Visualizing High Dimensional and High Frequency Electrical Biosignals

Andrés F. Duque
Department of Mathematics & Statistics
Utah State University
andres.dc@aggiemail.usu.edu

Guy Wolf
Department of Mathematics & Statistics
Université de Montréal; Mila – Quebec AI Institute
guy.wolf@umontreal.ca

Kevin R. Moon
Department of Mathematics & Statistics
Utah State University
kevin.moon@usu.edu

Abstract

Manifold learning techniques for dynamical systems and time series have been effective at learning a low-dimensional representation for many applications. However, they are often insufficient for visualizing the global and local structure of the data. Here we present DIG[1] (Dynamical Information Geometry), a visualization method for multivariate time series data that extracts an information geometry from a diffusion framework. We implement a new group of distances in the context of diffusion operators, which may be useful to reveal structure in the data that is not accessible by the commonly used diffusion distances. We then apply our visualization tool to electrical biosignals: EEG data for visualizing sleep stages, as well as Electromyographic (EMG) data to visualize hand gestures.

1 Introduction

Manifold learning techniques are very effective at studying high dimensional data. The principal assumption behind manifold learning is that high dimensional data often encapsulate redundant information. In these cases, the data have an extrinsic dimensionality that is artificially high, while its intrinsic structure is well-modeled as a low-dimensional manifold plus noise. Following the same line of reasoning, dynamical systems and time series can be regarded as processes governed by few underlying parameters, confined in a low-dimensional manifold [2, 3].

For example, electroencephalographic (EEG) measurements can be considered in this analytical framework. The measurements are taken from different parts of the brain over time, resulting in a multivariate time series in a high dimensional space. It is known that these time series are highly correlated with each other, suggesting that there is a low-dimensional representation of the intrinsic dynamics of the brain that explain a broad spectrum of physical and psychological phenomena such as sleep stages. Additionally, it can be very useful for researchers to achieve meaningful visual representations of this phenomena in two or three dimensions to better understand the overall shape and finer patterns within the data.

We present DIG (dynamical information geometry), a dimensionality reduction tool that is designed for visualizing the inherent low-dimensional structure present in high-dimensional dynamical processes. DIG uses a diffusion framework adapted to dynamical processes, followed by an embedding of a group of information distances applied to the diffusion operator. This embedding is noise resilient

[1] DIG was initially presented by the same authors in the 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP). Here we have modified the original paper presenting DIG [1] to fit the format of this conference in addition to applying DIG to a new dataset. The description of the method in this paper is largely the same as in [1].

and presents a faithful visualization of the true structure at both local and global scales with respect to time and the overall structure of the data. We demonstrate DIG on high-dimensional EEG and Electromyographic (EMG) data.

1.1 Background

Many dimensionality reduction methods have been used for visualization [4, 5, 6, 7, 8, 9, 10]. Principal components analysis (PCA) [9] and t-distributed stochastic neighborhood embedding (t-SNE) [4] are two of the most commonly used methods for visualization. However, these and other methods are inadequate in many applications as they tend to favor one aspect of the data at the expense of the other. For example, PCA typically shows the large scale global structure of the data while neglecting the finer, local structure in the first two or three dimensions. In contrast, t-SNE focuses on the local structure and often distorts the global structure, potentially leading to misinterpretations [11]. Both methods also fail to denoise the data for visualization.

Diffusion maps (DM) is a popular nonlinear dimensionality reduction technique that effectively denoises the data while capturing both local and global structure [12]. DM has been extended to dynamical systems previously [13, 14, 15, 16]. In particular, Talmon and Coifman [14, 15] introduced an approach called empirical intrinsic geometry (EIG) that builds a diffusion geometry using a noise resilient distance, resulting in a noise-free embedding that captures the true structure of the underlying process. However, DM and its extensions like EIG are not optimized for visualization as the learned structure of the data is encoded in higher dimensions. Recently, PHATE was introduced in [17] that uses an information distance to present the information learned in DM in low dimensions for visualization. In this work, we introduce a new visualization method DIG that is well-suited for visualizing high-dimensional dynamical processes by preserving an information distance between the diffusion probabilities constructed from a noise resilient distance. This results in a visualization method that represents the true structure of the underlying dynamical process.

