Active data collection for efficient estimation and comparison of sensory processing models

Christopher DiMattina[1,3] & Kechen Zhang[2,3]

Department of Electrical Engineering & Computer Science[1], Case Western Reserve University, Cleveland, OH 44106

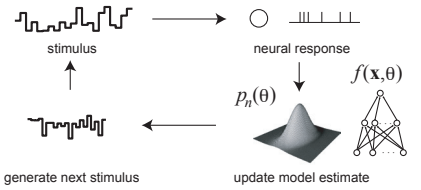
Departments of Biomedical Engineering[2] & Neuroscience[3], The Johns Hopkins University School of Medicine, Baltimore, MD 21205

290.13

Introduction

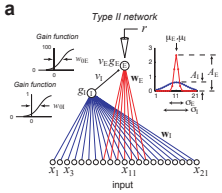
Previous work [1] has demonstrated that applying active data collection methods from the statistical theory of optimal experimental design (OED) [2] can greatly reduce the number of stimuli needed to accurately estimate the parameters of a generalized linear model, for instance the linear STRF commonly utilized in sensory neurophysiology. In this study we extend these ideas and demonstrate an OED procedure for estimating and comparing multiple competing nonlinear neural network models which generalize the linear STRF. We demonstrate that OED may be necessary for accurate estimation of nonlinear models like neural networks due to the possibility of continous parameter confounding in these models [3]. We then introduce a novel two-stage procedure for model estimation and comparison in online experiments and illustrate this procedure in the context of a hypothetical neurophysiology experiment for characterizing a nonlinear sensory neuron which encodes a particular feature conjunction.

Active data collection



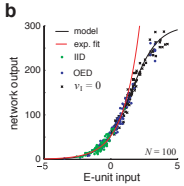
In the OED paradigm, the goal is to reliably estimate the parameters θ of an assumed underlying sensory processing model $f(x,\theta)$ which gives rise to noisy data. This can be done by iteratively choosing new stimuli x_{n+1} which minimize the expected spread or uncertainty of the subsequent posterior density $p_{n+1}(\theta)$. OED has been shown in previous studies [1] to drastically reduce the number of stimuli needed to estimate a single generalized linear model. Our goal is to extend OED to allow one to estimate and compare multiple nonlinear models.

OED may be necessary to recover model parameters

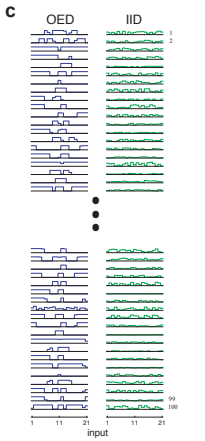


a) A simple center-surround neural network having a single inhibitory interneuron (I) which projects to the output (E) unit.

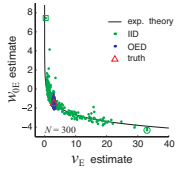
b) Random stimuli (Panel (c), 'IID') drive E-unit responses (green dots) over a limited range of the E-unit gain function (black curve) which can be fit by an exponential (red curve). In contrast, when the I-unit is removed or optimally designed stimuli (Panel (c), 'OED') are used, the E-unit is driven over its the full range of its gain function (black crosses, blue dots).



c) Optimally designed stimulus set chosen adaptively ('OED') and random stimuli ('IID').

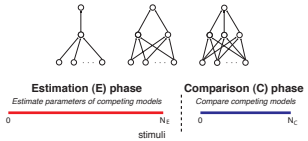


d) Estimates of two parameters of the model in (a) attained from training with both IID and OED datasets shown in (c). We see that the estimates attained from IID data (green dots) lie scattered along the confounding curve predicted from the fact that the output gain may be approximated by an exponential function. In contrast, we see that the estimates attained from OED data (blue dots) correctly recover the truth (red triangle).



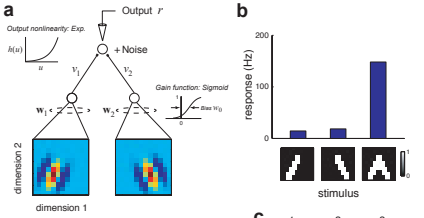
For many nonlinear models like neural networks it may be impossible to recover the true model parameters without active data collection due to the possibility of a continuum of parameter values giving rise to nearly identical input-output mappings [3].

Two-phase procedure



We utilize a two-phase OED procedure to estimate and compare competing nonlinear models on-line.

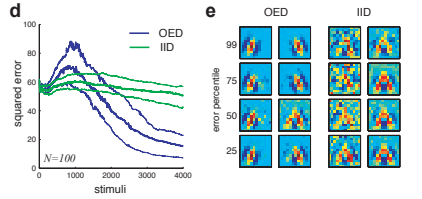
Estimating multiple neural networks on-line



a) The true nonlinear network. Our goal is to estimate the number of hidden units and the network parameters

b) The network in (a) performs nonlinear feature detection.

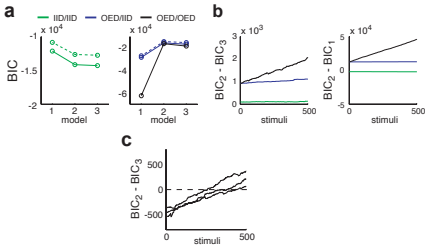
c) *Top:* OED stimuli for networks having 1, 2, 3 hidden units. *Bottom:* IID stimuli.



d) Expected estimation error for true model as a function of number of stimuli for both OED (blue) and IID (green) data collection strategies. Thin lines indicate middle 50 percent of data.

e) Recovered input weights (true values in (a)) of two subunit model at various levels of estimation accuracy for OED and IID strategies.

Comparing competing neural networks on-line



a) Relative likelihood (Bayes Information Criterion or BIC) of neural network models having 1, 2 and 3 hidden units before (dashed lines) and after (solid lines) comparison (C) phase. We see that OED C-phase stimuli generated to best discriminate the three models increase the relative likelihood of the true model (2) (solid black line), whereas IID C-phase stimuli do not increase the relative likelihood of the true model (solid green line, solid blue line).

b) Model likelihood differences during the comparison phase as a function of stimulus number. Note that OED during the comparison phase drastically increases the evidence in favor of the true model (black curves) while IID stimuli do not (green, blue curves).

c) In some experiments, an initially incorrect model choice (3 hidden unit model) after the E-phase can be corrected by collecting OED data during the C-phase.

We see from this and the previous panel that our proposed two-phase OED procedure does a much better job of accurately estimating the parameters of the true model and allowing us to correctly determine the number of subunits in the true model.

Conclusion

We recommend using active data collection to estimate and compare multiple competing nonlinear models in on-line sensory neurophysiology.

References

[1] Lewi, Butera & Paninski (2009). *Neural Comp., 21(3),* 619-687.
[2] Atkinson & Donev (1992). *Optim. Exper. Design*: Oxford Univ. Press
[3] DiMattina & Zhang (2009). *Neural Comp.,* In press.