Published in final edited form as:
Comput Vis Image Underst. 2013 September 1; 117(9): 1138–1146. doi:10.1016/j.cviu.2012.11.011.

Statistical Shape Model for Manifold Regularization: Gleason grading of prostate histology

Rachel Sparks[a,b] and **Anant Madabhushi**[b]

Rachel Sparks: rsparks@eden.rutgers.edu; Anant Madabhushi: axm788@case.edu

[a]Department of Biomedical Engineering, Rutgers University, Piscataway, NJ, 08854

[b]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, 44106

Abstract

Gleason patterns of prostate cancer histopathology, characterized primarily by morphological and architectural attributes of histological structures (glands and nuclei), have been found to be highly correlated with disease aggressiveness and patient outcome. Gleason patterns 4 and 5 are highly correlated with more aggressive disease and poorer patient outcome, while Gleason patterns 1–3 tend to reflect more favorable patient outcome. Because Gleason grading is done manually by a pathologist visually examining glass (or digital) slides subtle morphologic and architectural differences of histological attributes, in addition to other factors, may result in grading errors and hence cause high inter-observer variability. Recently some researchers have proposed computerized decision support systems to automatically grade Gleason patterns by using features pertaining to nuclear architecture, gland morphology, as well as tissue texture. Automated characterization of gland morphology has been shown to distinguish between intermediate Gleason patterns 3 and 4 with high accuracy. Manifold learning (ML) schemes attempt to generate a low dimensional manifold representation of a higher dimensional feature space while simultaneously preserving nonlinear relationships between object instances. Classification can then be performed in the low dimensional space with high accuracy. However ML is sensitive to the samples contained in the dataset; changes in the dataset may alter the manifold structure. In this paper we present a manifold regularization technique to constrain the low dimensional manifold to a specific range of possible manifold shapes, the range being determined via a statistical shape model of manifolds (SSMM). In this work we demonstrate applications of the SSMM in (1) identifying samples on the manifold which contain noise, defined as those samples which deviate from the SSMM, and (2) accurate out-of-sample extrapolation (OSE) of newly acquired samples onto a manifold constrained by the SSMM. We demonstrate these applications of the SSMM in the context of distinguish between Gleason patterns 3 and 4 using glandular morphologic features in a prostate histopathology dataset of 58 patient studies. Identifying and eliminating noisy samples from the manifold via the SSMM results in a statistically significant improvement in area under the receiver operator characteristic curve (AUC), 0.832 ± 0.048 with removal of noisy samples compared to a AUC of 0.779 ± 0.075 without removal of samples. The use of the SSMM for OSE of newly acquired glands also shows statistically significant improvement in AUC, 0.834 ± 0.051 with the SSMM compared to 0.779 ± 0.054 without the SSMM. Similar results were observed for the synthetic Swiss Roll and Helix datasets.

Keywords

Manifold Learning; Statistical Shape Models; Regularization; Prostate Histology; Gleason Grading

1. Introduction

Blinded needle sextant biopsy is the current gold standard for prostate cancer (CaP) diagnosis; each biopsy yields 12–18 needle cores which are then analyzed under a microscope by a pathologist [1, 2]. If CaP is identified, a pathologist will then assign a Gleason score to the biopsy samples, determined as a summation of the two most prevalent Gleason patterns which range from 1 to 5, hence, Gleason score has a range of 2–10 [3]. Low Gleason patterns (1–3) are characterized by a coherent spatial architecture with distinct gland lumen surrounded by cell nuclei [3, 4]. For higher Gleason patterns (4 and 5), the arrangement and morphology of histological structures begins to breakdown with gland lumen becoming indistinct and crowded with an increase in the concentration of cell nuclei. The most dominant Gleason patterns seen on needle biopsies are patterns 3 and 4 [5]. Accurately distinguishing intermediate Gleason patterns 3 and 4 manually is a difficult problem; previous studies having reported an inter-observer agreement between pathologists of 0.47–0.64 (reflecting low to moderate agreement) [6, 7]. Gleason score aids in determining the course of treatment, patients with less aggressive CaP (Gleason score 6 and under) may be enrolled in active surveillance programs while patients with more aggressive CaP (Gleason score 7 and above) will undergo radiation therapy or surgery [8].

Availability of digital prostate histology samples [9] has led to the development of computer assisted decision support tools which allow for quantification of subtle morphologic changes in prostate tissue and may potentially allow for better and more reproducible discrimination between Gleason patterns [10–19]. Previous attempts at building computer aided decision support tools for Gleason scoring have employed the use of image texture [10, 14, 18], arrangement and morphology of nuclei [12, 13, 15, 19], or morphology of glands [16, 17]. However, a large number of features are typically necessary to accurately perform Gleason grading of histology images, resulting in a high dimensional feature space [12, 15]. The high dimensional feature space may have more dimensions than samples in the dataset, referred to as the curse of dimensionality, which makes classification infeasible [20].

Dimensionality reduction offers a way to overcome the curse of dimensionality by constructing a low dimensional space in which to perform classification while not compromising object-class relationships. Manifold learning (ML) refers to a class of nonlinear dimensionality reduction methods that aim to learn a low dimensional embedding space that can preserve subtle relationships between samples in the high dimensional space [21–23]. Previous applications of ML to histopathology datasets have demonstrated that the low dimensional embedding space is better suited to classification compared to the original high dimensional space [11, 17, 24].

ML finds a low dimensional manifold representation $\mathcal{M}$ from a dataset **O** which preserves relationships between samples in **O**. Most ML methods assume that **O** is contained in a high dimensional space $\mathbb{R}^D$ [21–23, 25]. Additionally, ML assumes that **O** is densely clustered and concentrated within a small region of $\mathbb{R}^D$. An example of dense clustering can be seen in Figure 1(a) which shows an example of the synthetic Swiss Roll dataset. In $\mathbb{R}^3$ the samples cluster along a $2D$ planar structure.

To calculate $\mathcal{M}$, ML techniques typically construct a dissimilarity matrix A which quantifies dissimilarity between samples in $\mathbb{R}^D$ [21–23]. For a dataset **O** containing N samples, A is a

$N \times N$ dissimilarity matrix defined such that $A(o_i, o_j)$ quantifies the differences between the samples $o_i, o_j \in \mathbf{O}$. Typically $A(o_i, o_j) = \psi(o_i, o_j)$ where $\psi(\cdot, \cdot)$ is a dissimilarity measure (e.g. heat kernel [11, 26], geodesic distance [22], Diffeomorphic Based Similarity [17]) which is dataset and feature set dependent. ML then calculates $\mathcal{M}$ to best preserve the relationships in A. ML techniques preserve relationships in A differently, some methods such as Local Linear Embedding (LLE) [21] try to preserve the local neighborhood relationships between samples. Isomaps [22] and Graph Embedding (GE) [23] find the best embedding space to preserve the global structure of A, albeit with different underlying algorithms.

ML schemes tend to be sensitive to the dataset considered, and changes in the dataset may cause changes to the learned manifold [27]. Consider a sample $o_i \in \mathbf{O}$ perturbed by some error ε; the new location for o_i would be $\hat{o}_i = o_i + \varepsilon$. A would have to be altered such that $\hat{A}(\hat{o}_i, o_j) = \psi(\hat{o}_i, o_j)$ for all o_j contained in $\mathbf{O}$, resulting in changes to $2N - 2$ elements in A. The manifold $\hat{\mathcal{M}}$ learned from $\hat{A}$ will reflect those changes. Hence even a small change in $\mathbf{O}$ may cause large changes to $\mathcal{M}$. Figure 2 demonstrates this phenomenon for a prostate histology dataset comprising 888 glands. Two manifolds were generated by applying ML to 90% of samples in the dataset (800 glands) such that for each manifold a different set of 88 samples were excluded. Each manifold has a distinct structure evident by (a) changes in the planar structure of the manifold and (b) changes in object-class relationships on the manifold, displayed as color differences between manifolds.

