Variable Importance in Nonlinear Kernels (VINK): Classification of Digitized Histopathology

Shoshana Ginsburg[1], Sahirzeeshan Ali[1], George Lee[2], Ajay Basavanhally[2], and Anant Madabhushi[1,⋆]

[1] Department of Biomedical Engineering, Case Western Reserve University, USA
[2] Department of Biomedical Engineering, Rutgers University, USA

Abstract. Quantitative histomorphometry is the process of modeling appearance of disease morphology on digitized histopathology images via image–based features (e.g., texture, graphs). Due to the *curse of dimensionality*, building classifiers with large numbers of features requires feature selection (which may require a large training set) or dimensionality reduction (DR). DR methods map the original high–dimensional features in terms of eigenvectors and eigenvalues, which limits the potential for feature transparency or interpretability. Although methods exist for variable selection and ranking on embeddings obtained via linear DR schemes (e.g., principal components analysis (PCA)), similar methods do not yet exist for nonlinear DR (NLDR) methods. In this work we present a simple yet elegant method for approximating the mapping between the data in the original feature space and the transformed data in the kernel PCA (KPCA) embedding space; this mapping provides the basis for quantification of variable importance in nonlinear kernels (VINK). We show how VINK can be implemented in conjunction with the popular Isomap and Laplacian eigenmap algorithms. VINK is evaluated in the contexts of three different problems in digital pathology: (1) predicting five year PSA failure following radical prostatectomy, (2) predicting Oncotype DX recurrence risk scores for ER+ breast cancers, and (3) distinguishing good and poor outcome p16+ oropharyngeal tumors. We demonstrate that subsets of features identified by VINK provide similar or better classification or regression performance compared to the original high dimensional feature sets.

1 Introduction

Due to the *curse of dimensionality* (COD), which precludes effective classification and prediction when the dimensionality of the feature space is high and the

⋆ Research reported in this publication was supported by the NSF Graduate Research Fellowship and the National Cancer Institute of the National Institutes of Health under Award Numbers R01CA136535-01, R01CA140772-01, R43EB015199-01, and R03CA143991-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 238–245, 2013.

sample size is small, stepwise feature selection algorithms or greedy techniques are often employed to reduce the number of features. However, these methods are computationally intensive and unstable, and they do not necessarily guarantee the optimal feature subset [1,2]. Another way to overcome the COD involves dimensionality reduction (DR) to transform high dimensional data via a linear (e.g., principal components analysis (PCA)) or nonlinear (e.g., kernel PCA) mapping to a low dimensional space, where classification can now be performed [3]. In the contexts of many biomedical problems nonlinear DR (NLDR) tends to provide better separation between target classes in the reduced dimensional space than linear DR schemes [4]. However, NLDR methods involve the eigen–decomposition of a kernel matrix rather than the data itself; consequently, the resulting variables are twice removed from their original physical meaning. The inability to interpret the features input to the classifier and thereby understand the underlying classification model is a significant limitation in the application of NLDR methods to biomedical problems. Consequently, there is a definite need for a variable ranking scheme to quantify the contributions of features in the high dimensional space to classification on the low dimensional embedding.

Recently we presented a variable ranking scheme to overcome this problem when PCA is used for DR. This variable importance for projections (VIP) score for PCA (PCA–VIP) quantifies the contributions of individual features to classification on a PCA embedding [5]. PCA–VIP exploits the mapping between the original feature space and the data in the PCA embedding space to quantify the contributions of individual features. The fact that this mapping is not explicitly defined in the context of NLDR limits the ability to extend PCA–VIP to NLDR methods.

In this work we present a simple yet elegant method for approximating the mapping between the data in the original feature space and the low–dimensional representation of the data provided by kernel PCA (KPCA); this mapping provides the basis for quantifying variable importance in nonlinear kernels (VINK). Since NLDR methods such as Isomap [6] and Laplacian eigenmap [7] can be formulated in terms of KPCA [3], VINK enables feature interpretation when these NLDR methods are used. VINK aims to quantify the contributions of individual features to classification or regression performance on a low–dimensional embedding, providing information about which features are useful and how useful they are. VINK may reveal that several similar, highly correlated features are all important, thereby identifying types of features that are useful for a specific classification task.