EEG signals have been embedded in low dimensional representations for detecting emotional states [18], pre-seizure states [3, 19] and sleep dynamics [16]. In the latter DM was implemented using several approaches to learn the distances. EIG was also applied to data including both respiratory and EEG signals [20]. Manifold learning methods for EMG have been employed to leverage classification accuracy in weight-lifting data [21]. The study via manifold learning of other biomedical signals such as ECG, can be found in [22, 23].

2 The DIG Algorithm

Here we extend principles of EIG and PHATE to dynamical systems to derive DIG. In this context, we present a family of information distances and derive some of their properties.

2.1 Diffusion with Dynamical Systems

To learn the local structure of dynamical systems we construct a matrix that encodes the local distances between data points. These local distances can be taken as an input in the PHATE algorithm to build a diffusion operator, from which information is extracted for visualization. To do this, we build upon the EIG framework [15, 14] which uses a state-space formalism (1)-(2):

$$\boldsymbol{z}_t = \boldsymbol{y}_t(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_t \tag{1}$$

$$d\theta_t^i = a^i(\theta_t^i)dt + dw_t^i, \;\; i = 1, \ldots, d. \tag{2}$$

The multivariate time series $\boldsymbol{z}_t$ represents the observed time series data while $\boldsymbol{\theta}_t$ represents the hidden (unobserved) states that drive the process. $\boldsymbol{z}_t$ can be viewed as a corrupted version of a clean process $\boldsymbol{y}_t$ that is driven by the hidden states, where the corruption $\boldsymbol{\xi}_t$ is a stationary process independent of $\boldsymbol{y}_t$. In general, we can view $\boldsymbol{y}_t$ as being drawn from a conditional pdf $p(\boldsymbol{y}|\boldsymbol{\theta})$. In the stochastic process (2), the unknown drift functions a^i are independent from θ^j, $j \neq i$. Therefore, we assume local independence between θ_t^i and θ_t^j, $\forall i \neq j$. The variables w_t^i are Brownian motions.

The pdf $p(\boldsymbol{z}|\boldsymbol{\theta})$ is a linear transformation of $p(\boldsymbol{y}|\boldsymbol{\theta})$ [15, 14]. Since the pdfs are unknown, we use histograms as their estimators. Each histogram $\boldsymbol{h_t} = (h_t^1, \ldots, h_t^{Nb})$ has Nb bins, and is built with the observations within a time window of length L_1, centered at $\boldsymbol{z_t}$. The expected value of the histograms,

e.g. $\mathbb{E}(h_t^j)$, is a linear transformation of $p(\boldsymbol{z}|\boldsymbol{\theta})$. Since the Mahalanobis distance is invariant under linear transformations, the following distance in the histograms space thus is noise resilient [14]: $d^2(\boldsymbol{z}_t, \boldsymbol{z}_s) = (\mathbb{E}(\boldsymbol{h}_t) - \mathbb{E}(\boldsymbol{h}_s))^T (\boldsymbol{C}_t + \boldsymbol{C}_s)^{-1} (\mathbb{E}(\boldsymbol{h}_t) - \mathbb{E}(\boldsymbol{h}_s))$, where C_t and C_s are the covariance matrices in the histograms space, in a time window of length L_2, centered at $\boldsymbol{h}_t$ and $\boldsymbol{h}_s$, respectively. Under certain assumptions, $d(\boldsymbol{z}_t, \boldsymbol{z}_s)$ is a good approximation of the distance between the underlying state variables [14]: $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_s\|^2 \approx d^2(\boldsymbol{z}_t, \boldsymbol{z}_s)$.

To learn the global relationships from the local information encoded by the Mahalanobis distances, we first apply the α-decay kernel from PHATE to construct affinities. The diffusion operator P is constructed by row-normalizing the resulting kernel matrix and global relationships are learned via the diffusion process.

