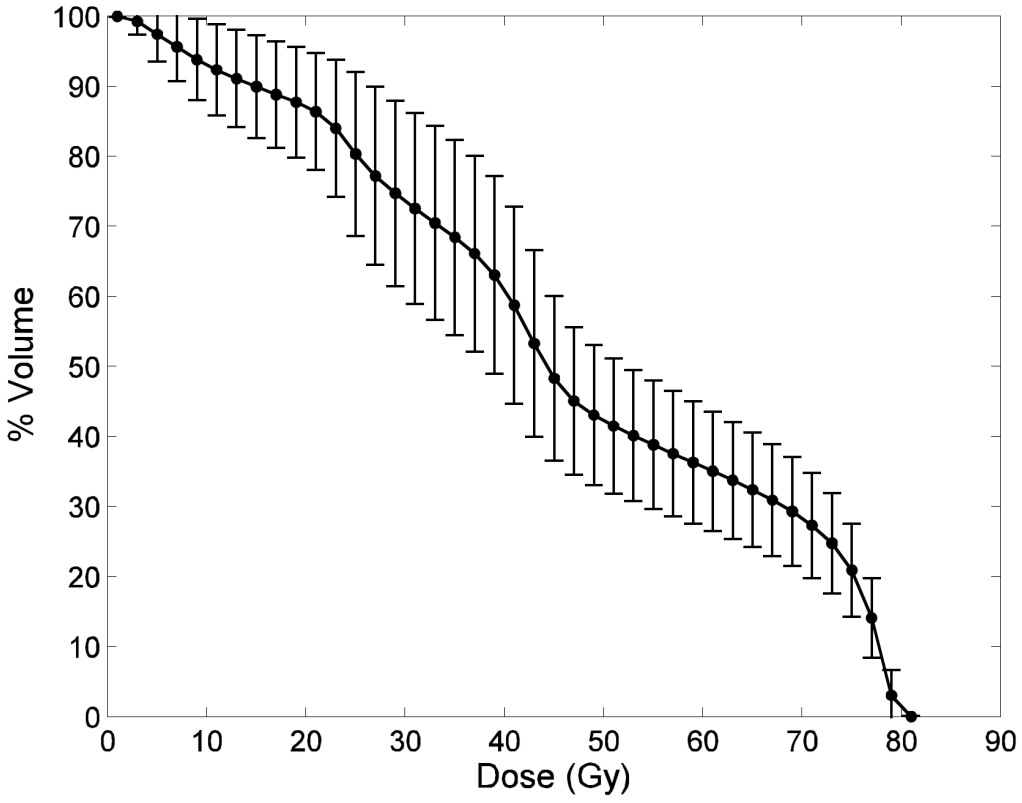

Figure 1(a). Weighted mean DVH of all 75.6 Gy patients. The error bars represent the standard deviation at each dose.

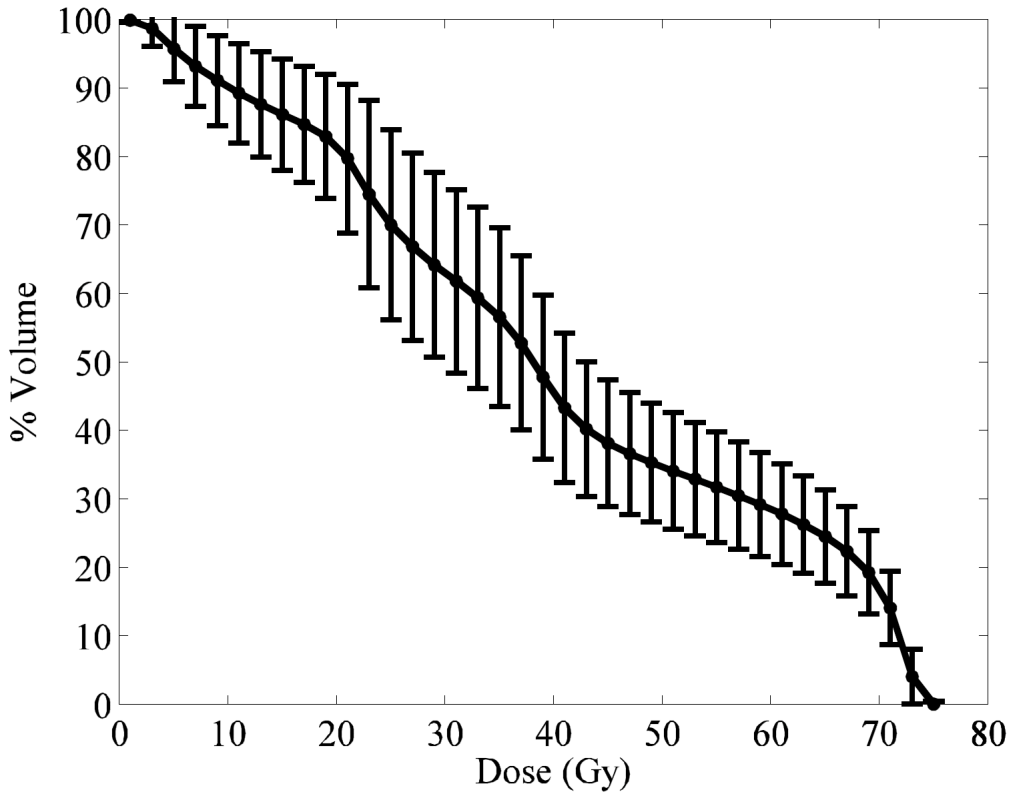

Figure 1(b). Weighted mean DVH of all 70.2 Gy patients. The error bars represent the standard deviation at each dose.

The DVH data for each patient is formed from bins of width 2 Gy; the 75.6 Gy patient data is comprised of 41 bins and the 70.2 Gy patient data is comprised of 38 bins. For compactness of notation and utility during computations we define a matrix **X.** The rows of **X** contain the mean-adjusted DVHs for each patient while the columns indicate the doses i.e. $X_{i\alpha}$ is the volume of patient *i* exposed to dose ≥ α. We also define a weight matrix **W**:

$$W_{ij} = \delta_{ij} \frac{w_i}{\sum_{i \in patients} w_i} \qquad (3)$$

Using these definitions we can write the variance-covariance matrix more compactly.

$$\mathbf{C} = \frac{N_{pt}}{N_{pt} - 1} \mathbf{X^{Tr} W X} \qquad (4)$$

The mean-adjusted variance-covariance matrix for patients treated to 75.6 Gy is shown in [2]. The correlated regions are clearly visible in the block structure.

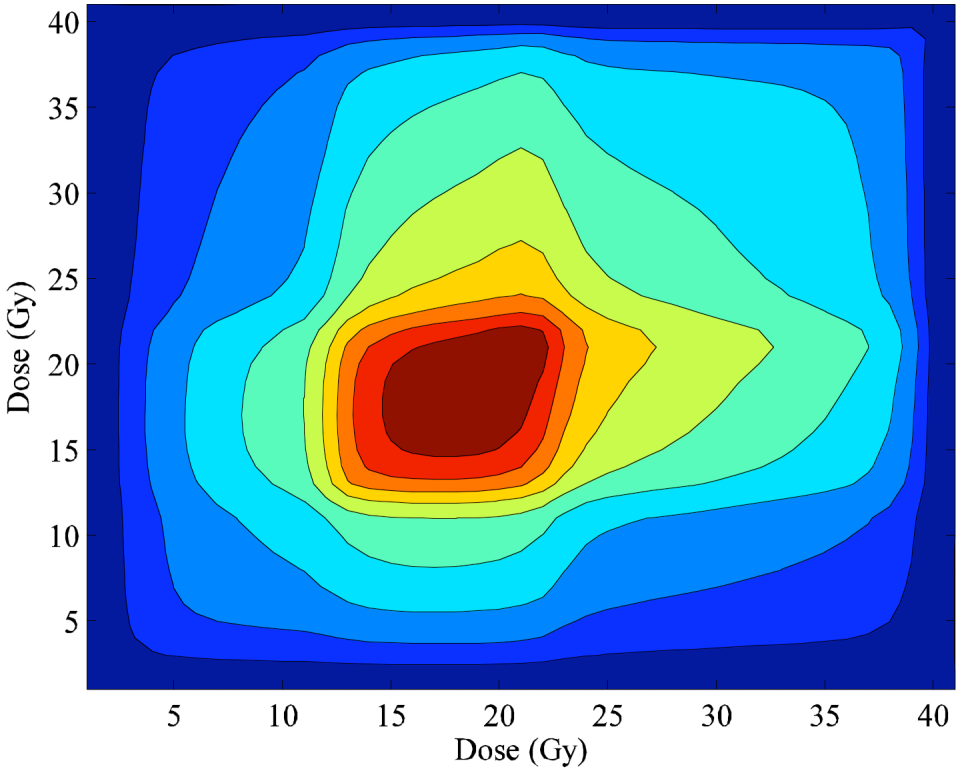

Figure (2). Variance-Covariance Matrix of %-Volume Exposed

The PCA diagonalizes the matrix **C** via an orthogonal transformation and produces a new coordinate system in which to represent the data, known as the principal components. To obtain the principal components we use the MATLAB statistical package function *pcacov* which returns the eigenvalues and normalized eigenvectors of the covariance matrix.

The eigenvalues of **C** represent the variance explained by each of these new variables. The eigenvalues are ranked in descending order and the rank of an eigenvalue is used to label its corresponding eigenvector. The eigenvectors define the principal axes; these are orthogonal and lie along the directions of sequentially maximal variance of the data after all higher variance components have been taken into account. The 38 and 41 bins of the 70.2 Gy and 75.6 Gy patients respectively yield 38 and 41 sets of eigenvalues and eigenvectors respectively.

We define **U** as the matrix whose columns are the eigenvectors of **C** and the matrix **S**=**XU**. The rows of **S** correspond to an individual patient's mean-adjusted DVH in the space of the principal components, while the columns of **S** contain the components of the patient DVHs along a particular PC. The projection of a DVH onto a principal component is its score on that component. Figures [3(a)-(b)] show a typical mean-adjusted DVH and the representation of this DVH in the space of principal components.

The score plot shown in [3(b)] is typical in that most of the scores beyond the first several are near zero indicating that only the first several principal components have utility for understanding the structure of the DVH. In our analysis of 3D-CRT DVH data, we retain the minimal set of modes required to account for > 99% of the total variance. We also include the contribution of a mode toward explaining outcome as an additional criterion for retention in the analysis. Once the data is decomposed by variance and the scores corresponding to each patient's projection onto the principal components are obtained, it remains to test each set of scores for correlation to bleeding.

