What is a Statistical Model? (McCullagh, [2002])

Overview.

In this work, Peter McCullagh creates a set of formalisms for reasoning about statistical models in category theoretic terms. His major contribution is a framework within which he uses the commutability condition of natural transformations to formalize a few statistical intuitions.

He begins by defining categories for statistical units, responses and covariates. He then defines statistical designs, model parameters, and sample spaces as functors over these categories. He builds on this structure to define statistical models as natural transformations between statistical design and parameter functors. The commutativity of these natural transformations enforces that parameter maps maintain their structure independant of sample space transformations and that the "meaning" of a parameter does not change in response to modifications of the statistical design.

Comments.

It's interesting to note that McCullagh chose to not formulate his product categories, such as the category of factorial designs, as monoidal categories. This seems like it would be a natural fit, especially since we can represent factorial designs as functions that accept a finite length vector, and finite length vector spaces form a compact closed category.

One of the larger issues with McCullagh's framework is that it fails to represent more involved statistical models. For example, McCullagh's construction enforces that the role of a parameter should not change in response to changes in sample size. However, in certain models like a non-parametric mixture model, parameter meanings do change in response to changes in the sample size or training dataset. Even in the simple case of a Lasso-penalized model, certain sample spaces can cause model parameters to disappear.

Furthermore, McCullagh's construction fails to provide any insight into one of the most common sources of absurd statistical models, which is overly complex or unrestricted models. For example, an n-degree polynomial is rarely a good fit for a dataset where the underlying process is linear or where we have only observed n data points. Researchers have spent a great deal of time exploring model complexity, and it may be valuable to express some of the key concepts, such as VC dimension (Vapnik and Chervonenkis [2]), in category theoretic terms.

REFERENCES

[1] McCullagh, P. (2002). What is a statistical model? *Annals of statistics*, pages 1225–1267.

[2] Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer.