A spatial modelling technique using principal component analysis

Muhamed Baraković[1]**, Peter Rogelj**[2]

[1]*University of Verona, Department of Biotechnology*
Ca' Vignal 1, Strada Le Grazie 15, 37134 Verona, Italy
[2]*University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies*
Glagoljaška 8, 6000 Koper, Slovenia
E-mail: muhamed.barakovic@studenti.univr.it

Abstract

The purpose of this study was to develop a methodology for the construction of models of interest to improve the choice of areas to radiate in the use of brachytherapy (BT). This work aims to propose a principal component model which is constructed from the data of different patients including medical images of arbitrary resolution and modality supplemented with delineations of radiation target (HR-CTV) structure, reconstructed applicator structure and eventual organs at risk (OAR) structures. The principal component model provides information about the spatial variability described by only a few parameters. It can be used to predict specific extreme situations in the scope of sufficient applicator radiation dose coverage in the target structure as well as radiation dose avoidance in OAR structures.

1 Introduction

A model that can define components of the cancer's shape can have a fundamental role for the technique that wants to cure the cancerous tissue using radiotherapy.

In this work we wanted to develop a methodology to obtain data statistically using past and present data available on patients suffering from cancer of the cervix. The information required by each patient includes BT planning of medical images, delineated structure HR-CTV, the rebuilt structure applicator BT, and organs (OAR) structures at risk. HR-CTV and OAR structures are in each 3D image bounded to each slice image, where the specific structure is present and, therefore, available as a set of closed planar contours. Position of BT applicators is defined by anatomy such that applicator structures enable to put the structures of all patients into the same reference coordinate system. This is the starting point for constructing our model.

Due to high complexity of potential target structures BT, we have selected the most common representation of the structures by binary images. It is important that HR-CTV and OAR structures cannot overlap and by joining HR-CTV and OAR structure binary images we obtain three-level images with OARs represemted by -1, background as 0 and HR-CTV by 1. This property was used to simultaneously model both structure types without increasing the amount of data in the principal component analysis (PCA).

Typically the significance of tumor distribution depend on the type of tumor. It can help in the development of cancer treatment and biopsy strategies and techniques [1, 2]. In the case of cervical cancer it is also important due to analysis of different types of BT applicator and their improvements.

In the following sections our approach to model the spatial configuration of cervical cancer is described first. We describe the proposed method of building the principal component model using synthetic data. We conclude with a discussion that includes the analysis of the benefits and limitations.

2 Principal component model

The principal component model provides information of the BT target spatial variability expressed by only a small number of parameters. The general idea is to be able to reconstruct any target configuration, i.e., position and extent of HR-CTV as well as OAR structures, by correctly setting the model parameters. As such the principal component model can be used to predict various target configurations, e.g., extreme situations in the scope of sufficient applicator radiation dose coverage in the target structure as well as radiation dose avoidance in OARs. Such situations may be crucial for testing real applicator efficiency. The principal component model tends to extract a minimal set of orthogonal components of spatial variations in the region of interest using the principal component analysis. PCA projects the data into a lower dimensional linear space such that the variance of the projected data is maximized, or equivalently, it is the linear projection that minimizes the mean squared distance between the data points and their projections. PCA provides a full set of components that enable perfect data reconstruction, however, it also orders the components according to their importance, i.e., according to their contribution to the data description. It turns out that majority of the components have low importance and only a small error is made when only a few most important components are used. In this case the important components can be computed more efficiently using singular value docomposition (SVD) [3].

