STATISTICAL INFERENCE WITH TREE-INDEXED MARKOV PROCESSES

by

Mohammad Khabbazian

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Electrical and Computer Enineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 11/10/2016

The dissertation is approved by the following members of the Final Oral Committee:
Karl Rohe, Assistant Professor, Statistics and Electirical and Computer Engineering
Cécile Ané, Associate Professor, Statistics and Botany
Rebecca Willett, Associate Professor, Electirical and Computer Engineering
Amir Assadi, Professor, Mathematics
John Gubner, Professor, Electirical and Computer Engineering

To my parents.

Acknowledgments

I would like to thank my PhD advisors, Professor Karl Rohe and Professor Cécile Ané, for all their help, directions, and encouragements throughout the course of my dissertation research. I would also like to thank Professor Amir Assadi for introducing me to the exciting field of machine learning and data science.

I am thankful to the other two members of my committee, Professor Rebecca Willett and Professor John Gubner, for providing insightful comments and suggests on my work. I greatly appreciate the time and effort they spent.

I would also like to thank Professor Shirin Malekpour and Professor Diane Rivard in the Mathematics department for their wonderful support during my teaching in the Wisconsin Collaboratory for Enhanced Learning center. Many thanks to all my friends in Madison who made my time here fun and exciting.

Last but not least, I would like to thank my parents and my wife for their love and support.

Contents

List of Tables

List of Figures

Abstract

My dissertation develops a combination of combinatorial and statistical approaches to extract dependencies in observed data and to model the most significant summary of such dependencies in computational and statistical friendly terms. In non-technical terms, it models dependencies in a complex system by tree topology. Moreover, to define probability distributions on the observed data, the central mathematical framework is stochastic processes on tree topology. The first part of the dissertation addresses the detection of shifts in trait evolution and the second part focuses on a novel variate of chain referral samplings for collecting samples from hidden populations.

The detection of evolutionary shifts in trait evolution from extant taxa is motivated by the study of convergent evolution, or to correlate shifts in traits with habitat changes or with changes in other phenotypes. My dissertation proposes a phylogenetic lasso method to study trait evolution from comparative data and detect past changes in the expected mean trait values. The new method uses the Ornstein-Uhlenbeck process, which can model a changing adaptive landscape of continuous traits on phylogenetic trees. The method is very fast, running in minutes for hundreds of species, and can handle multiple continuous traits. Moreover, it proposes a phylogenetic Bayesian information criterion that accounts for the phylogenetic correlation between species, as well as for the complexity of estimating an unknown number of shifts at unknown locations in the phylogeny. This criterion does not suffer model overfitting and has high precision, so it offers a conservative alternative to other information criteria.

Respondent-driven sampling (RDS) is a type of chain referral sampling popular for sampling hidden populations. As such, even under the ideal sampling assumptions, the performance of RDS is restricted by the underlying social network: if the network is divided into weakly connected communities, then RDS is likely to oversample one of these communities. In order to diminish the referral bottlenecks between communities, we propose anti-cluster RDS (AC-RDS), an adjustment to the standard RDS implementation. Using a Markov process on the referral tree, we construct and study the Markov transition matrix for AC-RDS. We show that if the underlying network is generated from the Stochastic Blockmodel with equal block size, then the transition matrix for AC-RDS has a smaller spectral gap and consequently faster mixing properties than the standard random walk model for RDS. In addition, it is shown that AC-RDS reduces the covariance of the samples in the referral tree compared to the standard RDS and consequently leads to the smaller variance and design effect.

Chapter 1

Introduction

“Et ignotas animum dimittit in artes. (And he turned his mind to unknown arts.)”

Ovid, Epigraph to A Portrait of the Artist as a Young Man

Detection of evolutionary shifts in Ornstein-Uhlenbeck models

Is is possible to detect dramatic changes that occurred in the past from observed data? Is it possible to have a statistical framework to assess and compare plausibility of various hypotheses about the past given present day data?

A group of related species is expected to have similar inherited characteristics (traits) due to their shared ancestors and evolutionary history. These evolutionary relationships can be summarized by a tree $\mathcal{T}$ equipped with a length function, "phylogenetic tree." The leaves of the phylogenetic tree $L(\mathcal{T})$ show the contemporary species, the internal nodes represent the extinct shared ancestors and the length function corresponds to time intervals. Let $y_i(t) : [0, T] \to R$ be the trait of the i^{th} species from time 0 to T and Y be the trait values at the leaves where $Y_i = y_i(T)$. A significant change in the average trait values of a clade, the group of leaves below an edge, from other species is a sign of adaptation of the clade.

The framework of a stochastic process on a tree can describe the evolution of quantitative traits (e.g. morphology) over time. Data from fossil lineages validate that the Ornstein-Uhlenbeck (OU) stochastic process, a generalization of the Brownian motion is an accurate model of the trait evolution over time (Hunt et al., 2008; Hopkins and Lidgard, 2012). The OU process takes into account both the variation and adaptation aspects of the trait evolution. In addition, the phylogenetic tree can be reconstructed from DNA sequences. The reconstruction is possible under certain assumptions such as modeling gene evolution by a continues-time Markov chain. This estimated phylogenetic tree is a similar mathematical object to hierarchical clustering trees; semi-labeled rooted binary trees.

In this framework, the traits at the leaves have a multivariate Gaussian distribution. Adaptation to a new condition manifests through a change in the expected trait values, EY_i, of species i in a clade. Such changes may be due to a shift in ecological niches on the corresponding edge and time.

Chapter 2 of this dissertation (Khabbazian, Kriebel, Rohe, and Ané, 2016b) demonstrates that the problem suffers from identifiability issues. Accepting the limitations and focusing on parsimonious solutions, Section 2.3 states the model selection problem as a linear regression model

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma_{(T,\alpha,\sigma^2)}),$$

where the design matrix X encodes the topology of the tree and adaptation rate; columns and rows correspond to the edges and leaves, respectively. The support set of the coefficient vector β represents the position of the shifts on the tree. The parsimonious restriction appears as an ℓ_0 constrain. I approximate the NP-hard estimation problem as an ℓ_1 regularized convex optimization. In Section 2.4, I show that non-parsimonious models correspond to sets of dependent columns of X. As a result, the solution to the convex optimization satisfies the parsimony condition. Furthermore, I provide the necessary and sufficient conditions under which the ℓ_1 regularized linear regression has a unique solution.

Moreover, this dissertation proposes a novel Bayesian information criterion (pBIC) for detecting a statistically-justifiable number of shifts on the tree in Section 2.5. It also brings to attention a linear time algorithm to compute the inverse of the covariance matrix $\Sigma^{-1}_{(\mathcal{T},\alpha,\sigma^2)}$ in Section 2.6. This work has been implemented as an R package ℓ1ou that is available open source at `https://github.com/khabbazian/l1ou`.

Novel Sampling Design for Respondent-driven Sampling

Referral sampling approaches are popular for sampling populations for which constructing a sampling frame is not possible, but the members are connected through a social network. Denote the network by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ and $\mathcal{E}$ represent the population and connections among members. Let function $y : \mathcal{V} \to R$ assign a characteristic to each member. Referral sampling starts from a seed node that refers a few members. As the procedure continues, participants refer more members. This referral process forms a directed tree $\mathcal{T}$ with $\texttt{root}(\mathcal{T})$ as the seed node. In the tree, nodes represent the participants and edges show the referral relationship. The referral process can be modeled by a “Markov chain indexed by a tree” (Benjamini and Peres, 1994). The model takes into account the multiple referrals and the

Markov property of the process. Under certain assumptions

$$\Pr\left(\nu \in \mathtt{child}(u) | u \in \mathcal{V}(\mathcal{T})\right) = P_{u\nu},$$

where P is the transition matrix of for example random walk on the social network.

Chapter 3 of dissertation (Khabbazian, Hanlon, Russek, and Rohe, 2016a) expresses the relationship between the structure of the underlying network and the sampling process as a projection of the trait function y and the spectral decomposition of the transition matrix. Define the inner product $\langle f, g\rangle_\pi = \sum_{u\in\mathcal{V}} f(u)g(u)\pi(u)$ and let f_j be the j^{th} eigenvector of the transition matrix under the defined inner product. Let $X_i, i = 1, 2, \cdots$ be a Markov chain with reversible transition matrix P. Then

$$\mathrm{Cov}(y(X_i), y(X_{i+t})) = \sum_{j=2}^{|V|} \langle y, f_j\rangle_\pi^2 \lambda_j^t,$$

where π is the stationary distribution.

In the spectral clustering literature, it is well known that the span of the leading non-trivial eigenvectors of Laplacian matrix corresponds to the community structure of a network. Roughly speaking, the decomposition demonstrates that if the trait of interest is correlated with the communities in the network, then the consecutive samples are highly correlated. Chapter 3 proposes a novel sampling procedure, "anti-cluster RDS" that incorporates triangles, 2-dimensional simplices, as a local structure of the social network to reduce covariance of the samples. Anti-cluster RDS is privacy preserving. I show that under certain conditions the AC-RDS reduces the covariance of the samples. More formally, let $(X_i)_{i=1}^n$ and $(X_i^{ac})_{i=1}^n$ be the samples collected by the standard and AC-RDS procedure respectively. If the network is generated from a certain class of stochastic blockmodels, then for all $i, i+t \in \{1, \cdots, n\}$, and $t \neq 0$,

$$\mathrm{Cov}(y(X_i^{ac}), y(X_{i+t}^{ac})) < \mathrm{Cov}(y(X_i), y(X_{i+t})).$$

As a consequence, the proposed sampling procedure collects more representative samples from the population compared to the standard method. The proof uses techniques from matrix concentration to show that the sampled network Laplacian is close the expected Laplacian matrix under the spectral norm. Additionally, it uses invariant subspace perturbation theorems to show that AC-RDS reduces samples covariance.

Chapter 2

Detection of Evolutionary Shifts in Ornstein-Uhlenbeck Models

Contribution

- I conducted all the simulations and result visualizations expect Figure A.6.
- I developed the theoretical basis presented in Section 2.4.
- I designed the 2-step lasso procedure to handle the whitening (decorrelation) of the data dependent on the unknown correlation level.
- I participated in the work/discussion to design the lasso procedure to handle multivariate data.
- I designed the lasso procedure to detect convergent regimes.
- I wrote the first draft of the manuscript.
- I implemented l1ou R package (https://github.com/khabbazian/l1ou).