Consider a large dataset $\mathbf{O}$ from which the manifold $\mathcal{M}$ is generated. In the absence of knowing the true manifold, $\mathcal{M}$ is the best manifold representation to capture the relationships between samples in the dataset. If we consider a subset $\hat{\mathbf{O}} \subset \mathbf{O}$ then $\hat{\mathbf{O}}$ can be used to create an alternative manifold $\hat{\mathcal{M}}$ which approximates $\mathcal{M}$. Manifold regularization constrains the structure of $\hat{\mathcal{M}}$ giving a better approximation of $\mathcal{M}$ and hence resulting in a better representation of the relationships between samples in $\hat{\mathbf{O}}$.

In this work we present a statistical shape model of manifolds (SSMMs) to perform manifold regularization. SSMM merges the theory of ensemble learning [28] with statistical shape models (SSMs) [29]. The theory behind ensemble learning is that an ensemble of weak classifiers has higher accuracy compared to any single weak classifier [30]. Similarly, consensus clustering takes a ensemble of weak clusterings of a dataset, obtained by applying an unstable clustering method such as k-mean clustering to a dataset multiple times, and combines the ensemble to generate a strong clustering of the dataset [31]. Viswanath *et. al.* [32] demonstrated that consensus embedding, obtained by generating an ensemble of manifolds from a single dataset, produced a more stable low dimensional manifold compared to any single manifold in the ensemble. In this work we hypothesize an ensemble of manifolds will have a more accurate representation of the manifold shape that any single manifold. The concept of SSMMs is that an ensemble of manifolds can be modeled with a SSM. SSMs have been proposed to model shape variation in anatomical structures [29]. In much the same way, we utilize a SSM to model which manifold shapes are statistically most likely to occur. The SSMM describes the maximum likelihood estimate (MLE) of the manifold shape and primary modes of variation for a series of different manifolds constructed by randomly selecting a subset of samples from a dataset. For a new, related dataset, the resulting manifold can be constrained to only the range of shapes dictated by the SSMM. Hence every sample on the new manifold is spatially and locally constrained to within 2 standard deviations of its location on the average manifold shape.

The SSMM can be utilized in several ways. (1) Regions on a new, related manifold which deviate from the SSMM can be identified. By identifying these regions, meaningful differences between the dataset and the SSMM may be determined. (2) Noisy samples on a

manifold can be identified based on their deviation from the SSMM. Removing these samples from the dataset may result in a more accurate low dimensional manifold, and hence improve classification accuracy. (3) A classifier can be trained on the SSMM which would allow for (a) classifier decision boundaries to be applied to a new, related manifold without retraining the classifier or (b) new, related samples could be projection onto the SSMM. The projection of newly acquired samples onto a previously calculated manifold can be performed by out-of-sample extrapolation (OSE) [33].

The remainder of the paper is organized as follows. Section 2 describes previous work in Gleason grading of prostate histology and manifold regularization. An overview of SSMM construction and its novel contributions are discussed in Section 3. In Section 4 an ensemble based theoretical framework for the SSMM is presented. Section 5 describes the methodology to construct a SSMM and its application to (a) outlier identification and (b) OSE of newly acquired samples onto the SSMM. Section 6 describes the experimental design and results for two synthetic datasets as well as a prostate histology dataset. Concluding remarks are presented in Section 7.

2. Previous Work

2.1. Automated Gleason Grading

Pathologists perform Gleason grading of prostate cancer tissue specimens via qualitative, visual evaluation of a tissue section previously stained with Hemotoxilyin and Eosin (H& E) [3]. The primary discriminating traits of Gleason patterns on histopathology are the difference in the arrangement and morphology of the nuclei and glands within a tissue sample [3, 4]. In devising automated pattern recognition methods for distinguishing different Gleason patterns on histopathology, the key questions to consider are (1) what is the best feature set to distinguish between Gleason patterns? and (2) what is the best method to reduce the dimensionality of the feature set prior to classification?

Jafari *et. al.* [10] characterized tissue patch texture via wavelet features and classified Gleason patterns with an accuracy of 97% for the best performing feature. Huang *et. al.* [14] characterized tissue patch texture via Fractal Dimension and achieved an accuracy of 95%. However, a limitation of these approaches were that the image patches were manually selected to obtain regions which contained only one tissue class on the digitized slide. DiFranco *et. al.* [18] characterized tissue patch texture for each color channel independently showing 90% accuracy classifying images on a per tile. Although tiles were automatically determined, tiles which contained more than one tissue class were removed from the dataset.

Structural features (as opposed to texture features) have also been explored by some researchers for automated categorization of Gleason patterns. Veltri *et. al.* [13] and Ali *et. al.* [19] showed that the quantitative characterization of the shape of individual nuclei on tissue microarrays can distinguish between Gleason patterns with high accuracy. In a preliminary study by Veltri *et. al.* [13] characterization of manually segmented nuclei were able to distinguish between Gleason pattern 3, 4, and 5 with 73–80% accuracy. Ali *et. al.* [19] automated the nuclear segmentation and classification steps in [13], yeilded an 84% accuracy on 80 tissue microarrays. Doyle *et. al.* [15] characterized manually selected image patches according to nuclear arrangement, reporting a predictive positive value of 76.0% in distinguishing between Gleason patterns 3, 4, and 5 within a multi-classification scheme.

In previous work we have shown that gland morphology, quantified by Diffeomorphic Based Similarity (DBS), is able to distinguish between Gleason 3 and 4 patterns with 88% accuracy in 58 patient studies [17]. DBS is calculated by constructing shape models for each

gland contained in a set of histology images and then quantifying the differences between shape models.

Tabesh *et. al.* [12] combined gland morphology, texture features, color channel variance, and nuclear arrangement to classify different Gleason patterns with 81.0% accuracy. Golugula *et. al.* [24] used proteomic data in conjunction with histology derived image features to distinguish between prostate cancer patients who following radical prostatectomy had biochemical recurrence within 5 years from patients who did not.

Most automated Gleason grading systems are described by a high dimensional feature space [12, 15, 17, 18, 24]. To perform accurate classification, the high dimensional feature space must be reduced to a lower dimensional space [20]. One approach to reduce the high dimensional feature space is to perform feature selection, thereby determining a small subset of the original feature space in which accurate classification can be performed [12, 15, 18]. Difranco *et. al.* [18] utilized a random forest feature selection algorithm. Doyle *et. al.* [15] utilized a cascaded classification approach to perform feature selection for a series of pairwise classification tasks. Feature selection schemes have the advantage of selecting those features that give the most accurate classification while discarding features, which may contain noise, that have relatively poorer classification accuracy [12, 15]. However, a limitation of these approaches is that the excluded features may contain important classification information, and their removal may diminish classification accuracy in some tasks [34].