2.2 Embedding Information Distances

Information is often extracted from the diffusion operator P by either eigendecomposition or by directly embedding the diffusion distances (e.g. via MDS). However, the former typically fails to provide a low-dimensional representation that is sufficient for visualization while the latter can result in unstable embeddings in some cases [17]. To overcome this, DIG extracts the information from the diffusion operator by embedding an information distance instead. We focus on a broad family of information distances that are parametrized by γ:

$$D_t^\gamma(\boldsymbol{z}_i, \boldsymbol{z}_j)^2 = \begin{cases} \sum_{m=1}^N \frac{(\log[P^t]_{mi} - \log[P^t]_{mj})^2}{\phi_0(m)}, & \gamma{=}1 \\ \sum_{m=1}^N \frac{([P^t]_{mi} - [P^t]_{mj})^2}{\phi_0(m)}, & \gamma{=}{-}1 \\ \sum_{m=1}^N \frac{2(([P^t]_{mi})^{\frac{1-\gamma}{2}} - ([P^t]_{mj})^{\frac{1-\gamma}{2}})^2}{(1-\gamma)\phi_0(m)}, & -1{<}\gamma{<}1. \end{cases} \tag{3}$$

The parameter γ controls the level of influence of the lower differences among probabilities in the overall distance. For example, the standard diffusion distances ($\gamma = -1$) are highly influenced by the highest absolute differences among probabilities. In contrast, the potential distances ($\gamma = 1$), which were used in PHATE, account for the relative differences between them. Thus the standard diffusion distances and the potential distances can be viewed as two extremes of a general class of distances over the diffusion geometry.

Algorithm 1 The DIG algorithm

Input: Data matrix X, neighborhood size k, locality scale α, time windows length L_1 and L_2, number of bins Nb, information parameter γ, desired embedding dimension m (usually 2 or 3 for visualization)

Output: The DIG embedding Y_m

1: $D \leftarrow$ compute pairwise distance matrix from X using the Mahalanobis distances in the histograms space
2: $K_{k,\alpha} \leftarrow$ compute local affinity matrix from D and σ_k
3: $P \leftarrow$ normalize $K_{k,\alpha}$ to form a Markov transition matrix (diffusion operator)
4: $t \leftarrow$ compute time scale via Von Neumann Entropy [17]
5: Diffuse P for t time steps to obtain P^t
6: $D_t^\gamma \leftarrow$ compute the information distance matrix in eq. 3 from P^t for the given γ
7: $Y' \leftarrow$ apply classical MDS to D_t^γ
8: $Y_m \leftarrow$ apply metric MDS to D_t^γ with Y' as an initialization

For $\gamma \in [-1, 1]$, the distance D_t^γ forms an M-divergence [24, 25]. Furthermore, when $\gamma = 0$, D_t^γ becomes proportional to the Hellinger distance, which is an f-divergence [26, 27]. f-divergences are directly related to the Fisher information and thus are well-suited for building an information geometry [28]. Therefore, f-divergences may be desirable for embedding the diffused probabilities.

We also consider another information distance based on f-divergences that has not been applied previously to diffusion operators. Since the rows of the diffusion matrix P are multinomial distributions, we can compute the geodesic distance between them using the Fisher information as the Riemannian metric [29]. This is an extension of the KL divergence or the Hellinger distance, both f-divergences, for distributions far apart from each other. These distances are defined as [30, 31]: $D(\boldsymbol{z}_i, \boldsymbol{z}_j) = 2\cos^{-1}\left(\sum_{m=1}^N \sqrt{[P^t]_{mi}[P^t]_{mj}}\right)$.

After the information distances have been obtained, DIG applies metric MDS to the information distances to obtain a low-dimensional representation. See Algorithm 1 for pseudocode summarizing DIG.

3 Experimental results

We now demonstrate DIG on two datasets. Figure 1 shows the results for EEG data from [32, 33]. The original data is sampled at 512Hz and labeled for every 30 second interval, within six sleep categories according to R&K rules (REM, Awake, S-1, S-2, S-3, S-4). Due to the lack of observations in some stages, we group S-1 with S-2, and S-3 with S-4. We band-filtered the data between 8-40 Hz, and down-sampled it to 128Hz. In Figure 2 we present DIG applied to EMG data labeled by hand gestures [34].

