Low-Rank Nonlinear Decoding of μ-ECoG from the Primary Auditory Cortex

Melikasadat Emami[1], **Mojtaba Sahraee Ardakan**[1], **Parthe Pandit**[1], **Alyson K. Fletcher**[2]
{emami,msahraee,parthepandit,akfletcher}@ucla.edu
[1]Department of Electrical and Computer Engineering, [2]Department of Statistics, University of California, Los Angeles,
420 Westwood Plaza, Los Angeles, CA 90095

Sundeep Rangan, srangan@nyu.edu
Department of Electrical and Computer Engineering, New York University,
2 MetroTech Center, Brooklyn, NY 11201

Michael Trumpis, Brinnae Bent, Chia-Han Chiang, Jonathan Viventi
{michael.trumpis,brinnae.bent,ken.chiang,j.viventi}@duke.edu
Department of Biomedical Engineering, Duke University
101 Science Drive Durham, NC 27708

Abstract

This paper considers the problem of neural decoding from parallel neural measurements systems such as micro-electrocorticography (μ-ECoG). In systems with large numbers of array elements at very high sampling rates, the dimension of the raw measurement data may be large. Learning neural decoders for this high-dimensional data can be challenging, particularly when the number of training samples is limited. To address this challenge, this work presents a novel neural network decoder with a low-rank structure in the first hidden layer. The low-rank constraints dramatically reduce the number of parameters in the decoder while still enabling a rich class of nonlinear decoder maps. The low-rank decoder is illustrated on μ-ECoG data from the primary auditory cortex (A1) of awake rats. This decoding problem is particularly challenging due to the complexity of neural responses in the auditory cortex and the presence of confounding signals in awake animals. It is shown that the proposed low-rank decoder significantly outperforms models using standard dimensionality reduction techniques such as principal component analysis (PCA).

Keywords: auditory decoding; neural networks; low-rank filter; dimensionality reduction

Introduction

Advancements in neural recording technologies, particularly calcium imaging and high-dimensional micro-electrocorticography (μ-ECoG), now enable measurements of tremendous numbers of neurons or brain regions in parallel (Chang, 2015; Fukushima, Chao, & Fujii, 2015; Stosiek, Garaschuk, Holthoff, & Konnerth, 2003). While these recordings offer the potential to observe neural activity at unprecedented level of detail, the high-dimensionality presents a fundamental challenge for learning neural decoding systems from data.

This dimensionality problem is particularly acute for the focus of this work, namely neural decoding of signals in the primary auditory signals from state-of-the-art μ-ECoG. Most importantly, in modern μ-ECoG systems, the dimensionality of the meaured responses often exceeds the number of training examples. For example, in the application we discuss below the responses from the μ-ECoG array (Insanally et al., 2016) for each stimuli consists of approximately 160 time samples across 61 electrodes, resulting in a raw feature dimension of $(160)(61) = 9760$. However, due to experimental limits on the duration of the experiments, there are less than 400 training examples. Moreover, responses in the primary auditory cortex are known to be complex (Zatorre, Belin, & Penhune, 2002; Młynarski & McDermott, 2018). Also, for awake animals, the responses may have confounding components from movements. Consequently, neural decoding systems must be sufficiently rich to enable nonlinear decoding and confounding signal rejection.

This work presents a novel approach for neural decoding from parallel neural measurements with a small number of parameters while being able to capture complex nonlinear relationships between the measurements and stimulus. The approach is based on a traditional neural network structure, but with two key novel properties: (1) A discrete-cosine transform (DCT) pre-processing stage used to reduce the sampling rate; and (2) An initial linear layer of the neural network with a low rank structure. We argue that both structures are well-justified based on the physical processes and can dramatically reduce the number of parameters. The method is demonstrated in neural decoding in the rat primary auditory cortex (A1) from a new high-dimensional μ-ECoG array (Insanally et al., 2016).