Dimensionality reduction methods learn a low dimensional embedding space which best preserves the original high dimensional feature space [11, 17, 24]. For instance Golugula *et. al.* [24] performed dimensionality reduction via supervised canonical correlation analysis to learn a low dimensional space in which patient classification was performed. Naik *et. al.* [11] demonstrated that GE is well suited for the preservation of a high dimensional feature space which characterized histological differences in texture, nuclear architecture, and gland morphology. Similarly, DBS features in conjunction with GE resulted in 88% classification accuracy for Gleason pattern 3 and 4 glands [17]. However, all of these schemes have utilized the full dataset to perform ML and then trained a classifier within the low dimensional embedding space. These methods are sensitive to noise in the high dimensional feature space as well as the samples considered when learning the low dimensional space. If newly acquired samples or samples which contain noise are included in these systems they will alter the low dimensional embedding space and may detrimentally affect classification performance. Manifold regularization can alleviate this problem by constraining the manifold shape to only shapes which are most likely to occur.

2.2. Manifold Regularization

ML is well known to be sensitive to the dataset considered, as well as noise and outliers contained within a dataset [26, 27]. Perturbations in the manifold structure may reduce classification performance in the low dimensional embedding space as object-class relationships may be obscured. Manifold regularization techniques have been proposed which impose additional constraints on ML to better preserve object-class relationships in the low dimensional space. For instance, Chang *et. al.* [27] proposed a weighted ML scheme, where outlier samples were assigned low weights, to reduce the effect outliers have on learning the manifold. Other manifold regularizers perform local smoothing on the learned manifold [35]. Manifold regularization techniques may add a smoothness constraint into the ML algorithm [26, 36]. All of these methods over smooth the manifold, as they reduce the effects of outliers which including meaningful information as well as noise.

Another type of regularization learns a consensus embedding (CE) from a set of manifolds. Hou *et. al.* [37] learned a set of manifolds by obtaining multiple views for each sample in the dataset and then generated a consensus manifold across the views. Other CE schemes have varied the parameters or samples considered to find a manifold set, and then generated a CE from the set [38, 39]. These methods rely on the manifolds in the set being independent, which may not be a valid assumption when generating manifold sets across ML parameters. Additionally, relationships between samples across the individual manifolds are not taken into account when determining a CE.

3. Brief Overview and Novel Contributions

A flowchart of the proposed SSMM methodology is displayed in Figure 3. Table 1 list the notation used throughout the paper. To construct the SSMM we (1) generate a set of manifolds $\mathbb{M}$ for a dataset **O**. For this task we divide the dataset **O** into K folds, and then generate $\mathbb{M}$ using a leave-one-fold-out scheme. (2) As manifolds in $\mathbb{M}$ will be misaligned, primarily due to rotational and translational differences, a Procrustes based registration scheme aligns all the manifolds in $\mathbb{M}$. (3) Calculate the MLE and primary modes of variation for $\mathbb{M}$. Once constructed the SSMM constrains a new manifold instance $\mathcal{M}$ of related datasets to only those shapes statistically most likely to occur resulting in the regularized manifold $\tilde{\mathcal{M}}$. In this work we demonstrate that the SSMM can (a) determine noisy samples by identifying samples which deviate from the SSMM, and (b) accurately perform OSE of newly acquired samples onto a manifold constrained by the SSMM.

The novel contributions of the SSMM are:

- A computerized decision support system which utilizes a SSMM based on the morphologic features of glands on prostate histopathology to automatically distinguish between Gleason patterns 3 and 4.
- A novel combination of SSMs and ensemble learning theory to generate a more accurate manifold representation of **O**.
- A novel method to generate $\mathbb{M}$, an set of K manifolds, containing all samples in **O**.
- A novel manifold registration to align all manifolds in $\mathbb{M}$. As each sample $o_i \in \mathbf{O}$ has a corresponding embedding location $y_{i,k}$ on the manifold $\mathcal{M}_k$, the registration algorithm minimizes the differences between $y_{i,k}$ for all $k \in \{1, \ldots, K\}$ and all $o_i \in \mathbf{O}$ via Procrustes registration [40].

4. Statistical Shape Model of Manifolds Theory

—We prove theoretically that SSMMs are appropriate to determine the MLE of a manifold shape. Specifically, we prove that constructing a SSMM from a set of manifolds is guaranteed to represent the underlying manifold structure at least as well as any manifold contained in the set. To perform the theoretical analysis we utilize the theory of ensemble learning [28].

A dataset of N samples is defined as $\mathbf{O} = \{o_1, \ldots, o_N\}$. A sample $o_i \in \mathbf{O}$ is defined as a point in a D-dimensional space $\mathbb{R}^D$.

Definition 1: A true manifold $\mathcal{M} \in \mathbb{R}^d$ is defined by a set of N true embedding locations $\mathcal{M} = \{x_1, \ldots, x_N\}$. Each true embedding location $x_i \in \mathbb{R}^d$ corresponds to a sample $o_i \in \mathbb{R}^D$ where $d \ll D$.

Definition 2: A manifold $\hat{\mathcal{M}}$ estimates $\mathcal{M}$ by a set of N embedding locations $\hat{\mathcal{M}} = \{\hat{x}_1, \ldots, \hat{x}_N\}$. Each embedding location $\hat{x}_i \in \mathbb{R}^d$ corresponds to a sample $o_i \in \mathbb{R}^D$ where $d \ll D$.

Definition 3: The manifold $\hat{\mathcal{M}}$ approximates $\mathcal{M}$ with an error $\varepsilon_{\hat{\mathcal{M}}}$ given as,

$$\varepsilon_{\hat{\mathcal{M}}} = E_i(\widehat{x_i} - x_i)^2, \quad (1)$$

where E_i is the mean over $i = \{1, \ldots, N\}$.

Proposition 1: Given a set of K independent, identically distributed manifolds $\mathbb{M} = \{\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_K\}$, a manifold $\bar{\mathcal{M}}$ exists such that $\bar{\mathcal{M}} \rightarrow \mathcal{M}$ as $K \rightarrow \infty$.

Proof: Each estimated manifold $\hat{\mathcal{M}}_k$ is defined by the embedding locations $\hat{x}_{i,k} \in \hat{\mathcal{M}}_k$: $i \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$. Assuming that each embedding location $\hat{x}_{i,k} \in \hat{\mathcal{M}}_k$: $i \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$ is normally distributed about x_i, the Central Limit Theorem states,

$$\lim_{K \rightarrow \infty} (E_k(\widehat{x_{i,k}}) - x_i) = 0, \quad (2)$$

where E_k is the mean over $k = \{1, \ldots, K\}$. Therefore $\bar{\mathcal{M}}$ exists and is defined $\bar{\mathcal{M}} = E_k(\hat{x}_{i,k})$: $i \in \{1, \ldots, N\}$.

The error between $\bar{\mathcal{M}}$ and $\mathcal{M}$ is defined as (similar to Equation 1),

$$\varepsilon_{\bar{\mathcal{M}}} = E_i(\bar{x}_i - x_i)^2, \quad (3)$$

where $\bar{x}_i = E_k(\hat{x}_{i,k})$. From Equation 1 the error over K embeddings is given as,

$$\varepsilon_{K,\bar{\mathcal{M}}} = E_k\left[\varepsilon_{\hat{\mathcal{M}}}\right] = E_k\left[E_i(\widehat{x_{i,k}} - x_i)^2\right]. \quad (4)$$

Proposition 2: Given K independent, identically distributed manifolds, $\hat{\mathcal{M}}_k \in \mathbb{M}$, $\varepsilon_{K,\bar{\mathcal{M}}} \geq \varepsilon_{\hat{\mathcal{M}}}$.

Proof: Comparing Equation 4 and Equation 3 in a manner analogous to Bagging [28] gives the proof. In Bagging, an ensemble classifier is constructed from a set of weak classifiers. Similarly, the $\bar{\mathcal{M}}$ obtained from the SSMM can be viewed as an ensemble embedding constructed from a set of weak embeddings $\mathbb{M}$. Hence the result follows.