Abstract

The detection of evolutionary shifts in trait evolution from extant taxa is motivated by the study of convergent evolution, or to correlate shifts in traits with habitat changes or with changes in other phenotypes. We propose here a phylogenetic lasso method to study trait evolution from comparative data and detect past changes in the expected mean trait values. We use the Ornstein-Uhlenbeck process, which can model a changing adaptive landscape over time and over lineages. Our method is very fast, running in minutes for hundreds of species, and can handle multiple traits. We also propose a phylogenetic Bayesian information criterion (pBIC) that accounts for the phylogenetic correlation between species, as well as for the complexity of estimating an unknown number of shifts at unknown locations in the phylogeny. This criterion does not suffer model overfitting and has high precision, so it offers a conservative alternative to other information criteria. Our re-analysis of Anolis lizard data suggests a more conservative scenario of morphological adaptation and convergence than previously proposed. Software is available on GitHub.

Keywords: phylogenetic comparative method, adaptation, convergent evolution, lasso, regularization, phylogenetic BIC, l1ou.

2.1 Introduction

Recent advances in DNA sequencing technology and phylogenetic methods enabled accurate reconstructions of the evolutionary relationships among very large groups of species, and opened new avenues to study phenotypic trait evolution. The inference of evolutionary trees with thousands of taxa or thousands of genes demands complex mathematical models and computational tools (see for instance Bininda-Emonds et al., 2007; Wickett et al., 2014). Likewise, the inference of phenotypic trait evolution on very large trees demands complex models that are capable of handling heterogeneity across a wide range of species. Hansen (1997) used an Ornstein-Uhlenbeck (OU) process to model the macroevolution of a phenotype subject to selection pressure towards an "optimal" value. This OU model was validated on a large number of fossil lineages (Hunt et al., 2008; Hopkins and Lidgard, 2012), as well as in cross-species comparative analyses (Harmon et al., 2010).

Hansen (1997) proposed to use heterogeneous OU models with different optimal phenotype values on different branches of the tree. These models can then be used to test various hypotheses about phenotypic adaptation (Butler and King, 2004). For instance, Scales et al. (2009) evaluated a small set of predefined hypotheses to place the various optima on the tree, to investigate whether fiber-type composition of a leg muscle in lizards is adaptive to the species predator escape strategy, or to its foraging strategy, or both. Mahler et al. (2013) also used OU models with varying optima, but without a preselected set of hypotheses for the number and placement of these optima (see also Ingram and Mahler, 2013; Ingram and Kai, 2014). To do so, they used a stepwise search among OU models to study how natural selection shaped the morphology of Caribbean Anolis lizards (Losos, 2009), and then correlated the phylogenetic placements of shifts in OU optima to habitat changes. Repeated evolution of similar phenotypes in similar environments was taken as evidence for a deterministic aspect of macroevolution.

Several methods were proposed for OU models with multiple optima on phylogenetic trees, to infer the number and the position of shifts in trait optimum without predefined hypotheses. This task is difficult both computationally and theoretically, due to the very large number of models to be evaluated and compared statistically. Uyeda and Harmon (2014) developed a Bayesian method, with a Monte-Carlo Markov chain implementation in the R package bayou. This method quantifies the uncertainty about the number of shifts and their phylogenetic placement. The results can vary quite heavily, however, depending on the prior distribution that the user needs to specify for the number of shifts. Ingram and Mahler (2013) developed a maximum likelihood method and a stepwise search algorithm, "surface", with possibly convergent shifts to the same optimum (see also Mahler and Ingram, 2014). Surface uses the Akaike information criterion (AIC) to select the number of shifts. In this setting however, Ho and Ané (2014) showed that AIC is biased towards model overfit and suggested using instead a modified Bayesian information criterion (mBIC, Zhang and Siegmund, 2007) to reduce the detection of false shifts. In addition to the theoretical difficulties of inferring the correct number of shifts, both bayou and surface can become computationally heavy with large trees, handling a maximum of a few hundred taxa.

We propose here a new method to detect shifts in phenotypic optima under the OU model on trees. The method, ℓ1ou, is based on the lasso (Tibshirani, 1996) and can handle extremely large phylogenetic trees with thousands of taxa. For example, analysis of sporangium shape from 886 moss (Bryophyta) species (Rose et al., 2015) takes only 220 minutes with our method, whereas surface did not complete after 6 weeks. As far as we know, it is the first time that a lasso-type method is proposed for phylogenetically structured data. We present our lasso-based methods in Section 2.2, along with choices to deal with identifiability issues and with a new phylogenetic-aware information criterion (pBIC) to do model selection. This section can be skipped at first and its technical details are presented in Section 2.5. In Section 2.7, we show using simulations that our ℓ1ou method is also more accurate and can take advantage of multiple traits to infer a more robust model. We then illustrate the method and its scalability on data from 100 *Anolis* lizard species and 4 traits (Section 2.8). We implemented the method in R, available at `https://github.com/khabbazian/l1ou`.

Although we focus on OU models with shifts in the optimal phenotype value, we recognize that many other types of heterogeneity might affect real data, especially at deep evolutionary scales. Changes in the rate of evolution was considered by others, mostly for Brownian motion models that exclude adaptation, to test prespecified hypotheses about where rate changes have taken place (Stack et al., 2011; O'Meara et al., 2006), or to detect the phylogenetic position and number of these rate changes (Eastman et al., 2011; Rabosky, 2014). Changes in the strength of selection towards the optimum value have also been proposed by Beaulieu et al. (2012), although simultaneously detecting shifts in several of these parameters was shown to be difficult. We also caution against a literal interpretation of OU model parameters, especially at deep phylogenetic scales. In particular, even if the "optimal value" is estimated to be constant within a given clade, this value may only reflect a broad adaptive zone, around which the true optimal value constantly fluctuates (Uyeda and Harmon, 2014). In this case, it is prudent to interpret α as a parameter for phylogenetic correlation, rather than a direct estimate of the selection strength.

2.2 Lasso-based Method for Shift Detection

The OU Model on a Phylogenetic Tree

We model the evolution of a continuous phenotypic trait $y(t)$ over time t with an Ornstein-Uhlenbeck (OU) process, defined by the following stochastic equation:

$$dy(t) = \alpha\big(\theta(t) - y(t)\big)dt + \sigma dB(t),$$

where $B(t)$ is the Brownian motion (BM). This process considers trait adaptation to the environment through the parameter $\theta(t)$, called the optimum value of the trait, and which may vary over time. The parameter $\alpha \geqslant 0$ is the rate of adaptation. Equivalently, the phylogenetic half-life, $\log(2)/\alpha$, is the amount of time it takes for the trait expected value to reach halfway to the optimum value. If $\alpha \approx 0$, or $\log(2)/\alpha$ is much larger than the time interval of interest (e.g. the tree height), then the expected value of $y(t)$ converges slowly to the optimum relative to the observed time period. In this case, $y(t)$ mostly varies around the ancestral state and the OU process reduces to a Brownian motion.

Throughout we assume a known phylogenetic tree for the species of interest. We also assume that this tree is rooted, binary and ultrametric. The OU process is assumed for the evolution of trait y along each branch of tree, independently for the two daughter branches of each node conditional on the trait value at that node. For simplicity and identifiability of the model parameters we assume that, although unknown, α and σ^2 are fixed across the tree but that the optimum value θ may vary across time and across branches in the tree.

We make further assumptions on changes in $\theta(t)$ because its estimation suffers from identifiability issues. Ho and Ané (2013); Ho and Ané (2014) showed that a relatively small variation in $\theta(t)$ cannot be distinguished with certainty from variation caused by the Brownian motion part of the process, even with an infinite number of present-day species if the tree height is bounded (for trees of growing height such as from the Yule process, see Ané et al., 2015; Adamczak and Miłoś, 2015; Bartoszek and Sagitov, 2015). Ho and Ané (2014) also showed that the exact location and number of changes in the optimum value, also called shifts, cannot be identified when these shifts are on the same branch. (see Figure 2.1, left). Given these restrictions, we assume that $\theta(t) = \theta_b$ is constant along branch b, so that θ is a piecewise constant function from the root to any species (leaf). In other words, we assume at most one shift on each branch, located at the beginning of the branch if present. This parsimonious model can still describe the effect of many shifts on each branch.

Even with this parsimonious assumption, the shift positions on a tree can still be unidentifiable. For example, Figure 2.1 (right) shows different shift placements that all correspond to the same grouping of taxa, and would all receive equal likelihoods. We explain below (and prove in Section 2.4, Theorem 2.5) that our method deals with this unidentifiability, and automatically returns a parsimonious model in terms on number of shifts and shift magnitudes (in absolute values).

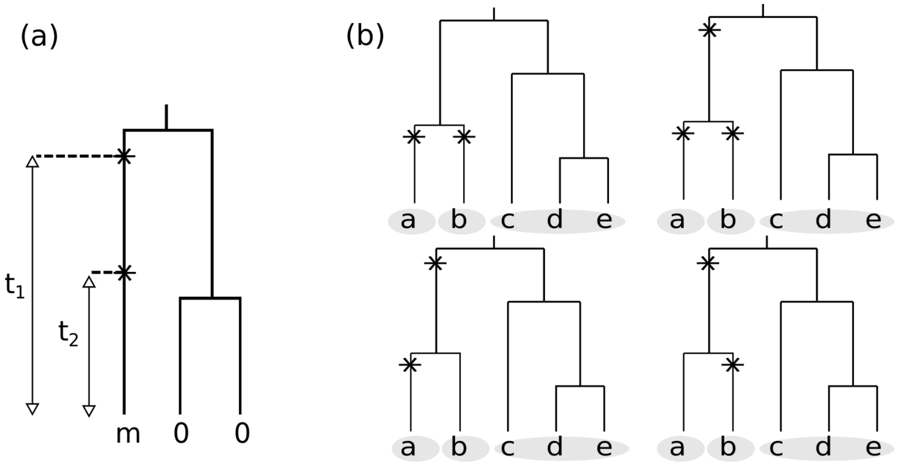


Figure 2.1: The number and position of shifts on a given branch cannot be identified. (a) On a single branch, one shift at age t_1 or one shift at age t_2 or two shifts at ages t_1 and t_2 lead to the exact same model with means $m, 0, 0$ at the leaves, provided that the shift magnitudes $\Delta\theta_i$ (at t_i) satisfy $(1-\exp(-\alpha t_1))\Delta\theta_1+(1-\exp(-\alpha t_2))\Delta\theta_2 = m$. (b) These 4 shift configurations generate the same model, with 3 clusters of tips sharing the same mean: $\{a\}, \{b\}$, and $\{c, d, e\}$. The top right configuration is not parsimonious and cannot be returned by ℓ1ou. The other 3 configurations are all parsimonious and may be returned by ℓ1ou depending on the data.

