II-46. Signatures of low-dimensional neural predictive manifolds

Stefano Recanatesi[1] STEFANO.RECANATESI@GMAIL.COM
Matthew Farrell[1] MSF9@UW.EDU
Guillaume Lajoie[2] LAJOIE@DMS.UMONTREAL.CA
Sophie Deneve[3] SOPHIE.DENEVE@ENS.FR
Mattia Rigotti[4] MRIGOTT@US.IBM.COM
Eric Shea-Brown[1] ETSB@UW.EDU

[1]University of Washington
[2]Universite de Montreal
[3]Ecole Normale Superieure
[4]IBM

Many of the recent advances of neural networks in time-dependent applications such as natural language processing hinge on the use of representations obtained by predictive models. This success seems to correlate with the emergence of low-dimensional representations of latent structure, when networks are trained to perform semantic relational tasks [1,2]. Motivated by the recent theoretical proposal that the hippocampus performs its role in sequential planning by organizing semantically related episodes in a relational network [3], we investigate the hypothesis that this organization results from learning a predictive representation of the world. Using an artificial recurrent neural network model trained with predictive learning on a simulated spatial navigation task, we show that network dynamics exhibit low dimensional but non-linearly transformed representations of sensory input statistics. These neural activations that are strongly reminiscent of the place-related neural activity that is experimentally observed in the hippocampus and in the entorhinal cortex [4,5]. We establish a link between place-related activity and the low-dimensional latent structure formed by predictive learning by using novel measures of representation dimensionality. More precisely, we show that this relationship can be explained by computing a *dimensionality gain* between linear algebraic and intrinsic dimensionalities of network activity. Furthermore, we provide theoretical arguments as to why predictive learning objectives necessarily imply the emergence of low-dimensional representations. We thus suggest a unifying mechanism to aid the interpretation and analysis of several experimental observations in a common framework.