The data of each image is reordered into a row vector and joined for all the patients into a matrix $\mathbf{X}_{P\times L}$, with P the number of patients and L the number of pixels in the image. Then the mean vector $\overline{\mathbf{X}}$ is computed and

subtracted from each data row to obtain the matrix $\mathbf{X}_0$ representing the zero-mean data variation. SVD decomposes $\mathbf{X}_0$ into three matrices; matrix $\mathbf{V}$ with orthogonal columns that represent principal components, diagonal matrix $\mathbf{S}$ with singular values that represent importance of the components, and matrix $\mathbf{U}$ providing component weights for reconstructing the input data:

$$\mathbf{X}_0 = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{1}$$

The efficient SVD implementations, e.g., Matlab *svds* function, enable computation of only a given number of principal components R, and as such provide approximate solutions:

$$\mathbf{X}_{0\,P\times L} \approx \mathbf{U}_{P\times R}\mathbf{S}_{R\times R}\mathbf{V}^T_{L\times R} \tag{2}$$

The obtained matrices $\mathbf{S}$ and $\mathbf{V}$ represent a principal component model of the HR-CTV and OAR structures, such that they of any patient can be represented with the R components, i.e., the columns of $\mathbf{V}$, with weights:

$$\mathbf{U}' = \mathbf{X}'_0\mathbf{V}\mathbf{S}^{-1} \tag{3}$$

where $\mathbf{X}'_0 = \mathbf{X}' - \overline{\mathbf{X}}$ represents deviation of the data from the average. Similarly, BT target data can also be simulated by manual setting component weights in $\mathbf{U}$, following equation (2) and adding the mean vector $\overline{\mathbf{X}}$. Component weights form a low dimensional linear space with a certain region around the origin that corresponds to realistic data variation. The limits of this realistic subspace can be estimated by analyzing large amount of data, i.e., large number of patients. Values at the border of the realistic subspace can be used in $\mathbf{U}$ to simulate specific extreme situations suitable for BT applicator analysis.

The simulated or reconstructed data that results from the principal component model, as well as principal components themselves, can be reordered back into 3D images. Due to interpolations and approximations the reconstructed structures are not presented only with values 1, -1 and 0 for target structures, OAR structures and surrounding respectively. Consequently, we recommend completing the reconstruction procedure with thresholding using thresholds -0.5 and 0.5.

3 Results

The principal component model was tested using a simulated dataset that we have created for this purpose. Note that the simulated structure images presented here do not realistically simulate the BT target configuration, however enable illustration of the concept and testing of its suitability for creating a realistic model.

The simulated data was generated using four random parameters where three of them were used to simulate variability of the HR-CTV structure and the additional one for the variability of one OAR structure. The HR-CTV structure was simulated as an ellipsoid with the three parameters representing the semi-axes lengths while its center was always in the applicator coordinate system origin. The OAR region was simulated as a sphere with the given parameter representing its radius, while its center

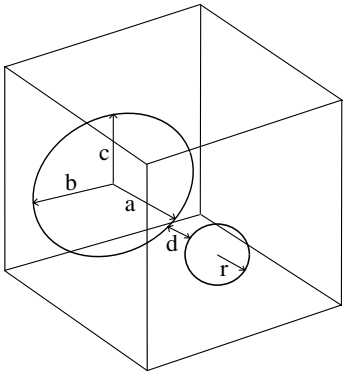

Figure 1: Illustration of the simulated dataset configuration. The HR-CTV was simulated with an ellipsoid and one additional OAR structure as a sphere with a constant distance d from the HR-CTV.

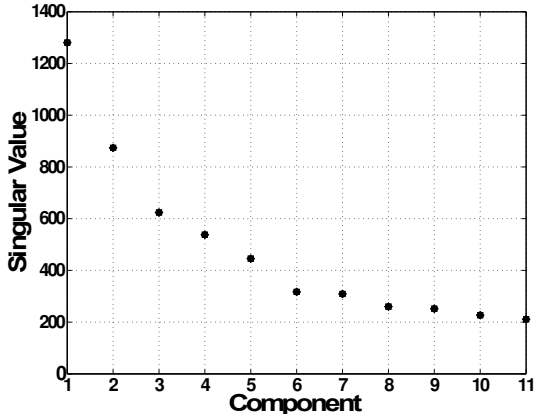

Figure 3: Singular values corresponding to the first 11 components; singular values represent the distribution of the dataset's energy.

was defined such that the distance between the edges of BT target and the OAR structure was costant. For the illustration see Figure 1. We used this solution in order to test the possibility to model rules, which represent interdependence of parameters and structures.