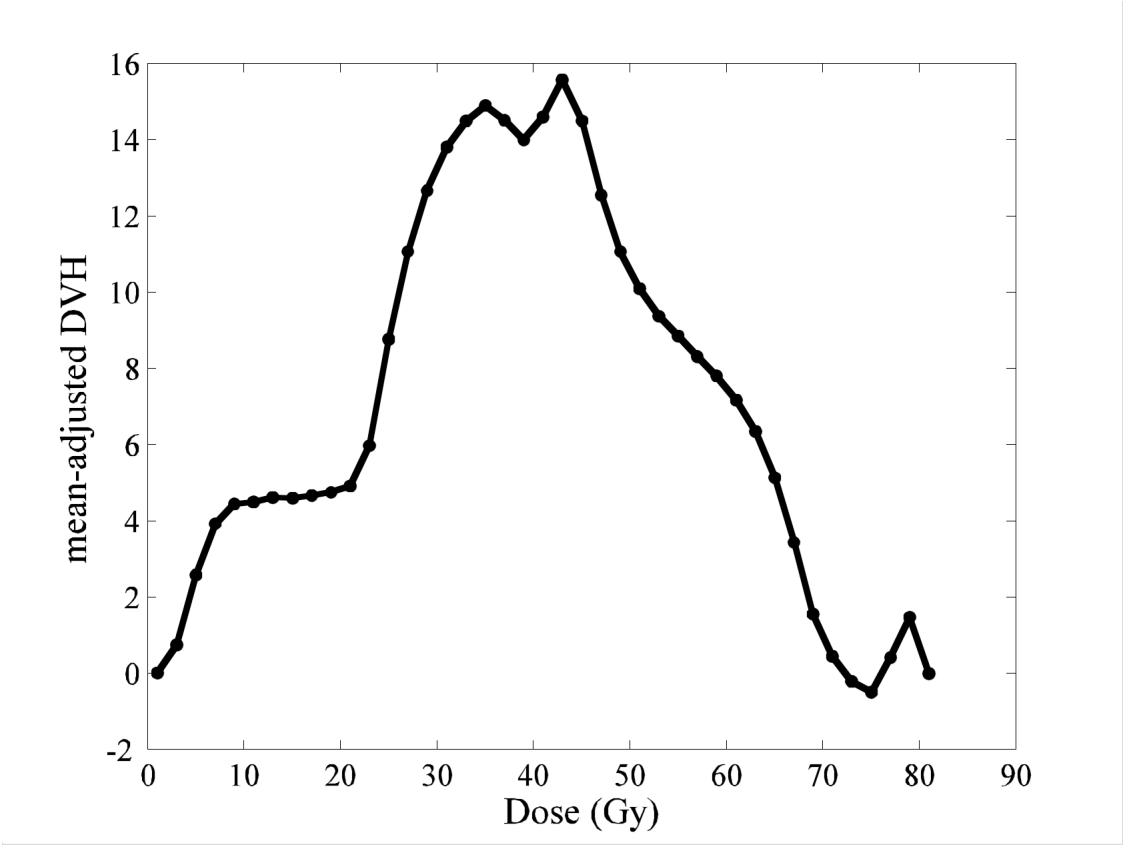

Figure 3(a) Typical Mean-Adjusted DVH for a 75.6 Gy patient.

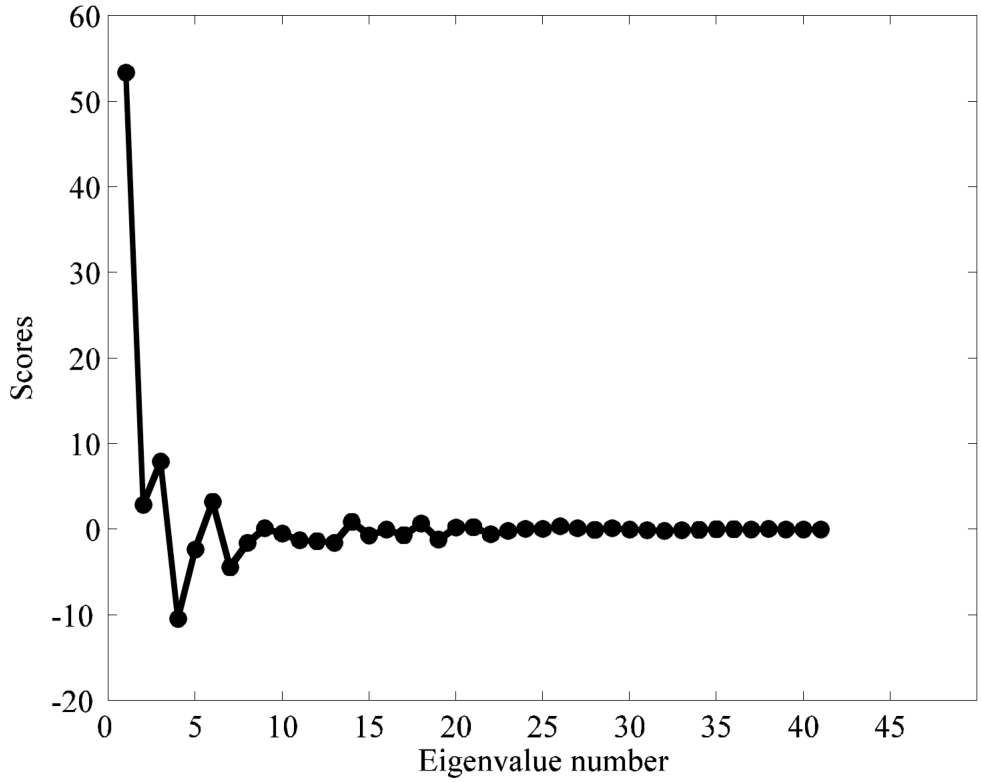

Figure 3(b) Scores associated with DVH in [3(a)].

1.2. Logistic Regression

Patient scores corresponding to specific principal components are individually tested for correlation to rectal bleeding by performing univariate logistic regression (Le 2003,Kleinbaum 1994, Prentice and Pike 1979, McCullagh and Nelder 1989). A step-forward procedure is used to build a multivariate model of all scores that contribute to the understanding of patient outcome. The bleeding variable for patient *i*, Y_i, is binary and has been coded 1 for bleeding patients and 0 for non-bleeding patients. A multivariate logistic model provides the probability of bleeding for each patient given the scores of the contributing PCs.

$$P(Y_i = 1) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_\alpha \beta_\alpha * S_{i\alpha}\right)}} \qquad (5)$$

$S_{i\alpha}$ is the score of patient *i* on PC α, the β's are the coefficients obtained from the logistic regression. Significance of the scores, or equivalently significance of the mode corresponding to the α^{th} principal component, is determined by the p-value of the corresponding β_α. A particular β_α is significant if the probability of obtaining the fit value

or greater of $|\beta_\alpha|$ is less than or equal to 0.05 under the null hypothesis: $\beta_\alpha = 0$. As a result of the logistic regression, a single number characterizing the bleeding information content of each patient's DVH is obtained:

$$S_i' = \sum_\alpha \beta_\alpha S_{i\alpha} \ . \qquad (6)$$

Here the sum is carried out over statistically significant components.

1.3 Receiver Operator Characteristic Analysis

The effectiveness of the numbers S'_i in distinguishing bleeding from non-bleeding patients in the data set can be obtained through a receiver operator characteristic (ROC) analysis (Le 2003, Hanley and McNeill 1982,1983). For each cutoff value the number of true positives, TP, is defined as the number of bleeding patients with a regression model value greater than or equal to the cutoff. The true positive rate, TPR, for a particular cutoff is the number of true positives divided by the total number of bleeding patients. Similarly the number of false positives, FP, is given by the number of non-bleeding patients with a regression model value greater than or equal to the cutoff value. The corresponding false positive rate, FPR, is given by the number of non-bleeding patients with a model value greater than or equal to the chosen cutoff divided by the total number of non-bleeding patients. In the ROC analysis, the true positive rate defined for each S'_i is plotted against the corresponding false positive rate.

The area under the ROC curve (AUC) is the probability of a random bleeding patient in the data having a value of S' greater than that of a random non-bleeding patient. An estimate of the error of the AUC is obtained using the method of Hanley and McNeill (1982, 1983). A consistency calculation for the AUC and its error is performed using a bootstrap analysis. A full set of bleeders and non-bleeders are randomly selected from their respective sample populations with replacement. The covariance matrix is formed from these data and a PCA is performed yielding a model of bleeding. The AUC is computed by counting the number of pairs for which the model value of a bleeder exceeds the model value of a non-bleeder divided by the product of the number of bleeders and non-bleeders for each bootstrap configuration i.e. the number of bleeder – non-bleeder pairs (Hanley and McNeill 1982). A distribution of areas is obtained from which the mean AUC is computed. For each bootstrap configuration the AUC is given by:

$$AUC = \frac{\sum_{i \in bleeders,\, j \in non-bleeders} \Theta(S_i - S_j)}{n_{bleeders} \times n_{non-bleeders}} \qquad (7)$$

Θ is the Heavyside function and S_i , as above, is the model value. The error is computed from the standard deviation of the bootstrap areas.

1.4 Varimax

A well-known limitation of PCA is that the higher order principal component eigenvectors are not always easily interpretable. While the eigenvectors of **C** are orthogonal, they are not necessarily sparse; a sparse mode is one with many elements near or equal to zero and just a few elements that are large. The eigenvectors oscillate with increasing frequency as their order increases leading to difficulty in obtaining unique volumes exposed. The Varimax rotation method (Harmon 1970) maximizes the sparseness of each of a group of modes by means of an orthogonal rotation. As a result of a Varimax rotation, a new set of modes is produced; each of these new modes will have few components with large amplitude and many with small or nearly zero amplitude. The Varimax procedure is performed on the subspace of PCs necessary to reconstruct the data to the determined accuracy and not just those correlated to outcome.

