35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

# Information-Theoretic Metric Learning: 2–D Linear Projections of Neural Data for Visualization

Austin J. Brockmeier, Luis G. Sanchez Giraldo, Matthew S. Emigh, Jihye Bae,
John S. Choi, Joseph T. Francis, and Jose C. Principe

*Abstract*— **Intracortical neural recordings are typically high-dimensional due to many electrodes, channels, or units and high sampling rates, making it very difficult to visually inspect differences among responses to various conditions. By representing the neural response in a low-dimensional space, a researcher can visually evaluate the amount of information the response carries about the conditions. We consider a linear projection to 2–D space that also parametrizes a metric between neural responses. The projection, and corresponding metric, should preserve class-relevant information pertaining to different behavior or stimuli. We find the projection as a solution to the information-theoretic optimization problem of maximizing the information between the projected data and the class labels. The method is applied to two datasets using different types of neural responses: motor cortex neuronal firing rates of a macaque during a center-out reaching task, and local field potentials in the somatosensory cortex of a rat during tactile stimulation of the forepaw. In both cases, projected data points preserve the natural topology of targets or peripheral touch sites. Using the learned metric on the neural responses increases the nearest-neighbor classification rate versus the original data; thus, the metric is tuned to distinguish among the conditions.**

## I. INTRODUCTION

Although neural recordings may be very high-dimensional, often stimuli are applied or the behavior is performed in 2–D or 3–D space. This is especially true for motor and tactile experiments. The similarity among the conditions may correspond to similarity among behaviors or stimuli, such as spatial organization of the targets in a reaching task or the location of touches in a somatosensory task. In these cases, it may be possible to find a low-dimensional representation of the neural responses. If this representation preserves the relationships among the conditions, then it can be used to help understand distinctions in the neural data between these conditions.

Alternatively, simply decoding the stimulus from the neural responses can also gauge the task-relevant information carried by the neural responses, such as in decoding the movement during a natural reaching task [1]. This is especially true if the stimulus exists in a continuous space. However, the classification rate alone is insufficient to determine how the neural response varies on a trial-by-trial basis.

A number of unsupervised methods [2] have been explored to analyze the similarity between trials and the evolution of the neural response during trials. Here we have the goal of finding a low-dimensional representation for visualization that preserves similarities among conditions. The low-dimensional representation is produced by a linear projection trained using just the discrete labels corresponding to different conditions. We explore this approach on two real datasets, and quantify the performance by using nearest-neighbor assignment as a classifier on the original and projected spaces.

### A. Learning Low-dimensional Representations

Previous research has been conducted on unsupervised methods for low-dimensional representations of neural data [3], [4]. While principal component analysis may seem appropriate for the task, the first two principal components often fail to produce useful projections of neural data [5].

Non-linear dimensionality reduction algorithms produce low-dimensional representations without supervision or knowledge of the temporal ordering within trials [6]. The representations produced by manifold learning are often tuned to either preserve local similarities in data, [7], [8], or to preserve global structure. Consequently, the choice of emphasizing either local or non-local structure will influence the projection, and no explicit mapping is found to apply to novel data.

Another approach is to train state-space models to explain temporal relationships within time-series data. State-space models can easily be applied to novel data. They enable analysis of the trial-wise variance by using a low-dimensional or discrete state variable to describe the temporal evolution of the neural response. Gaussian process factor analysis [9] has been used on neural responses relating to motor planning and execution. The approach assumes all of the trials have temporal trajectories that are captured in a low-dimensional space, and the covariance of these trajectories can be described by a Gaussian kernel.

Using hidden Markov models to capture the temporal dynamics with discrete states has also proven useful for neural data analysis [10], [11], [12] A combination of state-space dynamics and a discrete state was shown to capture population responses [13]. In any case, as with purely unsupervised models, there is no guarantee that a state-space

This work was supported in part by the Univ. Florida Graduate School Fellowship and DARPA Contract N66001-10-C-2008.

A. J. Brockmeier, L. G. Sanchez Giraldo, M. S. Emigh, J. Bae, and J. C. Príncipe are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: {ajbrockmeier,principe}@cnel.ufl.edu)

J. S. Choi and J. T. Francis are with the Department of Physiology and Pharmacology, State University of New York Downstate Medical School, Brooklyn, NY 11203 USA.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

model representation, either continuous or discrete, is useful in distinguishing different conditions.

