Latent space modeling of seismic data: An overview

BRADLEY C. WALLET, MARCILIO C. DE MATOS, and J. TIMOTHY KWIATKOWSKI, The University of Oklahoma
YOSCEL SUAREZ, Chesapeake Energy

Modeling of seismic data takes two forms: those based on physical or geological (phenomenological) models and those that are data-driven (probabilistic) models. In the phenomenological approach, physical and geologic models are tied to seismic data either through geologic analogs or principles of structural deformation and sedimentary deposition. The results are then compared to the observed data, and the model is iterated as necessary to improve agreement. In contrast, probabilistic modeling looks at patterns in the data. The data could include raw seismic observations or seismic attributes. Probabilities can then be assigned to observations or potential observations; however, many common techniques such as neural networks and clustering do not explicitly take this step.

While phenomenological modeling has the strength that it is tightly linked to geology, it is a hypothesis-driven approach and lacks the flexibility of more exploratory, data-driven methods. Furthermore, the probabilistic approach has attractive properties in that it can provide a strong quantitative assessment of reservoir uncertainty. However, probabilistic modeling of seismic data can become mathematically infeasible due to high dimensionality that introduces high levels of statistical uncertainty.

Wallet and Marfurt (2008) observed that each seismic attribute can be represented as a dimension of a d-dimensional space in which the data reside where d is the number of attributes. A common method for dealing with high-dimensional data is to reduce the dimension using a linear projection (Guo et al., 2006). In modeling spectral decomposition of seismic data, the attribute space may be several hundred dimensions. In the concrete example presented later in this paper, we model 16-sample vertical windows of seismic traces. Each of these 60-ms waveforms, represented by a 16-dimensional vector, resides in a 16-dimensional attribute space. Throughout the rest of this paper, we will refer to waveforms, and these should be understood to be observations in attribute space.

The approach discussed in this paper is to model the data as a lower-dimensional manifold representation of a latent space embedded in attribute space. A manifold can be thought of a space that can be approximately mapped into a Euclidean space. For instance, a one-dimensional manifold can take the form of a line or a curve that can be straightened out to form a one-dimensional Euclidean space. A latent space is a lower-dimensional manifold embedded in attribute space that approximately contains the vast majority of the probability mass. If a latent space model is correct, virtually all observed waveforms in a seismic data set should fall close to the latent space. Figure 1 shows a pedagogical example of a one-dimensional latent space in a two-dimensional attribute space. In this case, the data should be mapped into the latent space by orthogonal projection with the manifold of the latent space corresponding with straightening the green line.

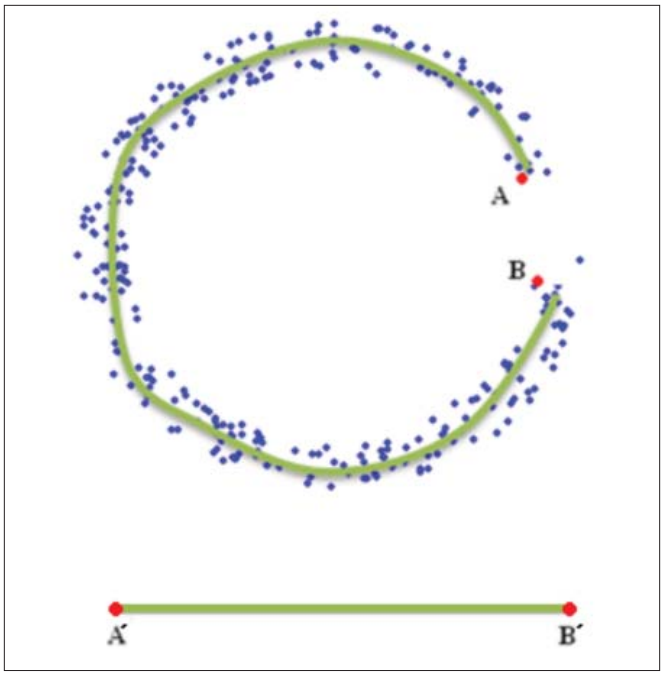

Figure 1. *A pedagogical example of a one-dimensional latent space manifold embedded in two dimensions. Note that while points A and B are relatively close in Euclidean distance in the attribute space, they are extremely distant when mapped into the latent space as points A' and B'. The green curve represents a possible latent space that might explain this data set. A point distant from this green line would be of very low probability and would be considered implausible.*