The Varimax Rotation matrix R_v is obtained by maximizing the combined variance of the retained modes, known as the simplicity, given by:

$$VMX(R) = \sum_{\beta \in subspace} \left[\frac{1}{K} \sum_{\alpha=1}^{K} \left(N_{\alpha\beta}\right)^4 - \left[\frac{1}{K} \sum_{\alpha=1}^{K} \left(N_{\alpha\beta}\right)^2 \right]^2 \right]. \qquad (8)$$

Here, **N=UR** is the eigenvector in the Varimax rotated space, **U** is the matrix of PCA eigenvectors given above, K is the dimension of the eigenvectors (i.e. 41 and 38 for 75.6 and 70.2 Gy patients respectively), and β indexes the mode in the retained subspace of PCA eigenvectors.

The simplicity is minimal and equal to zero for a set of vectors that has uniform extent over its support, i.e. all components of the eigenvectors are equal in magnitude. It is maximal for eigenvectors each containing a single non-zero element. The Varimax Rotation produces orthogonal modes as can be verified by considering the orthogonality relations for the matrices U and R_v.

However it can be shown that a result of the Varimax Rotation is to reintroduce correlations into the Varimax scores which are the projection of the data onto the Varimax modes. This may become problematic when using multivariate logistic regression for outcome analysis as the explanatory or independent variables are no longer uncorrelated possibly confounding their effect on outcome. We will discuss the effects of this issue on our analysis in the results section of this paper.

1.5 Partitioning the bleeding from non-bleeding patients

If the model derived from PCA is to be of clinical utility we must consider how to partition the bleeding patients from the non-bleeding patients. A specific numerical value, S'_c, or cutoff, separating bleeding from non-bleeding patients must be determined. This requires an analysis of the expected cost of misclassifying a patient (Bradley 1997). For example, incorrect classification of a non-bleeder as a bleeder might change his treatment by underdosing the PTV to avoid complications while incorrectly classifying a bleeder as a non-bleeder might result in overdosing the PTV to enhance tumor control resulting in severe complication. To determine the cost of such actions would require the

contemplation of medical factors beyond the scope of this paper and in general would not be a trivial task.

It is instructive to consider a cost function which is the expected cost per patient of false positives (FP) and false negatives (FN) combined (Bradley 1997).

$$Cost = p(FP) * C(FP) + p(FN) * C(FN). \qquad (9)$$

The quantities p(FP) and p(FN) are the probabilities of a false positive and a false negative respectively. A patient whose model value, S'_i, is greater than or equal to the cutoff is considered to have a positive test value. Conversely if the patient's model value is less than the cutoff, he is considered to have a negative test value. A false positive model value corresponds to a model value for a patient who tests positive for bleeding given the chosen cutoff but is in fact a non-bleeding patient. A false negative similarly corresponds to a model value for a patient who tests negative for bleeding but is in fact a bleeding patient. The cost function may be re-expressed in terms of the true positive rate, TPR, the false positive rate, FPR, defined in section III, and the proportion of bleeding in the patient population denoted by *prop*. The proportion of bleeding is given by the fraction, appropriately weighted, of bleeding patients in the data set.

$$Cost = FPR * (1 - prop) * C(FP) + (1 - TPR) * prop * C(FN). \qquad (10)$$

We also consider the predictive values and accuracy of this test. The positive predictive value (Le 2003), PPV, of the test, is the probability that a patient is a bleeder conditioned on a positive test result. The positive predictive value is a function of the chosen cutoff and is the weighted fraction of all positives that are true positives (TP) as determined by that cutoff. The negative predictive value (Le 2003), NPV, is the probability that a patient is a non-bleeder conditioned on a negative test result; it is the weighted fraction of true negatives (TN) determined from that cutoff. The test accuracy or efficiency for a chosen cutoff is given by

$$\eta = TPR \times prop + (1 - FPR) \times (1 - prop) \qquad (11)$$

and is the proportion of correctly classified patients. The misclassification rate is the complement 1- η.

An optimal cutoff is determined by minimizing the cost. In the simple case of equal costs for false positive and false negative test results, the minimum cost analysis for obtaining an optimal cutoff for the model reduces to a minimum misclassification analysis of the model.

2. Results

2.1 PCA

We analyze the DVH data representing the % rectal wall volume exposed to a dose $\geq \alpha$ for the patients having received 75.6 Gy. The proportion of bleeding patients is 16.5%. The weight associated with a bleeding patient is 1.0556 and the weight associated with a non-bleeding patient is 2.3133. The spectrum of eigenvalues obtained by diagonalizing the variance-covariance matrix for these data ranked by the size of the eigenvalue is shown in [4].

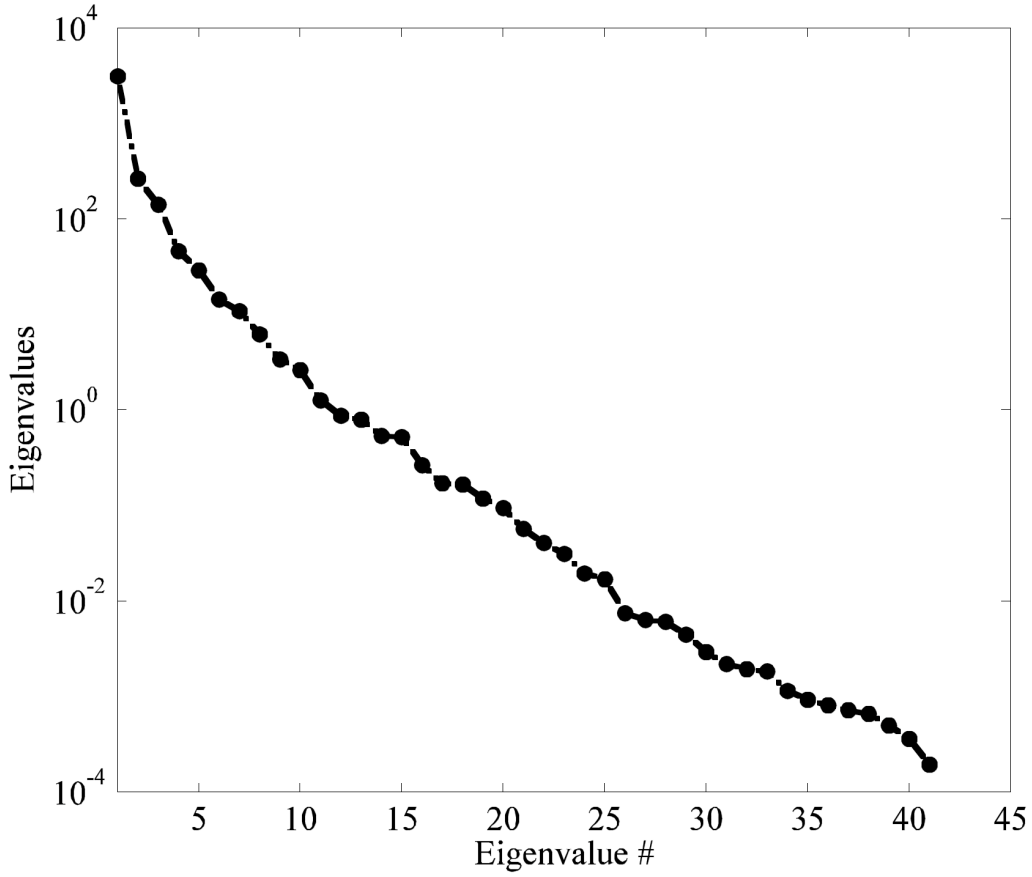

Figure (4). Eigenvalue spectrum of variance-covariance matrix for 75.6 Gy patient data.

The first eight eigenvalues account for greater than 99.5% of the variance with the 1st mode explaining 85.6% of the total variance in the DVH data. We restrict our analysis to only these modes. When each of these eight modes is univariately tested for correlation to rectal bleeding, modes 1, 6, and 7 attain significance. Under multivariate analysis only modes 1 and 7 survive. Mode 7 only explains 0.3% of the DVH data however the logistic regression analysis for this mode clearly indicates it is significant with regards to outcome. Thus we obtain a two-mode model of rectal bleeding for the 75.6 Gy patients. Modes 1 and 7 are shown in [5(a)] and [5(b)] respectively.

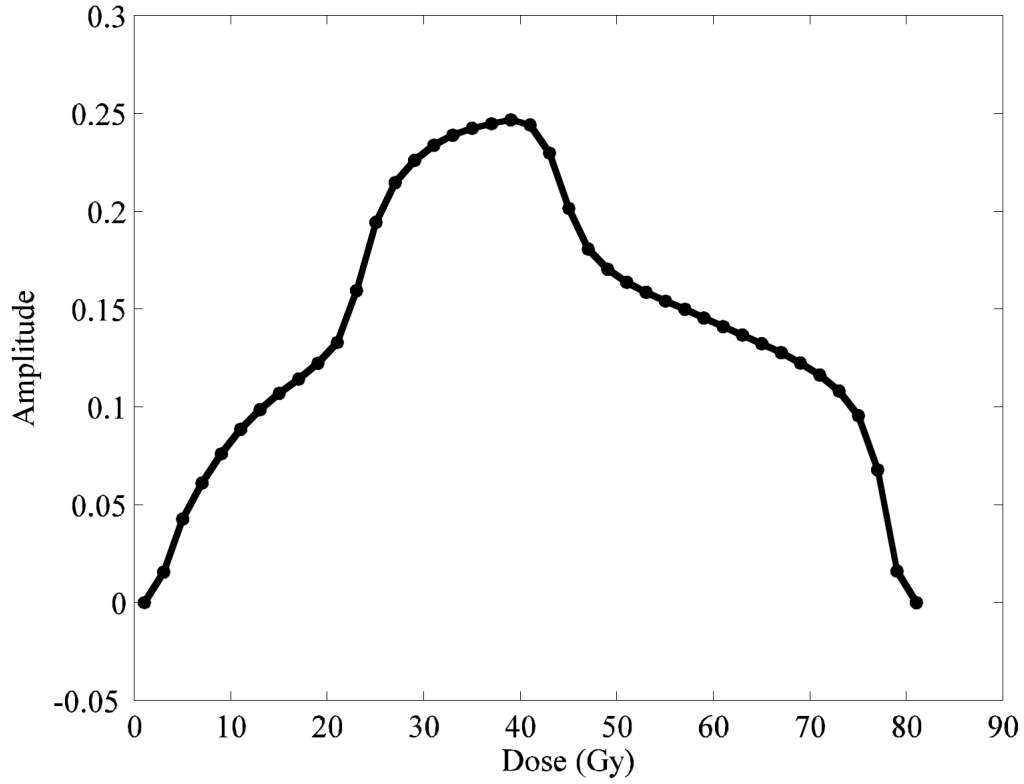

Figure 5(a). Principal Component Analysis eigenvector belonging to largest eigenvalue for 75.6 Gy patient data.