We consider the case when known labels are used in training the low-dimensional representation. A classic example of this is Fisher discriminant analysis [14], [15]. The dimensionality reduction can be posed as a metric-learning problem [16]. The goal of metric learning is to parametrize a distance function (through a projection) such that examples from the sample class are deemed close and examples from different classes are considered far apart. Note also that *no explicit classifier* is used in constructing the projection, that is, the proposed algorithm does not rely on a particular classifier or the classification error. Instead, the algorithm explored here [17] solves the metric-learning problem using information-theoretic quantities [18], [19], and a nearest-neighbor assignment is performed post-hoc. We compare against local Fisher discriminant analysis (LFDA) [15], a state-of-the-art method with an analytic solution based on a generalized eigenvalue problem.

## II. METHOD

### A. Neural data representation

Multi-electrode arrays implanted into the cortex can provide both local field potentials (LFPs) and spike trains corresponding to series of neuronal action potentials. (Here the spike trains are quantized to an instantaneous firing rate using non-overlapping fixed-width bins.) For both LFPs and firing rates, we consider a single sample from each trial as the concatenated response of all the selected channels/neurons for the entire trial. Let $\mathbf{x}_i \in \mathbb{R}^d$ denote the combined population response for the $i$th trial, $i \in \{1, \ldots, n\}$. Let $l_i \in \{1, \ldots, L\}$ denote the label corresponding to a certain condition or stimulus for the $i$th trial. We wish to find a linear projection $\mathbf{y}_i = \mathbf{A}^{\mathrm{T}}\mathbf{x}_i \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{d\times p}, p \ll d$ such that the projected points $\{\mathbf{y}_i\}$ can be used to classify and visualize the neural responses to different conditions. As discussed, learning this projection for classification is referred to as metric learning.

### B. Information-Theoretic Metric Learning

Given a set of points and labels $\{(\mathbf{x}_i, l_i)\}_{i=1}^N$, we seek to learn a positive semidefinite matrix $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, that parametrizes a Mahalanobis distance between two samples as $d(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}(\mathbf{x}-\mathbf{x}')}$. In terms of the projected samples $\mathbf{y} = \mathbf{A}^{\mathrm{T}}\mathbf{x}$ and $\mathbf{y}' = \mathbf{A}^{\mathrm{T}}\mathbf{x}'$, the metric is Euclidean $d(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{y}-\mathbf{y}')^{\mathrm{T}}(\mathbf{y}-\mathbf{y}')}$. Our goal is to find a parametrization matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$ such that the conditional entropy $S_\alpha(L|Y)$ of the labels $\{l_i\}$ given the projected samples $\{\mathbf{y}_i\}$ is minimized. (Here we use $p = 2$ so the projected data can be visualized.) We refer to this problem as *conditional entropy metric learning (CEML)*, and it can be posed as the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{A}\in\mathbb{R}^{d\times p}}{\text{minimize}} \; S_\alpha(L|Y) \\ &\text{subject to} \quad \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = p, \end{aligned} \tag{1}$$

where the trace constraint prevents the solution from growing unbounded.

Ideally, minimizing the conditional entropy $S_\alpha(L|Y)$ would require knowing the distributions of $Y$ and $L$. In practice however, these distributions are unknown and the only available information is provided by a sample $\{(\mathbf{x}_i, l_i)\}_{i=1}^N$. A common approach to this problem is to estimate the entropy of the data in a two-stage approach. First, the density of the data is estimated using methods such as Parzen windows; the approximated entropy is then computed by plugging this estimate into the entropy definition. The disadvantage of this approach is requiring the solution to a rather difficult problem (density estimation) before the desired quantity can be obtained.

The authors of [19], [17] propose an alternative method to circumvent the above two-stage process and obtain a differentiable quantity that is amenable for optimization. Instead of computing an estimator of entropy, the authors propose a quantity that exposes similar properties to Renyi's $\alpha$-order entropy and is based on the data.