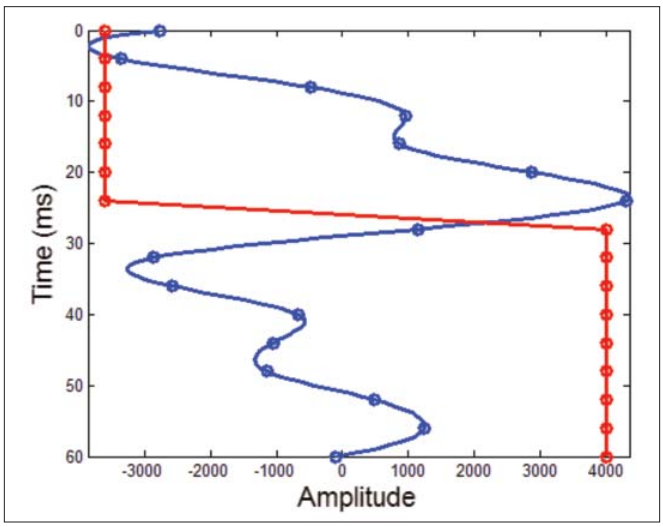

Figure 2. *Seismic waveforms (or vectors) representing two points in the set of possible seismic waveforms of 16 samples in length. The blue segment is taken from the real data set and is entirely plausible while the red segment is artificially created and is highly improbable.*

Figure 2 shows the motivation behind our effort. This figure shows two points in attribute space (waveforms) in the set of all possible waveforms. One of these waveforms