Method for One Trait (Univariate Case)

Shift Detection as a Linear Model Selection Problem

Under our assumption that there exists at most one shift at the beginning of any given branch, the trait values at leaves follow this linear model (see Section 2.3 for the full derivation):

$$Y = \beta_0 \mathbb{1} + X^{(\alpha)}\beta + \epsilon \tag{2.1}$$

where β_0 is an overall mean ($\mathbb{1}$ is a vector of ones). The β coefficients contain the magnitude of the shifts in selection optimum, i.e. changes in θ values, one for each branch b in the tree: $\beta_b = \theta_b - \theta_{p(b)}$ where $p(b)$ is the parent of b. The non-zero elements in β correspond to the set of branches where θ changes, that is, the shift positions. Following Rabosky et al. (2014), we call this set of branches with shifts a "shift configuration". The design matrix $X^{(\alpha)}$ has n rows (number of taxa) and p columns (number of branches), and depends on α. Define a_b to be the age of b's parent node, that is, the distance from the parent node to its descendant species. For taxon i and branch b,

$$X_{ib}^{(\alpha)} = \begin{cases} 1 - e^{-\alpha a_b} & \text{if } b \text{ is on the path from the root to taxon } i \\ 0 & \text{if taxon } i \text{ is not a descendant of } b \end{cases}$$

(see Section 2.3 for details). Correlations due to shared evolutionary history are captured in the error ϵ that follows a centered normal distribution with covariance $\Sigma^{(\alpha)}$ derived from the OU model:

$$\Sigma_{ij}^{(\alpha)} = \begin{cases} \sigma^2 e^{-\alpha d_{ij}}\,(1 - e^{-2\alpha t_{ij}})/(2\alpha), & \text{if the root value is fixed} \\ \sigma^2 e^{-\alpha d_{ij}}\,/(2\alpha), & \text{if the root value has the stationary distribution} \end{cases} \tag{2.2}$$

where t_{ij} is the evolutionary time shared between species i and j, and d_{ij} is their tree distance.

The linear regression (2.1) cannot be solved with ordinary least squares for several reasons. First, $X^{(\alpha)}$ has more columns (branches with potential shifts) than rows (species with observations). Second, the columns in $X^{(\alpha)}$ are highly correlated, in particular because one shift on a given branch is equivalent to two shifts of equal magnitudes located on each of the two daughter branches. Finally, the predictors in $X^{(\alpha)}$ depend on the unknown adaptation rate, α. However, if we restrict the set of hypothetical shifts and if we reduce $X^{(\alpha)}$ to these branches accordingly, then (2.1) may have a least-squares solution. We show that it is indeed the case if the shift configuration is 'identifiable', that is, if every hypothesized shift is 'visible' from at least one taxon (more formally, see Section 2.4). The main problem is then to select the shift configuration that best fits the data, among all the identifiable shift configurations.

Regularization with Lasso

To tackle the challenges outlined above, which come from the high-dimension nature of the problem, the typical assumption is that only a relatively small subset of predictors (here, shifts) describes the response. In other words we assume that β is sparse, or that most shift magnitudes are 0. A common way to achieve this is to consider the lasso problem (Tibshirani, 1996) whose solution $\hat{\beta}$ minimizes the following ℓ_1-penalized least square criterion:

$$\frac{1}{2}\|Y - \hat{\beta}_0 \mathbb{1} - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1, \tag{2.3}$$

where λ is a tuning parameter and the ℓ_1 norm of the shift magnitudes is simply the sum of their absolute values: $\ell_1(\hat{\beta}) = \sum_b |\hat{\beta}_b|$. This penalty term causes many estimated shifts in $\hat{\beta}$ to be zero, which leads to selecting the most relevant features. By varying the tuning parameter λ from zero to ∞, we increase the weight of the penalty and obtain $\hat{\beta}$'s with support of size n shifts (no penalty) to zero shifts (extreme penalty). Compared to an ℓ_2 penalty in ridge regression, for instance, the ℓ_1 penalty has the advantage of sparsity: where the estimated shifts are $\hat{\beta}_b = 0$ exactly on many branches.

The theory of the lasso is well explored (for instance Bühlmann and Van De Geer, 2011; Eldar and Kutyniok, 2012). To guarantee statistical selection consistency, small prediction error and uniqueness of the estimate, various sufficient conditions were introduced on the sparsity of the coefficient vector and coherency of the design matrix (e.g. Van De Geer et al., 2009). For instance, Zhao and Yu (2006) showed that if (1) X satisfies the 'irrepresentable condition', (2) ϵ contains independent random variables with finite variance, and (3) λ is chosen to have the appropriate scale, then with high probability, the non-zero elements of $\hat{\beta}$ are identical to the non-zero elements of the true β. These results allow for p to grow asymptotically faster than n, so long as the number of non-zeros in β grows slower than n. Furthermore, different methods based on convex optimization, combinatorial, and greedy algorithms were proposed to compute the exact or approximate solution. Efron et al. (2004) showed an intuitive connection between the lasso and stepwise selection solutions. They proposed the fast LARS algorithm to find the lasso estimates $\hat{\beta}$ that minimize (2.3) at every value of λ.

We now rewrite model (2.1) to derive an appropriate ℓ_1 penalty so as to estimate a parsimonious shift configuration, and to account for phylogenetic correlation. If this correlation were ignored, a straight ℓ_1 penalty would bias shift detection in favor of large clades in the tree, for which similarity might otherwise be explained by common ancestry. We first consider the case when α is known, which implies that $X := X^{(\alpha)}$ and the phylogenetic covariance $\Sigma := \Sigma^{(\alpha)}$ are known. To remove phylogenetic correlation we consider $\Sigma^{-1/2}Y$, whose components are uncorrelated, but whose mean is $\Sigma^{-1/2}(\beta_0\mathbb{1} + X\beta)$. Therefore, our lasso estimate is the solution β that minimizes the following ℓ_1-penalized criterion:

$$\frac{1}{2}\|\Sigma^{-1/2}(Y - \beta_0\mathbb{1} - X\beta)\|_2^2 + \lambda\|\beta\|_1. \tag{2.4}$$

Throughout the document, this will be referred to as the *phylogenetic lasso*. We use the R package lars to solve this optimization problem for all values of the tuning parameter λ (see Figure 2.2 for an example) (Efron et al., 2004). An extra model selection phase is then required to find the appropriate λ and the corresponding estimated number of shifts.

Under some mild conditions and for every λ, we prove in Theorem 2.5 that there is a unique solution $\hat{\beta}$ minimizing (2.4), and that the support of $\hat{\beta}$ is an identifiable shift configuration. Furthermore, in Section 2.6 we explain a linear algorithm to calculate $\Sigma^{-1/2}$ efficiently in linear time. This algorithm is based on the method proposed by Stone (2011).

Model Selection for the Number of Shifts

In traditional models with uncorrelated errors, tuning the penalty weight λ is typically done with tools such as cross-validation, minimum expected information loss (AIC), or maximum model posterior probability (e.g. BIC, Schwarz, 1978). In our problem, cross-validation is not appropriate since leaving out some taxa may erode small clades with a shift, taking away part of the signal of interest. In surface

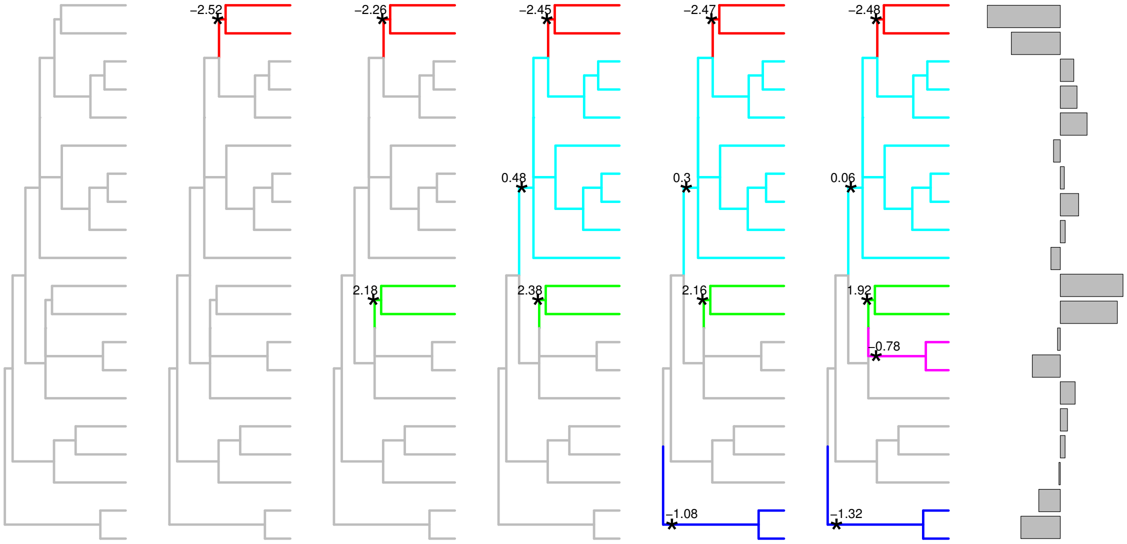


Figure 2.2: Example of our lasso solution path. The number of estimated shifts depends on the penalty parameter, with 0 to 5 estimated shifts as λ decreases progressively from infinity to $\lambda = 3.07$ (one estimated shift $\beta_b \neq 0$), $\lambda = 3.31, 2.64, 1.92$ and 1.73 (5 estimated shifts). The shift configurations are shown from left to right. Each estimated shift is indicated by a star and by its magnitude $\hat{\beta}_b$. Decreasing λ further would further increase the number of estimated shifts (at $\lambda = 1.09, 0.93$ etc.) The sample data are shown with the bar graph.

the following criterion is used:

$$\text{AIC}_c(\mathcal{M}_k) = -2 \log \text{lik}(\mathcal{M}_k) + 2p + \frac{2p(p+1)}{nm - p - 1}$$

where $\mathcal{M}_k$ is the hypothesis that there are k shifts, $\text{lik}(\mathcal{M}_k)$ is the maximum likelihood of the best k-shift configuration, and m is number of traits, all assumed to share the same shift configuration. Here $p = k + m(k+3)$ is the number of parameters, counting the position of each shift as one parameter, and $k+3$ parameters specific to each trait (shift magnitudes, β_0, α and σ). Ho and Ané (2014) showed that minimizing AIC leads to strong model overfitting, however. Therefore, we adapt BIC to better estimate the model posterior probability in the situation when errors are phylogenetically correlated. The traditional BIC score of $\mathcal{M}_k$ can be defined as

$$\mathbf{BIC}(\mathcal{M}_k) = -2\log \mathrm{lik}(\mathcal{M}_k) + (k + m(k+3))\log(n),$$

where again each shift location is counted as a parameter and $k+3$ parameters are specific to each trait.