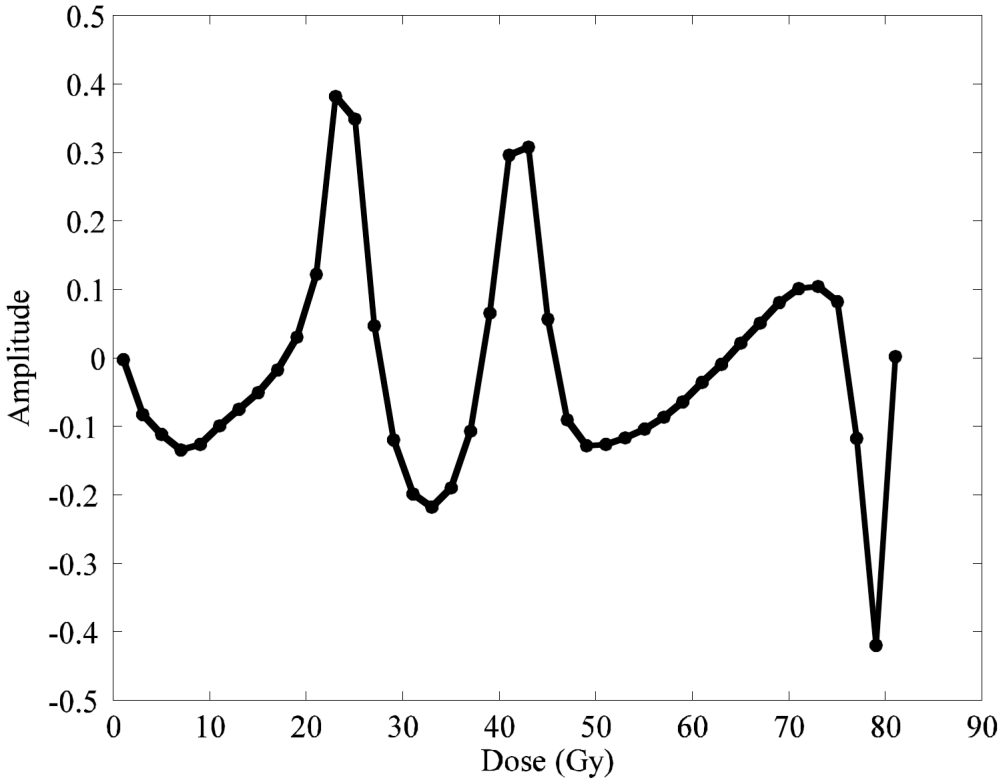

Figure 5(b). Principal Component Analysis eigenvector belonging to 7th largest eigenvalue for 75.6 Gy patient data.

The mode-1 eigenvector follows the shape of the variance in the % rectal wall volumes very closely and can be interpreted as a positive fluctuation of the % volume at all doses for bleeding patients versus non-bleeding patients. The information on outcome contained in this mode is redundant with the mean dose. Indeed a logistic regression model with both mean dose and the scores obtained from mode-1 input as explanatory variables indicates that they contain overlapping information with respect outcome. Individually each of these variables, mean dose and PC-1, under logistic regression analysis attains a p-value $<< 0.05$ while combined they attain p-values $>> 0.05$. The reason for this is almost certainly due to the fact that the mean dose is the integral over the DVH with respect to dose which, for discrete DVH data, can be viewed as the unweighted sum of volumes-exposed. The DVH is constrained at low dose to be near 100% and at high dose it must go to 0%. All of the variance in the integral of the DVH must stem from the middle range doses. This is precisely the dose region that the first mode emphasizes; the volumes with the largest variance and covariance. This result is also reminiscent of previous work showing that the mean value of the bleeding patient

DVHs is shifted relative to the mean value of the non-bleeding patient DVHs (Skwarchuk 2000, Jackson 2001).

As seen in [5(b)], mode-7, on the other hand, is more complicated and does not readily lend itself to interpretation. It shows strong fluctuations near 23 - 25 Gy, 41 - 43 Gy, and 79 Gy. These comprise of two distinctive groupings. Analysis shows that the volumesV23, V24, V41, and V43 are all highly correlated, with the correlation between V23 and V24 > 0.9, and the correlation between V41 and V43 > 0.9. The inter-group correlation, the correlation between V23 and V24 with V41 or V43, is > 0.8. This is obtained by computing the linear correlation coefficient.

$$\rho_{ij} = \frac{Cov_{ij}}{\sqrt{Var_i \bullet Var_j}} \qquad (12)$$

The numerator, Cov_{ij}, represents the correlation, the weighted inner product, of the PCA scores corresponding to modes i and j and Var_i represents the variance of the score corresponding to mode j. The high inter-group correlation between these volumes is a result of the beam geometry. The volume V79 has a relatively small variance and is not as highly correlated with the other volumes; the correlation coefficient between V79 and the V23-V24 is ≈ 0.2 and the correlation of V79 to V41-V43 is ≈ 0.25.

There are several possible explanations for the structure of mode-7. One is that this PCA mode is correlated with outcome due to the influence of V79. Mode-7 is a high-order mode, explaining only a small fraction of the variance. V79 has a relatively small variance and it is not highly correlated with any of the other volumes as can be observed in [2]. One expects the low order, high variance, PCA modes, e.g. mode-1, to be dominated by high-variance, strongly correlated volumes. The lower order modes, e.g. mode-7, by the definition of PCA, should contain low-variance volumes and account for variance not already accounted for in the previous modes. Earlier investigation of this data has shown that the volume exposed to high dose, $\approx$ V79, is strongly correlated with outcome(Jackson 2001). Understanding the role of the other volumes in this mode is more difficult. This mode could be revealing a separate contribution to outcome fromV42-V43 or V23-V25, in addition to V79. What is clear is that while one might obtain or infer insights into the information contained in the DVH from the principal component analysis, the procedure does not allow for a definitive statement concerning dose-volumes and their unique contribution to outcome when high-order modes attain significance.

Despite this difficulty we can construct a model for PCA mode 1 and mode 7, $S_i^{'}=\beta_1 S_{i,1}+\beta_7 S_{i,7}$. The ROC curve of this model is shown in [6]. The AUC for this model, computed directly from the TPR, FPR curve, is 0.80 ± 0.05. A ranking of the S'_i is shown in [7]. The tendency for bleeding patients is to have positive values; the non-bleeding patients are more evenly distributed but are relatively rare at the highest model values.

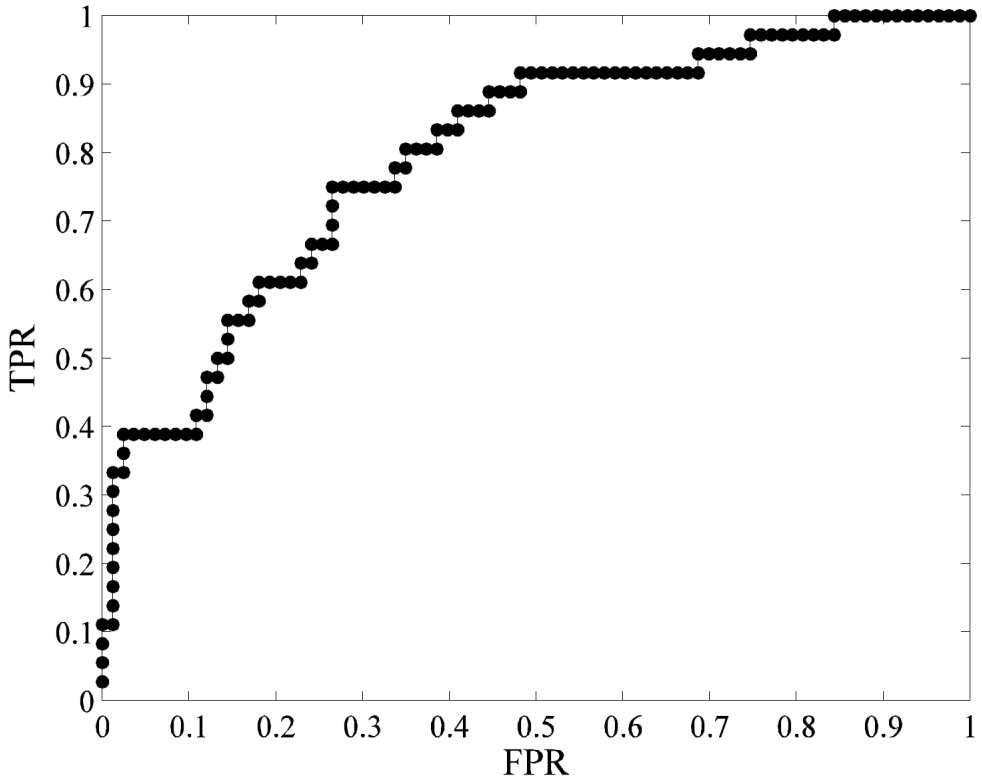

Figure 6. Receiver operator characteristic curve – True Positive Rate (TPR) versus False Positive Rate (FPR) for 75.6 Gy patient data using PCA modes 1 and 7. The area under the curve is 0.80.