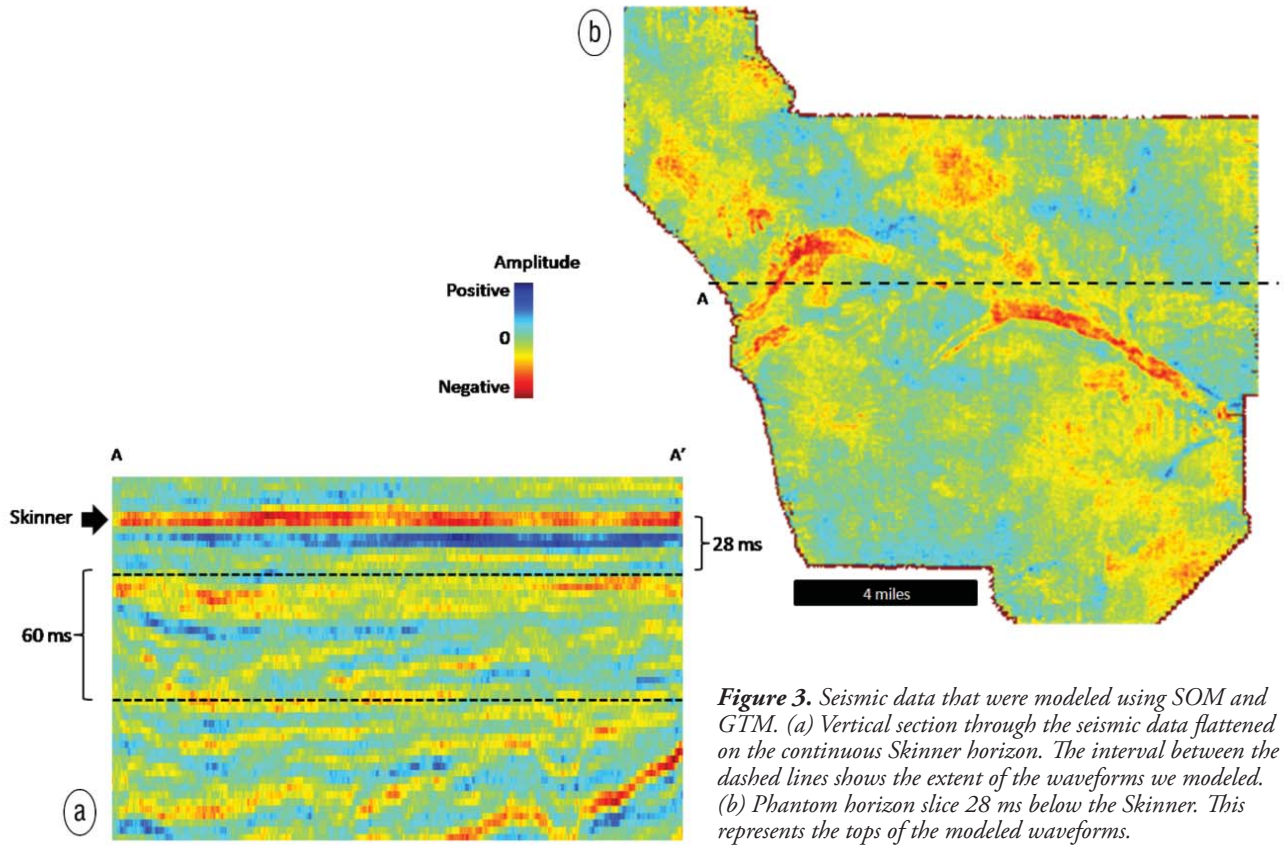

Figure 3. *Seismic data that were modeled using SOM and GTM. (a) Vertical section through the seismic data flattened on the continuous Skinner horizon. The interval between the dashed lines shows the extent of the waveforms we modeled. (b) Phantom horizon slice 28 ms below the Skinner. This represents the tops of the modeled waveforms.*

is an actual observation while the other is artificially created to be improbable. In this concrete example, it would be improbable since seismic data are band-limited, and the sudden discontinuity would appear outside of the seismic frequency band. Our approach seeks to model the regions of attribute space that are associated with plausible observations while not modeling regions associated with implausible observations. In keeping with our general philosophy of letting the data speak for themselves, the space of plausible observations, the latent space, is garnered from the data set being modeled. Hence, what is improbable is estimated based upon the data.

In summary, the goal of latent space modeling is to take our attribute data and map them into a lower-dimensional space. Once the data are mapped in this way, they may be visualized as an image or used for further processing such as in pattern recognition or facies analysis. To keep things simple in this paper, we focus upon concrete methods for estimating or learning these latent spaces from a data set. In statistics and computer science, the process of gaining information from a data set is often referred to as learning, and we will use this term to describe estimating the latent space from a data set. We will define these methods without giving details of their mathematics or implementation; then, we will discuss their relative merits and show some results of their application to seismic data.

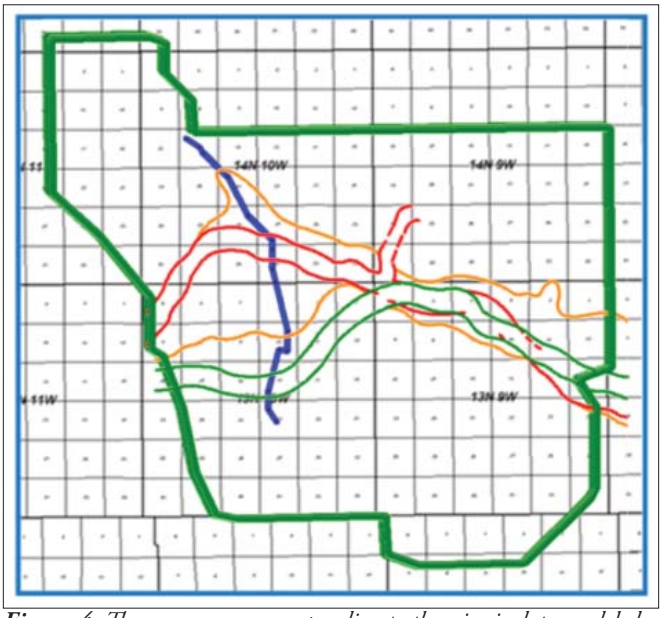

Figure 4. *The survey area corresponding to the seismic data modeled in this paper contains a series of incised channels with varying characteristics. These channels have been previously mapped using a combination of well logs, seismic interpretation, and seismic attribute analysis. A detailed discussion of the various channels present in this system can be found in Suarez et al. (2008).*

We will discuss three methods of learning the latent space corresponding to the embedded manifold: self organizing maps (SOM), generative topographical maps (GTM), and diffusion maps. These methods differ in their assumptions, strengths, and weaknesses. SOM learn the latent space by simultaneously clustering the data and ordering the clusters