In Section 2.5, we show that a phylogenetic correction must be applied to better approximate the marginal probability that the true model has k shifts, leading to the following phylogenetic BIC for $m = 1$ trait:

$$\mathbf{pBIC}(\mathcal{M}_k) = -2\log \mathrm{lik}(\mathcal{M}_k) + 2k\log(2n-3) + 2\log(n) + \log\det\left(X_{\mathcal{M}_k}^{(\hat{\alpha})\prime} \nu \Sigma^{(\hat{\alpha})^{-1}} X_{\mathcal{M}_k}^{(\hat{\alpha})}\right) \tag{2.5}$$

where $X_{\mathcal{M}_k}^{(\hat{\alpha})}$ is the matrix $X^{(\alpha)}$ reduced to the columns corresponding to the k estimated branches with a shift but expanded with a column of ones to include the intercept, and ν is the observed trait variance. Informally, $2k\log(2n-3)$ is the penalty term for the shift positions and comes from approximating twice the log of the number of configurations with k shifts, when the tree grows ($n \to \infty$). The penalty for the shift magnitudes and the intercept is captured by the last term, which appears when these parameters are integrated out with a non-informative flat prior. Interestingly, this penalty is not a simple function of the number of parameters. The determinant term depends on α and more importantly, on the location of the shifts through the structure of $X_{\mathcal{M}_k}^{(\alpha)}$. For instance, if α is infinite and if the configuration has 2 shifts that separate the taxa into 3 distinct groups of sizes n_1, n_2 and n_3, then the last penalty term is proportional to $\log(n_1) + \log(n_2) + \log(n_3)$, just like in the modified BIC proposed by Ho and Ané (2014). These numbers of taxa n_i are the effective sample sizes for the intercept and shift values, i.e. the number of observation that effectively provide information on these parameters (when $\alpha = \infty$). This last penalty term generalizes the effective sample size proposed in Ané (2008), to an OU phylogenetic model with any number of shifts.

While pBIC is written here specifically for an OU process, it can easily be applied to any process with k shifts in the mean and any phylogenetic correlation structure, such as a BM process with jumps. To do so, $X^{(\hat{\alpha})}_{\mathcal{M}_k}$ in (2.5) needs to be the design matrix controlling how shift coefficients affect the species means, $\Sigma^{(\hat{\alpha})}$ the estimated phylogenetic covariance, and $2\log(n)$ needs to be replaced by $p\log(n)$ where p is the number of parameters for the phylogenetic covariance structure, including σ^2.

For multiple traits, $2k\log(2n-3)$ appears only once to penalize the shift configuration shared by all traits, but each trait contributes its own $2\log(n)$ and determinant terms to penalize the trait-specific shift magnitudes, β_0, α and σ.

In order to choose λ, we compute the information criterion (BIC or pBIC) for each shift configuration found by the lasso solution path, and then we pick the few top solutions (and their associated λ). While our phylogenetic lasso assumes a fixed α in (2.4), α is then optimized during the likelihood and pBIC (or BIC) evaluation of each shift configuration found by lasso. The columns of the design matrix in (2.4) can be correlated, causing the lasso to pick groups with redundant shifts. To drop these shifts, we add an extra "backward selection" step: any shift whose removal improves the information criterion is dropped. This backward procedure is only performed for the best few models in the solution path to obtain the final estimated model.

Dealing with Unknown Phylogenetic Covariance

Our prior assumption that the adaptation rate α is known is not realistic. So we repeat the procedure twice, once with a conservative starting value for α, and then again with an estimate of α informed by the shift configuration found in the first round (see the outline below with all steps).

We assume in the first round that $\alpha \approx 0$, which leads to the greatest level of phylogenetic correlation, that of a Brownian motion. This is conservative because similarity among all species of a clade might be explained by shared ancestry, rather than a shift at the base of the clade. However, $X^{(\alpha)}$ in (2.1) is degenerate when $\alpha = 0$ (absence of adaptation to the shifts), so we consider its linear approximation when α is small. Its non-zero terms are $1 - e^{-\alpha a_b} \sim \alpha a_b$ and this approximation is most accurate for young branches (young age a_b). Therefore, for our first round with $\alpha \approx 0$ we rewrite (2.1) as follows:

$$Y = \widetilde{X}\widetilde{\beta} + \epsilon, \tag{2.6}$$

where $\widetilde{X}_{ib} = a_b$ if taxon i is a descendant of b, $\widetilde{X}_{ib} = 0$ otherwise, and $\widetilde{\beta} = \alpha\beta$. The phylogenetic covariance for ϵ is assumed to be $\Sigma^{(0)}$ from the BM. The phylogenetic lasso (2.4) is solved in this first round using $\widetilde{X}$ and $\Sigma^{(0)}$. As already noted by Hansen (1997), this multipeaked OU process with $\alpha \approx 0$ corresponds to a BM model with regime-specific trends, with the trend coefficients estimated by $\widetilde{\beta}$ here.

Recall that α is estimated through maximum likelihood during the pBIC (or BIC) evaluation, separately on each candidate configuration, when tuning the lasso penalty λ to do model selection. This is performed with a linear time algorithm in the R package phylolm v2.2 (Ho and Ané, 2014). We then use $\hat{\alpha}$ estimated from the best shift configuration selected in the first round, as input to the phylogenetic lasso (2.4) for a second round. Simulations show that this second round improves the final estimates of the shift positions. We summarize below these various steps of our method, which we call ℓ1ou+IC, where IC is any information criterion (e.g. pBIC).

1. Find the solution path of the phylogenetic lasso (2.4) for $\alpha = 0$ (BM covariance), using the linear approximation for $X^{(\alpha)}$.

2. Calculate $\hat{\alpha}$, $\hat{\beta}$ that maximize the likelihood then calculate IC for each candidate configuration on the path from step 1 (and some simpler configurations, see previous section). Retain the configuration with the best IC.

3. Solve the phylogenetic lasso (2.4) using $\alpha = \hat{\alpha}$ from the configuration found in step 2.

4. Repeat step 2 but on the path of candidate configurations found in step 3.

5. Retain the shift locations, $\hat{\alpha}$ and $\hat{\beta}$ from the configuration with the best IC among those found in steps 2 and 4.

Detecting Convergent Regimes

An adaptation of the phylogenetic lasso can determine if some shifts converge to the same optimal value in multiple parts of the tree, as might be expected if different clades share a similar environment. After shift locations have been estimated by $\ell 1$ou, convergent evolution can be detected by minimizing the following criterion

$$\frac{1}{2}\|\Sigma^{-1/2}(Y - \beta_0\mathbb{1} - X\beta)\|_2^2 + \lambda\|\mathbf{M}\beta\|_1. \tag{2.7}$$

This differs from the phylogenetic lasso (2.4) because it penalizes linear combinations of shift magnitudes, $\mathbf{M}\beta$. $\mathbf{M}$ is built so that each row captures the difference in optimal value between two regimes in the tree. To detect convergence among the first two shifts for example, if the configuration estimated by $\ell 1$ou was as in the left tree of Figure 2.4, $\mathbf{M}$ would include a row with entries $(1 - e^{-\alpha a_{b_1}})$ and $-(1 - e^{-\alpha a_{b_2}})$ in the columns corresponding shifts 1 and 2 respectively (a_b is the age of a shift on branch b), and 0 entries otherwise. In general, $\mathbf{M}$ has at most $k(k-1)/2$ rows if k shifts were detected by $\ell 1$ou, but could have fewer rows because we do not need to test for a convergence that would remove a single shift. Tibshirani and Taylor (2011) provide a fast solution path algorithm to solve the generalized lasso for an arbitrary $\mathbf{M}$, implemented in the R package genlasso. An information criterion can then be used to select the best model (or λ) along the solution path. For pBIC, the design matrix $X^{(\alpha)}_{\mathcal{M}_k}$ is reduced to the convergent model with one column per distinct optimal value. This pBIC formulation is heuristic here (like AIC$_c$ or BIC) as our derivation of (2.5) assumed independent shifts.

Method for Multiple Traits (Multivariate Case)

Using multiple traits should increase the power and increase the method's robustness to detect shifts. An easy way to analyze multiple traits is to reduce the data to just a few dimensions, such as with principle component analysis (PCA), and separately analyze the first few dimensions that explain most of the variance. Revell (2009) demonstrated that PCA is misleading for phylogenetic data and proposed phylogenetic PCA (pPCA) instead, which assumes a BM covariance among taxa. Recently, Uyeda et al. (2015) showed that both standard PCA and pPCA are biased, in that the top principal components (PC) are most influenced by the traits varying early in the tree. This bias suggests that false shifts might be detected near the root of the tree if ℓ1ou (or other shift detection methods) are used on the first few PC axes. Indeed, this was confirmed in our simulations (see Section 2.7 and Figure 2.8).

To extend our ℓ1ou method to multiple traits, we assume that traits shifted at the same time in the past, on the same branches in the tree. In other words, we group the shift magnitudes for all traits on a given branch together, and we seek to estimate a model where either all shifts in a group are 0 (none of the traits shifted on that branch) or most of the shifts in a group are not 0 (many of the traits shifted on that branch). More formally, we assume (like in surface) that the m traits arose from independent OU processes, each with its own α and σ^2 parameters, but with shifts on a shared set of branches. We write the m observed traits in a long vector $\mathbf{Y}$ of size nm by stacking each trait on top of one another, and we collect the trait-specific adaptation and variance rates in vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$. We also write the shift magnitudes as a long vector $\boldsymbol{\beta}$ by stacking the coefficient of each trait (β_{jb} for trait j on branch b) on top of one another, and we similarly stack the intercepts for all traits into a vector $\boldsymbol{\beta}_0$ of size m. The multivariate response model becomes

$$\mathbf{Y} = \mathbb{1}\boldsymbol{\beta}_0 + \mathbf{X}^{(\boldsymbol{\alpha})}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}^{(\boldsymbol{\alpha})}$ is a block diagonal matrix of size $mn \times mp$ with $X^{(\alpha_j)}$ for trait j on the diagonal, and $\mathbb{1}$ is similarly block diagonal with $\mathbb{1}$ as diagonal terms. The errors $\boldsymbol{\epsilon}$ are assumed to be phylogenetically correlated with variance $\Sigma^{(\alpha_j)}$ for trait j, but independent across traits. It means that, conditional on knowing the true shifts, residual variation ($\boldsymbol{\epsilon}$) is uncorrelated between traits. If shifts are unknown however, traits are correlated because they shift on the same branches. So in fact, we assume that all the between-trait correlation (as could be estimated with straight Pearson correlation coefficients) is due to correlation between shifts.