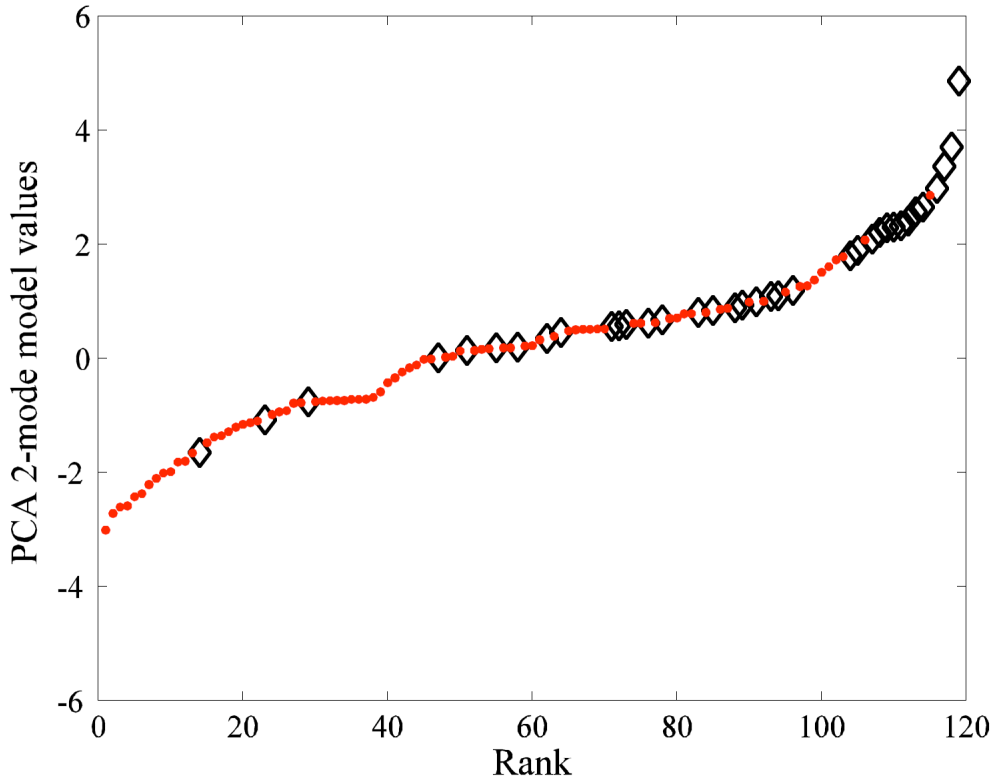

Figure 7. Ranked model value from combined 1st and 7th principal component. Bleeding patients are represented by open diamonds, non-bleeding patients are represented by filled circles. The minimum cost cutoff corresponds to 2.1 on the vertical scale.

2.2 Cost Analysis

Applying the cost analysis method we find the optimal cutoff for the model, We consider the case where the costs of false positives and false negatives are set equal to unity. In this case the total cost (10) is identically equal to the misclassification rate (11).

The cost for 75.6 Gy patients is minimized when the FPR = 0.012 and the TPR = 0.33. This corresponds to a value of $S_c^{'}$=2.1 for the two component PCA model. The (PPV,NPV) pair obtained at this value of $S_c^{'}$ is (0.85,0.88). For these values of FPR and TPR the efficiency of the test is 88%; and hence the misclassification rate is 12%. The inability of the model to more accurately classify patient outcome is largely due to the entanglement of the bleeding and non-bleeding patient DVH data and is intrinsic to the data set. The TPR is only 33% at the cutoff. This analysis minimizes the false positives as one would expect for a relatively low prevalence complication.

An estimate of the errors in this analysis is obtained using a leave-one-out cross-validation. Each patient in turn is left out of the construction of the variance-covariance

matrix and PCA analysis is performed. The left-out patient's value for the model is then predicted. Using all the predicted values a new ROC is formed with an AUC = 0.78. This is well within the stated error of the original AUC. The cost analysis is used on the cross-validated data to produce a new cutoff. We determine the FPR, TPR, PVP, and NPV values corresponding to this cutoff as well as the misclassification rate. The calculated values corresponding to the cross-validated data are essentially unchanged from the estimates based on the original data set.

Figure [8] shows the Cost, TPR, FPR, PPV, and NPV for this model. The cost initially decreases with FPR and flattens at model values $\geq$ 1.8.The relative rates of decrease of FPR and TPR control the cost. In [7] it is noticed that as the cutoff value increases, fewer non-bleeders are incorrectly classified. When the model value S'=1.8 is attained, all but one non-bleeder is correctly classified. Further increase in the cutoff results in a rapid decrease in the TPR and a slow decrease in the FPR until it becomes zero. The PPV is bound from below by the proportion of bleeders which in this case is equal to 0.165. The NPV is bound from below by the proportion of non-bleeders which is equal to 0.835. As stated earlier, the chosen cutoff is dependent on the relative costs associated with false positives and false negatives. The proportion of bleeding patients to non-bleeding patients, i.e. prop to (1-prop), is about 0.2 indicating the ratio of false negatives to false positive costs would need to become large to have an impact on the chosen cutoff.

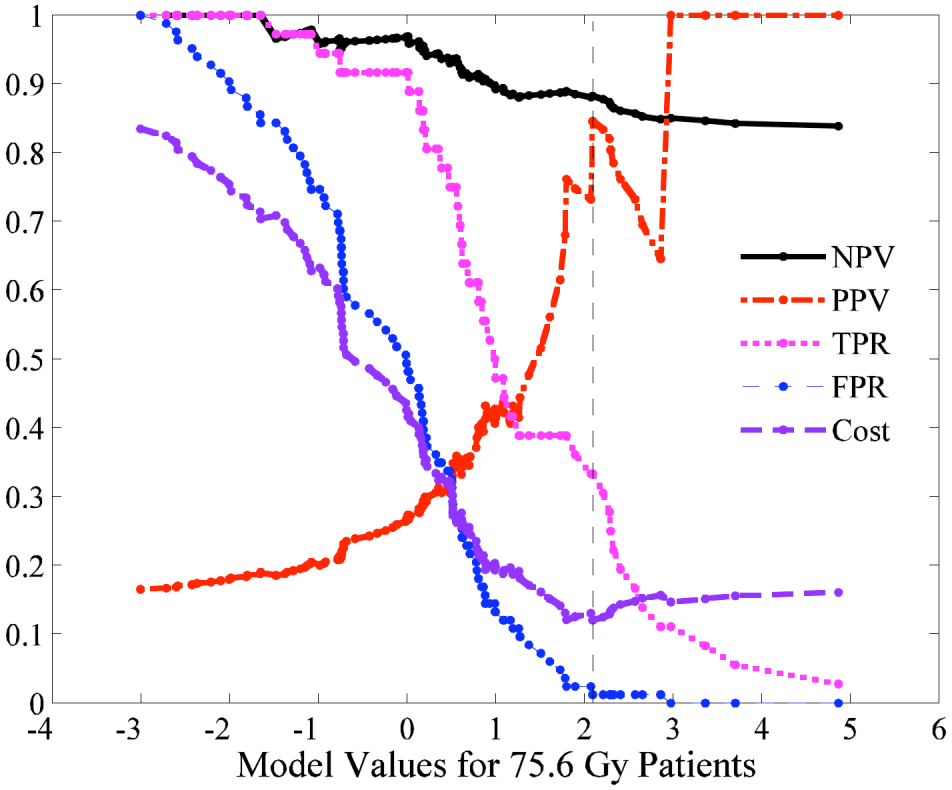

Figure 8. Cost analysis for 75.6 Gy patient data. The vertical dotted line indicates the minimum cost values for TPR, FPR, PVP, and NPV. The dots represent actual values while the lines serve as a guide. The vertical line indicates the minimum cost values of the displayed quantities.

2 .3 Varimax Rotation

Applying the Varimax Rotation to the 1st eight retained modes, we find three Varimax modes are significant with respect to outcome. These modes are labeled mode-2, mode-5, and mode-6 due to their order on return from the Varimax routine; as explained, the Varimax procedure introduces correlation between the modes thus the individual variance of the modes no longer plays a unique role in identifying them. Modes-2, 5 and 6 are shown in [9 (a)-(c)].The ROC curve for a model produced from these three modes, [10], yields the same AUC as the ordinary PCA modes, 0.80 ± 0.05. The utility of Varimax modes is that only small regions of each mode are large. We notice there is no appreciable amplitude for doses < 40 Gy in any of the modes. Varimax mode-6 clearly represents a contribution from the 77 - 79 Gy dose bins to rectal bleeding. Mode-2 is peaked near 47- 49 Gy and mode-5 near 73 Gy indicating contributions from those doses respectively are increasing the probability of bleeding. The minimum cost analysis yields TPR=0.25, FPR=0.012, $S'_c = 2.5$. The PPV and NPV for this model are 0.80 and 0.87 respectively The misclassification rate is approximately 13%. The Varimax procedure incurs an expense of an additional degree of freedom as compared with a simple application of the PCA method for the 75.6 Gy patient data.

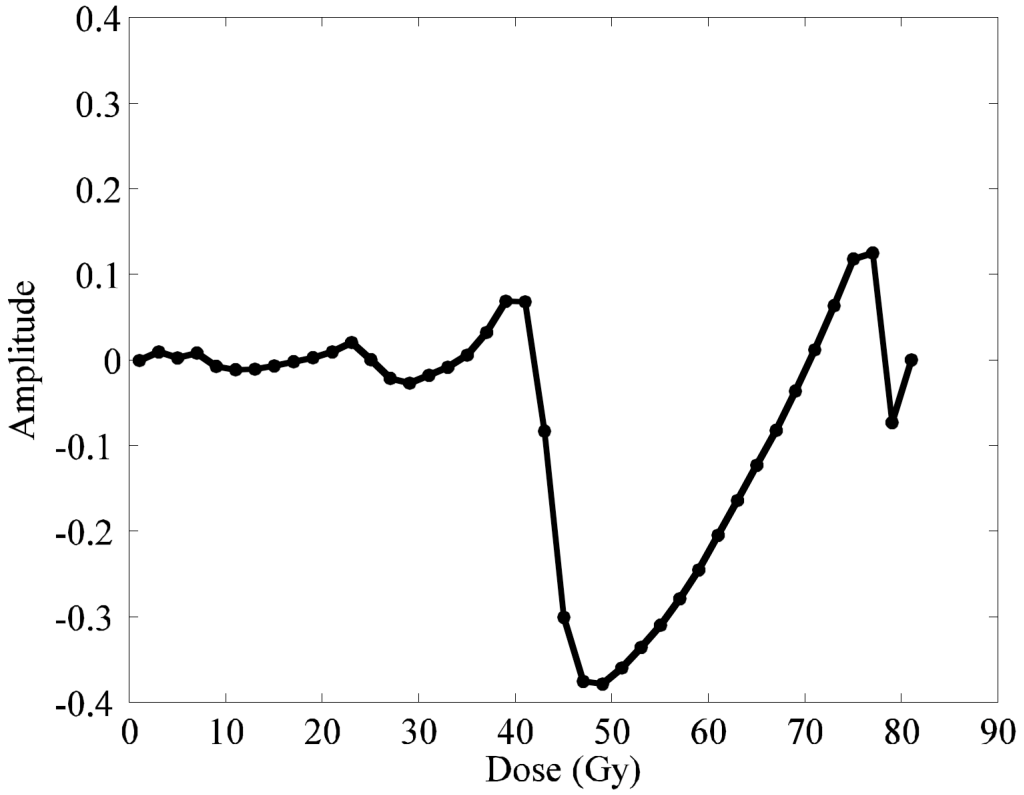

Figure 9(a). Varimax mode 2 for 75.6 Gy patient data showing strong amplitude in the 47-47 Gy region of dose.