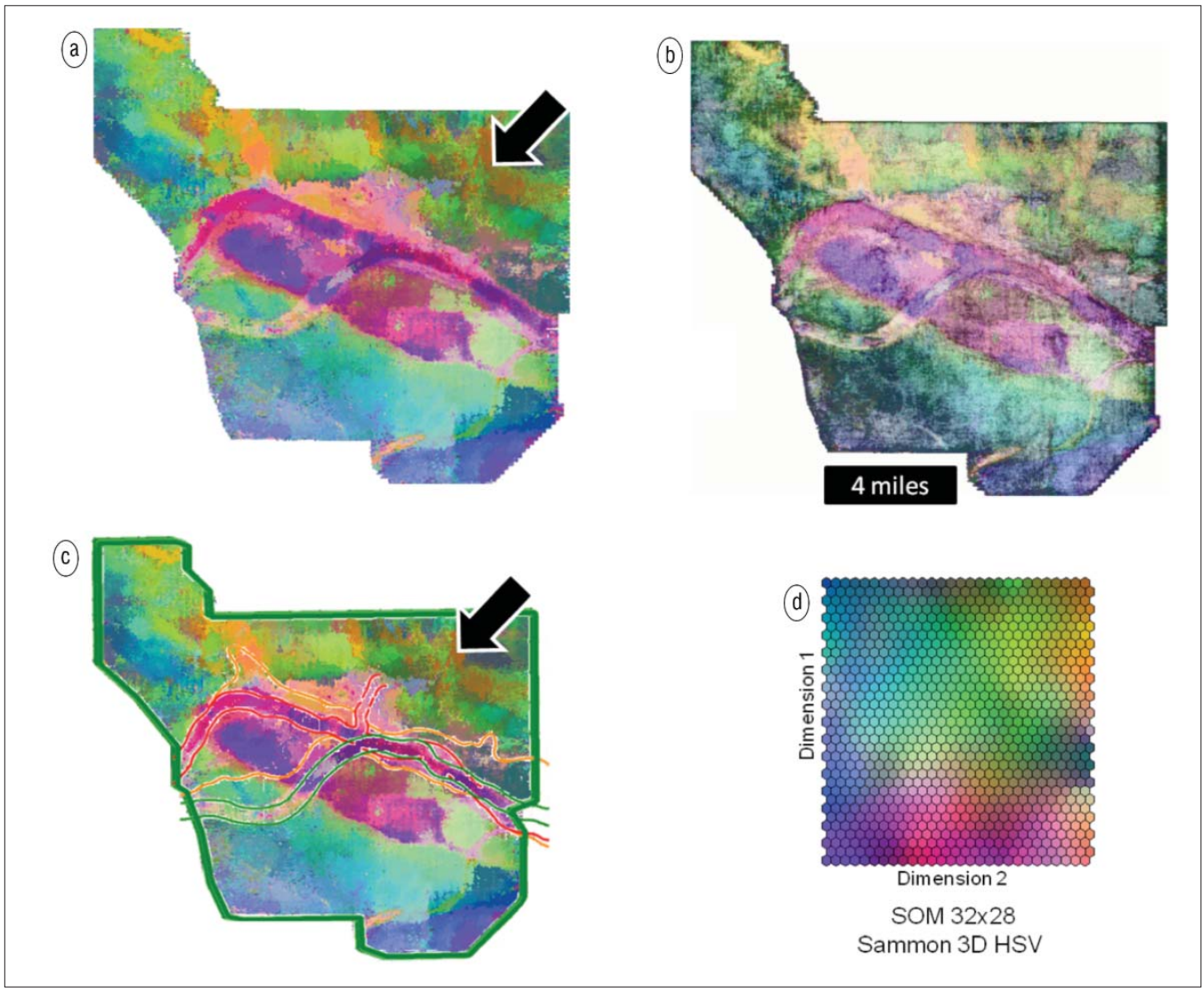

Figure 5. *Images related to running SOM upon the data set. The new images agree with the previous interpretation while providing additional richness that promises to aid in future analysis. Additionally, the black arrows denote what we interpret to be a fan feature not previously seen. (a) The output SOM with the latent space mapped vectors displayed as an image. (b) The SOM image blended with a coherency image. (c) The interpretation shown in Figure 4 overlain on the SOM image. (d) The color map used in the display of the SOM image (de Matos et al., 2009). Each hexagon represents a cluster with images (a) through (c) being colored with this color map.*

along the manifold. GTM learn the latent space by estimating a maximum likelihood solution of a constrained probability density function (PDF). Diffusion maps learn the latent space by learning the manifold defined by principal interpoint connectivities.