Yuan and Lin (2006) proposed the group lasso to generalize the lasso when there are predefined groups of coefficients. Here, each branch b in the tree corresponds to a group of coefficients: $(\beta_{jb})_{j \leqslant m}$ across traits. To capture the trend that all coefficients in a group are 0 (or not) together, the group lasso uses the ℓ_1 penalty over groups, rather than over individual coefficients:

$$\sum_{\text{branch } b} \|\beta_{\cdot,b}\|_2 = \sum_{\text{branch } b} \Big(\sum_{\text{trait } j} \beta_{jb}^2 \Big)^{1/2}.$$

The $\|\beta_{\cdot,b}\|_2$ acts as an ℓ_1 penalty on the group of shifts on branch b. This group contains all of the shifts on branch b, for every one of the traits. Because it acts as an ℓ_1 penalty on the group, this penalty selects groups (here branches) to be either entirely zero or entirely non-zero. In the special case when there is only one trait, this penalty reduces to the earlier ℓ_1 penalty: $\sum_b |\beta_b|$. Using this group penalty, we consider the following multivariate phylogenetic lasso

$$\min_{\beta} \frac{1}{2} \|\mathbf{\Sigma}^{-1/2}(\mathbf{Y} - \mathbb{1}\boldsymbol{\beta}_0 - \mathbf{X}^{(\alpha)}\beta)\|_2^2 + \lambda \sum_{b} \|\beta_{\cdot,b}\|_2, \qquad (2.8)$$

where $\mathbf{\Sigma} := \mathbf{\Sigma}^{(\alpha)}$ is block diagonal with $\Sigma^{(\alpha_j)}$ on its diagonal. We used the R package grplasso for solving this group lasso step. Unlike LARS, the search for the λ values where the shift configuration changes is done using a grid search, which can be slower. We then select λ and the associated shift configuration using the same ℓ1ou procedure as before, simply replacing (2.4) by (2.8) in the lasso steps 1 and 3.

Bootstrap Support for Shifts

To quantify uncertainty in the detected shifts, we use an adapted bootstrap procedure, borrowing ideas from Freckleton and Harvey (2006) (see also Pennell et al., 2015).

1. Use ℓ1ou to estimate $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$. For each trait j, compute $\Sigma_j^{-1/2}$ and $\Sigma_j^{1/2}$ in linear time, where $\Sigma_j = \Sigma_j^{(\hat{\alpha}_j)}$ is the phylogenetic correlation for trait j. Then compute the vector of residuals for trait j: $R_j = \Sigma_j^{-1/2}(Y_j - X^{(\hat{\alpha}_j)}\hat{\beta}_{j\cdot})$.

2. Repeat a very large number of times (B times) the following. For each trait j, sample from R_j with replacement to create a bootstrap sample of n residuals $\widetilde{R}_j$. Use ℓ1ou to estimate the shift configuration and shift magnitudes from the bootstrap data with $\widetilde{Y}_j = X^{(\hat{\alpha}_j)}\hat{\beta}_{j\cdot} + \Sigma_j^{1/2}\widetilde{R}_j$,

3. For each branch, calculate the bootstrap support for a shift on that branch as the proportion of bootstrap iterations when a shift was detected on that branch for one of more traits.

This procedure is expected to be conservative, because shifts that are undetected in step 1 cannot receive high bootstrap support. An undetected shift would just contribute one large residual, which would be re-sampled and 'scattered' throughout the tree in the bootstrap resampling step 2. Note that the bootstrap results from step 2 could be summarized more thoroughly in step 3. For instance, on each branch with a estimated shift, a bootstrap confidence could be obtained for the magnitude of this shift.

2.3 The Regression Model

We prove here the linear model formulation (1) for an OU model with shifts in the optimal value, given our assumptions that the tree is ultrametric (every path from the root to any leaf has the same length T), and that any shift occurs at the beginning of a branch. With this assumption, any shift configuration corresponds exactly to a subset of branches in the phylogenetic tree. Recall that θ_b denotes the optimum value on branch b, and a_b denotes the age of the branch, i.e. the age of its parent node where a shift might occur. We will also denote $p(b)$ the parent edge of b, so that a shift on branch b corresponds to a case when $\beta_b = \Delta\theta_b = \theta_{p(b)} - \theta_b \neq 0$. We further denote $a_{c(b)}$ the age of the child node of b, so that $a_b - a_{c(b)} = t_b$ is the length of branch b.

Let y_0 be the trait at the root and Y_i be the trait of taxon i. It follows from the OU model that given y_0, the values Y_i at the leaves are normally distributed, with variance given by (2) and with mean affected by θ values as follows (Hansen, 1997;

Butler and King, 2004; Beaulieu et al., 2012):

$$\mathrm{E}(Y_i) = y_0 e^{-\alpha T} + \sum_{b \in \mathrm{path(root},i)} \left(e^{-\alpha a_{c(b)}} - e^{-\alpha a_b}\right) \theta_b \tag{2.9}$$

where, along the path from the root to taxon i, $a_b = T$ necessarily for the first branch b connected to the root and $a_{c(b)} = 0$ for the last branch on that path, the external branch to taxon i.

To show equation how (2.9) leads to (1), we rewrite θ_b in terms of the ancestral optimum value θ_0 and of shifts that occurred on b and on earlier branches b':

$$\theta_b = \theta_{p(b)} + \Delta\theta_b = \theta_0 + \sum_{b' \preccurlyeq b} \Delta\theta_{b'}$$

where $b' \preccurlyeq b$ means that b' is on the path from the root to b (or $b' = b$), or in other words, that b is a descendant of or equal to b'. We can now rewrite (2.9):

$$\begin{aligned}
\mathrm{E}(Y_i) &= y_0 e^{-\alpha T} + \sum_{b \in \mathrm{path(root},i)} (e^{-\alpha a_{c(b)}} - e^{-\alpha a_b})\theta_0 \\
&\quad + \sum_{b \in \mathrm{path(root},i)} \sum_{b' \preccurlyeq b} (e^{-\alpha a_{c(b)}} - e^{-\alpha a_b})\Delta\theta_{b'} \\
&= y_0 e^{-\alpha T} + (1 - e^{-\alpha T})\theta_0 + \sum_{b' \in \mathrm{path(root},i)} \sum_{b; b' \preccurlyeq b \preccurlyeq i} (e^{-\alpha a_{c(b)}} - e^{-\alpha a_b})\Delta\theta_{b'} \\
&= y_0 e^{-\alpha T} + (1 - e^{-\alpha T})\theta_0 + \sum_{b' \in \mathrm{path(root},i)} (1 - e^{-\alpha a_{b'}})\Delta\theta_{b'} \qquad (2.10)
\end{aligned}$$

Finally, (2.10) becomes (1) if we define $\beta_b = \Delta\theta_b$ and $\beta_0 = y_0 e^{-\alpha T} + (1 - e^{-\alpha T})\theta_0$, a weighted mean of the ancestral state and ancestral optimal value. This coefficient β_0 is shared by all species because we assumed an ultrametric tree, with a shared time T from the root to all tips.

If the tree is not ultrametric and contains extinct species, then the intercept $\beta_0\mathbf{1}$ in (1) needs to be replaced by a vector with entry $\beta_{0,i} = y_0 e^{-\alpha t_{ii}} + (1 - e^{-\alpha t_{ii}})\theta_0$

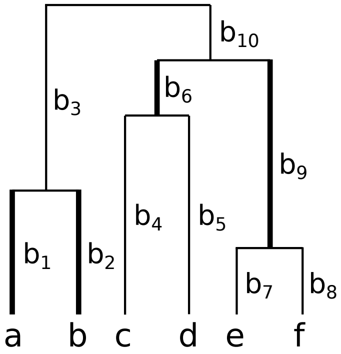


Figure 2.3: The set of bold branches is a minimal cut-set of this rooted tree.

for species i. In this case, the ancestral state y_0 and the ancestral optimum θ_0 are identifiable from each other.

2.4 Identifiability and Uniqueness of Lasso Estimation

We prove here that for every λ and under some mild conditions, the phylogenetic lasso criterion (4) is minimized at a unique $\hat{\beta}$, whose support is an identifiable shift configuration. We first formalize the definition of an identifiable configuration, loosely characterized by requiring that each shift is visible from the leaves.

Definition 2.1 (Cut-set and minimal cut-set). *A subset $\mathcal{B}$ of branches in the phylogenetic tree $\mathcal{T}$ is called a "cut-set" of $\mathcal{T}$ if all the paths from the root to the leaves have at least one branch in $\mathcal{B}$. In other words, removing the branches in $\mathcal{B}$ from $\mathcal{T}$ "cuts" all the leaves from the root, which is then not visible from the leaves. $\mathcal{B}$ is a "minimal cut-set" if, for every branch b in $\mathcal{B}$, the set difference $\mathcal{B}\backslash\{b\}$ is not a cut-set. If so, then the path from the root to any leaf is cut by exactly one branch in $\mathcal{B}$. Figure 2.3 gives an example of a cut-set, that is minimal.*

Definition 2.2 (Identifiable shift configuration). *A set of shift branches $\mathcal{B}$ in tree $\mathcal{T}$ is called an "identifiable shift configuration" if for every branch b with a shift, the other* *branches in $\mathcal{B}$ do not cut the subtree rooted at* b*. More formally,* $\mathcal{B}$ *is identifiable if for every branch* $b \in \mathcal{B}$*, the subset of branches* $\mathcal{B}' = \{b' \in \mathcal{B}; b \prec b'\}$ *is not a cut-set of the subtree rooted at the child node of* b*.*

Remark 2.3. *Definition 2.2 requires that the number of shifts, i.e. the number of branches in* $\mathcal{B}$*, be strictly less than the number of leaves.*

The link between linear model (1) and our graph definition above is the following.