5. Construction of Statistical Shape Manifold Model (SSMM)

In this section we present our methodology for constructing a SSMM. In Section 5.1 we provide an overview of ML. We then describe our novel K fold algorithm to calculate $\mathbb{M}$ in Section 5.2. Section 5.3 presents the Procrustes based registration of manifolds in $\mathbb{M}$. In Section 5.4 we discuss SSMM construction and Section 5.5 describes the fitting of $\hat{\mathcal{M}}$ to the SSMM. Finally we present two novel algorithms for (a) noise identification and removal in Section 5.6 and (b) OSE of newly acquired samples onto the SSMM in Section 5.7.

5.1. Review of Manifold Learning

5.1.1. Graph Embedding (GE)—In this work, we implemented the ML scheme GE [23] to perform nonlinear dimensionality reduction as it has few parameters to optimize over (only γ an empirically determined scaling term) and is relatively computationally efficient. GE learns a manifold estimate $\hat{\mathcal{M}}$ for a dataset $\mathbf{O}_T \in \mathbb{R}^D$. $\hat{\mathcal{M}}$ is described by a set of embedding locations $\hat{\mathbf{y}} \in \mathbb{R}^n$ where $n \ll D$. $\hat{\mathbf{y}}$ is obtained by performing the eigenvalue decomposition (EVD),

$$W\hat{\mathbf{y}}' = \hat{\lambda}L\hat{\mathbf{y}}', \quad (5)$$

where $W(a, b) = e^{-A(a,b)/\gamma}$, γ is an empirically determined scaling term, and L is the diagonal matrix $L(a, a) = \Sigma_b W(a, b)$. $\hat{\mathbf{y}}$ is defined as the n eigenvectors which correspond to the top n eigenvalues in $\hat{\lambda}$.

5.1.2. Out-of-Sample Extrapolation (OSE)—A sample not in the original dataset, i.e. $o_k \notin \mathbf{O}_T$, will not have a corresponding embedding location in $\hat{\mathbf{y}}$. To calculate the embedding location $\hat{y}_k$ the dissimilarity matrix A and the EVD would have to be recomputed to include o_k in $\mathbf{O}_T$. Repeating ML for every new sample acquired is computationally infeasible [33]. The aim of OSE is to determine embedding locations $\tilde{\mathbf{y}}$ for samples in a newly acquired dataset defined as $\mathbf{O}_R$.

The Nsyström Method (NM) is a OSE algorithm which estimates $\tilde{\mathbf{y}}$ as a weighted sum of the known embeddings $\hat{\mathbf{y}}$ [33]. Given $\hat{\mathbf{y}}$ for $\mathbf{O}_T$ generated in Section 5.1.1, $\tilde{\mathbf{y}}$ for $\mathbf{O}_R$ are calculated as,

$$\tilde{y}_{i,d} = \frac{1}{\hat{\lambda}_d} \sum_{j=1}^{S} \hat{y}_{j,d} W(o_j, o_i), \quad (6)$$

where $d \in \{1, \ldots, n\}$ is the dth embedding dimension corresponding to the dth largest eigenvalue $\hat{\lambda}_d$. OSE does not alter the underlying relationships contained in $\hat{\mathbf{y}}$. Furthermore the samples contained in $\mathbf{O}_R$ in no way alter or affect the relationships contained in $\hat{\mathbf{y}}$.

5.2. Construction of the Manifold Set

A set of K manifolds $\mathbb{M} = \{\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_K\}$ are obtained from a dataset of N samples defined as $\mathbf{O} = \{o_1, \ldots, o_N\}$. $\mathbb{M}$ is generated utilizing a K fold scheme via the following steps:

1. Samples in $\mathbf{O}$ are randomly divided into K equal partitions such that $\mathbf{O} = \{\mathbf{O}_1 \cup \ldots \cup \mathbf{O}_K\}$.
2. Testing and training sets are obtained via a leave one fold out scheme. A testing set is defined as $\mathbf{O}_{R,k} = \mathbf{O}_k$: $k \in \{1, \ldots, K\}$ and the corresponding training set is defined as $\mathbf{O}_{T,k} \in \mathbf{O}_{R,k} = \mathbf{O}$.
3. Each training set $\mathbf{O}_{T,k}$ is utilized to find $\hat{\mathbf{y}}_k$ which defines $\hat{\mathcal{M}}_k$ via GE as described in Section 5.1.1. The samples in $\mathbf{O}_{T,k}$ are then used to determine the structure of the manifold $\hat{\mathcal{M}}_k$.
4. Each test set $\mathbf{O}_{R,k}$ is extrapolated into the manifold $\hat{\mathcal{M}}_k$ to determine $\tilde{\mathbf{y}}_k$ via NM as described in Section 5.1.2.

5. Training and testing sets are combined to determine $\mathbf{y}_k = \{\hat{\mathbf{y}}_k, \check{\mathbf{y}}_k\}$. This combination allows for point correspondence between manifolds in $\mathbb{M}$ to be estimated.

In this work $K = 10$ was chosen to construct $\mathbb{M}$, and the steps described above were performed 5 times for a total of 50 constituent manifolds in $\mathbb{M}$. GE and NM were chosen for experiments showcased in this work, but it is worth noting any ML [21–23, 25] and OSE [33, 41] scheme can be used to construct $\mathbb{M}$.

5.3. Manifold Alignment via Procrustes Based Registration

Manifolds contained in $\mathbb{M}$ may not align, the algorithm for ML preserves pairwise relationships between samples but may not preserve the global relationship of samples in the low dimensional embedding space. Procrustes registration is applied to align all manifolds in $\mathbb{M}$ [40]. Procrustes registration can be performed since there are point correspondences between all manifolds in $\mathbb{M}$ as each sample in $\mathbf{O}$ has a location on every manifold in $\mathbb{M}$.

A reference manifold $\mathcal{M}_a$: $a \in \{1, \ldots, K\}$ is randomly selected. All other manifolds, $\mathcal{M}_b$: $b \neq a$ are registered to $\mathcal{M}_a$ by minimizing,

$$\breve{\mathcal{M}}_b = \min_{T^{a,b}} \left(\sum_i \| y_{i,a} - T^{a,b}(y_{i,b}) \| \right), \quad (7)$$

where $y_{i,a}$ is a embedding location in $\mathcal{M}_a$ for a sample o_i and $y_{i,b}$ is a embedding location in $\mathcal{M}_b$ for a sample o_i. The transform $T^{a,b}$ selected was a rigid transform, to take into account scale and rotational differences between $\mathcal{M}_a$ and $\mathcal{M}_b$. $\|\cdot\|$ denotes the L2-norm. Registration is performed for all $\mathcal{M} \in \mathbb{M}$ to obtain the aligned set of manifolds $\breve{\mathbb{M}}$.

5.4. Statistical Shape Manifold Model (SSMM)

Once all manifolds are aligned the statistical properties of the manifold set can be determined. The SSMM is defined via the mean and principal modes of variation for $\breve{\mathbb{M}}$. The mean of $\breve{\mathbb{M}}$ is calculated by,

$$\overline{\mathcal{M}} = \frac{1}{K} \sum_k \breve{y}_{i,k} : \forall \breve{y}_{i,k} \in \breve{\mathcal{M}}_k. \quad (8)$$

The principal modes of variation for the manifold defined as P are obtained by performing PCA on $\breve{\mathbb{M}}$ [29]. Only the P corresponding to the top 95% of variance in the sample locations $y_{i,k}$ for all $k \in \{1, \ldots, K\}$ are retrained to constrain the SSMM to those shapes within 2 standard deviations of the mean shape.