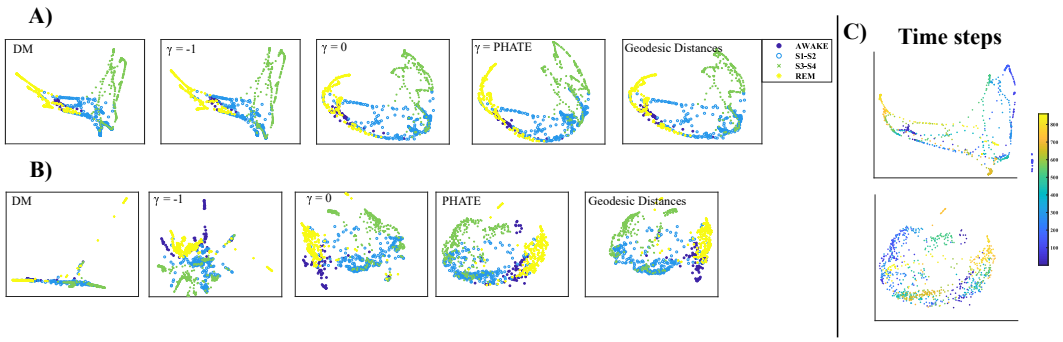

Figure 1: Impact of γ on the visualizations of EEG data from [32, 33]. **(A)** Embeddings using different values of γ colored by sleep stages. The DM embedding is obtained by the eigendecomposition of P^t, while the $\gamma = -1$ uses diffusion distances embedded with MDS. **(B)** For comparison, we also use an alternative distance to the Mahalanobis distance in the histogram space. Assuming that the data within time windows of length L_1 centered at z_t follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma_t)$, we can compute the geodesic distance between different time windows of data using the Fisher information as the Riemannian metric [31] as follows: $d^2(\boldsymbol{z}_t, \boldsymbol{z}_s) = \frac{1}{2}\sum_{i=1}^{N} \ln(\lambda_i)$,, where λ_i are the roots of, $|\Sigma_t - \lambda\Sigma_s| = 0$. The diffusion operator can then be obtained using this distance as input to the α-decay kernel. Note that the embedding is noisier using this distance, suggesting the Mahalanobis distance is effective at denoising. **(C)** Embeddings using the two histogram distances colored by time progression. The top embedding shows clearly how the subject alternates between different sleep stages.

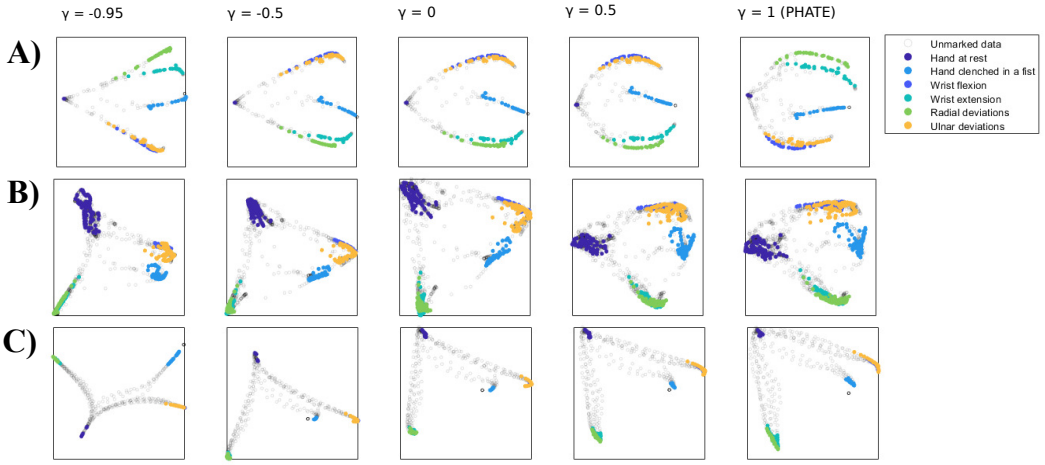

Figure 2: Embeddings of EMG data colored by hand gestures for different values of γ. **(A)** Time windows $L_1 = 100$ and $L_2 = 5$. DIG achieves a good separation among hand gestures. $\gamma = 1$ seems the best for distinguishing classes. **(B)** Time windows $L_1 = 40$ and $L_2 = 5$. A lower L_1 allows for a greater number of observations, but typically presents a noisier embedding. **(C)** Time windows $L_1 = 100$ and $L_2 = 10$. The global structure of the process is better captured in this embedding, where the progression of the unmarked data shows how the subject goes from the hand at rest to the given hand gesture, and then back to the resting position.