We evaluate the ability of VINK to identify a subset of features that provide classification or regression performance similar to or better than the entire high dimensional feature set in three distinct problem domains. In each case, tumor appearance of disease morphology on digitized histopathology images is modeled by nuclear, architectural, morphological, and textural features. The implicit assumption is that the morphology and architectural arrangement of tumor cell nuclei or glands encode information about the tumor's current status and prognosis. We demonstrate that VINK is able to successfully identify high–performing features

when either KPCA with a Gaussian kernel (Gaussian KPCA) or Isomap and Laplacian eigenmap variants of KPCA are used for NLDR. By revealing the specific features that contribute most to the embeddings, VINK could have a significant impact on adoption of these NLDR schemes for biomedical applications.

2 NLDR Methods as Variants of Kernel PCA

2.1 Review of Kernel PCA

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be a matrix of n m–dimensional observations, and let $\mathbf{y} \in \mathbb{R}^n$ denote a vector of outcome variables. PCA [8] attempts to find a linear transformation to maximize the variance in $\mathbf{X}$ and applies this transformation to obtain the most uncorrelated features. The orthogonal eigenvectors of $\mathbf{X}$ express the variance in the data, and the h eigenvectors that comprise most of the variance in the data are the principal components. Thus, PCA forms the following model:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\mathsf{T}, \tag{1}$$

where $\mathbf{T}$ is made up of the h principal component vectors $\mathbf{t}_i$, $i \in \{1, ..., h\}$, as columns and $\mathbf{P}^\mathsf{T}$ is comprised of the h loading vectors $\mathbf{p}_i$ as rows.

Given a series of observations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, KPCA computes the principal components of $\mathbf{\Phi}(\mathbf{x}_1), \mathbf{\Phi}(\mathbf{x}_2), ..., \mathbf{\Phi}(\mathbf{x}_n)$. The nonlinear function $\mathbf{\Phi}$ need not be explicitly defined; only the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with elements $k_{ij} = \mathbf{\Phi}(\mathbf{x}_i)^\mathsf{T}\mathbf{\Phi}(\mathbf{x}_j)$ needs to be known. This kernel matrix may represent a Gaussian kernel, a radial basis function kernel, or a similarity or dissimilarity matrix. Diagnolization of the kernel matrix gives rise to a low–dimensional embedding of the data. In general, KPCA forms the following model:

$$\mathbf{K} = \mathbf{T}\mathbf{Q}^\mathsf{T}, \tag{2}$$

where $\mathbf{T}$ comprises the principal components of $\mathbf{K}$ and $\mathbf{Q}$ contains the KPCA loading vectors.

2.2 Isomap and Laplacian Eigenmap as Variants of KPCA

Ham et al. [3] demonstrated that NLDR methods such as Isomap [6] and Laplacian eigenmap [7], which preserve the local neighborhood structure in data while globally mapping a manifold, can be formulated in terms of KPCA. More specifically, they showed that KPCA using a kernel matrix that contains the geodesic distances between points is equivalent to Isomap, and KPCA using the kernel matrix $\mathbf{K} = \mathbf{L}^\dagger$, where $\dagger$ denotes the pseudo–inverse and $\mathbf{L}$ is defined as

$$\mathbf{L}_{ij} = \begin{cases} \sum_{i \neq j} e^{-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\sigma^2}} & \text{if } i = j \\ -e^{-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\sigma^2}} & \text{if } i \neq j \\ 0 & \text{else} \end{cases},$$

is equivalent to Laplacian eigenmap [3].

3 Variable Importance in Projections (VIP)

3.1 VIP for PCA

A variable importance in projections (VIP) scoring system was introduced in [9] to quantify the contributions of individual features to regression on a partial least squares embedding and was later extended to quantify the contributions of features to classification on a PCA embedding [5]. The VIP for PCA is computed for each feature as follows:

$$\pi_j = \sqrt{m \frac{\sum_{i=1}^{h} b_i^2 \mathbf{t}_i^\mathsf{T} \mathbf{t}_i \left(\frac{p_{ji}}{||\mathbf{p}_i||}\right)^2}{\sum_{i=1}^{h} b_i^2 \mathbf{t}_i^\mathsf{T} \mathbf{t}_i}}, \tag{3}$$

where m is the number of features in the original, high–dimensional feature space and the b_i are the coefficients that solve the regression equation

$$\mathbf{y} = \mathbf{T}\mathbf{b}^\mathsf{T}, \tag{4}$$

which correlates the transformed data with the outcome vector $\mathbf{y}$. The degree to which a feature contributes to classification in the PCA transformed space is directly proportional to the square of its VIP score π. Thus, features with VIP scores near 0 have little predictive power, and the features with the highest VIP scores contribute the most to class discrimination in the PCA embedding space.