5.5. Constraining a New Manifold Instance to the SSMM

A new manifold $\dot{\mathcal{M}}$ is obtained by applying GE to $\mathbf{O}$. $\dot{\mathcal{M}}$ is constrained to only likely shapes as defined by the SSMM obtained in Section 5.4.

$$\tilde{\mathcal{M}} = T^{a,K+1}(\overline{\mathcal{M}} + P * b), \quad (9)$$

where b controls the shape of $\dot{\mathcal{M}}$ and $T^{a,K+1}$ is a rigid transformation between the SSMM and $\dot{\mathcal{M}}$. b is found via a linear least squares fit between the SSMM and $\dot{\mathcal{M}}$ and is constrained to $\dot{\mathcal{M}} \pm 2\sigma$ to limit the SSMM to only those shapes statistically most likely to occur [29].

5.6. Application of SSMM to Identify Noisy Samples

The SSMM can aid in the identification of samples which contain noise. The algorithm *FilterManifold* assumes samples which contain noise are those samples which deviate most from the SSMM.

A dataset contains N samples defined as $Q = \{q_1, \ldots, q_N\}$. The following algorithm can be used to identify the samples which contain noise Q_n and the samples which do not contain noise Q_c within Q given a user defined threshold τ. The value assigned to τ is dataset specific as sample variation across datasets may vary. In this work τ was chosen such that 5% of the samples in the dataset were excluded.

Algorithm

FilterManifold

Input: Q, τ

Output: $\mathcal{M}_c$

begin

1. Obtain $\tilde{\mathcal{M}}$ from Q via application of the SSMM.
2. Obtain $\hat{\mathcal{M}}$ from Q by GE (Eq. 5).
3. Calculate $\epsilon(q_i) = \|\hat{y}_i - \tilde{y}_i\|$.
4. Obtain $Q_n = q_i : q_i \in Q, \epsilon(q_i) \geq \tau$.
5. Obtain $Q_c : Q_c \cap Q_n = \varnothing$.
6. Obtain $\mathcal{M}_c$ for Q_c via GE (Eq. 5)

end

5.7. Application of SSMM to OSE

The SSMM can be utilized for robust OSE, by generating a more accurate manifold representation of a dataset. The algorithm *OSE-SSMM* demonstrates how the SSMM can be used for this purpose.

A dataset Q is divided into training samples Q_{tr} and testing samples Q_{te} such that $Q_{tr} \cap Q_{te} = \varnothing$. To find a set of testing embeddings $\mathcal{M}_{te,c}$ for a filtered manifold we apply the following algorithm,

Algorithm

OSE-SSMM

Input: Q_{tr}, Q_{te}, τ

Output: $\mathcal{M}_{te,c}$

begin

1. Obtain $\mathcal{M}_{tr,c}$ for Q_{tr} via *FilterManifold*.
2. Obtain $\mathcal{M}_{te,c}$ for Q_{te} via NM (Eq. 6) with $\mathcal{M}_{tr,c}$ as the training manifold.

end

6. Experimental Design and Results

6.1. Dataset Description

6.1.1. Synthetic Datasets—Two synthetic datasets, Swiss Roll and Helix, described in Table 2 were utilized to demonstrate the application of SSMM to manifold regularization. The Swiss Roll is a $2D$ planar manifold divided into two classes which exists in a $3D$ space. The Helix is a $1D$ circular manifold divided into six classes which exists in a $3D$ space. The benefit of both datasets is that the high dimensional $3D$ space and the low dimensional $2D$ embedding space may be visualized. Gaussian noise was added to 5% of samples within each dataset where the standard deviation of the noise was set equal to 15% of the standard deviation of samples in the dataset. The dissimilarity measures for both datasets are reported in Table 2.

6.1.2. Prostate Histopathology—Prostate needle core tissue biopsies were obtained from 58 patients. Biopsies were stained with H & E and digitized at 40× optical magnification using an Aeperio scanner. An expert pathologist selected regions of interest (ROIs) on each biopsy. In total 120 ROIs were selected across. Each ROI was assigned a Gleason pattern of either BE, G3, or G4. All glands contained within each ROI were manually segmented to obtain a total of 888 glands from BE ($N = 93$), G3 ($N = 748$), and G4 ($N = 47$) ROIs. For this set of experiments only G3 and G4 glands were considered during classification. DBS was the dissimilarity measure utilized to quantify morphologic differences between glands [17].

6.2. Evaluation Measures

6.2.1. Silhouette Index (SI)—SI is a measure of how well samples cluster by class label [42] with 1 corresponding to perfect clustering by class and −1 corresponding to no clustering by class. SI is calculated as, $\eta^{SI} = \sum_{i=1}^{N} \frac{G(i)-C(i)}{\max[C(i),G(i)]}$ where $C(i) = \Sigma_{j,l_j=l_i} \|\tilde{y}_i - \tilde{y}_j\|$ and $G(i) = \Sigma_{j,l_j \neq l_i} \|\tilde{y}_i - \tilde{y}_j\|$.

6.2.2. Area Under the Receiver Operator Characteristic (ROC) Curve (AUC)—A probabilistic boosting tree (PBT) classifier [43] was trained and evaluated using a 5×2 cross validation scheme [44]. For each of the 5 runs, the dataset was divided into 2 folds such that all samples from a single patient were contained in the same fold and all folds maintained class balance. The PBT classifier assigns a probability value to each sample of belonging to the positive class. Altering the threshold level of the PBT classifier allows for the construction of a ROC Curve. For each ROC Curve the area under the curve (AUC) is calculated.

6.3. Experiment 1: Application of SSMM to Filtered Manifold Learning

For each dataset Q in Table 2, a manifold $\mathcal{M}$ was calculated from Q using GE as described in Section 5.1.1. Similarly a filtered manifold $\mathcal{M}_c$ was found by *FilterManifold* as described in Section 5.6. The measures described in Section 6.2 were used to evaluate $\mathcal{M}$ and $\mathcal{M}_c$. A Student's t-test was calculated to determine the statistical significance between $\mathcal{M}$ and $\mathcal{M}_c$ for each evaluation measure described in Section 6.2.

Experimental results for all datasets are reported in Table 3. Across all datasets $\mathcal{M}_c$ outperforms $\mathcal{M}$ in terms of SI and AUC. In the prostate histology dataset these increases in SI and AUC were statistically significant ($p \leq 0.1$). Hence $\mathcal{M}_c$ is better able to preserve object-class relationships in the datasets evaluated.

For the synthetic datasets, changes in SI and AUC are not always statistically significant. However, as may be noted in Figure 1 (d) $\mathcal{M}_c$ is a closer approximation to the true embedding (Figure 1 (b)) than compared to $\mathcal{M}$ (Figure 1 (c)). In Figure 1 the samples are colored according to their location on the true embedding to aid in visualization.

6.4. Experiment 2: Application of SSMM to Filtered OSE

For each dataset Q in Table 2, a training set Q_{tr} and a testing set Q_{te} were defined so that Q_{te} is 10% of Q and $Q_{tr} \cup Q_{te} = \varnothing$. Q_{tr} and Q_{te} were used to construct an original manifold $\mathcal{M}_{te}$ and filtered manifold $\mathcal{M}_{te,c}$. $\mathcal{M}_{te}$ is generated by applying GE as described in Section 5.1.1 and then applying NM as described in Section 5.1.2 to Q_{te} where $\mathcal{M}_{tr}$ is the training manifold. The filtered manifold $\mathcal{M}_{te,c}$ is calculated by *OSE-SSMM* as described in Section 5.7. The measures described in Section 6.2 were used to evaluate $\mathcal{M}_{te}$ and $\mathcal{M}_{te,c}$. A Student's t-test was calculated to determine the statistical significance between $\mathcal{M}_{te}$ and $\mathcal{M}_{te,c}$ for each evaluation measure described in Section 6.2.