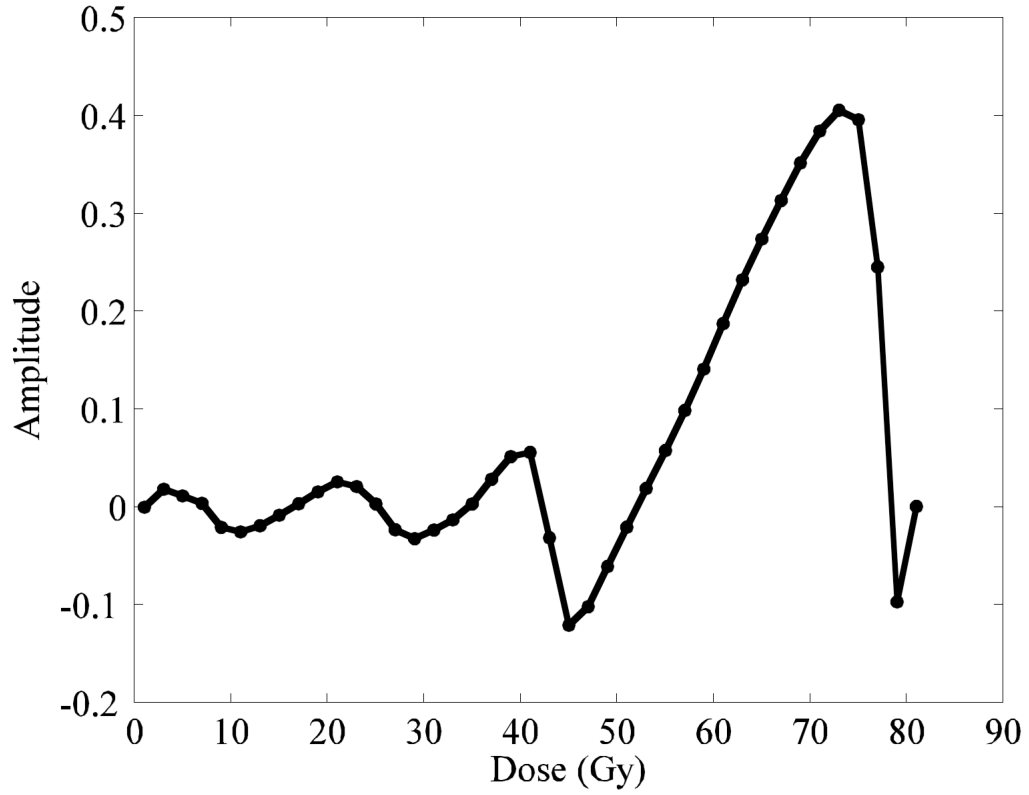

Figure 9(b). Varimax mode 5 for 75.6 Gy patient data showing strong amplitude in the 73 Gy dose region.

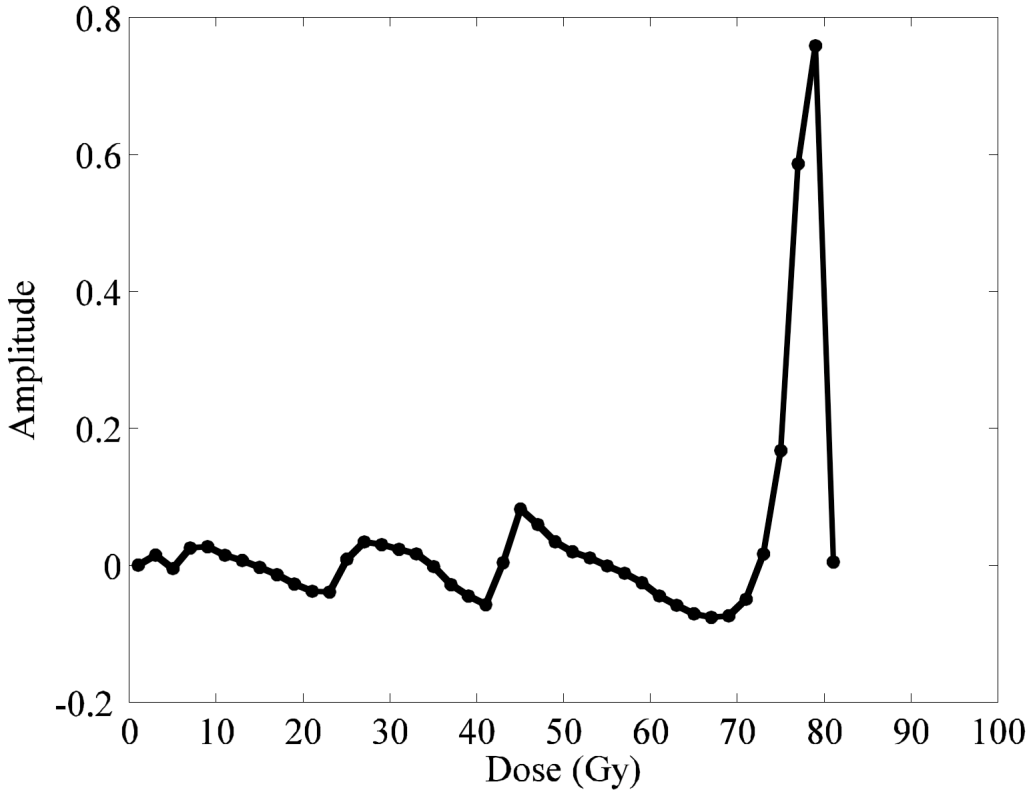

Figure 9(c). Varimax mode 6 for the 75.6 Gy patient data showing strong amplitude in the 77 - 79 Gy dose region.

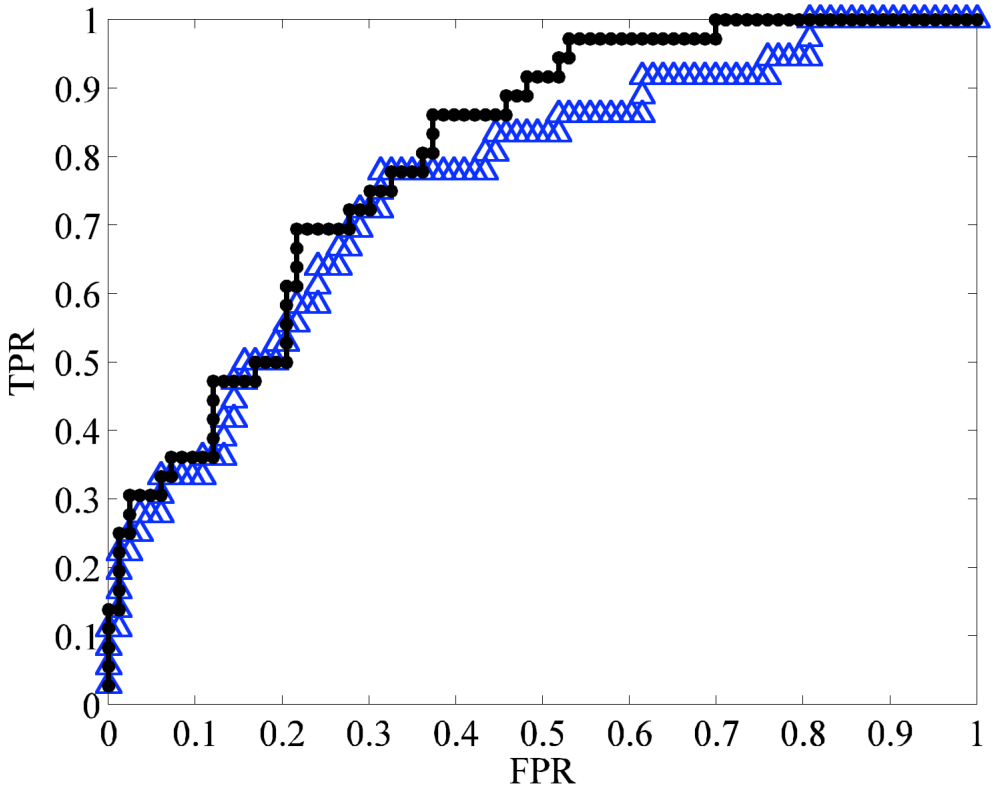

Figure 10. Receiver Operator Characteristic Curves for 75.6 Gy data set: 3-mode Varimax model (filled circles) AUC=0.80, 2-mode Varimax model (triangles) AUC=0.76.

We address the correlation between the Varimax scores by computing their linear correlation coefficient.. The relevant correlations are: ρ_{26}=-0.42, ρ_{25}=-0.93, ρ_{56}=0.53 where the subscripts refer to the Varimax modes. Clearly the correlation between scores 2 and 5 is so great that there is redundant information included in the model when both modes are kept and one can question whether they both belong in the model. In point of fact V73 is highly correlated with V47. Mode 5, the least significant under univariate analysis, enters the multivariate model last and only becomes significant in the presence of modes 2 and 6. If it is removed from the model we obtain an AUC = 0.76±0.05. This AUC is not discernable from the full 3-mode model given the errors. The correlation between mode-2 and mode-6 is < 0.5 and we may reasonably use these two modes in the model. Thus the 2-mode model constructed from Varimax mode 2 in combination with Varimax mode 6 could reasonably be used instead of the full 3-mode model although we may be losing some separation power in distinguishing bleeders from non-bleeders. Using the minimum cost analysis, the 2-mode model produces a PPV = 0.79, and a NPV = 0.87. At the minimum cost point the misclassification rate is 14%; the TPR and FPR is 22% and 1.2% respectively. The score values produced from the two-mode Varimax model for each patient are shown in [11].