Self organizing maps

Self organizing maps (SOM) learn the latent space by a recursive clustering algorithm. An initial manifold is selected and uniformly populated with cluster centers. The observed waveforms are then recursively entered into the model in a random manner. Each observation is mapped to a neighborhood of closest clusters defined by point-to-cluster distances, and the clusters are subsequently updated, thereby pulling the latent space to better fit the data. It is superior to the commonly used k-means algorithm as it assigns the ordered clusters which can be used with an ordered color map (Coleou et al., 2003), and it is this ordering that justifies categorizing SOM as a method of latent space modeling.

While the clusters themselves are defined in the original n-dimensional space, they are mapped into a lower-dimensional, typically one- or two-dimensional, latent space. Each waveform can be mapped into the latent space according to its nearest clusters. Since the representation is defined by this set of clusters, SOM is a form of vector quantization.

SOM has a number of strengths. It is easy to implement. Furthermore, it is relatively computationally inexpensive both in terms of memory and processing. It is also well understood, and it has been the subject of a considerable body of research and commercial software development. However, it does have a number of weaknesses. The most obvious of these is related to the initialization. The resultant model depends upon the initial conditions and the order in which the data are incorporated into the model. Furthermore, while SOM learns the

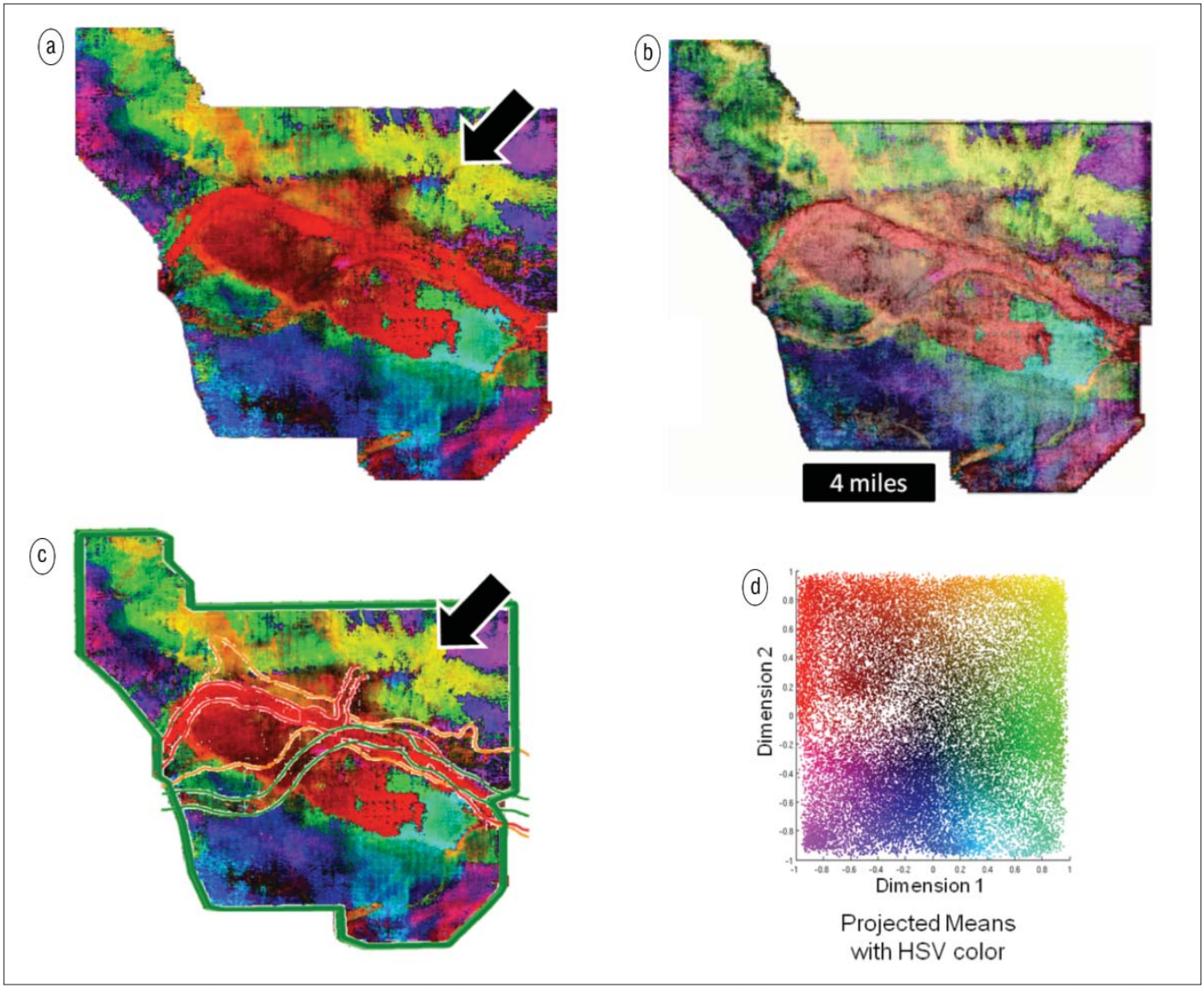

Figure 6. *Images related to running GTM upon the data set. The new images agree with the previous interpretation while providing additional richness that promises to aid in future analysis. Additionally, the black arrows denote what we interpret to be a fan feature not previously seen. (a) The output GTM with the latent space mapped vectors displayed as an image. (b) The GTM image blended with a coherency image. (c) The interpretation shown in Figure 4 overlain on the GTM image. (d) The color map used in the display of the SOM image. Each dot represents a cluster with the latent space thus expressed in a discrete two-dimensional space.*

latent space, there is no provision for learning the dimension of this space. Additionally, the theoretical basis for SOM is weaker than for other methods. For instance, while multiple iterations through the data set can be performed, there is no proof of convergence. Finally, though different starting conditions will result in different models, there is not an obvious criterion for model comparison.

Generative topological maps