Previous work

Despite advancements in machine learning tools, traditional methods are still common in auditory decoding (Glaser, Chowdhury, Perich, Miller, & Kording, 2017). Some of these methods consider both linear and non-linear mapping of the neural responses to the auditory spectrogram (Pasley et al., 2012). Linear neural decoders like support vector machines (SVM) have also been widely discussed to classify behavioral responses using population activity (Francis et al., 2018). As in (de Cheveigné et al., 2018) other methods like canonical corre-

lation analysis (CCA) have also been used as linear models to measure the correlation between the stimulus and response as a goodness of fit after transforming them.

Multi-layer neural networks showed remarkable success in feature extraction and classification in machine vision and speech processing (Yamins & DiCarlo, 2016). Since auditory signals arrive to the cortex after having been passed through a number of sensory processing areas, these networks are appealing to model the responses in auditory cortex (Hackett, 2011).

There is also a large body of literature in dimensionality reduction methods for high-dimensional neural recordings (Cunningham & Byron, 2014; Mazzucato, Fontanini, & La Camera, 2016; Williamson et al., 2016; Sadtler et al., 2014). The methods are largely based on the unlabeled data and attempt to find a low-dimensional latent representation that can capture the bulk of the signal variance. Neural decoders can then be trained on the low-dimensional representation to reduce the number of parameters. As we will see in the results section below, our method can outperform these dimensionality reduction-based techniques since the proposed method operates on the labeled data and, in essence, find the directions of variance that are best tuned for the neural decoding task.

Model Description

We consider the problem of decoding stimuli from d-dimensional neural responses recorded from some area of the brain. Such responses can arise from any parallel measurements system including responses measured by an ECoG microelectrode array with d channels, calcium traces from d neurons, or signals recorded by the recently developed Neuropixel probes (Callaway & Garg, 2017). Let $X^i \in \mathbb{R}^{d\times T}$ be the response to some stimulus y^i recorded in a time window of length T after the stimulus is applied. Given N input-output sample pairs $\{(X^1,y^1),\ldots,(X^N,y^N)\}$, the neural decoding problem is to learn a decoder that can estimate the stimulus y from a new response X. Depending on whether the stimuli y is discrete or continuous-valued, the decoding problem can be viewed either as a classification or regression.

The key challenge in this decoding problem is the potential high-dimensionality of the input to the decoder X. Since the response X has $p = dT$ features, even linear classification or regression would require $O(dT)$ parameters. This number of parameters may easily exceed the number of trials N on which the decoder can be trained. Thus, some form of dimensionality reduction or structure on the decoder is required.

To address this challenge, we propose a novel low-rank neural network structure to reduce the number of parameters while still enabling rich nonlinear maps from the response to the stimulus estimate. Here, we present the model for a regression problem with a scalar target y. However, the same model can be used for classification or multi-target regression with minor modifications. Figure 1 shows the structure of the model proposed for decoding multidimensional neural processes. The first stage preprocesses the data by passing each of the T time samples of the d components through a discrete cosine transform (DCT). To low-pass filter the signal, only the first F coefficients in the frequency domain are retained, hence reducing the dimension from $d\times T$ to $d\times F$. The low-pass filtering is well-justified assuming that the neural responses to the stimuli are typically band-limited.

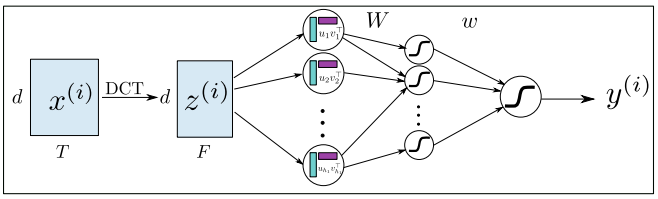

Figure 1: Schematic of the model used for decoding of a multidimensional neural process.