Lemma 2.4. *For a vector of shift magnitudes* β*, let* $\mathcal{B}$ *be its support, that is, the set of branches* b *such that* $\beta_b \neq 0$*. Then* $\mathcal{B}$ *is an identifiable shift configuration if and only if the columns of* $X^{(\alpha)}$ *corresponding the* $\mathcal{B}$ *are linearly independent.*

To prove this lemma and the main theorem below, it is useful to decompose

$$X^{(\alpha)} = ZD^{(\alpha)}$$

where Z contains the topology information and $D^{(\alpha)}$ contains all the dependence of $X^{(\alpha)}$ on α. Z is defined as the matrix of the same size as $X^{(\alpha)}$ with one row per taxon i and one column per branch b, with $Z_{ib} = 1$ if i is a descendant of b and $Z_{ib} = 0$ otherwise. To consider the intercept $Z_{\cdot 0} = 1$. $D^{(\alpha)}$ is defined as the diagonal matrix with one row and column per branch, with diagonal entry

$$D_b^{(\alpha)} = 1 - e^{-\alpha a_b}$$

for branch b and $D_0^{(\alpha)}$ represents the intercept. For example, the tree in Figure 2.3

corresponds to

$$
Z = \begin{array}{c}
\\ \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \\ \mathrm{d} \\ \mathrm{e} \\ \mathrm{f}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc} & b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7 & b_8 & b_9 & b_{10} \end{array} \\
\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1
\end{pmatrix}
\end{array}.
$$

Proof of Lemma 2.4. We first note that the column of $X^{(\alpha)}$ corresponding to a branch b is just a rescaled version of the $0/1$ column of Z corresponding to the same branch b, rescaled by $D_b^{(\alpha)} \neq 0$. So the columns of $X^{(\alpha)}$ corresponding to $\mathcal{B}$ are linearly independent if and only if the columns of Z corresponding to $\mathcal{B}$ are linearly independent.

First suppose that $\mathcal{B}$ is not identifiable. Then there exists a branch b_0 in $\mathcal{B}$ and descendant branches $\mathcal{B}'$ in $\mathcal{B}$ such that $\mathcal{B}'$ is a cut-set of the subtree rooted at the child of b_0. Without loss of generality, we can assume that $\mathcal{B}'$ is a minimal such cut-set. In other words, every path from b_0 to a descendant taxon i goes through exactly one branch in $\mathcal{B}'$. Because of the definition of Z, we have that

$$Z_{\cdot b_0} = \sum_{b \in \mathcal{B}'} Z_{\cdot b},$$

which implies that the columns $Z_{\cdot b}$ for $b \in \mathcal{B}' \cup \{b_0\} \subset \mathcal{B}$ are linearly dependent.

Now, suppose that the columns of Z corresponding to branches in $\mathcal{B}$ are linearly dependent. Let $\mathcal{B}_0$ be the set of minimal branches in $\mathcal{B}$, that is, branches b_0 that have no ancestor $b \in \mathcal{B}$. Let b_0 be one such edge in $\mathcal{B}_0$ and let $\mathcal{B}'$ be the set of descendant of b_0 in $\mathcal{B}$: $\mathcal{B}' = \{b \in \mathcal{B}; b_0 \prec b\}$. Then the columns of Z corresponding to b_0 and $\mathcal{B}'$ must be linearly dependent, because any other edge in $\mathcal{B}$ has a disjoint set of descendant taxa than b_0 or any other edge in $\mathcal{B}'$ does.

Hence, there exist non-zero scalars $(\delta_b)_{b \in \{b_0\} \cup \mathcal{B}'}$ such that $\delta_{b_0} Z_{ib_0} + \sum_{b \in \mathcal{B}'} \delta_b Z_{ib} =$

0 for every taxon i. For a descendant i of b_0 we have $Z_{ib_0} = 1$, so there must be an edge $b \in \mathcal{B}'$ such that $Z_{ib} = 1$, implying that b is an ancestor of i. Therefore $\mathcal{B}'$ is a cut-set of the subtree rooted at b_0, so $\{b_0\} \cup \mathcal{B}'$ (and $\mathcal{B}$) is an unidentifiable shift placement.

□

We can now state our result about the uniqueness and identifiability of the lasso solution, under some condition that depends on α.

Theorem 2.5. *Let $\mathcal{T}$ be a rooted, ultrametric tree. For a branch b, denote $\mathcal{T}_b$ the induced subtree rooted at the child node of b. Suppose that for every branch b_0, every minimal cut-set $\mathcal{B}$ of $\mathcal{T}_{b_0}$, and every arbitrary signs $s_b \in \{-1, +1\}$ for $b \in \mathcal{B}$, we have*

$$\sum_{b \in \mathcal{B}} s_b \frac{D_{b_0}^{(\alpha)}}{D_b^{(\alpha)}} = \sum_{b \in \mathcal{B}} s_b \frac{1 - e^{-\alpha a_{b_0}}}{1 - e^{-\alpha a_b}} \neq 1. \tag{2.11}$$

Then the lasso problem (3) and the phylogenetic lasso (4) each have a unique minimum, which is an identifiable shift configuration.

Remark 2.6. *Condition* (2.11) *depends on both α and the structure of the phylogeny. As α tends to infinity,* (2.11) *reduces to $\sum_{b \in \mathcal{B}} s_b \neq 1$, which breaks down for any odd set of branches $\mathcal{B}$ greater than 2. However, if we assume a bounded adaptation rate α, condition* (2.11) *holds generically for all trees, that is, with probability 1 if we consider that branch lengths are generated from some continuous distribution.*

Remark 2.7. *Condition* (2.11) *does not hold at the root of the tree, if we consider an extra root edge b_0 of length 0 (to represent the intercept) and $\mathcal{B}$ is composed of the two branches b_{out} and b_{in} connected to the root, because $a_{b_0} = a_{b_{\text{out}}} = a_{b_{\text{in}}}$. Intuitively, a shift at the base of the outgroup clade cannot be distinguised from a shift at the base of the ingroup clade, without a further outgroup. To restore uniqueness of the solution, we consider that b_{out} and b_{in} are in fact a single branch in the unrooted tree, and we remove one of them arbitrarily, say b_{out}, from the design matrix. However, any shift detected on the other branch b_{in} needs to be interpreted cautiously, because this shift could actually be located on b_{out} and* *the data has no information to bear on its exact placement. If the intercept were penalized like the shifts are, this problem would not occur, but the solution could vary with the data measurement scale.*

Proof of Theorem 2.5. In what follows, we will say that β is "ℓ_1-optimal" if it minimizes (3) for the lasso problem or (4) for the phylogenetic lasso problem, for some intercept value β_0. We will then say that its shift configuration is an "ℓ_1-optimal shift configuration".

We first show that an unidentifiable shift configuration cannot be ℓ_1-optimal, under condition (2.11). Consider an unidentifiable shift configuration $\mathcal{S}$. So there exists $b_0 \in \mathcal{S}$ and $\mathcal{B} \subset \mathcal{S}$ such that $\mathcal{B}$ is a cut-set of $\mathcal{T}_{b_0}$. Without loss of generality we can assume that $\mathcal{B}$ is a minimal cut-set of $\mathcal{T}_{b_0}$, so that $Z_{\cdot b_0} = \sum_{b \in \mathcal{B}} Z_{\cdot b}$. Let $\mathcal{B}^+ := \mathcal{B} \cup \{b_0\}$. Denote the corresponding shift values by $\beta_{\mathcal{B}^+}$. Necessarily, $\beta_b \neq 0$ for all $b \in \mathcal{B}^+$.

We now construct β' with the same shift configuration as β and with the same fit to the data but with a smaller lasso penalty, showing that β cannot be ℓ_1-optimal. For any branch b not in $\mathcal{B}^+$, we define $\beta'_b = \beta_b$. For branch b_0, we let $\beta'_{b_0} = \beta_{b_0} + \epsilon$ where ϵ is a small value to be determined later. For b in $\mathcal{B}$ we define $\beta'_b = \beta_b - \epsilon\, D^{(\alpha)}_{b_0}/D^{(\alpha)}_b$. This choice ensures that $ZD^{(\alpha)}\beta' = ZD^{(\alpha)}\beta$, and so β' provides the same fit to the data whether we consider the original lasso (3) or the phylogenetic lasso (4), because $X^{(\alpha)}\beta' = X^{(\alpha)}\beta$. We now need to choose ϵ in such a way as to reduce the lasso penalty. Because $\beta'_b = \beta_b$ for b not in $\mathcal{B}^+$, we only need to reduce the ℓ_1 penalty associated with branches in $\mathcal{B}^+$:

$$\|\beta_{\mathcal{B}^+}\|_1 = \sum_{b \in \mathcal{B}^+} |\beta_b| = \sum_{b \in \mathcal{B}^+} \operatorname{sign}(\beta_b)\, \beta_b.$$

Let us first assume that $|\epsilon|$ is small enough so that $\operatorname{sign}(\beta'_b) = \operatorname{sign}(\beta_b)$ for all

$b \in \mathcal{B}^+$. Then

$$\begin{aligned}\|\beta'_{\mathcal{B}^+}\|_1 &= \sum_{b \in \mathcal{B}^+} \operatorname{sign}(\beta'_b)\,\beta'_b = \sum_{b \in \mathcal{B}^+} \operatorname{sign}(\beta_b)\,(\beta_b + (\beta'_b - \beta_b)) \\ &= \|\beta_{\mathcal{B}^+}\|_1 + \left(\operatorname{sign}(\beta_{b_0}) - \sum_{b \in \mathcal{B}} \operatorname{sign}(\beta_b)\frac{D^{(\alpha)}_{b_0}}{D^{(\alpha)}_{b}}\right)\epsilon.\end{aligned}$$

Condition (2.11) implies that $\operatorname{sign}(\beta_{b_0}) - \sum_{b \in \mathcal{B}} \operatorname{sign}(\beta_b) D^{(\alpha)}_{b_0}/D^{(\alpha)}_{b}$ cannot be 0, hence we can choose ϵ of the appropriate sign to obtain $\|\beta'_{\mathcal{B}^+}\|_1 < \|\beta_{\mathcal{B}^+}\|_1$, and conclude that β is not ℓ_1-optimal.

