Integrating genetic, phenotypic and geographic data in ecologcial and evolutionary studies

The spatial mixture approach

Guillot, Gilles

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Integrating genetic, phenotypic and geographic data in ecologgcial and evolutionary studies. The spatial mixture approach.

Gilles GUILLOT

Mathematical Modelling anf Informatics, Technical University of Denmark, Copenhagen, Denmark

Background

Species delimitation is of interest in conservation biology (delimitation and management of endangered species), epidemiology (detection of new pathogens) and evolutionary biology to describe, quantify and understand mechanisms of speciation. Methodological advances in evolutionary biology have led to methods for species delimitation solely based on the variation of key genetic markers [e.g. DNA barcoding, Luo et al., 2011]. Limits of these single-marker approaches are made evident by conflicts between different genes in a multi-marker approach [Rodríguez et al., 2010, Turmelle et al., 2011] or between genetic and phenotypic markers [Nesi et al., 2011]. In this context of species or population delimitation, phenotypic data still remain of interest together with genetic markers.

Phenotypes such as size and/or shape of morphological structures are the product of numerous interacting nuclear genes [Klingenberg et al., 2001] and, as such, can provide a global estimate of the divergence between units. Furthermore, by being the target of selection, morphological variation can provide precious insights on the selection pattern contributing to shape the units. In the case of fossil lineages, it may even be the only information available to identify evolutionary and systematic units [Néraudeau, 2011, Girard and Renaud, 2011].

Method

Datasets considered

- n individuals sampled at sites $\mathbf{s} = (s_i)_{i=1,...,n}$ (where s_i is the two-dimensional spatial coordinate of individual i),
- phenotypic variables denoted $\mathbf{y} = (y_{ij})_{\substack{i=1,...,n \\ j=1,...,q}}$
- genetic markers denoted $\mathbf{z} = (z_{ij})_{\substack{i=1,...,n \cdot \\ j=1,...,l}}$

Any combination of phenotypic and genetic data, including situations where only phenotypic or only genetic data are available as well as situations when each individual is observed through its own combination of phenotypic and genetic markers. As it will be shown below, our approach also encompasses the case where sampling locations are missing (or considered to be irrelevant). The only constraint that we impose at this stage is that if spatial coordinates are used, they must be available for all individuals. We assume that each individual sampled belongs to one of K different clusters and that variation in the data can be captured by cluster-specific location and scale parameters.

Prior and Likelihood Model for Phenotypic Variables

Denoting by p_i the cluster membership of individual i ($p_i \in \{1, ..., K\}$), we assume that, conditionally on $p_i = k$, y_{ij} is drawn from a parametric distribution with cluster-specific parameters. Independence is assumed within and across clusters conditionally on cluster membership. This means in particular that there is no residual dependence between variables not captured by cluster memberships. Implications of this assumption are discussed later. Although most of the analysis that follows would be valid for all families of continuous distribution, we assume in the following that the y values arise from a normal distribution. Each cluster is therefore characterized by a mean μ_{kj} and a variance σ^2_{kj} and our model is a mixture of multivariate independent normal distributions [Frühwirth-Schnatter, 2006]. Following a common practice in Bayesian analysis [Gelman et al., 2004], we use the natural conjugate prior family on $(\mu_{kj}, 1/\sigma^2_{kj})$ for each cluster k and variable j. Namely, we assume that the precision $1/\sigma^2_{kj}$ (i.e. inverse variance) follows a Gamma distribution $\mathcal{G}(\alpha, \beta)$ (α shape, β rate parameter) and that, conditionally on σ_{kj}, the mean μ_{kj} has a normal distribution with mean ξ and variance σ^2_{kj}/κ. In the specification above, α, β, ξ and κ are hyper-parameters. Details about their choice are discussed in the appendix and in the supplementary material.

Prior and Likelihood Model for Genetic data

We assume here a mixture of multinomial distributions. Denoting frequency of allele a at locus l in cluster k by f_{kla}, for diploid genotype data we assume that

$$\pi(z_{ij} = \{a, b\} | p_i = k) = 2 f_{kla} f_{klb} \quad \text{whenever} \quad a \neq b \tag{1}$$

$$\text{and } \pi(z_{ij} = \{a, a\} | p_i = k) = f^2_{kla}. \tag{2}$$

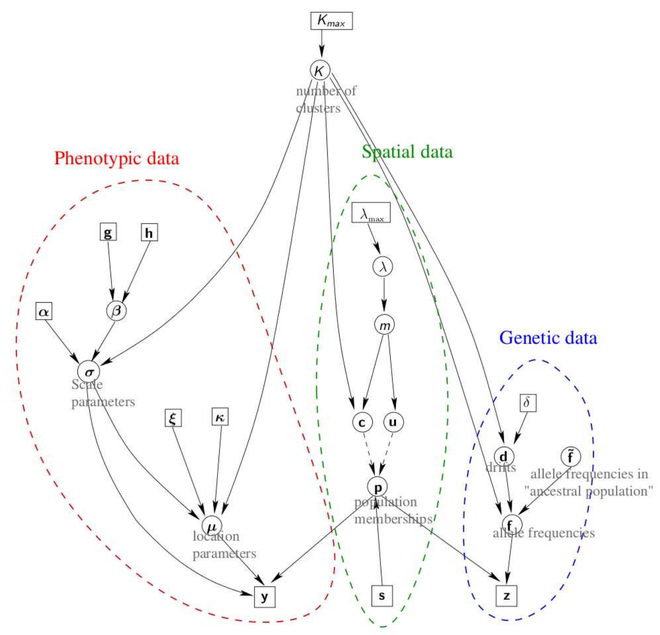
While for haploid data, we have

$$\pi(z_{ij} = a | p_i = k) = f_{kla} \tag{3}$$

We assume independence of the various loci within and across clusters conditionally on cluster memberships. We assume that allele frequencies $f_{kl\cdot}$ have a Dirichlet distribution. Independence of the vectors $f_{kl\cdot}$ is assumed across loci. Regarding the dependence structure across clusters, we consider either independence (referred to as Uncorrelated Frequency Model or UFM) or an alternative model (referred to as Correlated Frequency Model or CFM) introduced by Balding and Nichols [1995, 1997].

