CS231B paper review – Arbel Harpak

Deep Learning for Population Genetic Inference / Sara Sheehan and Yun Song

---

Computing the likelihood of complex population genetic model using genomic variation data from a sample of individuals remains largely intractable, even after decades of research. A few existing approaches assume a simple demographic model and infer demography jointly with modes of selection (also simple, often with a selected/neutral dichotomy). More often, existing approaches separate both the data—into neutral markers and selection candidates—and the inference, by performing sequential inference of demography and selection separately. In recent years, demographic inference has advanced to include non-parametric inference (with PSMC as the pioneering example) and thereby achieves ever higher resolution.

In their recent work, Sheehan and Song suggest a deep learning algorithm for joint inference of (genome-wise local) selection and demography (that acts to shape variation in the genome globally). Sheehan and Song train and test their algorithm with simulated data, and later apply the algorithm to a sample of 179 genomes of flies from Zambia. The work assumes to be a 3-epoch demographic history of a single panmictic population. The loci, mode and mean intensity of positive selection are assumed to each be taken from a limited set of values; the intensity is sampled from a prior centered at the respective mean for each data point that involves selection.

For simulations, the authors used baseline demographic parameters (baseline effective population size and epoch length) and realistic recombination and mutation rates, all previously inferred by other studies or using external software (e.g. PSMC). They assume a prior distribution on the scaling factors of the epochs' effective population sizes (N_e), and simulate four modes of selection—hard sweeps, soft sweeps from standing variation, heterozygote advantage balancing selection and neutrality—with various selection coefficients (s). The selection coefficients examined are relatively high ($N_e s$ ranges between 1000 and 10,000, compatible with work by Elyashiv et al. 2016).

The input layer of the network is consisted of a selection of test statistics all based on statistics obtained from variation data in the sample alone. The output layer (logistic activation function) represents the three continuous demographic parameters and the discrete selection modes and discrete selection parameter.

The performance of demographic history inference is overall good, with small relative errors on the three epochs. Similar to other demographic inference methods, the more ancient epochs tended to have less

variable estimates; but surprisingly the bias (from the real simulated values) was higher for ancient epochs. The performance of selection inference is also good overall, with over 92% accuracy for each of the selection modes, except for hard sweeps (83%). Hard sweeps were often classified as neutral, and the authors presume this is mostly due to incomplete sweeps.

On the Zambia flies data, Sheehan and Song infer a demography which is very similar to that obtained with PSMC on the same data, and the one inferred by Thornton and Andolfatto (2006). As for selection, they find roughly tenfold more regions that are classified as hard sweeps than regions classified as soft sweeps, despite the lack of power to detect hard sweeps discussed above. Focusing on regions classified as selected with high probability (output of the last logistic layer) they find that the hard sweep top regions harbor many genes involved in chromatin assembly/disassembly. They also find high-probability hard sweeps in the *fic domain-containing protein*, associated with light detection and and visual behavior, and in the *charybde* gene, which is associated with negative regulation of growth. Interesting high-probability regions in the soft sweep category include many genes related to the transcription machinery, and a few genes associated with pheromone detection.

For importance scoring—which here corresponds to the task of identifying informative genetic variation summary statistics—the authors use two complementary approaches: permutation testing and perturbation. In the permutation approach, the simulated values of the input statistic on the test set are permutated. One than uses the reduction in accuracy as a measure of importance.

The analyses of the importance scores in both approaches outlines many interesting facets of how patterns of genetic diversity inform us on evolutionary dynamics. Many of the findings corroborate previous intuitions that are accepted as rules of thumb in the field. For example, singletons are most informative about recent demography and selection, presumably because they occur solely on terminal branches of genealogies (Field et al. 2016). Tajima's D, a summary statistic often used directly as a test for selection, seems to indeed be very informative. Pairwise LD statistics are also highly informative for selection (Tajima 1983), and some IBS patterns are highly informative for demography but not for selection (Carmi et al. 2014). The raw SFS entries remain uninformative, but one must bare in mind that this is true in the presence of a plethora of SFS summary statistics that include the use of these SFS entries.

Sheehan and Song opted to use very little regularization in their algorithm. They use L2 regularization with a single regularization parameter, and only explore it through cross-validation very briefly.

The deep learning approach elegantly overcomes difficulties in the inference of both demography and selection in recent years, as well as difficulties in their joint inference. However, some difficulties remain to limit the overall utility of the analysis. For example, Sheehan and Song use previously inferred, constant recombination rate and ratio of recombination rate to mutation rate. However, they show that their performance is very sensitive to this choice, and speculate that it is highly sensitive to variation in recombination rate, which is known to play a big role in Drosophila.

The importance scoring performed by Sheehan and Song is of high interest to the field. This is perhaps the most systematic examination of the marginal utility of various genetic variation summaries to inference of evolutionary dynamics. One drawback of the approach the authors take pertains to the equal footing of rawer statistics and expert summary statistics. For example, when they examining the marginal importance of the raw SFS alongside its summaries (e.g. Tajima's D) is somewhat unfair; it doesn't really tell us about how informative the SFS as a whole, or some subset of its entries for inference, because the expert summaries already capitalize on this same piece of information. This is an important question, for example, for understanding the effect of sample size on power in population genetic inference.

In some cases, Sheehan and Song's results are very unintuitive and remain rather mysterious. This is mostly due to the inherently limited interpretability of deep learning algorithms. For example, the authors find that—even for the inference of demographic parameters—using all sites for inference outperforms limiting the inference to neutral regions (see table 1). This contradicts basic population genetics intuition and virtually all previous inference schemes.

This work is both pioneering and very comprehensive in many respects. For example, the large range of genetic variation summaries considered, at different proximities to the focal (selected) site. In other respects—like the limited demographic models that could be explored with the existing software, caveats in interpretability and the treatment of variation in molecular rates—there's ample room for future research.

Overall, Sheehan and Song's work paves the way for deep learning in population genetics. It does very well in making the assumptions, advantages and limitations of deep learning—most importantly, those that are domain specific—very clear and accessible to attract other population geneticists to follow. This is perhaps the most important contribution of the paper.