After the low-pass filtering, the resulting frequency-domain matrix $Z_0 \in \mathbb{R}^{d\times F}$ is passed through a neural network with two hidden layers and one output layer,

$$Z_{1j} = u_j^\top Z_0 v_j + b_{1j}, \quad j = 1,\ldots,h_1, \tag{1}$$

$$Z_{2j} = \sigma(w_{2j}^\top Z_1 + b_{2j}), \quad j = 1,\ldots,h_2, \tag{2}$$

$$\hat{y} = \sigma(w_3^\top Z_2 + b_3), \tag{3}$$

where $\sigma(t) = 1/(1+e^{-t})$ is the sigmoid function. The key novel feature of this network is in the first layer (1), where each hidden unit Z_{1j} is computed from inner product of the input Z_0 with a rank one matrix $v_j u_j^\top$. The second hidden layer (2) and output layer (3) are mostly standard. The only slightly non-standard component is that, in the output, we have assumed that the stimuli y is bounded as scaled to a range $y \in (0,1)$ so that we can use a sigmoid output.

The main motivation of the rank one structure (1) is to reduce the number of parameters. A standard fully connected layer would require $Fd+1$ parameters for each hidden unit, requiring a total of $h_1(Fd+1)$ parameters. In contrast, the rank one layer (1) uses only $h_1(F+d+1)$ parameters. We will see in the results section that this savings can be considerable.

The low rank structure can be justified, at least heuristically, under the assumption of a low rank structure of the neural responses. Specifically suppose that the frequency-domain neural responses, Z_0, are approximately given by,

$$Z_{0,if} \approx \sum_{k=1}^{h_1} \alpha_k u_{ki} v_{kf}, \quad i = 1,\ldots,d, \quad f = 1,\ldots,F, \tag{4}$$

where $\alpha = (\alpha_1,\ldots,\alpha_{h_1})$ are some latent variables caused by the stimuli, y, and u_{ki} and v_{kf} are, respectively, the responses of the latent variable α_k over the measurement channel index i and frequency index f. Under this assumption, a natural way to estimate the stimulus y, is to first estimate the vector of latent variables α from Z_0 and then estimate y from the vector α. Now, we can write (4) as $Z_0 = G(\alpha)$ where $G(\cdot)$ is a linear map. The (regularized) least squares estimate for α given Z_0 is then given by $\hat{\alpha} = (G^T G + \gamma I)^{-1} G^T(Z_0)$ for some

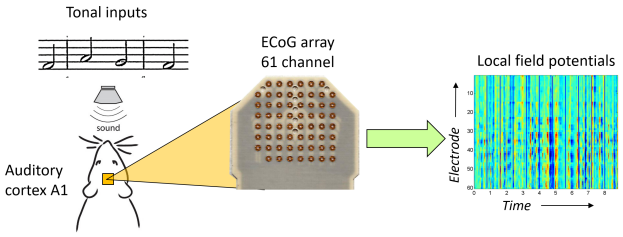

Figure 2: Schematic of the experiment setup.

regularization level γ. Due to the separability structure (4), it is easily verified that each estimate $\hat{\alpha}_k$ will be of the form,

$$\hat{\alpha}_k = \sum_{j=1}^{h_1} W_{2,kj} u_j^\top Z_0 v_j + b_{2k},$$

for some weights $W_{2,kj}$ and b_{2k}. Hence, the first layers (1) and (2) of the proposed neural network can be interpreted as recovering the latent variables under a linear low-rank output model.

Results

μECoG data from auditory cortex

We evaluate the performance of our model using *in vivo* μECoG recordings of A1 area of auditory cortex in moving rodents. Signals are recorded from a high resolution μECoG array with electrodes with $420\mu m$ spacing. The electrodes were arranged in an 8 x 8 grid where three corner electrodes were omitted (Insanally et al., 2016). In each experiment, single frequency tones with different frequencies are played for $50ms$ every second and the responses are recorded. Figure 2 shows the experiment setup and the electrode array. Recorded signals are then down-sampled to $2000Hz$ for further processing. There are a total of 390 tones played in each experiment.

