Supplementary Material

I. Proof of Proposition 1

The generalized empirical quantization error is formulated as

$$\ell_{\mathbf{P}}(\mathbb{X}, \mathbb{Y}, h_{\mathcal{G}}) = \sum_{i=1}^{N}\sum_{k=1}^{K} p_{i,k}\Big[||\mathbf{x}_i - h_{\mathcal{G}}(\mathbf{y}_k)||_2^2 + \sigma \log p_{i,k}\Big], \tag{1}$$

where $\sigma > 0$ is a regularization parameter.

Mean shift clustering [1] is a non-parametric model for locating modes of a density function. Let $\mathbb{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ be modes of $f(\mathbb{C})$ that we aim to find. The kernel density estimator is used to estimate the density function from $\mathbb{X}$, for any $\mathbf{c}_k \in \mathbb{C}$, given by

$$f(\mathbf{c}_k|\mathbb{X}) = \frac{1}{N}\sum_{i=1}^{N}(\pi\sigma)^{-D/2}\exp\left(-\frac{||\mathbf{x}_i - \mathbf{c}_k||^2}{\sigma}\right),$$

where Gaussian kernel with covariance matrix $\frac{1}{2}\sigma\mathbf{I}$ is used and $\sigma > 0$. Since data points $\mathbb{X}$ are independent and identically distributed, we have a joint density function over $\mathbb{C}$ given by

$$f(\mathbb{C}|\mathbb{X}) = \prod_{k=1}^{K} f(\mathbf{c}_k|\mathbb{X}).$$

An optimization problem is naturally formed to find the modes by maximizing the joint density function with respect to $\mathbb{C}$ as

$$\max_{\mathbb{C}} \sum_{k=1}^{K} \log\left[\sum_{i=1}^{N}\exp\left(-\frac{||\mathbf{x}_i - \mathbf{c}_k||^2}{\sigma}\right)\right]. \tag{2}$$

On the other hand, we introduce the left stochastic matrix $\mathbf{P} \in \mathcal{P}_c$, where

$$\mathcal{P}_c = \left\{p_{i,k}|\sum_{i=1}^{N} p_{i,k} = 1, p_{i,k} \geq 0, \forall i, k\right\}.$$

Let $\mathbf{c}_k = h_{\mathcal{G}}(y_k), \forall k$ and $\mathbf{p}_k = [p_{1,k}, \ldots, p_{N,k}]^T$. For each $k = 1, \ldots, K$, minimizing equation of generalized empirical quantization error can be reformulated as the following optimization problem

$$\begin{aligned}\min_{\mathbf{p}_k} \quad & \sum_{i=1}^{N} p_{i,k}\Big[||\mathbf{x}_i - \mathbf{c}_k||^2 + \sigma \log p_{i,k}\Big] \\ \text{s.t.} \quad & \sum_{i=1}^{N} p_{i,k} = 1, p_{i,k} \geq 0, \forall i.\end{aligned}$$

By Lagrange duality theorem [2], the KKT conditions for optimal solution $\mathbf{p}_k$ are given by

$$||\mathbf{x}_i - \mathbf{c}_k||^2 + \sigma(1 + \log p_{i,k}) + \alpha = 0, \forall k,$$
$$\sum_{i=1}^{N} p_{i,k} = 1, p_{i,k} \geq 0 \forall i, k.$$

By combining above two equalities, we have the optimal solution as

$$p_{i,k} = \frac{\exp\left(-||\mathbf{x}_i - \mathbf{c}_k||^2/\sigma\right)}{\sum_{i=1}^{N} \exp\left(-||\mathbf{x}_i - \mathbf{c}_k||^2/\sigma\right)}, \forall i, k \tag{3}$$

and the objective function to be minimized is given by

$$-\sigma \log \left[\sum_{i=1}^{N} \exp\left(-\frac{||\mathbf{x}_i - \mathbf{c}_k||^2}{\sigma}\right)\right], \forall k.$$

By substituting (3) into (1), it becomes Problem (2). The proof is completed.