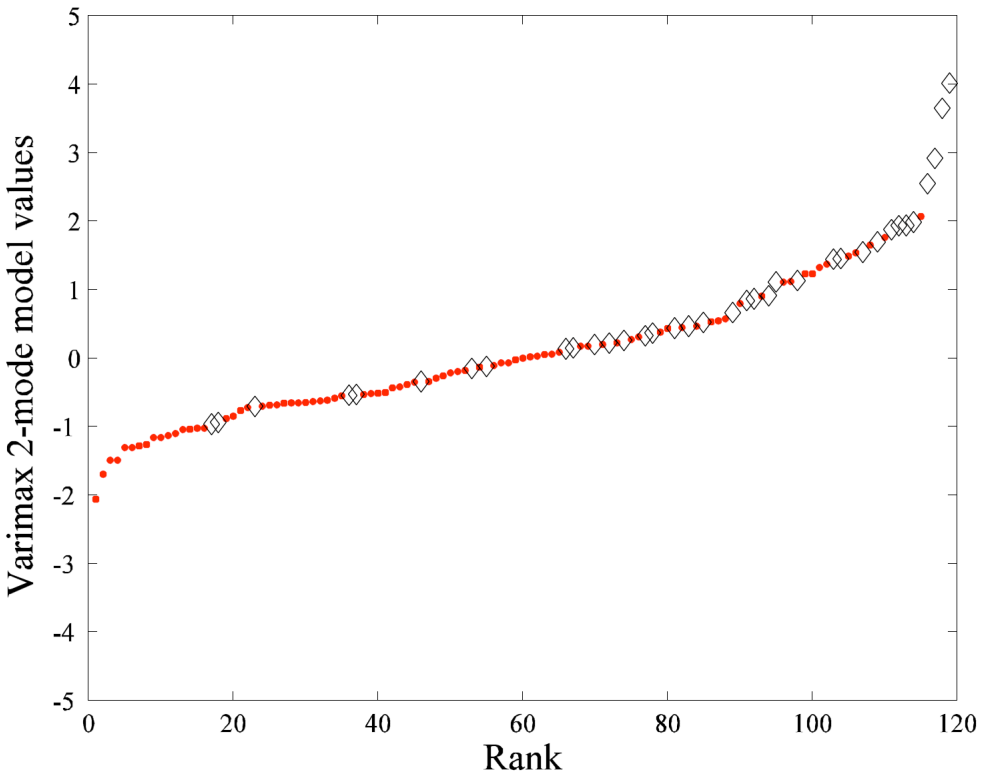

Figure 11. Ranked Varimax model coefficients for the 75.6 Gy patients: 2-mode model. Bleeders are represented by open diamonds. Non-bleeders are represented by filled circles.

The 2-and 3-mode models do not differ greatly in their predictive capability. In such cases we can question whether the information in the discarded mode is pertinent to outcome and if the explanatory value of the model has become compromised. Without more data we are not able to make more of a statement as to whether the 2- or 3-mode model is more useful in partitioning the outcome.

In addition to the % volume exposed for patients treated to 75.6 Gy, we have analyzed the % volume exposed data for patients treated to 70.2 Gy. These patients comprise a much smaller data set than the 75.6 Gy patients. A total of 52 patients are included with only 13 bleeding patients. The non-bleeding patients carry a weight of 5.3333 while the weight of a bleeding patient is 1.0. For these data the first eight modes account for 99.75% of the total variance, however only the first PC correlates significantly with rectal bleeding. This mode is shown in [12] and much like the 75.6 Gy data it reflects a net positive % volume fluctuation of bleeders with respect to non-bleeders at all doses. This mode explains nearly 87% of the variance for the 70.2 Gy patients' DVH data. In a similar fashion to the 1st PCA mode of the 75.6 Gy analysis, this mode contains information with respect to outcome analogous to the mean dose. An ROC

analysis of the model composed of only this one mode yields an AUC = 0.73 ± 0.09, [13]. Figure [14] shows the ranked 70.2 model values using the single significant mode. Given the distribution of the bleeders, their weights and the very low prevalence for bleeding, 6%, the cost function fails to find a suitable cutoff in the data.

Applying the Varimax procedure we find that a single Varimax mode peaked near 39 Gy is significant for correlation to rectal bleeding. This mode is shown in [15]. An ROC analysis of this mode gives an AUC = 0.76 ± 0.09. Only a single Varimax mode is contributing to outcome, confounding is therefore not a problem for this patient data set.

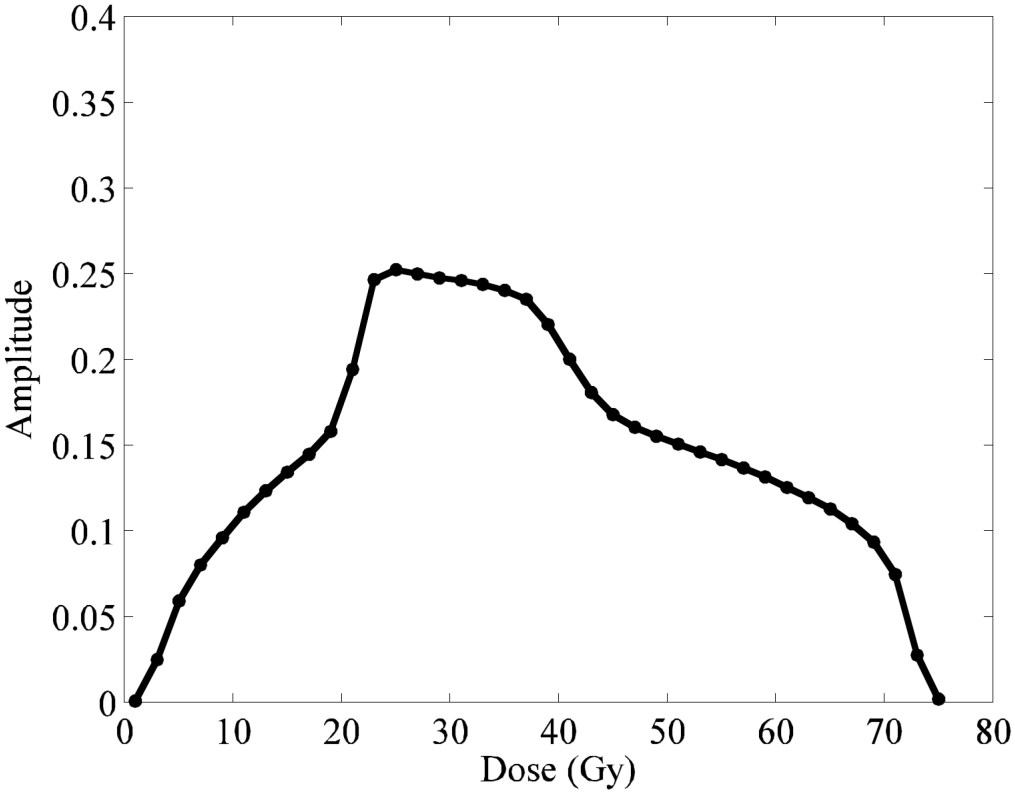

Figure 12. PCA mode-1 corresponding to the largest eigenvalue for the 70.2 Gy data set.

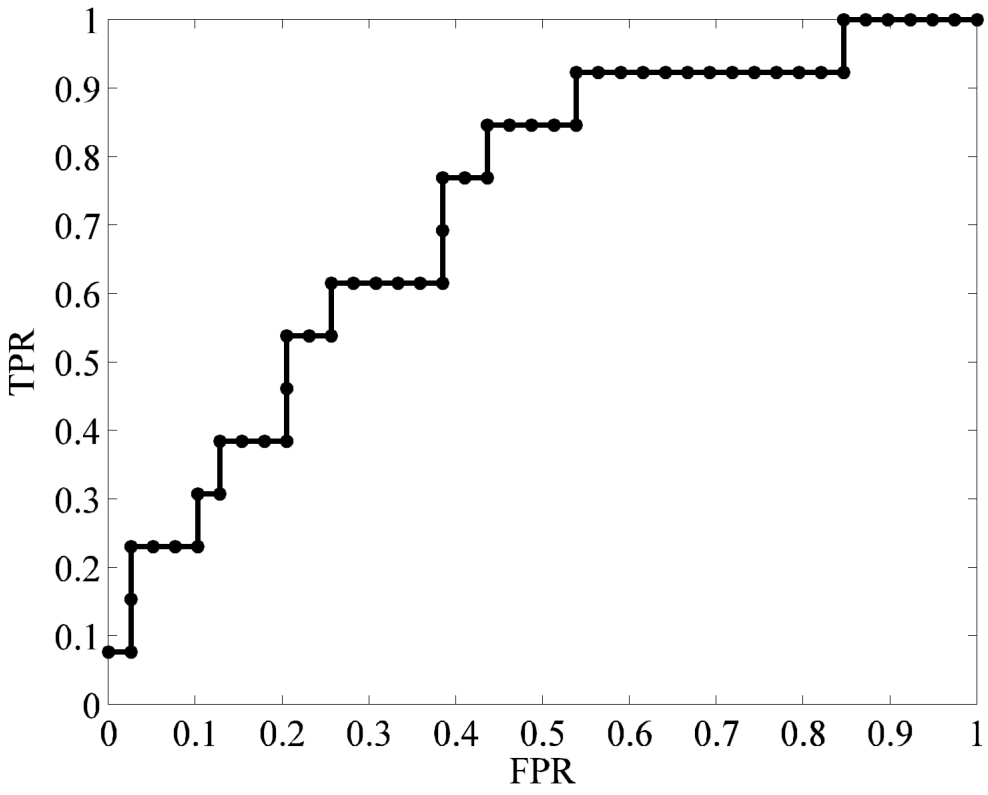

Figure 13. ROC curve for 70.2 Gy patient set based on model from largest PC – AUC=0.73.

