Variations on a Common STRUCTURE: New Algorithms for a Valuable Model

John Novembre
Department of Human Genetics, University of Chicago, Chicago, Illinois 60637

In this commentary, John Novembre discusses tools for the study of population structure, including the novel, fast algorithm fastSTRUCTURE presented by Raj *et al.* in the June issue of GENETICS.

Over the last 14 years, a highly influential tool for the study of population structure has been the admixture model of Pritchard, Stephens, and Donnelly (the "PSD" model, Pritchard *et al.* 2000) and its associated inference software STRUCTURE. In its basic form, this model does not explicitly consider the effects of mutation, drift, selection, or linkage. Nor is it a dynamical model, as it does not explictly have any temporal component. Viewed simply, it is just the Hardy–Weinberg model with two wrinkles: (1) subpopulations that differ in allele frequency, and (2) individuals whose genetic ancestry can be admixed, *i.e.*, an individual can inherit alleles from more than one of the multiple subpopulations according to a probability vector of "admixture proportions."

This simple model of population structure has proven incredibly useful. The original paper (also published in *GENETICS*) has garnered over 10,000 citations, and if one considers the impact of subsequent papers that elaborate on this approach directly (Falush *et al.* 2003, 2007; Hubisz *et al.* 2009) or use related approaches (*e.g.*, Dawson and Belkhir 2001; Anderson and Thompson 2002; Corander *et al.* 2003; Wilson and Rannala 2003; Huelsenbeck and Andolfatto 2007), the sum impact is truly remarkable. Interestingly, the same underlying model has played an independent and important role for text classification and mining, where it is known as the latent Dirichlet allocation (LDA) model (see Blei *et al.* 2003, which alone has over 8,000 citations). For biologists, the broad impact of the PSD model stems from the suprisingly large number of questions in evolutionary biology and ecology in which a simple assessment of population structure proves to be an insightful exercise. As an example, consider how useful it is to identify genetic subpopulations when managing a species or studying its historical biogeography. Also, consider how useful it is in a disease mapping study to be able to identify hybrid individuals, to infer the source of tissues of uncertain origins, or to assess case/control population stratification. The PSD model and its close relatives have been important in all of these applications.

While the PSD model is relatively simple, performing inference with it poses substantial computational challenges. To understand the crux of the problem, remember the original PSD paper sought to infer the vector of ancestry proportions for each individual (a vector q_i for individual i) assuming individuals are completely unlabeled with respect to origins. That is, any understanding of the subpopulation allele frequencies and the mixture proportions will need to be discerned from the genotype data alone. If one knew the population source of each allele in each individual, then the problem would be straightforward, but these are "missing data" or *latent variables*. The PSD paper approached the problem from a Bayesian perspective—it aimed to sample from the posterior on the q_i's by integrating out the uncertainty in the unobserved subpopulation allele frequencies and allelic source variables. This integration is in a highly dimensional space (with dimensions proportional to the product of sample size, the number of loci, and the number of alleles per locus) and thus it is computationally impractical to carry out the integration exactly. Thankfully, a Gibbs sampling approach allows approximation of the integral, and this is used in the algorithm that underlies the resulting software STRUCTURE.

The last 10 years have seen great strides in the scale of our observations of genetic variation, and this has been a blessing for our learning but a curse for computation.

doi: 10.1534/genetics.114.166264
Address for correspondence: Department of Human Genetics, University of Chicago, 920 E 58th Street Chicago, IL 60637. E-mail: jnovembre@uchicago.edu

STRUCTURE worked well for analyzing complete datasets until the advent of large SNP genotyping arrays with hundreds of thousands of SNPs in the mid-2000s (reviewed in Novembre and Ramachandran 2011). At that scale of data, the Gibbs sampler is simply too slow to be practically applied, and many researchers turned to using alternative approaches such as applying principal component analysis (PCA) to genotype data (Price *et al.* 2006). There are theoretic reasons why using PCA (and other forms of factor analysis) can provide insight to admixture proportions (Patterson *et al.* 2006; McVean 2009; Engelhardt and Stephens 2010), but the PSD model still is appealing as a probabilistic model-based approach for inference with admixed samples. For example, it explicitly considers how alleles within a genotype are a binomial sample from underlying subpopulation allele frequencies, rather than implicitly treating them as continuous variables (as PCA does), and thus can compute measures of uncertainty in an appropriate way.

To address the computational challenges that arise from applying the PSD model to SNP data, two groups recognized that the likelihood function underlying the model is amenable to efficient optimization techniques such that one can obtain *maximum likelihood* estimates of the ancestry proportions and allele frequencies. Tang *et al.* (2005) developed an EM algorithm distributed in their FRAPPE software, and another team (that I worked with) leveraged tools from convex optimization theory to develop the ADMIXTURE software (Alexander *et al.* 2009).

One partial drawback of the likelihood approaches is the inability to use Bayesian priors that favor the most "sensible" parameter estimates. For example, biological intuition suggests one should favor solutions in which each individual's ancestry is drawn from one or at most a few populations rather than many. Similarily one might favor solutions where the allele frequencies in all the subpopulations are similar to one another (*e.g.*, for populations that are weakly differentiated, such as $F_{ST} < 0.1$). A strict optimization of the likelihood in the PSD model does not produce such solutions unless the dataset is large; maximum likelihood may suffer from symptoms of overfitting (*e.g.*, erroneously inferring small proportions of ancestry from many populations to improve model fit) or from poorly estimated allele frequencies. Stated generally, in many high-dimensional inference problems, maximum likelihood solutions can benefit from regularization/penalization steps that are akin to imposing priors. For example, as an improvement to ADMIXTURE, Alexander and Lange (2011) introduced a penalized likelihood function that mimics the way a Dirichlet prior can create sparseness in the admixture coefficients and found it reduced biases substantially.