Experimental results for all datasets are reported in Table 4. For the histopathology dataset $\mathcal{M}_{te,c}$ outperforms $\mathcal{M}_{te}$ in terms of SI and AUC. The synthetic datasets, the Swiss Roll and Helix, do not show a significant improvement in performance.

7. Concluding Remarks

In this paper we presented a statistical shape model of manifolds (SSMM) which is a novel integration of statistical shape models (SSMs) with ensemble learning for regularizing low dimensional data representations of high dimensional spaces. New, related manifolds may then be constrained by the SSMM to only those shapes statistically most likely to occur.

The SSMM may be utilized for several applications including (a) identification of noisy samples based on their deviation from the SSMM. Removing these samples from the dataset may result in higher area under the receiver operator characteristic (ROC) curve (AUC). (b) A classifier could be trained on the SSMM allowing for (i) classifier decision boundaries to be applied to a new related manifold without retraining the classifier or (ii) new, related samples to be classified by projection of the samples onto the SSMM. (c) identification of regions on a new, related manifold which deviate the SSMM. Identifying these regions may aid in determining meaningful differences between the dataset and SSMM.

To construct the SSMM we (1) generate a set of manifolds $\mathbb{M}$ for a dataset **O**, (2) align manifolds in $\mathbb{M}$, and (3) calculate the maximum likelihood estimate (MLE) of the manifold shape and its primary modes of variation. The SSMM allows for constraining a new, related manifold instance to only those shapes statistically most likely to occur.

We have demonstrated in this work that SSMM can improve AUC in the context of Gleason grading of prostate histopathology utilized quantitative morphologic features of glands. For the dataset considered, the tissue samples corresponded to either Gleason pattern 3 or pattern 4. Improvements in AUC via the SSMM were demonstrated for two applications:(a) We demonstrated that outlier samples within a manifold can be identified as those samples which deviate from the SSMM via *FilterManifold*. Removal of outlier samples increased AUC and SI. (b) We demonstrated via *OSE-SSMM* that manifold regularization by the SSMM improves SI and AUC when performing OSE on never before seen samples onto the SSMM.

In future work we intend to explore the ability of the SSMM to identify regions of a new, related manifold which deviate from the SSMM. These regions will then be further investigated to determine subtle difference between the dataset and the SSMM. Secondly,

we plan on investigating the effects of dataset size on the SSMM by evaluating how accurately the mean manifold shape and primary modes of variation of the manifold shape are represented for SSMMs trained on different dataset sizes.

Acknowledgments

This work was made possible via grants from the National Cancer Institute (Grant Nos. R01CA136535-01, R01CA14077201, and R21CA167811), and the Department of Defense (W81XWH-11-1-0179). We would like to thank Dr. J. E. Tomaszewski from the University at Buffalo School of Medicine and Biomedical Sciences as well as Drs. M. D. Feldman and N. Shi from the Hospital of the University of Pennsylvania for providing prostate histology imagery and annotations.

References

1. Epstein JI, Walsh PC, Carmichael M, Brendler CB. Pathologic and clinical findings to predict tumor extent of nonpalpable (stage t1 c) prostate cancer. JAMA: The Journal of the American Medical Association. 1994; 271:368–374. [PubMed: 7506797]
2. Bostwick D. Grading prostate cancer. American Journal of Clinical Pathology. 1994; 102:S38–56. [PubMed: 7524306]
3. Gleason DF. Classification of prostatic carcinomas. Cancer Chemotherapy Reports. 1966; 50:125–128.
4. Epstein JI, Allsbrook WC, Amin MB, Egevad LL. The 2005 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. American Journal of Surgical Pathology. 2005; 29:1228–1242. [PubMed: 16096414]
5. Epstein JI. An update of the gleason grading system. The Journal of Urology. 2010; 183:433–440. [PubMed: 20006878]
6. Epstein PCWJI, Sanfilippo F. Clinical and cost impact of second-opinion pathology: Review of prostate biopsies prior to radical prostatectomy. The American Journal of Surgical Pathology. 1996; 20:851–857. [PubMed: 8669533]
7. Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist. Human Pathology. 2001; 32:81–88. [PubMed: 11172299]
8. Thompson I, Thrasher J, Aus G, Burnett A, Canby-Hagino E, Cookson M, D'Amico A, Dmochowski R, Eton D, Forman J, Goldenberg S, Hernandez J, Higano C, Kraus S, Moul J, Tangen C. Guideline for the management of clinically localized prostate cancer: 2007 update. The Journal of Urology. 2007; 177:2106–2131. [PubMed: 17509297]
9. Madabhushi A. Digital pathology image analysis: Oppurtunities and challenges. Imaging in Medicine. 2009; 1:7–10.
10. Jafari-Khouzani K, Soltanian-Zadeh H. Multiwavelet grading of pathological images of prostate. IEEE Transactions on Biomedical Engineering. 2003; 50:697–704. [PubMed: 12814236]
11. Naik, S.; Doyle, S.; Madabhushi, A.; Tomaszeweski, J.; Feldman, M. Gland segmentation and gleason grading of prostate histology by integrating low-, high-level and domain specific information. Workshop on Microscopic Image Analysis with Applications in Biology;
12. Tabesh A, Teverovskiy M, Pang H, Kumar V, Verbel D, Kotsianti A, Saidi O. Multifeature prostate cancer diagnosis and Gleason grading of histological images. IEEE Transaction on Medical Imaging. 2007; 26:1366–1378.
13. Veltri RW, Marlow C, Khan MA, Miller MC, Epstein JI, Partin AW. Significant variations in nuclear structure occur between and within gleason grading patterns 3, 4, and 5 determined by digital image analysis. The Prostate. 2007; 67:1202–1210. [PubMed: 17525934]
14. Huang PW, Lee CH. Automatic classification for pathological prostate images based on fractal analysis. IEEE Transactions on Medical Imaging. 2009; 28:1037–1050. [PubMed: 19164082]
15. Doyle S, Monaco J, Feldman M, Tomaszewski J, Madabhushi A. An active learning based classification strategy for the minority class problem: application to histopathology annotation. BMC Bioinformatics. 2011; 12:424. [PubMed: 22034914]