Let $\mathbf{K}$ be the matrix representing the distance between samples transformed by a Gaussian function, with user parameter $\sigma$,

$$K_{ij} = \frac{1}{n}\exp\left(-\frac{(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}(\mathbf{x}_i-\mathbf{x}_j)}{2\sigma^2}\right), \tag{2}$$

and $\mathbf{L}$ be the matrix of class co-occurrences where $L_{ij} = \frac{1}{n}$ if $l_i = l_j$ and zero otherwise. The proposed conditional entropy of order alpha can be computed as:

$$S_\alpha(L|Y) = S_\alpha\left(n\mathbf{K}\circ\mathbf{L}\right) - S_\alpha(\mathbf{K}) \tag{3}$$

where $S_\alpha(\mathbf{B}) = \frac{1}{1-\alpha}\log\left(\mathrm{tr}\mathbf{B}^\alpha\right)$ and $\circ$ denotes the Hadamard product. Notice that $\mathbf{B}^\alpha$ is a matrix function for which we can use the spectral theorem to compute the gradient of (3) at $\mathbf{A}$ as:

$$\nabla_{\mathbf{A}} S_\alpha(L|Y) = \mathbf{X}^{\mathrm{T}}\left(\mathbf{P} - \mathrm{diag}(\mathbf{P}\mathbf{1})\right)\mathbf{X}\mathbf{A} \tag{4}$$

where

$$\begin{aligned} \mathbf{P} &= \mathbf{K}\circ(n\mathbf{L}\circ\nabla S_\alpha\left(n\mathbf{K}\circ\mathbf{L}\right) - \nabla S_\alpha(\mathbf{K})), && (5)\\ \mathbf{X} &= (\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_N)^{\mathrm{T}}, && (6)\\ \nabla S_\alpha(\mathbf{B}) &= \frac{\alpha}{(1-\alpha)\mathrm{tr}(\mathbf{B}^\alpha)}\mathbf{U}\Lambda^{\alpha-1}\mathbf{U}^{\mathrm{T}}, && (7)\\ \mathbf{B} &= \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} : \text{eigen-decomposition of } \mathbf{B}, && (8) \end{aligned}$$

and $\mathbf{1}$ is a $n \times 1$ vector of ones. We can use (4) to search for $\mathbf{A}$ iteratively using gradient descent, conjugate gradient, or any other method using the gradient information. Because the performance surface has local optima, initialization of $\mathbf{A}$ is important. We explore using random Gaussian matrix or using the analytic solution obtained by LFDA [15] as initialization; another option is to try multiple restarts and choose the projection that minimizes the conditional entropy $S_\alpha(L|Y)$. A more sophisticated algorithm would improve performance.

## III. Neural recordings

All animal procedures were approved by the SUNY Downstate Medical Center IACUC and conformed to National Institutes of Health guidelines.

### A. Motor Cortex During Reaching Task

A female bonnet macaque was trained to perform an 8 target center-out reaching task [20]. After the monkey became proficient at the task, a 96-channel micro-electrode array was implanted in the motor cortex (M1). Recorded firing rates from 185 units are binned into 100ms bins with 7 bins per reach trial, yielding a 1295-dimensional vector for each trial. Here we use 178 successful reach trials from one session.

### B. Cortical Somatotopy of Rat Forepaw

Cortical LFPs were recorded during natural tactile stimulation (light thwacks of forepaw digits and palm) of a female Long-Evans rat under anesthesia. The rat was anesthetized with isofluorane, and a 32-channel Michigan Probes electrode array was inserted into the hand region of primary somatosensory cortex (S1). The array had 8 contacts on 4 shanks. Another array was inserted into VPL region of the thalamus, but the signals are not used here. The LFPs were filtered with cutoffs (5Hz, 300Hz) and sampled at a rate of 1220.7Hz. The signals were digitally filtered using a 3rd order Butterworth high-pass filter with cutoff of 4Hz and notch filters at 60Hz and its first 5 harmonics.

The experimental procedure involved delivering 225 tactile touches to the rat's forepaw at 9 sites (4 digits and 5 sites on the palm) using a motorized probe. For each location, the probe was positioned 4mm above the surface of the skin and momentarily pressed down for 150ms; this was repeated 25 times at random intervals. For analysis, 170ms (208 time indexes) of the 32 channel LFP response was used; this yields a 6656-dimensional vector for each touch.

## IV. Results

### A. Motor Cortex During Reaching Task

For the reaching task experiment described in section III-A, the dimension of the vectors is greater than the number of trials. So PCA is performed on the collection of trials. The first 130 components are kept, and the components are normalized and decorrelated. The normalized components are then used as inputs to the metric learning problem, (1), (3), and (2), with an entropy order of $\alpha = 1.01$ and $\sigma = 5\sqrt{2}$. Gradient descent is run with a stepsize of 0.002 for 500 iterations.