We now turn to show that the solution $\hat{\beta}$ is unique. Suppose there are two solutions $\beta^{(0)} \neq \beta^{(1)}$ with (phylogenetic) lasso criterion c^*. Suppose first that $X^{(\alpha)}\beta^{(0)} \neq X^{(\alpha)}\beta^{(1)}$. For $0 < \delta < 1$, consider $\beta^{(\delta)} = (1-\delta)\beta^{(0)} + \delta\beta^{(1)}$. Necessarily, the lasso criterion (3) (or phylogenetic lasso (4)) evaluated at $\beta^{(\delta)}$ is less than the weighted mean of the criterion evaluated at $\beta^{(0)}$ and $\beta^{(1)}$, $(1-\delta)c^* + \delta c^* = c^*$, because $\|a + \delta b\|_2^2$ is a strictly convex function of δ. We get a contradiction then, because $\beta^{(\delta)}$ would minimize the lasso criterion even further than either $\beta^{(0)}$ or $\beta^{(1)}$. Therefore $X^{(\alpha)}\beta^{(0)} = X^{(\alpha)}\beta^{(1)}$. By the same argument we get that $\|\beta^{(0)}\|_1 = \|\beta^{(1)}\|_1 = \|\beta^{(\delta)}\|_1$ for every $0 < \delta < 1$. In particular, $\beta^{(\delta)}$ is another solution to the (phylogenetic) lasso. We now choose any $\delta \in (0, 1)$ such that $\beta^{(\delta)}_b \neq 0$ for every branch b that satisfies $\beta^{(0)}_b \neq \beta^{(1)}_b$. Since $X^{(\alpha)}(\beta^{(0)} - \beta^{(1)}) = 0$ and because of Lemma 2.4, the shift configuration corresponding to $\beta^{(0)} - \beta^{(1)}$ is unidentifiable. Therefore $\beta^{(\delta)}$ also corresponds to an unidentifiable shift configuration, which contradicts the ℓ_1-optimality of $\beta^{(\delta)}$ and consequently contradicts $\beta^{(0)} \neq \beta^{(1)}$. □

2.5 Derivation of a Phylogenetic BIC

In this section, we seek to approximate the posterior probability of model $\mathcal{M}_k$ that the true configuration has k shifts and prove the rationale for (5). For this, we approximate the Bayes factor

$$\frac{\mathcal{P}(\mathcal{M}_k|Y)}{\mathcal{P}(\mathcal{M}_{-1}|Y)}$$

where $\mathcal{M}_{-1}$ is a model with no shifts and no intercept ($\beta = 0$ and $\beta_0 = 0$). Equivalently, under a uniform prior on k, we seek to approximate

$$\frac{\mathcal{P}\left(Y|\mathcal{M}_k\right)}{\mathcal{P}\left(Y|\mathcal{M}_{-1}\right)}.$$

Let $B^{(k)}$ denote the set of all configurations of k shifts, i.e. the set of all subsets $\mathcal{B}$ of k branches. For sake of simplicity we assume here that α and σ^2 are known. For the time being, we fix a hypothesized configuration $\mathcal{B} \in B^{(k)}$. To simplify notations, we let $\widetilde{X}_{\mathcal{B}}$ be the design matrix $X^{(\alpha)}$ reduced to the columns corresponding to $\mathcal{B}$ and expanded with a first column of ones to include the intercept.

We also denote by $\widetilde{\beta}$ the vector made by the intercept and the non-zero shifts $(\beta_b)_{b\in\mathcal{B}}$. Then we have

$$Y = \widetilde{X}_{\mathcal{B}}\widetilde{\beta} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \Sigma^{(\alpha)})$ with $\Sigma^{(\alpha)}$ given in (2). Under model $\mathcal{M}_{-1}$, the log-likelihood of data Y is

$$\log l_{\mathcal{M}_{-1}} = -\frac{1}{2}\log\det(\Sigma) - \frac{1}{2}Y^T\Sigma^{-1}Y.$$

Note that the proof below holds for any shift model where $\widetilde{X}_{\mathcal{B}}$ is the appropriate design matrix relating the model coefficients $\widetilde{\beta}$ (one per branch $b \in \mathcal{B}$) to the species means, and where Σ is the appropriate phylogenetic covariance matrix.

Under model $\mathcal{M}_k$ and configuration $\mathcal{B}$, the log-likelihood becomes

$$\begin{aligned}\log l_{\mathcal{B}} &= \log l_{\mathcal{M}_{-1}} - \frac{1}{2}\left(Y - \widetilde{X}_{\mathcal{B}}\widetilde{\beta}\right)^T \Sigma^{-1}\left(Y - \widetilde{X}_{\mathcal{B}}\widetilde{\beta}\right)\\ &= \log l_{\mathcal{M}_{-1}} - \frac{1}{2}\left\{Y^T\Sigma^{-1}Y - 2\widetilde{\beta}^T\widetilde{X}_{\mathcal{B}}^T\Sigma^{-1}Y + \widetilde{\beta}^T\widetilde{X}_{\mathcal{B}}^T\Sigma^{-1}\widetilde{X}_{\mathcal{B}}\widetilde{\beta}\right\}.\end{aligned}$$

We now choose a flat uniform (improper) prior on β_0 and on each β_b, $b \in \mathcal{B}$. We also choose a uniform prior probability on each $\mathcal{B}$ in $B^{(k)}$. With these choices, the marginal likelihood of model $\mathcal{M}_k$ is

$$l_{\mathcal{M}_k} = \frac{1}{|B^{(k)}|} \sum_{\mathcal{B} \in B^{(k)}} \int_{\widetilde{\beta} \in \mathbb{R}^{k+1}} \exp\left(\log l_{\mathcal{M}_{-1}} - \frac{1}{2} \left\{ Y^T \Sigma^{-1} Y - 2\widetilde{\beta}^T \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} Y + \widetilde{\beta}^T \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} \widetilde{\beta} \right\} \right) d\widetilde{\beta} \,.$$

Based on Lemma 2.8 below, we can solve the integral exactly and we get

$$l_{\mathcal{M}_k} = \frac{1}{|B^{(k)}|} \sum_{\mathcal{B} \in B^{(k)}} \frac{1}{\sqrt{|\widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}}|}} \exp\left(\log l_{\mathcal{M}_{-1}} - \frac{1}{2} Y^T \Sigma^{-1} Y + \frac{1}{2} Y^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} \left(\widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} \right)^{-1} \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} Y \right).$$

For a given $\mathcal{B}$, the maximum likelihood estimate of $\widetilde{\beta}$ is $\hat{\beta}_{\mathcal{B}} = \left(\widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} \right)^{-1} \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} Y$. Hence we can simplify $Y^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} (\widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}})^{-1} \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} Y = \hat{\beta}_{\mathcal{B}}^T \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} \hat{\beta}_{\mathcal{B}} = \hat{\beta}_{\mathcal{B}}^T \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} Y$. Therefore $-Y^T \Sigma^{-1} Y + \hat{\beta}_{\mathcal{B}}^T \widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}} \hat{\beta}_{\mathcal{B}} = -(Y - \widetilde{X}_{\mathcal{B}} \hat{\beta}_{\mathcal{B}})^T \Sigma^{-1} (Y - \widetilde{X}_{\mathcal{B}} \hat{\beta}_{\mathcal{B}})$ and the maximum likelihood value appears in the marginal likelihood:

$$l_{\mathcal{M}_k} = \frac{1}{|B^{(k)}|} \sum_{\mathcal{B} \in B^{(k)}} \frac{\hat{l}_{\mathcal{B}}}{\sqrt{|\widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}}|}}.$$

Let $\hat{\mathcal{B}}$ be the shift configuration in $B^{(k)}$ that maximizes the following penalized log likelihood

$$\mathrm{pll}_{\mathcal{B}} = \log \hat{l}_{\mathcal{B}} - \frac{1}{2} \log |\widetilde{X}_{\mathcal{B}}^T \Sigma^{-1} \widetilde{X}_{\mathcal{B}}|$$

To simplify notations, define $\hat{X} = \widetilde{X}_{\hat{\mathcal{B}}}$, and $\hat{\beta} = \hat{\beta}_{\hat{\mathcal{B}}}$. Therefore,

$$l_{\mathcal{M}_k} = \frac{1}{|B^{(k)}|} \exp\left(\mathrm{pll}_{\hat{\mathcal{B}}} \right) \sum_{\mathcal{B} \in B^{(k)}} \exp\left(\mathrm{pll}_{\mathcal{B}} - \mathrm{pll}_{\hat{\mathcal{B}}} \right).$$

The first term $|B^{(k)}| = \binom{2n-3}{k}$ because there are $2n-3$ branches in an unrooted binary tree. It can be approximated by $\log |B^{(k)}| = \log((2n-3) \cdots (2n-2-k)) - \log(k!) \sim$

$k \log(2n-3)$ as n goes to infinity and for any fixed k. The last term is a sum, where one value is exactly 1 and all others are smaller. We assume that this sum remains relatively small compared to the other terms. Zhang and Siegmund (2007) showed this to be true in a similar situation for time series. We then get that

$$-2 \log l_{\mathcal{M}_k} \approx -2 \log l_{\hat{\mathcal{B}}} + 2k \log(2n-3) + \log |\hat{X}^T \Sigma^{-1} \hat{X}|.$$

To cover the general case when α and σ^2 are unknown, we add a penalty $2\log(n)$ for these 2 parameters, which then gives (5). To make pBIC scale invariant (when σ^2 is unknown), we standardize Y to have variance 1 prior to applying (5). Equivalently up to a constant, the estimated Σ matrix is divided by the observed variance of Y. Also note that the term $2k\log(2n-3)$ comes from $|B^{(k)}|$, that is, to penalize the choice of the shift positions. The penalty for the shift magnitudes is all in the determinant term.

Lemma 2.8. *Assume* M *be a symmetric, invertible* $n \times n$ *matrix For all vectors* $\beta, a \in \mathbb{R}^n$

$$\int_{\beta \in \mathbb{R}^n} \exp\left(\beta^T a - \frac{1}{2}\beta^T M \beta\right) d\beta = (2\pi)^{\frac{n}{2}} |M|^{-\frac{1}{2}} \exp\left(a^T M^{-1} a/2\right).$$

Proof. Simply write $2\beta^T a - \beta^T M \beta$ as $a^T M^{-1} a - (\beta - M^{-1}a)^T M (\beta - M^{-1} a)$, then integrating β is straightforward. □

2.6 Fast Algorithm for the Square Root Covariance and Its Inverse

Let Σ be the covariance matrix from a BM model on a tree with n tips. For taxon i and j we have

$$\Sigma_{ij} = \sigma^2 t_{ij} \, . \tag{2.12}$$

Note that the OU covariance (2) can be written in this form, if the tree is ultrametric (i.e. all the tips are at the same distance from the root) and provided that we use transformed branch lengths. We further assume that any polytomy in the tree is arbitrarily resolved with branches of length 0.