A principal component model was generated from a dataset of 400 simulated 3D images with $100 \times 100 \times 50$ voxels. The simulation parameters were selected randomly in the following ranges: a $\in [40, 73]$, b $\in [35, 59]$, c $\in [35, 49]$, r $\in [15, 20]$ and d $= 5$. The computation of the principal component model was restricted to 11 principal components. The mean image $\overline{\mathbf{X}}$ and the components are illustrated in Figure 2. The singular values that represent the distribution of the dataset's energy among the principal components indicate that the component energy gradually decreases with the component number, see Figure 3. However, although not all of the energy was considered, the reconstructed images did not differ considerably from the images from the training set as shown in Figure 4, where a randomly selected input structure image is compared with its reconstructed approximations obtained using three and eleven principal components. We can notice minor differences even when reconstructing from three components only.

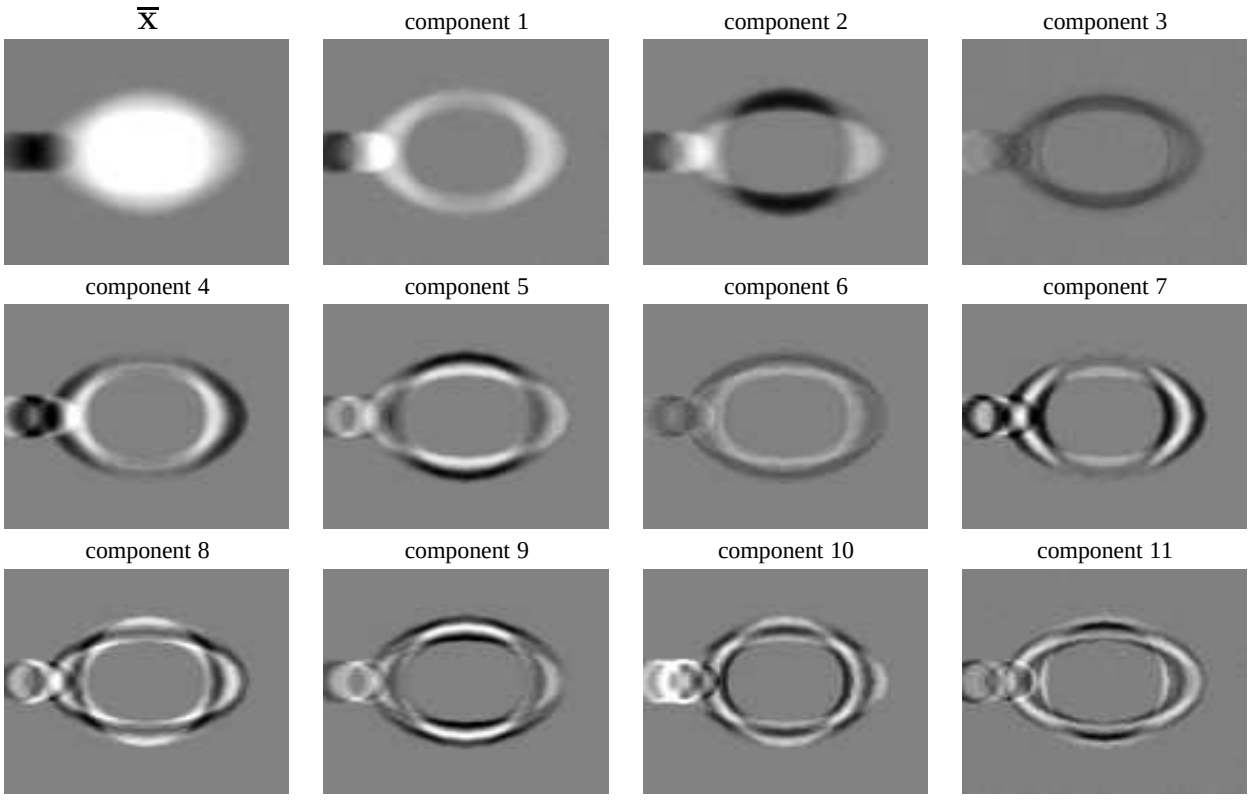

Figure 2: Components of the simulated dataset (central slices only). The mean image $\overline{\mathbf{X}}$ is presented in a scale from -1 (black) to +1 (white) and components with a scale from -0.01 (black) to +0.01 (white).