When using all 178 trials, *CEML* is able to find a projection that separates the reach trials into discrete clusters, each corresponding to a different target. A typical projection with samples labeled by target is shown in Fig. 1. Also shown are the target directions and corresponding target index numbers. Clearly the projection preserves the relative arrangement of the target placement.

Unfortunately, this level of separation corresponds to an overfit 2–D linear projection. In order to test this, we randomly partitioned the trials into training and testing sets; 2/3 of each target's trials were used for training and the remainder for testing. Metric learning was performed using only the training set, and the test-set samples were classified by their nearest neighbor (using Euclidean distance) in the training set. To increase classification rate and avoid overfitting, only the first 32 principal components were kept (the same as the next dataset), $\sigma$ was lowered to $\sqrt{2}$, and the step size was increased to 0.1. We compared the initialization of $A$ with random entries versus using the LFDA projection. The nearest-neighbor classification was also performed on the original data and the PCA-preprocessed data. The mean and standard deviation of the classification rate for 30 Monte Carlo divisions of the dataset are shown in Fig. 2.

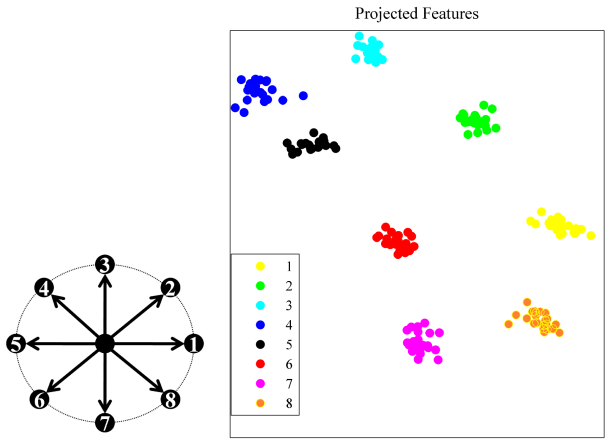


Fig. 1. (Left) Target orientation for the center-out reach task. (Right) The neural responses for all the reach trials projected into a 2–D space and colored by the reach target for each trial. The points for the same reach target are well clustered, and clusters for neighboring targets appear as neighbors in projected space: preserving the original target arrangement.

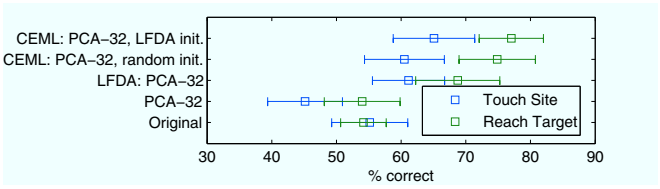


Fig. 2. Nearest neighbor prediction performance for both datasets: 2/3 training and 1/3 testing. Metric-learning is able to increase the classification rate by 10% and 20% versus the original space for the two datasets. Error bars show $\pm 1$ standard deviation across 30 divisions of the datasets.

### B. Cortical Somatotopy of Rat Forepaw

The same procedure described in the preceding section was performed on the LFPs recorded from S1 during natural touch of the forepaw, as in section III-B. Parameters were the same for both the visualization and the classification: 32 PCA components, $\sigma = \sqrt{2}$, and step size of 0.1. A typical projection with points labeled by the touch site is shown in Fig. 3. The nearest-neighbor classification results across 30-run Monte Carlo test are shown in Fig. 2.

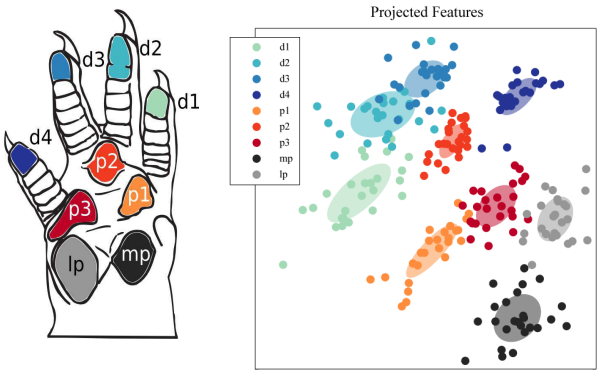


Fig. 3. (Left) The rat forepaw labeled with anatomical abbreviation and color coded. (Right) The neural responses for all the touch trials projected into a 2–D space and colored by the touch site. Each data point is a projection of the LFP response vector to a natural touch. A 2–D Gaussian distribution was fitted to each sites's points, and one standard deviation of this distribution is shown as an ellipse. The relative arrangement of the ellipses preserves the topology of the touch sites on the rat's forepaw.