3.2 Variable Importance for Nonlinear Kernels (VINK)

The primary contribution of this work is the extension of the VIP scheme to KPCA and thus to Isomap and Laplacian eigenmap. This extension is non–trivial because an essential component of the VIP score is the the fraction $\left(\frac{p_{ji}}{||\mathbf{p}_i||}\right)^2$, which reveals how much the j^{th} feature contributes to the i^{th} principal component in the low–dimensional embedding. Computing this term requires knowledge of the mapping $\mathbf{P}$ that relates the transformed data $\mathbf{T}$ to the original feature space (see Equation (1)). Whereas in PCA the matrix $\mathbf{P}$ is comprised of the loading vectors, in KPCA the transformed data is not directly related to the original data, but only to the kernel matrix. Furthermore, the mapping $\mathbf{\Phi} : \mathbf{X} \rightarrow \mathbf{K}$ that relates the kernel matrix to the original data is not necessarily computed; rather, $\mathbf{\Phi}$ is only implicitly defined, as is evident in Equation (6):

$$\mathbf{X} = \mathbf{K}\mathbf{\Phi}^\mathsf{T}. \tag{5}$$

Although knowledge of the mapping $\mathbf{P}$ is a prerequisite for implementing the VIP scheme, there is no closed–form solution for $\mathbf{P}$. Consequently, we suggest that the mapping $\mathbf{P}$ be approximated as follows. Combining equations (2) and (6) yields

$$\mathbf{X} = \mathbf{T}(\mathbf{\Phi}\mathbf{Q})^\mathsf{T}. \tag{6}$$

Note that $\mathbf{\Phi Q}$ represents the mapping that relates the transformed data to the original data. Consequently, rather than estimating $\mathbf{\Phi}$, we only compute the matrix $\mathbf{P} = \mathbf{\Phi Q}$. Because $\mathbf{T}$ is often singular, it follows from equation (2) that

$$\mathbf{P}^{\mathsf{T}} \approx \mathbf{T}^{\dagger}\mathbf{X}. \tag{7}$$

Obtaining an approximation $\mathbf{P}' = \mathbf{T}^{\dagger}\mathbf{X}$ of the mapping $\mathbf{P}$ provides the link for extending VIP to VINK, allowing for the calculation of VINK scores using the VIP formulation in Equation (3).

3.3 General Framework for Evaluating VINK

Evaluation of VINK in conjunction with PCA, Gaussian KPCA, Isomap, and Laplacian eigenmap involved the following steps:

1. **Perform dimensionality reduction**: Embed the high–dimensional data in a low–dimensional space.
2. **Identify most contributory features**: Compute VIP and VINK scores for each feature, and identify features with the highest VIP and VINK scores.
3. **Compare performance of top features with all features**: Use the k features with the highest VIP or VINK scores to embed the data in a low–dimensional space, where classification or regression is performed. The performance of models using these k selected features—referred to as $C_{\Lambda}^{k^*}$ for $\Lambda \in \mathcal{C}, \mathcal{C} = \{PCA, KPCA, Iso, LE\}$—is compared to models using all m features: $C_{\Lambda}^{m}\ \forall \Lambda \in \mathcal{C}$. Regression and classification performance were evaluated by mean squared error (MSE) and area under the receiver operating characteristic (ROC) curve (AUC), respectively.

To evaluate VIP and VINK scores associated with each feature, we bootstrap–sampled 75% of the data to construct the low–dimensional embeddings in order to obtain VIP and VINK scores; the perfomance of the classifiers was evaluated on the remaining 25% of the data. This randomized bootstrapping process was repeated 30 times, and the average VIP and VINK scores and performance measures were obtained.