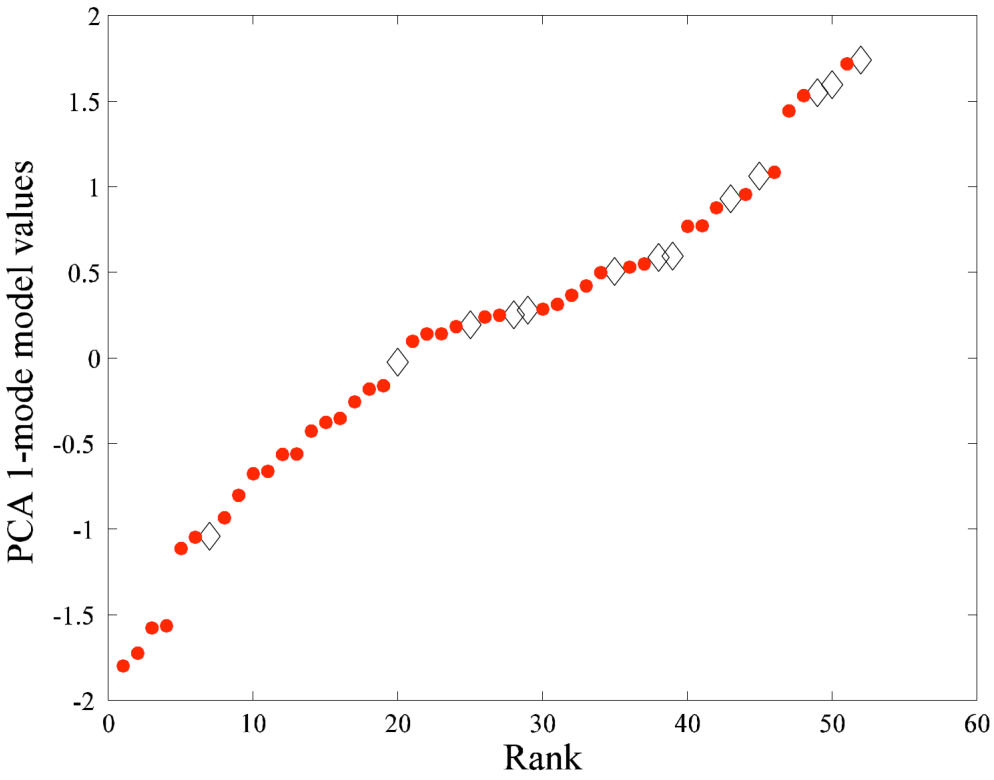

Figure 14. Ranked model value from 1st principal component for 70.2 Gy data. Bleeding patients are represented by open diamonds, non-bleeding patients are represented by closed circles.

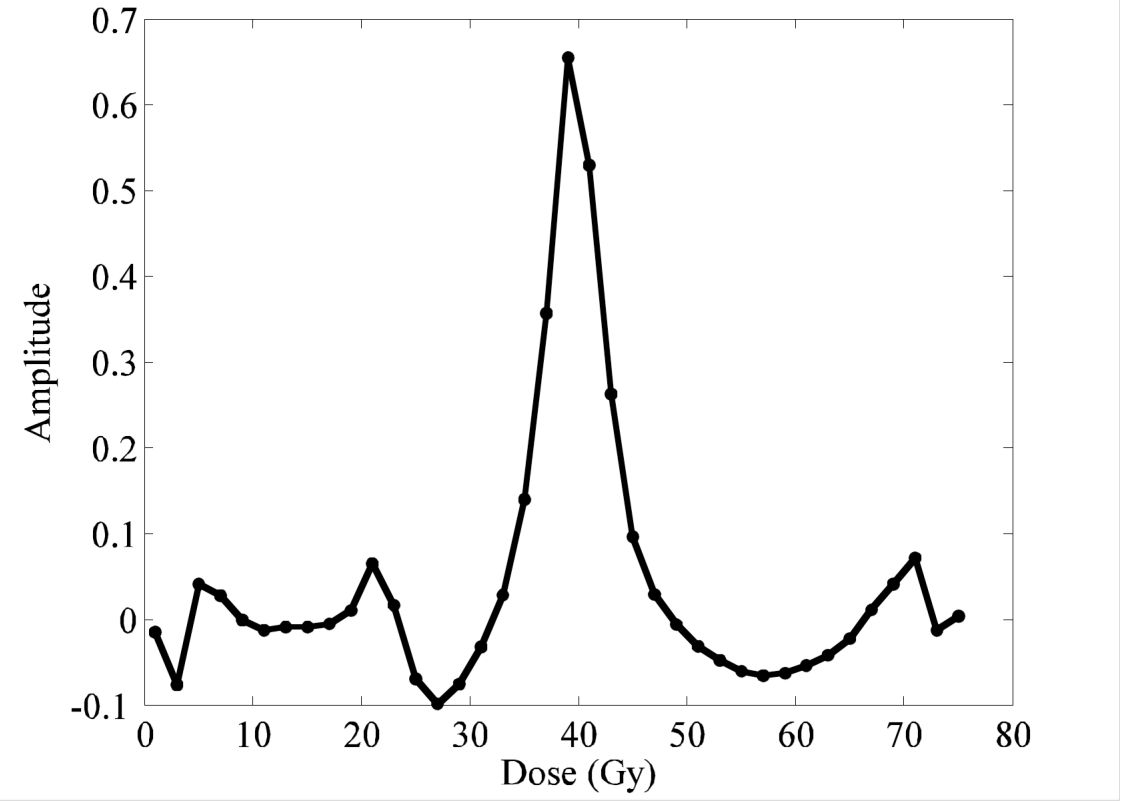

Figure 15. Varimax mode 6 for 70.2 Gy patient data.

3. Discussion and Conclusion

The PCA clearly shows us the strong correlations among the DVH variables, and emphasizes the difficulty of extracting unique dose volume combinations from clinical data such as ours. Additional techniques such as Varimax can only partially resolve these problems. Combination of our data with additional clinical data from complimentary treatments with a different pattern of correlations among dose volume variables may be the only way to fully resolve these issues.

The PCA analysis presented in this work on the MSKCC 3D-CRT patient data shows that a reduction in the number of degrees of freedom needed to describe outcome from the dose-volume histograms can be attained. One can use this reduction to form simple models relating outcome to the DVH. The 1st eigenvalue in the 75.6 Gy and 70.2 Gy data show common characteristics: a positive fluctuation across the entire DVH (see [5(a)] and [12]) and is related to variance in the mean dose. Higher order modes have more problematic interpretation as they oscillate wildly (see e.g. [5(b)]). Although the modes are uncorrelated with each other, and outcome models can be built from them, they do not allow us to clearly deduce unique DVH constraints for clinical use; although insight

into information contained in the DVHs may be inferred. We find that the variance in the DVH and outcome are not necessarily simply related. This is evident by the fact that the 1st PC is strongly related to outcome but one does not find additional correlation with outcome, in the 75.6 Gy data, in the subsequent modes until mode-7. Additionally we find that PC mode-6 is borderline and attains significance, in addition to modes 1 and 7, when particular patient data is excluded in a leave-one-out analysis or when a bootstrap analysis is performed. An analysis of the entirety of the data set, as presented, does not include mode-6 in the multivariate model.

The Varimax rotation brings us one step closer to our overall goal of deducing unique DVH constraints for clinical use. The Varimax modes are obtained by an orthogonal rotation of the PCA basis functions but now the support of each mode is maximally concentrated in one region of dose. Unfortunately the Varimax modes are once again correlated with each other; careful inspection of these modes is necessary to assess their independent value.

To deduce DVH constraints for use in treatment planning, one further step is required, a cost analysis of the consequences of imposing candidate DVH constraints. The cost analysis presented here (for purposes of illustrating our method) assumed that the cost of false positives and false negatives was equal. In reality a full analysis of these costs must take into account the likelihood and consequence of underdosing the tumor, in addition to the likelihood of rectal bleeding.

In this analysis, we used prostate DVHs, where the treatment plans are standardized and have applied each of our analyses to DVHs with one prescription dose. In the case of patients treated to different prescription doses or where the DVHs do not have the same regularities, for example Lung DVHs from non-small-cell lung cancer patients, we do not expect the PCA and Varimax analyses to give such concentration of variance in individual modes.

Acknowledgments

Funded in part by National Cancer Institute grant #PO1 – CA – 59017

One of the authors, J.D. Bauer, was supported by the Ruth Kirschstein National Research Service Award.

This work was performed in part under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

References

Bauer J D, Jackson A, Skwarchuk M, Zelefsky M, Leibel S, and Ling C 2004 Principal Component Analysis as Applied to Volume Effects in Treatment Outcomes for Patients with Prostate Cancer Treated with 3D-CRT *Medical Physics* **31** 1910

Bradley Andrew P 1997 The use of the area under the ROC curve in the evaluation of machine learning algorithms *Pattern Recognition* **30 1145-1159**

Dawson L A, Biersack M, Lockwood G, Math M, Eisbruch Avraham, Lawrence Theodore S , and Ten Haken, Randall K 2005 Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation *Int J Radiat Oncol Biol Phys* **62** 829-837

Hanley James A. and McNeil Barbara 1982 The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve *Radiology* **143** 29-36

Hanley James A. and McNeil Barbara 1983 The method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases *Radiology* **148** 839-843

Harman, Harry H. 1970 *Modern Factor Analysis, 2nd Ed. Rev.* University of Chicago Press, Chicago

Jackson A., Skwarchuk M W, Zelefsky M J, Cowen D M , Venkatraman E S, Levegrun S, Burman C M, Kutcher G J, Fuks Z, Liebel S A , and Ling C C 2001 Late rectal bleeding after conformal radiotherapy of prostate cancer. II Volume effects and dose-volume histograms *Int J Radiat Oncol Biol Phys* **49** 685-698

Joliffe Ian T 1995 Rotation of Principal Components: Choice of Normalization Constraints *Journal of Applied Statistics* **22** 29-35

Kleinbaum David G 2004 *Logistic Regression, A Self-Learning Text* Springer NY

Krzanowski W J 2000 *Principles of Multivariate Analysis A User's Perspective, Revised ed.* Oxford University Press Oxford

Lawton C A, Won M, Pilepic MV, et al. 1991 Long term treatment sequelae following external beam irradiation for adenocarcinoma of the prostate: Analysis of RTOG studies 7506 & 7706 *Int J Radiat Oncol Biol Phys*. **21** 935-939

Le Chap T 2003 *Introductory Biostatistics* Wiley Hoboken

The Mathworks Inc MATLAB Natick MA

McCullagh P and Nelder J A 1989 *Generalized Linear Models, 2nd Edition* Chapman and Hall London

Prentice R.L. and Pyke R 1979 Logistic disease incidence models and case-control studies *Biometrika* **66** 403-411

Skwarchuk M W, Jackson A, Zelefsky MJ, Venkatraman ES, Cowen DM, Levegrun S, Burman C M, Fuks Z, Leibel SA, Ling C C 2000 Late rectal toxicity after conformal radiotherapy of prostate cancer (I): multivariate analysis and dose-response *Int J Radiat Oncol Biol Phys.* 2000 **47,** 103-113