## V. Discussion

The arrangement of the projected points for each dataset is strikingly similar to the underlying arrangements of the reach targets and the touch sites. Again, *CEML* has no explicit knowledge of the underlying similarity between conditions—only the discrete labels. Thus, this similarity is present in the neural responses and is preserved by the linear projection. This is understandable assuming that similar conditions have similar neural responses, such as the motor cortex where a neuron's firing rate smoothly covaries with movement direction [1].

In both datasets the projected data are clearly separated by condition. This means the data was also separable in the high-dimensional space. Here the experiments were special due to a natural 2–D representation for the different conditions. In general, separation in 2–D may not be achieved, and in those cases the classification performance would be reduced. Metric learning can also be performed in a higher-dimensional space to improve classification, but here we pursued the hybrid objective such that the metric learning also projects the data into a 2-dimensional space for easy visualization.

## VI. Conclusions

Here we are motivated by the question: would a linear projection be able to capture the similarity of neural responses of individual trials to similar but distinct conditions? We have found that a projection trained using only the condition labels preserves the similarities in the environmental space for both motor and somatosensory cortex data. To find the linear projection, a conditional entropy optimization problem is posed and solved using estimators based on the data without requiring a probabalistic model. The conditional entropy metric-learning approach seems apt for investigating the relationships between neural responses.

## Acknowledgment

Thanks to Pratik Chhatbar and Brandi Marsh for collecting the reaching task data.

## References

[1] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, no. 4771, pp. 1416–1419, Sept. 1986.

[2] M. Churchland, B. Yu, M. Sahani, and K. Shenoy, "Techniques for extracting single-trial activity patterns from large-scale neural recordings," *Current Opinion in Neurobiology*, vol. 17, no. 5, pp. 609–618, 2007.

[3] M. M. Churchland, *et al.*, "Stimulus onset quenches neural variability: a widespread cortical phenomenon," *Nature Neuroscience*, vol. 13, no. 3, pp. 369–378, Feb. 2010.

[4] A. J. Brockmeier, E. Kriminger, J. C. Sanchez, and J. C. Principe, "Latent state visualization of neural firing rates," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, May 2011, pp. 144–147.

[5] B. Cowley, M. Kaufman, M. Churchland, S. Ryu, K. Shenoy, and B. Yu, "Datahigh: Graphical user interface for visualizing and interacting with high-dimensional neural activity," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, Sept. 2012, pp. 4607–4610.

[6] J. Sammon, J.W., "A nonlinear mapping for data structure analysis," *Computers, IEEE Transactions on*, vol. C-18, no. 5, pp. 401–409, May 1969.

[7] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, p. 2323, 2000.

[8] J. B. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity," *Journal of Neurophysiology*, vol. 102, no. 1, pp. 614–635, July 2009.

[10] G. Radons, J. Becker, B. Dülfer, and J. Krüger, "Analysis, classification, and coding of multielectrode spike trains with hidden Markov models," *Biological Cybernetics*, vol. 71, pp. 359–373, 1994.

[11] E. Seidemann, I. Meilijson, M. Abeles, H. Bergman, and E. Vaadia, "Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task," *Journal of Neuroscience*, vol. 16, no. 2, p. 752, 1996.

[12] C. Kemere, G. Santhanam, B. Yu, A. Afshar, S. Ryu, T. Meng, and K. Shenoy, "Detecting neural-state transitions using hidden Markov models for motor cortical prostheses," *Journal of Neurophysiology*, vol. 100, no. 4, p. 2441, 2008.

[13] B. Petreska, B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Dynamical segmentation of single trials from population neural data," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[14] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990.

[15] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *The Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[16] E. Xing, A. Ng, M. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side constraints," *Advances in Neural Information Processing Systems*, vol. 15, 2003.

[17] L. G. Sanchez Giraldo and J. C. Principe, "Information theoretic learning with infinitely divisible kernels," `arXiv:1301.3551 [cs.LG]`, January 2013.

[18] J. C. Principe, *Information theoretic learning: Rényi's entropy and kernel perspectives*. Springer, 2010.

[19] L. G. Sanchez Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," `arXiv:1211.2459 [cs.LG]`, November 2012.

[20] J. Bae, L. Giraldo, P. Chhatbar, J. Francis, J. Sanchez, and J. Principe, "Stochastic kernel temporal difference for reinforcement learning," in *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, Sept. 2011.