Table 1. Summary of histopathology datasets and corresponding analyses

Application	Analysis	Features	Objective
Prostate cancer (40)	Classification	100 morphological 51 architectural 52 texture 39 graph–based	Predict PSA failure
Breast cancer (118)	Regression	25 graph–based 25 nuclear	Predict Oncotype DX
Oropharyngeal cancer (65)	Classification	7 graph–based	Identify progressors

4 Experimental Results and Discussion

4.1 Predicting PSA Failure Following Radical Prostatectomy

Texture and graph–based features extracted from digitized histopathology images (see Table 1, Figure 1) were used to predict five year biochemical recurrence following radical prostatectomy in men with prostate cancer [10]. When C^5_{PCA}, C^5_{KPCA}, and C^5_{LE} are used to construct low–dimensional representations of the data, AUC remains the same or decreases slightly compared to C^m_{PCA}, C^m_{KPCA}, and C^m_{LE} (see Figure 2(a)). The fact that classification performance is not adversely affected by reducing the number of features from 242 to 5 suggests that the features associated with high VINK scores are indeed the primary contributors to class discrimination. We note that C^5_{Iso} provides a substantial increase in AUC compared to both C^m_{Iso} and the other DR methods. Regardless of which DR method is used, the features with the highest VINK scores are the morphological and architectural features describing gland proximity, a hallmark attribute within the Gleason grading framework [10].

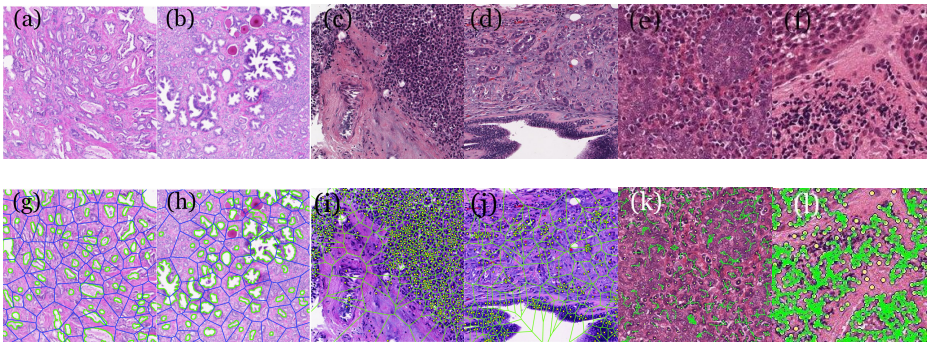

Fig. 1. Representative histologic images from resected prostates for men (a) with and (b) without five year biochemical recurrence; ER+ breast cancer corresponding to (c) high and (d) low Oncotype DX scores; and a (e) progressor and (f) non-progressor p16+ oropharyngeal cancer. (g)-(l) Graph–based representations using glands and nuclei as vertices for panels (a)-(f).

4.2 Predicting Oncotype DX Scores for ER+ Breast Cancers

Fifty features describing nuclear arrangement on digitized histopathology (see Table 1, Figure 1) were used to predict the Oncotype DX (ODx) risk of distant recurrence post surgery for ER+ breast cancer [11]. MSE remains the same for C^k_{PCA}, C^k_{KPCA}, and C^k_{LE} for $k = m$ and $k = 5$, although MSE increases slightly for C^5_{Iso} over C^m_{Iso} (see Figure 2(b)). The fact that feature reduction hardly impacts regression performance suggests that features associated with

high VINK scores are indeed the primary contributors to discrimination between high and low ODx scores in the embedding spaces.

It is highly significant that the top features are identical regardless of whether PCA, Isomap, or LE is used for DR. The top features included the disorder of Delaunay triangle areas and edge lengths and the disorder of minimum–spanning–tree triangle edge lengths. The importance of quantitative measures of nuclear disorder is not surprising since graph triangles tend to be more uniformly sized on low grade cancers (grade is known to be correlated to ODx score), whereas nuclear arrangement is more irregular in higher grade cancers (see Figures 1(i)-(j)).