16. Tambasco M, Costello BM, Kouznetsov A, Yau A, Magliocco AM. Quantifying the architectural complexity of microscopic images of histology specimens. Micron. 2009; 40:486–494. [PubMed: 19171487]
17. Sparks, R.; Madabhushi, A. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI); p. 658-665.
18. DiFranco MD, Hurley GO, Kay EW, Watson RWG, Cunningham P. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. Computerized Medical Imaging and Graphics. 2011; 35:629–645. [PubMed: 21269807]
19. Ali, S.; Madabhushi, A. Active contour for overlap resolution using watershed based initialization (acorew): Applications to histopathology. IEEE International Symposium on Biomedical Imaging (ISBI); p. 614-617.
20. Duda, RO.; Hart, PE.; Stork, DG. Pattern Classification (2nd Edition). 2. Wiley-Interscience; 2001.
21. Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000; 290:2323–2326. [PubMed: 11125150]
22. Tenenbaum J, de Silvia V, Langford J. A global framework for nonlinear dimensionality reduction. Science. 2000; 290:2319–2323. [PubMed: 11125149]
23. Shi J, Malik J. Normalized cuts and image segmentation, Pattern Analysis and Machine Intelligence. IEEE Transactions on. 2000; 22:888–905.
24. Golugula A, Lee G, Master S, Feldman M, Tomaszewski J, Speicher D, Madabhushi A. Supervised regularized canonical correlation analysis: Integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery. BMC Bioinformatics. 2011; 12:483. [PubMed: 22182303]
25. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Compuation. 2003; 15:1373–1396.
26. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research. 2006; 7:2399–2434.
27. Chang H, Yeung D. Robust locally linear embedding. Pattern Recognition. 2006; 39:1053–1065.
28. Breiman L. Bagging predictors. Machine Learning. 1996; 24:123–140.
29. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models their training and application. Computer Vision & Image Understanding. 1995; 61:38–59.
30. Dietterich, TG. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems; Springer-Verlag; 2000. p. 1-15.
31. Fred AL, Jain AK. Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27:835–850. [PubMed: 15943417]
32. Viswanath S, Madabhushi A. Consensus embedding: theory, algorithms and application to segmentation and classification of biomedical data. BMC Bioinformatics. 2012; 13:26. [PubMed: 22316103]
33. Bengio Y, Paiement J, Vincent P, Delalleau O, Roux NL, Ouimet M. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. Advances in Neural Information Processing Systems. :177–184.
34. Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics. 2004; 20:2429–2437. [PubMed: 15087314]
35. Park, JH.; Zhang, Z.; Zha, H.; Kasturi, R. Local smoothing for manifold learning. Computer Vision and Pattern Recognition, 2004. CVPR 2004; Proceedings of the 2004 IEEE Computer Society Conference on, volume 2; p. II–452-II–459.
36. Geng, B.; Xu, C.; Tao, D.; Yang, L.; Hua, X-S. Ensemble manifold regularization. 2009. p. 2396-2402.
37. Hou C, Zhang C, Wu Y, Nie F. Multiple view semi-supervised dimensionality reduction. Pattern Recognition. 2010; 43:720–730.
38. Jia J, Xiao X, Liu B, Jiao L. Bagging-based spectral clustering ensemble selection. Pattern Recognition Letters. 2011; 32:1456–1467.

39. Tiwari P, Rosen M, Madabhushi A. Consensus-locally linear embedding (c-lle): Application to prostate cancer detection on magnetic resonance spectroscopy. Medical Image Computing and Computer-Assisted Intervention MICCAI 2008. 2008; 5242:330–338.
40. Wang, C.; Mahadevan, S. Manifold alignment using procrustes analysis. Proceedings of the 25th international conference on Machine learning; p. 1120-1127.
41. Fowlkes C, Belongie S, Chung F, Malik J. Spectral grouping using the Nyström method. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004; 28:214–225. [PubMed: 15376896]
42. Rousseeuw P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational Applied Mathematics. 1987; 20:53–65.
43. Tu, Z. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. Computer Vision, 2005. ICCV 2005; Tenth IEEE International Conference on; p. 1589-1596.
44. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation. 1998; 10:1895–1923. [PubMed: 9744903]

Highlights

- Statistical shape model of manifold (SSMM) constructs an ensemble of manifolds.
- SSMM constrains the shape of a manifold resulting in a more stable low dimensional space.
- The SSMM improves area under the receiver operator characteristic (ROC) curve (AUC) in the context of Gleason grading of prostate histopathology.
- The SSMM can project new samples into a low dimensional space with high accuracy.

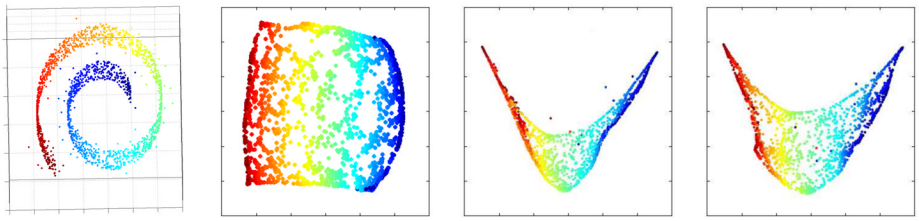

Figure 1.

(a) Original $3D$ Swiss Roll dataset with Gaussian noise added to 2% of samples in the datasest. (b) $2D$ manifold $\mathcal{M}$ in the absence of noise. This manifold structure best preserves the relationships between samples in the original high dimensional space. (c) Manifold $\hat{\mathcal{M}}$ found by applying ML to a dataset containing noise and (d) the manifold $\bar{\mathcal{M}}$ found by regularization of $\hat{\mathcal{M}}$ using the SSMM.

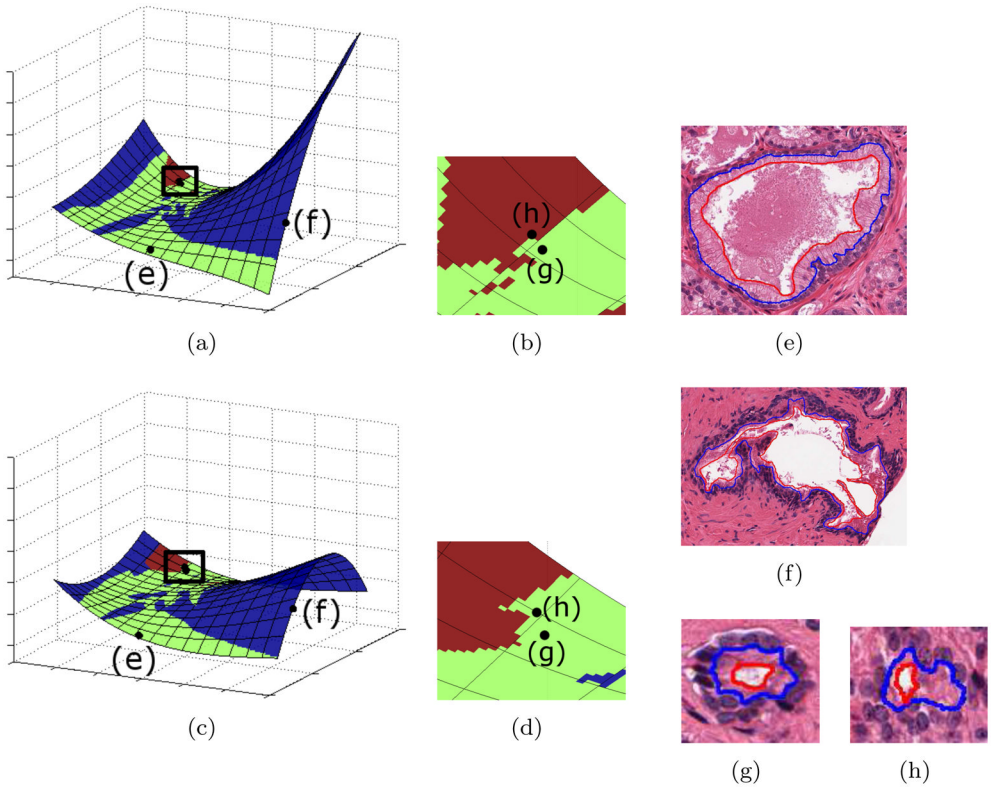

Figure 2.