II. Proof of Proposition 2

Let $\mathbf{P} \in \mathcal{P}_r$ be a right stochastic matrix where $\mathcal{P}_r = \left\{\sum_{k=1}^{K} p_{i,k} = 1, p_{i,k} \geq 0, \forall i, k\right\}$, and $\mathbb{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ where $\mathbf{c}_k = h_{\mathcal{G}}(\mathbf{y}_k)$, and $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K] \in \mathbb{R}^{D \times K}$. Following the proof of Proposition 1, we have the optimal solution of minimizing (1) with respect to $\mathbf{P}$ as

$$p_{i,k} = \frac{\exp\left(-||\mathbf{x}_i - \mathbf{c}_k||^2/\sigma\right)}{\sum_{k=1}^{K} \exp\left(-||\mathbf{x}_i - \mathbf{c}_k||^2/\sigma\right)}, \forall i, k, \tag{4}$$

where the optimal solution $\mathbf{c}$ is achieved by solving the following optimization problem

$$\max_{\mathbb{C}} \sigma \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \exp\left(-\frac{||\mathbf{x}_i - \mathbf{c}_k||^2}{\sigma}\right)\right]. \tag{5}$$

For a fixed σ, the above objective function is equivalent to the maximum likelihood function of $\mathbb{X}$ which are i.i.d. drawn from the mixture of Gaussian distributions consisting of K components each which is a Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{c}_k, \frac{1}{2}\sigma\mathbf{I})$ and discrete uniform distribution for each component, i.e. $\frac{1}{K}$.

Proof of Theorem 1

Let $\{\mathbf{W}_\ell, \mathbf{C}_\ell, \mathbf{P}_\ell\}$ be a solution obtained in the ℓth iteration. By Algorithm 1, at the $(\ell+1)$th iteration, since each subproblem is solved exactly, we have

$$\varrho(\mathbf{W}_\ell, \mathbf{C}_\ell, \mathbf{P}_\ell) \geq \varrho(\mathbf{W}_\ell, \mathbf{C}_\ell, \mathbf{P}_{\ell+1}) \geq \varrho(\mathbf{W}_{\ell+1}, \mathbf{C}_\ell, \mathbf{P}_{\ell+1})$$
$$\geq \varrho(\mathbf{W}_{\ell+1}, \mathbf{C}_{\ell+1}, \mathbf{P}_{\ell+1}).$$

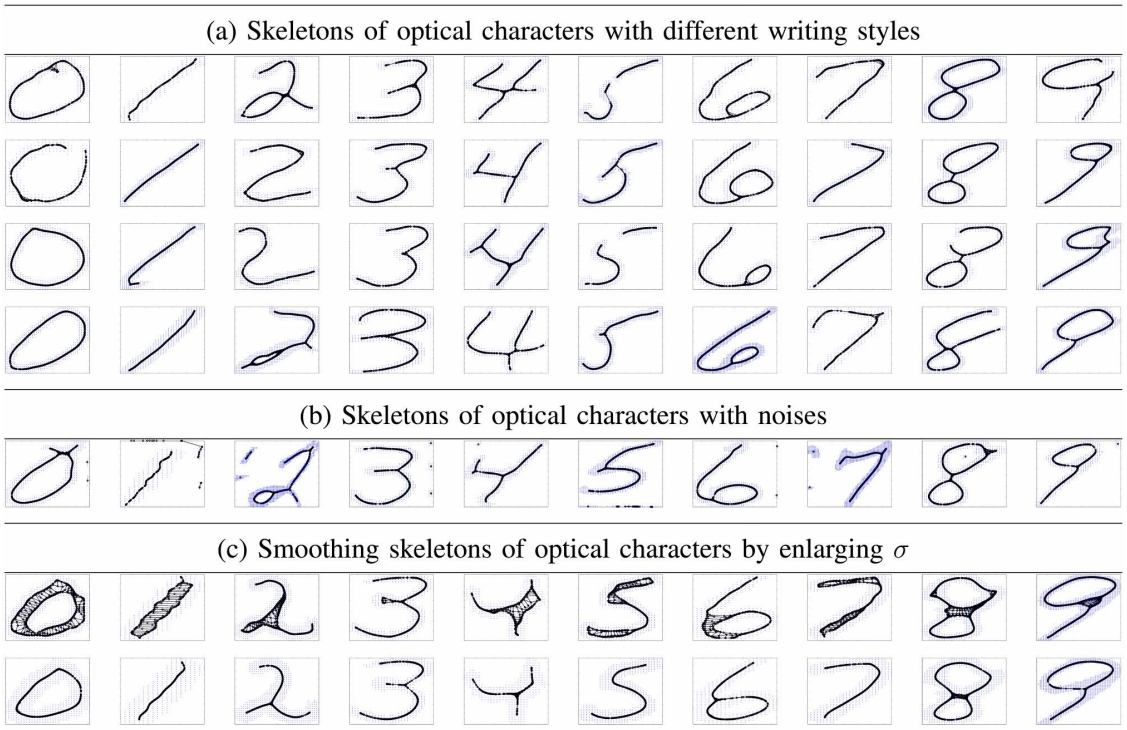


Fig. 1. Results of the proposed RPG-ℓ_1 in finding the smoothing skeletons of optical characters: (a) the learned smoothing skeletons of optical characters with different writing styles; (b) the learned smoothing skeletons of optical characters with noises; (c) the adjusted smoothing skeletons of optical characters by enlarging σ properly.