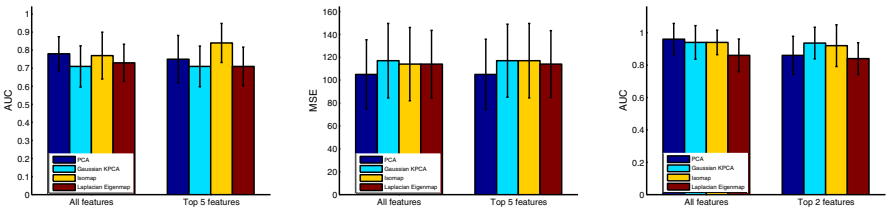

Fig. 2. AUC values for (a) predicting five year PSA failure following radical prostatectomy and (c) identifying progressors among oropharyngeal cancers and (b) MSE in predicting Oncotype DX scores for ER+ breast cancers.

4.3 Identification of Progressors Among Oropharyngeal Cancers

Seven features computed via a cell cluster graph (see Figure 1) were used to predict progression of p16+ oropharyngeal tumors [12]. AUC values are somewhat lower for C^2_Λ than for C^m_Λ $\forall \Lambda \in \mathcal{C}$ (see Figure 2(c)). Because this dataset contained only seven features, the COD does not appear to have as acutely affected this dataset as much as the prostate and breast cancer datasets. Since all features are relevant and have near–average VINK scores, classification performance is excellent when all seven features are used. Consequently, eliminating the five features that contribute least to classification on the embeddings leads to some deterioration in classification performance.

5 Concluding Remarks

In this paper we presented variable importance for nonlinear kernels to quantify the contributions of individual features to classification or regression in kernel PCA embedding spaces. We showed that VINK can be implemented in conjunction with Isomap and Laplacian eigenmap, two popular nonlinear dimensionality reduction schemes. In the contexts of three different problems involving digital pathology and quantitative histomorphometry, we demonstrated that VINK succeeds in identifying high–dimensional features that perform almost as well

as the entire high–dimensional feature set. In particular, nuclear disorder was found to be predictive of Oncotype DX recurrence risk of ER+ breast cancer, and proximities among glands were found to be important for predicting PSA failure five years post radical prostatectomy. The fact that the top features selected by VINK were the same for several DR methods serves to corroborate their importance and attests to the robustness of VINK in identifying important features.

References

1. Saeys, Y., Inza, I., Larranaga, P.: A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics 23, 2507–2517 (2007)
2. Yan, H., et al.: Correntropy Based Feature Selection using Binary Projection. Pattern Recogn. 44, 2834–2842 (2011)
3. Ham, J., et al.: A Kernel View of the Dimensionality Reduction of Manifolds. Max Planck Institute for Biological Cybernetics, Technical Report No. TR-110 (2002)
4. Shi, J., Luo, Z.: Nonlinear Dimensionality Reduction of Gene Expression Data for Visualization and Clustering Analysis of Cancer Tissue Samples. Computers Biol. Med. 40, 723–732 (2010)
5. Ginsburg, S., Tiwari, P., Kurhanewicz, J., Madabhushi, A.: Variable Ranking with PCA: Finding Multiparametric MR Imaging Markers for Prostate Cancer Diagnosis and Grading. In: Madabhushi, A., Dowling, J., Huisman, H., Barratt, D. (eds.) Prostate Cancer Imaging 2011. LNCS, vol. 6963, pp. 146–157. Springer, Heidelberg (2011)
6. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290, 2319–2323 (2000)
7. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Comput. 15, 1373–1396 (2003)
8. Esbensen, K.: Multivariate Data Analysis—In Practice: An Introduction to Multivariate Data Analysis and Experimental Design. CAMO, Norway (2004)
9. Chong, I.G., Jun, C.H.: Performance of Some Variable Selection Methods when Multicollinearity is Present. Chemometr. Intell. Lab 78, 103–112 (2005)
10. Golugula, A., et al.: Supervised Regularized Canonical Correlation Analysis: Integrating Histologic and Proteomic Measurements for Predicting Biochemical Recurrence Following Prostate Surgery. BMC Bioinformatics 12, 483–495 (2011)
11. Basavanhally, A., et al.: Multi–Field–of–View Framework for Distinguishing Tumor Grade in ER+ Breast Cancer from Entire Histopathology Slides. IEEE Trans. Biomed. Eng. (Epub ahead of print) (PMID: 23392336)
12. Ali, S., et al.: Cell Cluster Graph for Prediction of Biochemical Recurrence in Prostate Cancer Patients from Tissue Microarrays. In: Proc. SPIE Medical Imaging: Digital Pathology (2013)