The experiments show how DIG is capable of discovering the underlying trajectories of the process, as well as separate different states. We also note that modifying the time window lengths provides different resolutions. For instance, the embeddings in Figure 2(**B**) give more insight into the local structure, while the global structure of the data is more evident in the embeddings in Figure 2(**C**).

References

[1] Andrés F Duque, Guy Wolf, and Kevin R Moon. Visualizing high dimensional dynamical processes. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019.

[2] R.-S. Lin, C.-B. Liu, M.-H. Yang, N. Ahuja, and S. Levinson. Learning nonlinear manifolds from time series. In *European Conference on Computer Vision*, pages 245–256. Springer, 2006.

[3] R. Talmon, S. Mallat, H. Zaveri, and R.R. Coifman. Manifold learning for latent variable inference in dynamical systems. *IEEE Transactions on Signal Processing*, 63(15):3843–3856, 2015.

[4] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[5] J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[6] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, and E.W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38, 2019.

[7] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[8] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, 2 edition, 2001.

[9] T.K. Moon and W.C. Stirling. *Mathematical methods and algorithms for signal processing*. Prentice Hall, 2000.

[10] L. van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, 10(66–71):13, 2009.

[11] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. *Distill*, 2016.

[12] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[13] W. Lian, R. Talmon, H. Zaveri, L. Carin, and R.R. Coifman. Multivariate time-series analysis and diffusion maps. *Signal Processing*, 116:13–28, 2015.

[14] R. Talmon and R.R. Coifman. Intrinsic modeling of stochastic dynamical systems using empirical geometry. *Applied and Computational Harmonic Analysis*, 39(1):138–160, 2015.

[15] R. Talmon and R.R. Coifman. Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proceedings of the National Academy of Sciences*, 110(31):12535–12540, 2013.

[16] P.L.C. Rodrigues, M. Congedo, and C. Jutten. Multivariate time-series analysis via manifold learning. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 573–577. IEEE, 2018.

[17] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.

[18] X.-W. Wang, D. Nie, and Bao-L. Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

[19] P. Ataee, A. Yazdani, S.K. Setarehdan, and H.A. Noubari. Manifold learning applied on eeg signal of the epileptic patients for detection of normal and pre-seizure states. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5489–5492. IEEE, 2007.

[20] H.-t. Wu, R. Talmon, and Y.-L. Lo. Assess sleep stage by modern signal processing techniques. *IEEE Transactions on Biomedical Engineering*, 62(4):1159–1168, 2014.

[21] Elnaz Lashgari and Uri Maoz. Electromyography classification during reach-to-grasp motion using manifold learning. *bioRxiv*, 2020.

[22] E Lashgari, M Jahed, and B Khalaj. Manifold learning for ecg arrhythmia recognition. In *2013 20th Iranian Conference on Biomedical Engineering (ICBME)*, pages 126–131. IEEE, 2013.

[23] Jeremias Sulam, Yaniv Romano, and Ronen Talmon. Dynamical system classification with diffusion embedding for ecg-based person identification. *Signal Processing*, 130:403–411, 2017.

[24] M. Salicrú and A.A. Pons. Sobre ciertas propiedades de la m-divergencia en análisis de datos. *Qüestiió: quaderns d'estadística i investigació operativa*, 9(4):251–256, 1985.

[25] M. Salicrú, A. Sanchez, J. Conde, and P.S. Sanchez. Entropy measures associated with K and M divergences. *Soochow Journal of Mathematics*, 21(3):291–298, 1995.

[26] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

[27] S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

[28] V. Berisha and A.O. Hero. Empirical non-parametric estimation of the fisher information. *IEEE Signal Processing Letters*, 22(7):988–992, 2014.

[29] S. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[30] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(Jan):129–163, 2005.

[31] C. Atkinson and A.F.S. Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 345–365, 1981.

[32] M.G. Terzano, L. Parrino, A. Smerieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep medicine*, 3(2):187–199, 2002.

[33] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, and H.E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[34] Sergey Lobov, Nadia Krilova, Innokentiy Kastalskiy, Victor B Kazantsev, and Valeri A Makarov. A human-computer interface based on electromyography command-proportional control. In *NEUROTECHNIX*, pages 57–64, 2016.