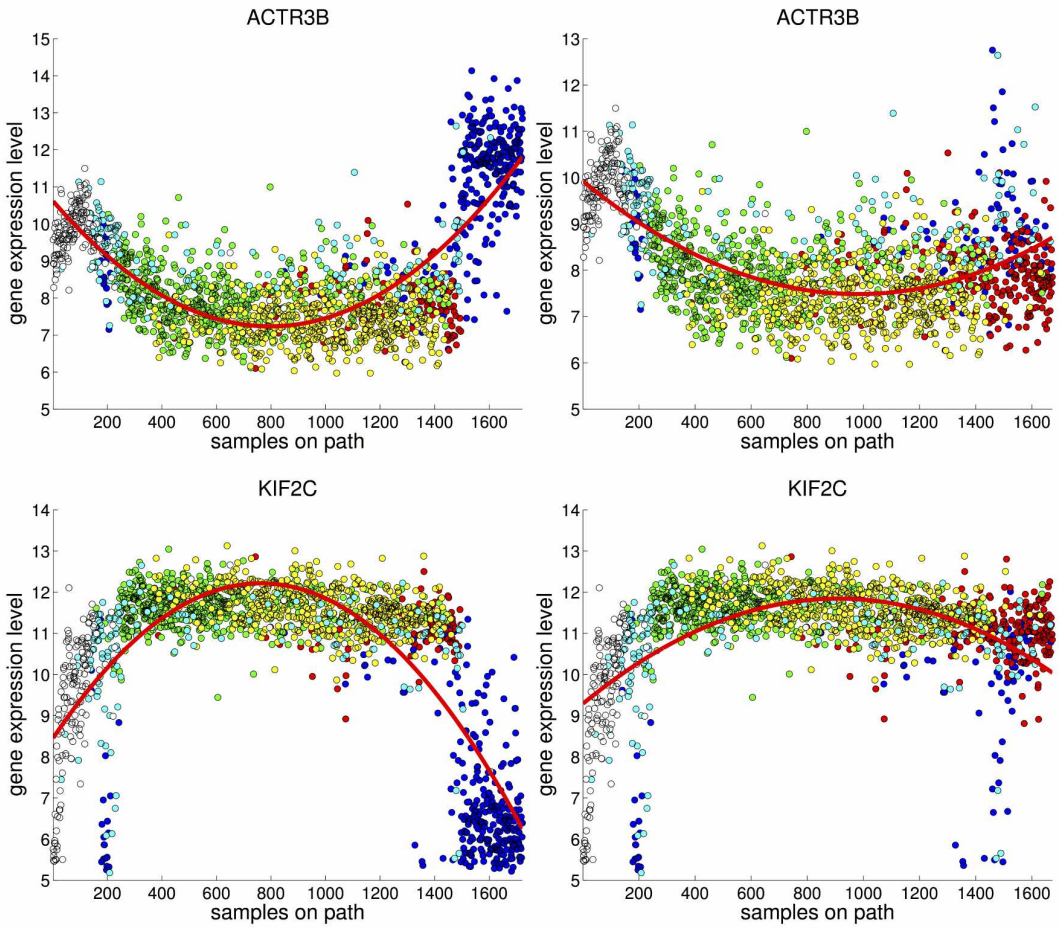
So sequence $\{\varrho_\ell\}$ is monotonically decreasing. Furthermore, function $\varrho(\mathbf{W}, \mathbf{C}, \mathbf{P})$ is lower-bounded, and then by Monotone Convergence Theorem, there exists $\varrho^* \geq -\gamma\sigma N \log N$, such that $\{\varrho_\ell\}$ converges to ϱ^*.