In the June issue of *Genetics*, Raj *et al.* (2014) present novel, fast algorithms that allow for elaborate Bayesian inference with the PSD model. The key innovation is that they attack the problem in a variational Bayes framework (for an introduction see Jordan *et al.* 1999). Variational Bayes avoids the difficult integration steps, which are typically computed using time-costly Gibbs samplers or Markov chain Monte Carlo techniques, by approximating the posterior in a strategic way. Importantly, an approximate posterior distribution (the "variational distribution") is constructed, which is mathematically simple to work with (*e.g.*, designed such that many terms factorize). It can be shown that by maximizing the variational distribution function with respect to the model parameters, one is maximimizing a lowerbound of the marginal likelihood, and thus finding parameters that fit the data well. The end result is that the challenging integrals of a standard Bayesian approach are replaced by functions that are easily computable, and only need to be optimized. In turn, the vast grab bag of tricks from numerical optimization (Nocedal and Wright 2006) can be used and parameter estimation can proceed quickly. In the text mining literature, variational methods have been used with success on the LDA model (Blei *et al.* 2003). Raj *et al.* (2014) report running times for their new variational algorithm, fastSTRUCTURE, that compete with ADMIXTURE (with small problems being solved roughly 10 times faster than STRUCTURE). Further, the run times are linear in the number of individuals, markers, and populations, so the approach will scale well to larger datasets. This speed comes at the expense of working with an approximation to the posterior, but in practice the resulting parameters are similar to those obtained in the full Bayesian inference.

A well-known, vexing problem for those using the PSD model has been how to appropriately choose the number of subpopulations (K) for the analysis and/or how to infer it directly from the data. As the number of parameters changes with K, this is a type of model selection problem, and several different approaches have been taken to attack it (*e.g.*, Pritchard *et al.* 2000; Evanno *et al.* 2005; Alexander and Lange 2011). Raj *et al.* (2014) find a cross-validation approach that is deployed in the ADMIXTURE software does not work as well for choosing K with fastSTRUCTURE and so then they develop two metrics (K^*_{ξ}, $K^*_{\varnothing^c}$) that can help establish a likely range for K when using the variational approach. While not perfect, these metrics allow a reasonable inference of K when a dataset is large and structure is strong, but in more weakly structured populations, the inference of K will continue to be problematic.

Raj *et al.* (2014) also found that a new logistic prior for allele frequencies (that replaces the standard F model used in earlier versions of STRUCTURE) is beneficial when teasing apart subtle structure in data. The elaboration of such priors could allow more detailed modeling of population history to be layered into the PSD model. One possibility would be to consider elaborate hierarchical priors, such as the tree-based prior on population frequencies developed by Pickrell and Pritchard (2012). Raj *et al.* (2014) found in their preliminary analyses that such priors did not improve model fit in their applications, but further exploration in this arena could prove fruitful.

As Raj *et al.* (2014) note, the PSD model is a coarse model of more complex populations. For this reason, interpreting the

results of inference under this model demands substantial care and critical thought to avoid pitfalls (*e.g.*, Anderson and Dunham 2008). The results from these simple models need to be viewed with an awareness of the complex evolutionary processes potentially shaping genetic variation in any dataset. Ideally, as our field matures, increasingly explicit and robust models of complex population history will be brought to bear on inference from genomic-scale data. It is humbling that even relatively simple models require much careful work and attention to computational detail, but new frameworks for inference, such as the variational Bayes used here, and new tools from numerical optimization, give hope for exciting progress.

Acknowledgments

Helpful comments on a draft were provided by Eric Anderson, Enrique Lessa, and Matthew Stephens.

Literature Cited

Alexander, D. H., and K. Lange, 2011 Enhancements to the admixture algorithm for individual ancestry estimation. BMC Bioinformatics 12: 246.

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Anderson, E. C., and E. A. Thompson, 2002 A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160: 1217–1229.

Anderson, E. C., and K. K. Dunham, 2008 The influence of family groups on inferences made with the program structure. Mol. Ecol. Resour. 8: 1219–1229.

Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003 Latent dirichlet allocation. J. Mach. Learn. Res. 3: 993–1022.

Corander, J., and P. Waldmann, and M. J. Sillanpää, 2003 Bayesian analysis of genetic differentiation between populations. Genetics 163: 367–374.

Dawson, K. J., and K. Belkhir, 2001 A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res. 78: 59–77.

Engelhardt, B. E., and M. Stephens, 2010 Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 6: e1001117.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14: 2611–2620.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Falush, D., M. Stephens, and J. K. Pritchard, 2007 Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol. Ecol. Notes 7: 574–578.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard, 2009 Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. 9: 1322–1332.

Huelsenbeck, J. P., and P. Andolfatto, 2007 Inference of population structure under a dirichlet process model. Genetics 175: 1787–1802.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, 1999 An introduction to variational methods for graphical models. Mach. Learn. 37: 183–233.

McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686.

Nocedal, J., and S. J. Wright, 2006 *Numerical Optimization*, Springer-Verlag, New York.

Novembre, J., and S. Ramachandran, 2011 Perspectives on human population structure at the cusp of the sequencing era. Annu. Rev. Genomics Hum. Genet. 12: 245–274.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Raj, A., M. Stephens, and J. Pritchard, 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197: 573–589.

Tang, H., J. Peng, P. Wang, and N. J. Risch, 2005 Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28: 289–301.

Wilson, G. A., and B. Rannala, 2003 Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163: 1177–1191.

Communicating editor: M. Johnston