Testing for genetic associations in arbitrarily structured populations

Lorenzo Lagos
March 31, 2017

This paper published in Nature tests for associations between traits and genetic markers in the presence of unobserved population structure. The reason why population structure can confound results is because it creates correlations between non-genetic factors and heterogeneity in genotype frequencies, which in turn generates spurious associations between some genetic markers and traits. It is interesting, and probably correct, that the authors never use the word causation. All the data they use is observational and there is no quasi-experimental variation being exploited. Instead, the general procedure seems to follow a "controlling for confounders" strategy.

The setup has unobserved population structure z, observed allele frequency $\pi(z)$, observed genetic markers x, and observed traits y. The most common approaches estimate $y|z$ to test for the association between x and y. This assumes that z is not correlated with non-genetic effects. Since the authors don't buy this assumption, their approach is to estimate $x|\pi(z)$ through logistic factor analysis (LFA) and then check whether $P(x|y, \pi(z)) = P(x|\pi(z))$. Rejecting this null hypothesis indicates an association between x and y controlling for z.

In my opinion, the neatest aspect of the authors' approach is the first stage LFA since it amounts to extracting the influence of the unobserved latent structures on genetic markers. I find this interesting because a common issue we have in causal inference is the presence of unobservables confounding our estimates. In this specific case, the unobserved structures only have an influence on traits through the alleles (which are observed). Hence, allele frequency can be used to uncover the effect of these structures on genetic markers. Unfortunately for me, this sort of situation in the social sciences is highly unlikely.