Generative topological maps (GTM), as the name implies, learn the latent space by fitting a probability density function (PDF) to observed waveforms. Like SOM, GTM starts with an initial latent space, uniformly populated with clusters. However, the clusters in GTM are themselves parametrically defined as multivariate Gaussian distributions. In this way, the PDF is a Gaussian mixture model. The initial model is then updated using an expectation-maximization (EM) algorithm. The clusters are constrained to a uniformly spaced grid which is projected onto a changing nonlinear manifold with the EM algorithm adjusting the manifold position in data space. The EM algorithm is an iterative optimization algorithm that is guaranteed to converge to a possibly local maximum point in the likelihood surface. In other words, the latent space is calculated in such a way as to maximize the likelihood of the data.

GTM was developed to address several weaknesses in SOM, and it thus has a number of notable strengths. Firstly, it is based upon Bayesian first principles, and it is proven to converge. Like with SOM, GTM can possibly converge to different solutions depending upon initialization. However, different GTM models may be assessed and compared based upon likelihood. Furthermore, since GTM is a generative model, it may be used to generate random observations that could be useful for geostatistical applications. Finally, since

GTM is based upon mixture models, it should be possible to formulate a Bayesian solution to the problem of latent space modeling.

GTM does have a number of weaknesses. Like SOM, the dimension of the latent space must be decided upon a priori. Furthermore, GTM is more computationally demanding than SOM in terms of memory and processing requirements. However, recursive forms of the EM algorithm have been formulated, and such an approach could easily be derived for GTM which could reduce the memory demands of GTM.

Diffusion maps

Diffusion mapping, also known as "spectral clustering", learns the latent space of the data based upon principal interpoint distances between the observations. Possible distance measures include cross-correlation, L1 (Manhattan distance), and L2 (Euclidean distance). Diffusion maps work by calculating the full matrix of interpoint similarities of a data set. This similarity matrix is then normalized to sum to one along each row. In this manner, the matrix thus corresponds to the diffusion probability matrix for random jumps between points, hence the name of this method. Eigenanalysis is then performed to determine the principal axes of this similarity matrix. A detailed discussion of the mathematics of this method may be found in a number of other sources (e.g. Coifman et al., 2005).

When using diffusion maps, the dimensionality of the manifold can be decided using the eigenvalues of the distance matrix which is a major strength of diffusion maps relative to the other methods. The other major strength of this method is that it is closed form and completely deterministic while the other methods are iterative and subject to their initialization conditions.