(a), (c) Two manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ generated by performing ML on quantitative morphologic features extracted from 800 prostate histopathology glands. The manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ were generated from two distinct datasets $\mathbf{O}_1$ and $\mathbf{O}_2$ such that 88 glands excluded from either $\mathbf{O}_1$ or $\mathbf{O}_2$. (b), (d) Manifold region enclosed by the black box in (a) and (c) respectively. Representative glands for (f) benign (BE), (e), (g) Gleason pattern 3 (G3), and (h) pattern 4 (G4) classes. A classifier trained in the reduced dimensional space allows for assignment of a single class to each region on the manifold, such that blue regions correspond to BE, green regions correspond to G3, and red regions correspond to G4. Differences between the manifolds can be seen in changes in global structure as well as class-object relationships on the manifold, which are evident by changes in region color. In the case of (h) a representative G4 gland, in one manifold ((c)) the gland was incorrectly projected on to the G3 class region.

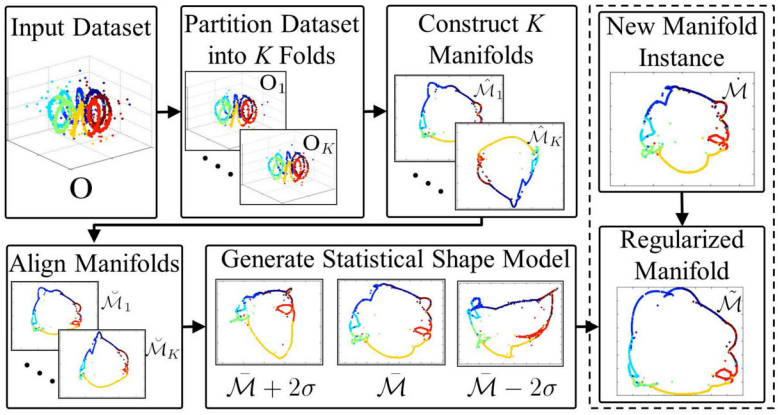

Figure 3.

Flowchart which describes the construction of the SSMM and its application to manifold regularization for the synthetic Helix dataset. SSMM construction consists of dividing the dataset **O** into K folds, denoted as $\{\mathbf{O}_1, \ldots, \mathbf{O}_K\}$. The K folds of **O** are utilized to find the manifold set $\mathbb{M} = \{ \check{\mathcal{M}}_1, \ldots, \check{\mathcal{M}}_K \}$. The manifolds in $\mathbb{M}$ are then aligned via Procrustes based registration scheme resulting in the aligned manifold set $\check{\mathbb{M}} = \{ \check{\mathcal{M}}_1, \ldots, \check{\mathcal{M}}_K \}$. The SSMM finds the MLE ($\bar{\mathcal{M}}$) and primary modes of variation for $\check{\mathbb{M}}$. Shown are the modes of variation corresponding to the statistical extremes of the model $\bar{\mathcal{M}} - 2\sigma$ and $\bar{\mathcal{M}} + 2\sigma$. Given a new manifold instance $\dot{\mathcal{M}}$ the SSMM constrains the structure to only those statistically likely to occur ($\bar{\mathcal{M}} \pm 2\sigma$). This results in the regularized manifold $\tilde{\mathcal{M}}$ which is a better approximation of the underlying relationships in **O** than any constituent manifold in $\mathbb{M}$. For the synthetic Helix dataset shown in this flowchart the ideal manifold is a $2D$ circular structure.

Table 1

Notation used in the paper.

Symbol	Description
$\mathbf{O}$	Dataset
o_i, o_j	Samples contained in $\mathbf{O}$
$\mathbf{O}_k$	kth fold of $\mathbf{O}$ for $k \in \{1, \ldots, K\}$
$\mathbb{R}^D$	High dimensional feature space
$\psi(\cdot, \cdot)$	Dissimilarity measure
A	Dissimilarity matrix defined as $\psi(o_i, o_j)$ evaluated for all $o_i, o_j \in \mathbf{O}$
γ	GE scaling term
$\mathbb{R}^n$	Low dimensional embedding space
$\mathbb{M}$	Ensemble manifold set
$\mathcal{M}_k$	kth manifold in $\mathbb{M}$
$\mathcal{Y}_k$	Embedding locations on $\mathcal{M}_k$
$y_{i,k}$	Embedding location for o_i on $\mathcal{M}_k$
$T^{a,b}$	Transformation to align $\mathcal{M}_b$ to $\mathcal{M}_a$
$\|\cdot\|$	L2 norm
$\breve{\mathbb{M}}$	Aligned manifold set
$\breve{\mathcal{M}}$	MLE for $\breve{\mathbb{M}}$
P	Primary modes of variation for $\breve{\mathbb{M}}$
$\dot{\mathcal{M}}$	New manifold instance
$\tilde{\mathcal{M}}$	Manifold constrained via the SSMM
Q	New dataset instance
Q_n	Samples which contain noise in Q
Q_c	Samples which do not contain noise in Q
τ	Threshold to determine sample deviation from the SSMM
$\mathcal{M}_c$	Manifold generated from Q_c
Q_{te}	Testing samples not contained in Q
$\mathcal{M}_{te,c}$	Manifold with samples in Q_{te} projected onto $\mathcal{M}_c$.

Table 2

Description of datasets and their dissimilarity measures.

Dataset	Sample Size	Dissimilarity Measure
Synthetic Swiss Roll [22]	3000	$A(o_i, o_j)\begin{cases} \|o_i - o_j\| & \text{if } \|o_i - o_j\| < \mathcal{N}, \\ 0 & \text{otherwise.} \end{cases}$ $\mathcal{N}$ is a neighborhood parameter.
Synthetic Helix [22]	3000	$A(o_i, o_j)\begin{cases} \|o_i - o_j\| & \text{if } \|o_i - o_j\| < \mathcal{N}, \\ 0 & \text{otherwise.} \end{cases}$ $\mathcal{N}$ is a neighborhood parameter.
Prostate Histology	888 (58 patients)	Diffeomorphic Based Similarity (DBS) [17]

Table 3

(a) SI and (b) AUC are reported for $\mathcal{M}$ and $\mathcal{M}_c$. The best value for each dataset is bolded. p-values are reported for a Student's t-test comparing $\mathcal{M}$ and $\mathcal{M}_c$.

(a)			
Dataset	$\mathcal{M}$	$\mathcal{M}_c$	***p-value***
Swiss Roll	0.56 ± 0.01	**0.57 ± 0.03**	0.063
Helix	0.44 ± 0.05	**0.47 ± 0.02**	0.138
Prostate	0.02 ± 0.01	**0.05 ± 0.03**	0.032

(b)			
Dataset	$\mathcal{M}$	$\mathcal{M}_c$	***p-value***
Swiss Roll	0.876 ± 0.067	**0.935 ± 0.065**	0.071
Helix	0.995 ± 0.002	**0.996 ± 0.002**	0.240
Prostate	0.779 ± 0.075	**0.832 ± 0.048**	0.073

Table 4

(a) SI and (b) AUC are reported for $\mathcal{M}_{te}$ and $\mathcal{M}_{te,c}$. The best value for each dataset is bolded. p-values are reported for a Student's t-test comparing $\mathcal{M}_{te}$ and $\mathcal{M}_{te,c}$.

(a)			
Dataset	$\mathcal{M}_{te}$	$\mathcal{M}_{te,c}$	***p-value***
Swiss Roll	0.57 ± 0.01	**0.58 ± 0.01**	0.061
Helix	0.47 ± 0.01	0.47 ± 0.01	0.77
Prostate	−0.04 ± 0.01	**−0.02 ± 0.02**	0.005

(b)			
Dataset	$\mathcal{M}_{te}$	$\mathcal{M}_{te,c}$	***p-value***
Swiss Roll	0.997 ± 0.003	**0.999 ± 0.002**	0.102
Helix	0.994 ± 0.002	**0.996 ± 0.002**	0.089
Prostate	0.779 ± 0.054	**0.834 ± 0.051**	0.032