Prior Models for Cluster Membership

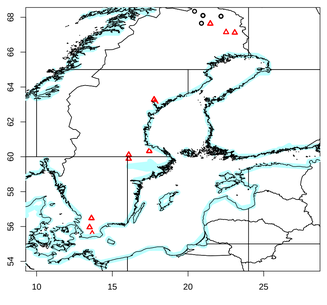
We consider the colored Poisson-Voronoi tessellation Møller and Stoyan [2009]. Loosely speaking, this model assumes that each cluster area in the geographic domain can be approximated by the union of a few polygons. The polygons are assumed to be centered around some points that are generated by a homogeneous Poisson process. Formally, we denote by $(u_1, ..., u_m)$ the realization of this Poisson process. These points in $\mathbb{R}^2$ induce a Voronoi tessellation into m subsets $\Delta_1, ..., \Delta_m$. The Voronoi tile associated with point u_i is defined as $\Delta_i = \{s \in \mathbb{R}^2, dist(s, u_i) < \text{dist}(s, u_j) \forall j \neq i\}$. Each tile receives a cluster membership c_i at random sampled independently from a uniform distribution on $\{1, ..., K\}$. Denoting by D_k the union of tiles with color k, the set $(D_1, ..., D_K)$ defines a tessellation in K subsets. This model is controlled by the intensity of the Poisson process λ (the average number of points per unit area) and the number of clusters K. We place a uniform prior on $[0, \lambda_{max}]$ and on $\{0, ..., K_{max}\}$ respectively.



Graph of proposed model. The parameters of interest to biologists are the number of clusters K, the vector $\mathbf{p}$ which encode the cluster memberships, and possibly allele frequencies $\mathbf{f}$, mean phenotypic values $\boldsymbol{\mu}$, phenotypic variance $\boldsymbol{\sigma^2}$ which quantify the genetic and phenotypic divergence between and within clusters. Other parameters can be viewed mostly as nuisance parameters.

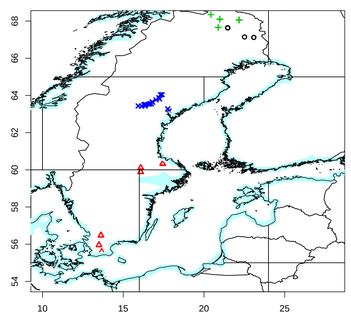
A Scandinavian bank vole dataset

The dataset consists of 182 individuals. These individuals were genotyped at 14 microsatellite loci [Lehanse, 2010].



Clustering with Phenotypic & Spatial

Morphometric clusters revealed only inter-specific differences between red-backed and bank voles.



Clustering with Genetic & Spatial

On the basis of microsatellite data, both inter- and intra-specific levels of differentiation emerged as separate clusters. The structure of genetic differentiation corroborates this interpretation. The inter-specific differentiation of the top North cluster from the rest of Sweden is indeed much stronger than the intra-specific differentiation among the bank vole populations from North-East, Central and South Sweden. Combining both data types allows us to interpret the complex phylogeographic structure of this species and helps to distinguish differences between true species and populations within a species.

References

D.J. Balding and R.A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.

D.J. Balding and R.A. Nichols. Significant genetic correlation among Caucasians at forensic DNA loci. *Heredity*, 78:583–589, 1997.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Series in Statistics. Springer, 2006.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004.

C. Girard and S. Renaud. The species concept in a long-extinct fossil group, the conodonts. *Comptes Rendus Palevol*, 10:107–115, 2011.

C. P. Klingenberg, L. J. Leamy, E. J. Routman, and J. M. Cheverud. Genetic architecture of mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics. *Genetics*, 157:785–802, 2001.

B. Lehanse. Étude génétique d'une zone de contact en suède entre deux lignées de campagnols roussâtres *Myodes glareolus*. Master's thesis, Université de Liège, 2010.

A. Luo, A. Zhang, S. Y.W. Ho, W. Xu, W. Shi, Cameron S.L., and C. Zhu. Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mamals. *BMC Genomics*, 12(84), 2011.

J. Møller and D. Stoyan. *Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings*, chapter Stochastic geometry and random tessellations. Springer, 2009.

Néraudeau. The species concept in palaeontology: Ontogeny, variability, evolution. *Comptes Rendus Palevol*, 10:71–75, 2011.

N. Nesi, E. Nakoumé, C. Cruaud, and A. Hassanin. DNA barcoding of African fruit bats (*Mammalia, Pteropodidae*). the mitochondrial genome does not provide a reliable discrimination between *Epomophorus gambianus* and *Micropteropus pusillus*. *Comptes Rendus Biologies*, 334:544–554, 2011.

F. Rodríguez, T. Pérez, S. E. Hammer, J. Albornoz, and A. Domínguez. Integrating phylogeographic patterns of microsatellite and mtDNA divergence to infer the evolutionary history of chamois (genus *Rupicapra)*. *BMC Evolutionary Biology*, 10:222, 2010.

A. S. Turmelle, T. H. Kunz, and M. D. Sorenson. A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. *Molecular Ecology*, 20:357–375, 2011.