Next, we prove that the sequence $\{\mathbf{W}_\ell, \mathbf{C}_\ell, \mathbf{P}_\ell\}$ generated by Algorithm 1 also converges. Due to the compactness of feasible sets $\mathbf{W}$ and $\mathbf{P}$, we have the sequence $\{\mathbf{W}_\ell, \mathbf{P}_\ell\}$ converges to $\{\mathbf{W}^*, \mathbf{P}^*\}$ as $\ell \to \infty$. Since $\mathbf{C}_\ell = \mathbf{X}\mathbf{P}(2\gamma^{-1}\mathbf{L} + \Lambda)^{-1}$, $\{\mathbf{Z}_\ell\}$ converges to $\mathbf{C}^* = \mathbf{X}\mathbf{P}^*(2\gamma^{-1}\mathbf{L}^* + \Lambda^*)^{-1}$, where $\mathbf{L}^* = \text{diag}(\mathbf{W}^*\mathbf{1}) - \mathbf{W}^*$ and $\Lambda^* = \text{diag}(\mathbf{1}^T\mathbf{P}^*)$.

III. More Experimental Results

A. More Results for Finding Skeletons of Optical Characters

Fig. 1 shows the results of the smooth skeletons on different character templates with various written styles. Skeletons are correctly identified as shown in Fig. 1(a). We also observed as shown in Fig. 1(b) that noises in the characters can also be tolerated and distinguished from the skeleton by the proposed method. A small percentage of characters do not achieve a smoothing effect as shown in Fig. 1(c). This problem can be effectively solved by increasing σ to 0.02



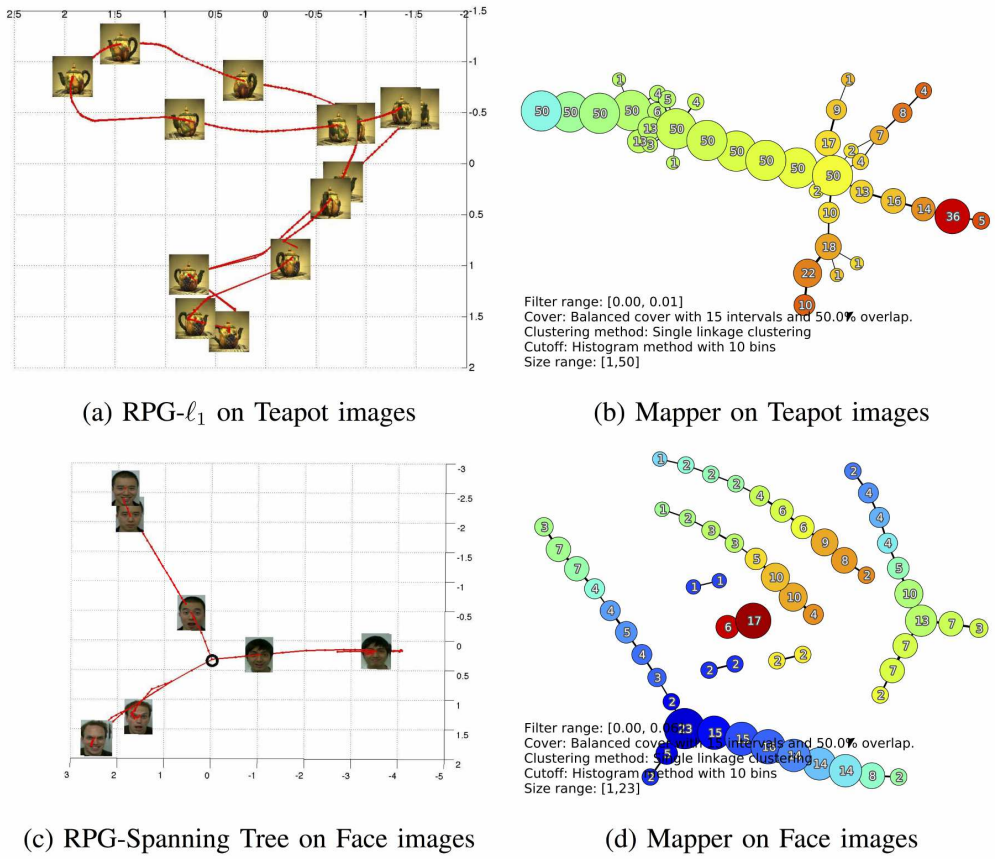
(a) normal to Basal (b) normal to Her2+

Fig. 2. Deep analysis on the progression path of breast cancer data learned by the proposed method. (a) the fitted curves of gene expression of one gene with respect to all samples on the path from normal to Basel by using quadratic function; (b) the fitted curves of gene expression of one gene with respect to all samples on the path from normal to Her2+ by using quadratic function.

as shown in the last row of Fig. 1. All these observations imply that our proposed method for learning an ℓ_1 graph can effectively deal with loops, bifurcations, self-intersections, and multiple disconnected components. Hence, it is a powerful tool for various problems in the scientific exploratory research.