Unfortunately, diffusion maps have two major drawbacks. The first is that it is extremely computationally demanding both in terms of processing and memory. The principal bottleneck in this regard is the need to perform an eigenvalue decomposition of an n x n matrix where n is the number of observed waveforms. For a typical data set, memory requirements to store this matrix could easily reach hundreds of gigabytes with computational requirements running to several multiples of this. Large decimation of the data set is thus often necessary. Unfortunately, this runs afoul of the second major drawback of this method. The mapped latent space is defined by the eigenvectors with each observed waveform used in the eigenvalue decomposition having a corresponding vector. Thus, observations that are not in the training data are not defined in the latent space model! Fortunately, a method for mapping arbitrary observations into this space has been developed though this can be complex to implement.

Application

We demonstrate the concept of latent space modeling by applying SOM and GTM to a seismic land survey acquired in the eastern part of the Anadarko Basin in central Oklahoma. We targeted our analysis upon the Middle Pennsylvanian Red Fork Formation. The interpretation challenge is to map a series of incised valleys (Suarez et al., 2008).

To define our attribute space, we begin by interpreting a horizon on the Upper Red Fork Formation. An image of the seismic amplitudes corresponding to 28 ms below this horizon can be found in Figure 3. An interpretation of this formation based upon seismic attribute analysis is shown in Figure 4. We then extracted 16 samples from each trace starting 28 ms below our interpreted horizon. These waveforms were then considered as a 16-dimensional attribute space. We then ran both SOM and GTM using implementations in Matlab upon this data set with the goal of learning a two-dimensional latent space, and we examined the results using a two-dimensional color mapping.

Most commercial applications of SOM use a one-dimensional latent space, with the waveforms plotted against a one dimensional "rainbow" color bar. Currently, there are no commercial implementations of GTM for seismic analysis. In our examples of both SOM and GTM, we use a two-dimensional latent space, and map the waveforms against a two-dimensional HSV color map.

Figure 5 and Figure 6 show resultant images for single runs of SOM and GTM, respectively. These figures also show the interpretation contained in Figure 4 overlain upon the images derived using the latent space projections. Examining these images shows that the previously interpreted features

are easily seen using both latent space methods. Within these features is visible additional richness that is likely to be useful in more complex analysis. In addition, both methods show what we interpret as a fan feature across the northern top of the primary channel.

Conclusions

Latent space modeling provides a useful tool for understanding and interpreting higher-dimensional data derived from seismic amplitude and attribute data. We show the application of two latent space modeling techniques, self organizing maps (SOM) and generative topographical maps (GTM). The results show that these methods can characterize depositional features that are not easily seen using conventional seismic attributes. In addition to highlighting details within the incised valleys and overbank deposits, we are also able to visualize what appears to be a corresponding fan. Like almost all attributes, latent space modeling methods are sensitive to input seismic data quality, and thereby suffer from acquisition footprint. The rich nature of the resulting projected data promises to be useful in well-log-constrained automated and semi-automated facies analysis.

While diffusion maps offer unique benefits, the method is currently too computationally intensive to handle seismic data volumes. However, Wallet and Perez (2009) show that diffusion maps are very effective in clustering well logs to form bed-set parameterization of parasequences. Currently, we are investigating using diffusion maps with a greatly decimated subset of the data as an initialization method for both SOM and GTM.

Suggested reading. "A grand tour of multispectral components: A tutorial" by Wallet and Marfurt (*TLE*, 2008). "Seismic color self organizing maps" by de Matos et al (*Proceedings of the International Congress of the Brazilian Geophysical Society*, 2009). "GTM: The generative topographic mapping" by Bishop et al. (*Neural Computation,* 1998). "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps" by Coifman et al. (*Proceedings* of The National Academy of Sciences, 2005). "Seismic attribute-assisted interpretation of channel geometries and infill lithology: A case study of Anadarko Basin Red Fork channels" by Suarez et al. (SEG 2008 *Expanded Abstracts*). "Unsupervised seismic facies classification: A review and comparison of techniques and implementation" by Coleou et al. (*TLE*, 2003). "Clustering bed sets from the Barnett Shale using diffusion map attributes" by Wallet and Perez (SEG 2009 *Expanded Abstracts*). TLE

Acknowledgments: Thanks to Chesapeake Energy for the use of their seismic data in research education. Additionally, we thank Kurt J. Marfurt for his advice and assistance in preparing this paper.

Corresponding author: bwallet@ou.edu