Decoder performance

To train our model and test its performance, we generate a dataset $\{(u^i, X^i)\}_{i=1}^{390}$. Each sample (u^i, X^i) consists of the frequency of the stimulus as the input u^i and a $80ms$ window extracted from the signals after the stimulus is applied as the response X^i. Since the sampling frequency is $2kHz$, each X^i is a $d \times T$ matrix with $d = 61$ channels and $T = 160$ time samples. The input frequencies are shifted and rescaled to fall inside the interval $[0.1, 0.9]$. Taking the F point DCT of the signal where $F = 256$, we choose the first 55 frequencies to reduce the dimensionality. We then pass the signal through a low-rank layer with 10 rank-one units. This layer is followed by a Dense layer with 4 hidden units and sigmoid activation. The output layer is a single linear unit with a sigmoid non-linearity which gives us the predicted frequency index. We have used ℓ_2 regularization with $\lambda = 0.001$ in learning the weights of both separable and fully connected layers. The model is trained on 66% of the whole dataset and evaluated on the remaining 34% as the test set. The goal is to estimate the index of the frequency as a regression problem and R-squared score is used as a measure of closeness of data to the fitted regression model.

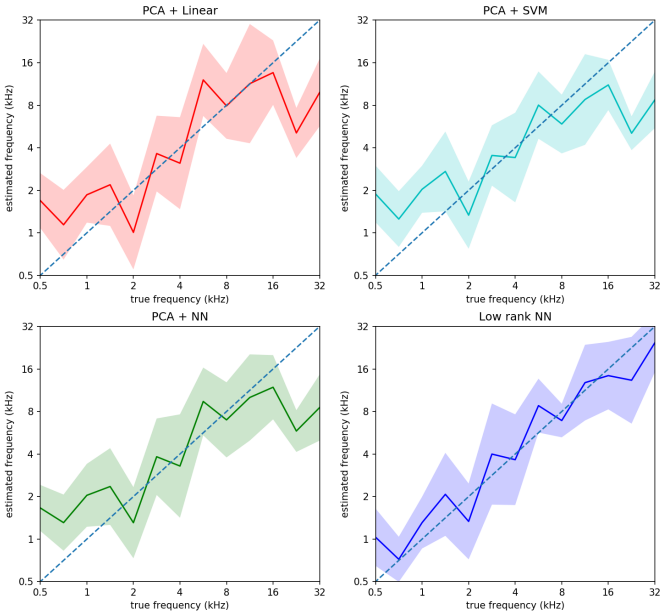

Figure 3: Plot of estimated log frequency ($\hat{f}_u$) of the stimulus against the true log frequency (f_u) for different models. Dashed line corresponds to the $\hat{f}_u = f_u$ line. The closer the estimated frequency curve to this line is, the performance of the model is better. The proposed low rank neural network model (bottom right) is performing better than the other three models.

We compare the performance of the proposed low-rank neural network with three commonly used models:

1. **PCA + linear:** top p principal components of the input are used for linear regression. There are a total of $p+1$ parameters in this model. We use both ℓ_1 and ℓ_2 regularizers.

2. **PCA + SVM:** top p principal components of the input are taken followed by a support vector regressor. There are a total of $p+1$ parameters in this model.

3. **PCA + NN:** top p principal components of the input are taken followed by a neural network with one hidden layer composed of n_h units. There are a total of $(p+2)n_h+1$ parameters in this model. We use ℓ_2 regularization for the weights.

For all three models we take top $p = 100$ principal components. Cross validation is used to tune the parameters. For SVM, both linear and radial basis function (RBF) kernel were tried and it was found that RBF gives better results.

Figure 3 shows the performance of all four models in estimating the stimulus frequency on the test dataset. Estimated frequency ($\hat{f}_u$) from each model with one standard deviation error is plotted against the true frequency (f_u). The dashed line shows the line $\hat{f}_u = f_u$, corresponding to a perfect model.