B. Deep Analysis of Breast Cancer Data

We also conducted deep analysis of gene expression on the progression path learned by the proposed method. As data from human tissue molecular profiling accumulates, advanced computational analyses that can provide insights into disease progression become increasingly



(a) RPG-ℓ_1 on Teapot images

(b) Mapper on Teapot images

(c) RPG-Spanning Tree on Face images

(d) Mapper on Face images

Fig. 3. Visualization results obtained by our proposed methods and Mapper.

important. The implementation of the proposed method can facilitate the derivation of interactive progression models that can enable the identification of the key molecular events associated with carcinogenesis and malignancy. To further demonstrate the importance of learning an explicit tree structure, we extract two longest paths by traversing the learned tree from one leaf node labeled as normal sample to the other leaf node labeled as either basal or HER2+, respectively. The path from normal to basal consists of 1722 samples, while the path from normal to HER2+ consists of 1673 samples. We then map onto the two extracted paths the expression data of the 50 genes used in PAM50 subtyping. Fig. 2(a) and (b) show the trends of gene expression variations of some selected genes (ANLN, FGFR4, ACTR3B, and KIF2C) along the two paths. These genes might be involved in the process of cancer development from luminal B to either basal or HER2+. For example, we can see that the expression level of the ACTR3B gene increases abruptly in the basal branch.

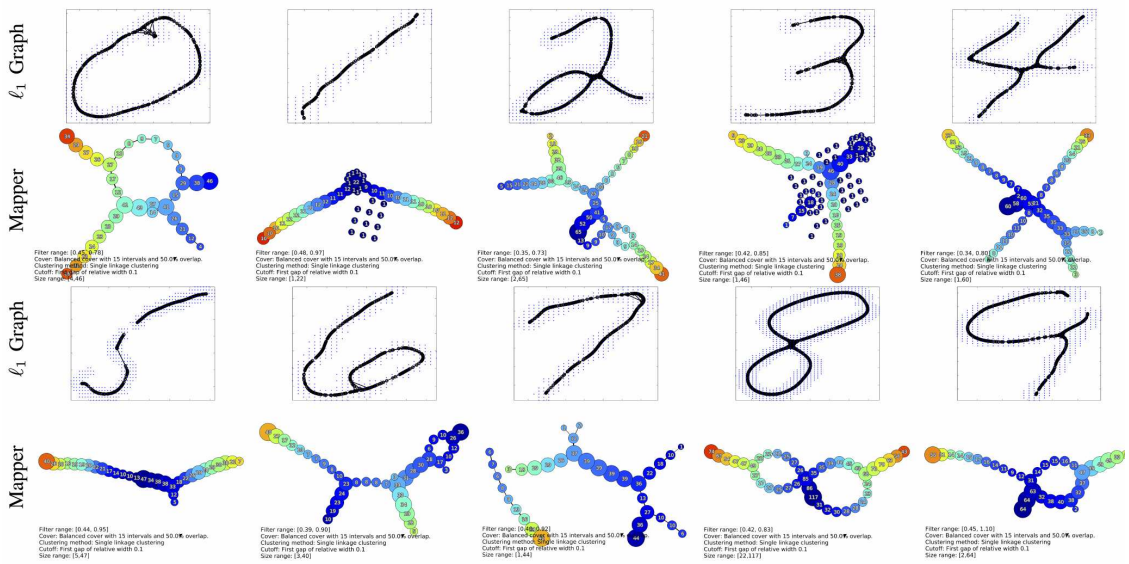


Fig. 4. Results of the proposed RPG-ℓ_1 and Mapper in finding the smoothing skeletons of optical characters.

C. More Comparing Results using Mapper

We conducted experiments by using Mapper, a topological data analysis tool[1], to learn underlying structures of given datasets. This method has been successfully applied to 3D object recognition [3] and breast cancer analysis [4]. The results of our proposed methods and Mapper on Teapot images and Facial expression images are shown in Fig. 3. Experimental results of skeletons of optical characters are shown in Fig. 4. Experimental results of breast cancer data using three methods are shown in Fig. 5. These results clearly demonstrate that our proposed methods can more correctly uncover the underlying structure of datasets by comparing with Mapper.