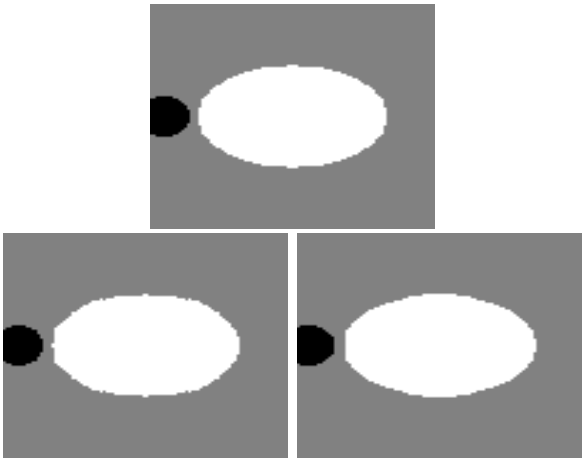

Figure 4: The central slice of an input structure image (top) and its reconstruction using 3 and 11 components (bottom left and right respectively).

If we observe the component weights (the values of matrix U), we can see that they are spread over a limited PCA subspace, see Figure 5, which corresponds to valid structure images. According to the shape of the subspace, we can conclude that component weights of valid images are not fully independent, although the components are orthogonal. By selecting weights manually, additional structure images can be simulated. If the selected weights are from the subspace of valid structure images, the simulated images follow the concepts of the input dataset, else the results may include major deviations as demonstrated in Figure 6.

The possibility to simulate structure images and have control over its validity offers good opportunity to generate specific synthetic images of the BT target region that represent extreme situations for BT applicator testing. In that case the principal component model should be created from real patient data and the test cases selected at the border of the populated PCA subregion.

The realistic model has not been created, yet, however we are looking forward to create it in collaboration with medical institutions that maintain large databases of their cervix cancer patients.

4 Discussion and conclusion

The applicator testing must take into account the BT target variability, e.g., by testing on diverse specific target configurations, which can correspond to real patients or obtained by modelling.

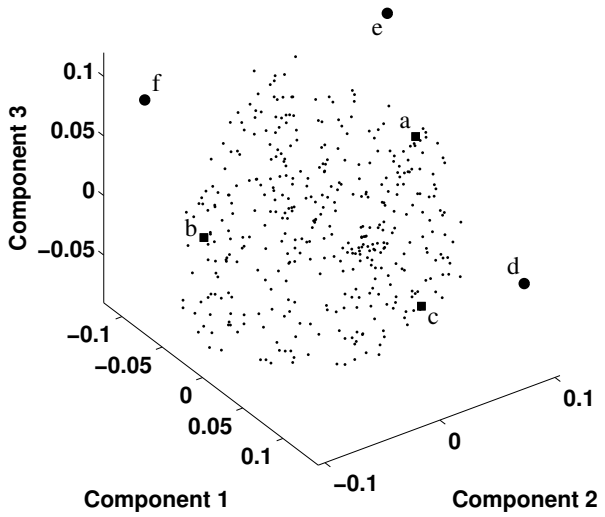

Figure 5: Weights for the first three components of the simulated structure images. The small dots correspond to images from the input dataset, squares and large dots represent selected values inside and outside the populated subspace for further simulations.