Table 1: R-squared score and RMSE of estimating log-frequency of the stimulus for different methods.

Method	R-squared score	RMSE
PCA + Linear	0.484	0.179
PCA + SVM	0.476	0.181
PCA + NN	0.510	0.174
Low-rank NN	0.761	0.121

Therefore, the distance of the prediction curve of each model to this line corresponds to the bias of the estimator and the error shades correspond to the variance of the estimator. The low-rank neural network is closest to the reference line, showing that it is performing better the other models. Table 1 summarizes the performance of each model in estimating the log-frequency of the stimulus in terms of the R-squared metric along with the root-mean-square errors (RMSE).

Conclusion

The problem of decoding multidimensional neural responses can be challenging due to high dimensionality of the data. In this work, we presented a neural network model with low-rank structure weights as the first hidden layer which significantly reduces the number of parameters compared to a fully connected network. We tested the model for decoding μECoG data recorded from A1 area of auditory cortex of awake rats. We compared the proposed model with some of the most widely used models for decoding neural signals. We showed that our model performs much better in predicting the frequency of the stimulus.

References

Callaway, E. M., & Garg, A. K. (2017). Brain technology: Neurons recorded en masse. *Nature*, *551*(7679), 172.

Chang, E. F. (2015). Towards large-scale, human-based, mesoscopic neurotechnologies. *Neuron*, *86*(1), 68–78.

Cunningham, J. P., & Byron, M. Y. (2014). Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, *17*(11), 1500.

de Cheveigné, A., Wong, D. D., Di Liberto, G. M., Hjortkjær, J., Slaney, M., & Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, *172*, 206–216.

Francis, N. A., Winkowski, D. E., Sheikhattar, A., Armengol, K., Babadi, B., & Kanold, P. O. (2018). Small networks encode decision-making in primary auditory cortex. *Neuron*, *97*(4), 885–897.

Fukushima, M., Chao, Z. C., & Fujii, N. (2015). Studying brain functions with mesoscopic measurements: Advances in electrocorticography for non-human primates. *Current opinion in neurobiology*, *32*, 124–131.

Glaser, J. I., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2017). Machine learning for neural decoding. *arXiv preprint arXiv:1708.00909*.

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hearing research*, *271*(1-2), 133–146.

Insanally, M., Trumpis, M., Wang, C., Chiang, C.-H., Woods, V., Palopoli-Trojani, K., ... Viventi, J. (2016). A low-cost, multiplexed μecog system for high-density recordings in freely moving rodents. *Journal of neural engineering*, *13*(2), 026030.

Jaderberg, M., Vedaldi, A., & Zisserman, A. (2014). Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*.

Mazzucato, L., Fontanini, A., & La Camera, G. (2016). Stimuli reduce the dimensionality of cortical activity. *Frontiers in systems neuroscience*, *10*, 11.

Młynarski, W., & McDermott, J. H. (2018). Learning midlevel auditory codes from natural sound statistics. *Neural computation*, *30*(3), 631–669.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., ... Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, *10*(1), e1001251.

Rigamonti, R., Sironi, A., Lepetit, V., & Fua, P. (2013). Learning separable filters. In *Computer vision and pattern recognition (cvpr), 2013 ieee conference on* (pp. 2754–2761).

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., ... Batista, A. P. (2014). Neural constraints on learning. *Nature*, *512*(7515), 423.

Stosiek, C., Garaschuk, O., Holthoff, K., & Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences*, *100*(12), 7319–7324.

Williamson, R. C., Cowley, B. R., Litwin-Kumar, A., Doiron, B., Kohn, A., Smith, M. A., & Byron, M. Y. (2016). Scaling properties of dimensionality reduction for neural populations and network models. *PLoS computational biology*, *12*(12), e1005141.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, *6*(1), 37–46.