[1]http://danifold.net/mapper/index.html

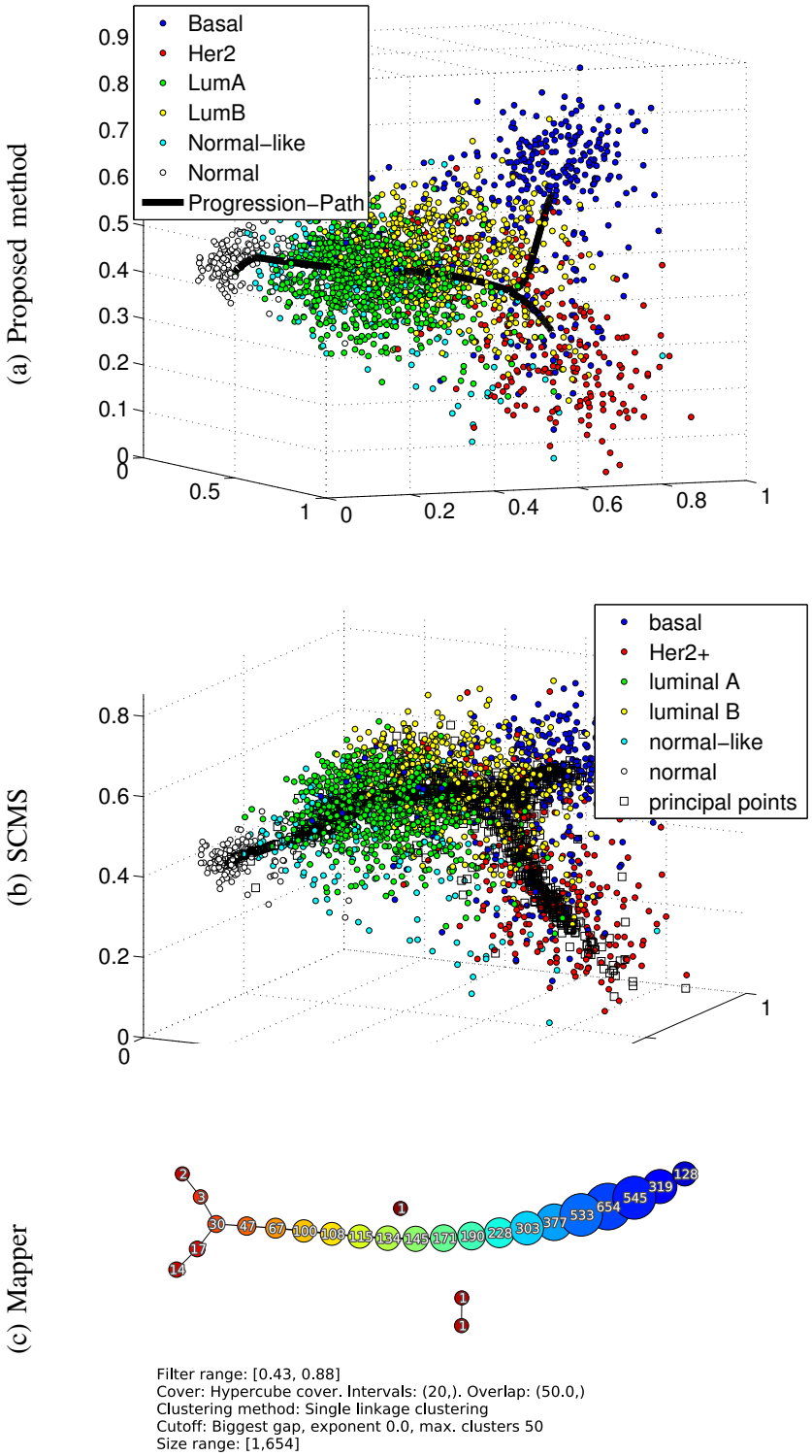


Fig. 5. Progression path of breast cancer data. (a) the tree structure learned by the proposed method for learning a spanning tree; (b) the principal points learned by SCMS; (c) the topological structure obtained by mapper.

REFERENCES

[1] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE T-PAMI*, vol. 17, no. 8, pp. 790–799, 1995.

[2] S. Boyd and L. Vandenberghe, *Conex Optimization*. Cambridge University Press, 2004.

[3] G. Singh, F. Mémoli, and G. E. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3d object recognition." in *Eurographics Symposium on Point-Based Graphics*, 2007.

[4] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 7265–7270, 2011.