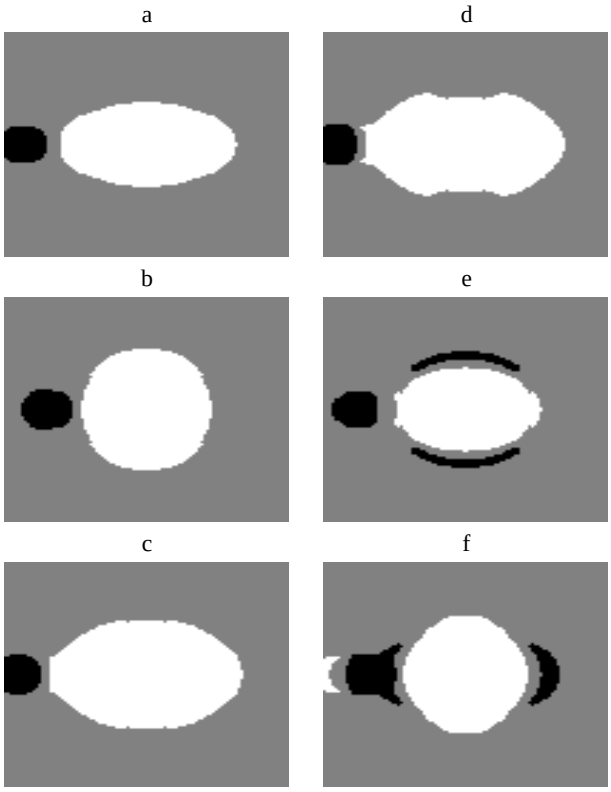

Figure 6: Central slices of simulated structure images using selected component weights from the subspace of valid structure images (left) and from other parts of the PCA space (right).

The proposed principal component model has advantages over using the real patients' data, because of the established control over the specificity of the cases, a possibility to simulate the non-existent cases and depersonalization.

The limitation of the principal model is in its high computational cost. Computation of all the PCA components would require enormous amount of memory, only the $\mathbf{V}$ matrix would have the size of 500k $\times$ 500k elements (assuming $2 \times 2 \times 2$ mm voxel size), which in float data format requires 1TB of memory. Using the SVD approach with computation of the most important components only, drastically reduces the memory requirements; in our simulated case matrix $\mathbf{V}$ occupied only 22MB. Such reduction of components is possible due to final thresholding, which is applicable due to binary nature of the structures. When the computational cost remains a problem, high efficient PCA solutions [4] or alternative structure representations could be used.

A principal component model of real cervix cancer has not been made, yet. A large number of patient datasets is required and in contrast to the spatial distribution model the OAR structures must be included. The preparation of such data is tedious due to non-standardized structure naming. However the benefits of such dataset are not only in the support of applicator development, but also in outcomes of further statistical analysis that could support clinical process, e.g., structure delineation or radiation planing, as well as making of clinical decisions.

To conclude, it may be widely accepted that reducing dose at organs of risk is difficult without reducing dose at large tumors [5], we believe that applicator improvements based on spatial modelling could provide better alternatives.

References

[1] M. B. Opell, J. Zeng, J. J. Bauer, R. R. Connelly, W. Zhang, I. A. Sesterhenn, S. K. Mun, J. W. Moul, and J. H. Lynch, "Investigating the distribution of prostate cancer using three-dimensional computer simulation." *Prostate Cancer Prostatic Dis*, vol. 5, no. 3, pp. 204–208, 2002.

[2] Y. Ou, D. Shen, J. Zeng, L. Sun, J. Moul, and C. Davatzikos, "Sampling the spatial patterns of cancer: optimized biopsy procedures for estimating prostate cancer volume and gleason score." *Med Image Anal*, vol. 13, no. 4, pp. 609–620, Aug 2009.

[3] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular Value Decomposition and Principal Component Analysis," *ArXiv Physics e-prints*, Aug. 2002.

[4] V. Zipunnikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu, "Multilevel functional principal component analysis for high-dimensional data," *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 852–873, 2011.

[5] R. Kim, A. F. Dragovic, and S. Shen, "Spatial distribution of hot spots for organs at risk with respect to the applicator: 3-d image-guided treatment planning of brachytherapy for cervical cancer," *International Journal of Radiation Oncology, Biology, Physics*, vol. 81, no. 2, S, p. S466, 2011.