The recursive algorithm below produces matrices $B = \Sigma^{1/2}$ and $D = \Sigma^{-1/2}$ with a interpretable mapping of rows and columns to internal nodes in the tree. More formally, D and B satisfy

$$BB^T = \Sigma, \quad B = (D^{-1})^T, \quad D^T \Sigma D = I_n, \quad \text{and } \mathbf{1}^T D = (0, \ldots, 0, a)$$

for some a. The first $n-1$ columns of D are Felsenstein's phylogenetic independent contrasts (1985). Each one corresponds to an internal node in the tree, with non-zero entries for taxa descending from that node only. If that node has 2 children, the entries of all taxa descending from a given child share the same sign (one child associated with positive entries and the other associated with negative entries). The n^{th} column of D is associated with the root branch above the root node, and the precision of the estimated ancestral state at the root.

Similarly, the first $n-1$ columns of B are each associated with a given internal node in the tree, with zero entries for any taxa not descending from that node. All entries at the taxa descendant of a given child of the node are equal (one child associated with positive entries and the other child associated with negative entries). The n^{th} column of B is associated with the root branch above the root node, and all its entries are equal.

The proof that the algorithm below satisfies all the claims above was sketched by Stone (2011). Our contribution here is to provide a general description of the algorithm, in particular the inverse to Felsenstein's algorithm to obtain B. We also provide an implementation of this algorithm available at `https://github.com/khabbazian/l1ou`.

Algorithm. Each step takes in a tree $\mathcal{T}_k$ with k leaves, F and G matrices of size $n \times k$, and D and B matrices of size $(n-k) \times (n-k)$. The result of each step is a reduced tree with $k-1$ leaves, reduced matrices F and G of size $n \times (k-1)$, and expanded matrices D and B of size $(n-k+1) \times (n-k+1)$. The algorithm is initialized with $k = n$, the original tree, empty matrices D and B and $F = G = I_n$. Each column in F and G corresponds to a leaf in the current tree (that is, an leaf or internal node in the original tree), and each row corresponds to a taxon in the original tree. We will show recursively that for each k, $F_{ji} > 0$ if taxon j is a descendant of node i, $F_{ji} = 0$ otherwise, that $\mathbf{1}^T F = (1, \ldots, 1)$, and that $G_{ji} = 1$ if j is a descendant of i and $G_{ji} = 0$ otherwise. These properties are obviously met at initialization. For each $k \geqslant 2$, the following steps are followed.

1. Choose a cherry in $\mathcal{T}_k$. Let i_1 and i_2 bet these 2 leaves, b_1 and b_2 the external branches leading to i_1 and i_2, t_1 and t_2 the lengths of b_1 and b_2. Also, let b be the branch that is parent to both b_1 and b_2, with length t in $\mathcal{T}_k$. Define $\mathcal{T}_{k-1}$ by removing branches b_1 and b_2 from $\mathcal{T}_k$, grafting a new leaf i_{new} to b, and changing the length of b' to

$$t_{\text{new}} = t + \left(t_1^{-1} + t_2^{-1}\right)^{-1}.$$

   Without loss of generality, we can assume that i_1 and i_2 are the indices of the columns in F and G that correspond to leaves i_1 and i_2.

2. The new matrix D is formed by adding to the current D the new column

$$D_{\cdot i_{\text{new}}} = \frac{1}{\sqrt{t_1 + t_2}} \left(F_{\cdot i_1} - F_{\cdot i_2}\right).$$

Because $F_{ji} > 0$ if j is a descendant of i and 0 otherwise, $D_{\cdot i_{\text{new}}}$ has positive entries equal to those in $F_{\cdot i_1}/\sqrt{t_1 + t_2}$ for descendants of i_1, negative entries equal to those in $-F_{\cdot i_2}/\sqrt{t_1 + t_2}$ for descendants of i_2, and 0 entries for all other taxa. Also, $\mathbf{1}^T D_{\cdot i_{\text{new}}} = 0$ necessarily because $\mathbf{1}^T F_{\cdot i_1} = \mathbf{1}^T F_{\cdot i_2} = 1$.

3. The new matrix B is formed by adding to the current B the new column

$$\frac{t_1}{\sqrt{t_1 + t_2}} G_{\cdot i_1} - \frac{t_2}{\sqrt{t_1 + t_2}} G_{\cdot i_2} \,.$$

   Because $G_{ji} = a > 0$ if j is a descendant of i and 0 otherwise, this new column has positive and equal entries for all descendants of i_1, and negative and equal for all descendants of i_2, and 0 entries for all other taxa.

4. The new matrix F is formed by replacing the columns $F_{\cdot i_1}$ and $F_{\cdot i_2}$ by a single column $F_{\cdot i_{\text{new}}}$ defined as:

$$F_{\cdot i_{\text{new}}} = \frac{1/t_1}{1/t_1 + 1/t_2} F_{\cdot i_1} + \frac{1/t_2}{1/t_1 + 1/t_2} F_{\cdot i_2} \,.$$

   Because i_1 and i_2 are sister, the descendants of i_{new} is the union of all descendants of i_1 and i_2, so $F_{j i_{\text{new}}} = 0$ if j is not descendant of i_{new}, and $F_{j i_{\text{new}}} > 0$ otherwise. Also, $\mathbf{1}^T F_{\cdot i_{\text{new}}} = 1$ because both $\mathbf{1}^T F_{\cdot i_1} = 1$ and $\mathbf{1}^T F_{\cdot i_2} = 1$.

5. The new matrix G is formed by replacing the columns $G_{\cdot i_1}$ and $G_{\cdot i_2}$ by a single column $G_{\cdot i_{\text{new}}}$ defined as:

$$G_{\cdot i_{\text{new}}} = G_{\cdot i_1} + G_{\cdot i_2} \,.$$

   By induction, $G_{j i_{\text{new}}} = 1$ if taxon j is a descendant of i_{new} and $G_{j i_{\text{new}}} = 0$ otherwise.

At the end when $k = 1$, no cherry can be picked (step 1). $\mathcal{T}_1$ has a single leaf i_{last} connected to a single branch of length t_{last}, and F and G both consist of a single column. At this stage, D is augmented by the column

$$\frac{1}{\sqrt{t_{\text{last}}}} F_{\cdot i_{\text{last}}}$$

whose entries are all positive, and B is augmented by the column

$$\sqrt{t_{\text{last}}}\, G_{\cdot i_{\text{last}}} = \sqrt{t_{\text{last}}}\, \mathbf{1}\,.$$

The new tree, F and G are then left empty.

2.7 Simulations

Shifts in One Trait

We used simulations to compare the accuracy of different methods: ℓ1ou combined with either pBIC, BIC or the same AIC_c as used in surface (using its forward phase only to focus on the shift configurations rather than the shift magnitudes), bayou, and the stepwise selection method proposed by Ho and Ané (2014). This stepwise method is capable of accepting various criteria, but we used here the 'mBIC' also proposed by Ho and Ané (2014). Bayou requires the user to choose a prior distribution for each model parameter, and the results are sensitive to this choice. We made choices based on the true parameters used to simulate the data: the number of shifts was given a conditional Poisson prior distribution with mean the true number of simulated shifts. A uniform prior was chosen for α and σ^2 on $[\alpha - 0.5, \alpha + 0.5]$ and $[\sigma^2 - 0.5, \sigma^2 + 0.5]$. An empirical Bayes approach was taken for the shift magnitudes as in Uyeda and Harmon (2014), with a centered normal prior distribution with standard deviation equal to twice that observed in the tip data. Since bayou is a Bayesian method, it returns a posterior distribution on shift configurations. To summarize this distribution, we took a liberal approach and said that a branch was detected to have a shift if the posterior probability of a shift on that branch was 0.10 or greater. For ℓ1ou methods, we used the random root

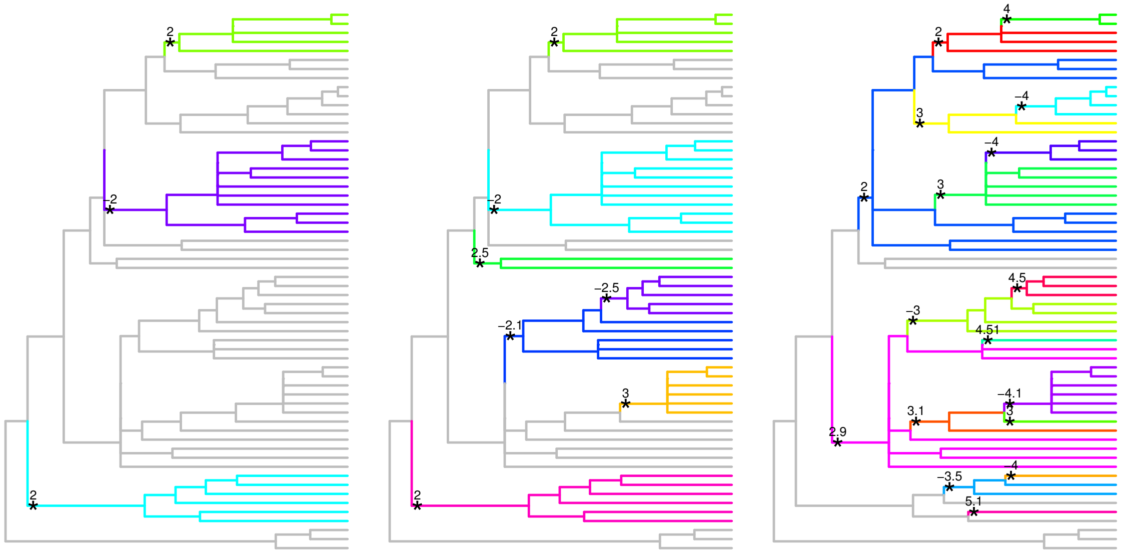


Figure 2.4: Tree with 60 taxa used in simulations to compare the accuracy of various methods. Data were simulated under the OU model with no shifts or with multiple shifts (left: 3, center: 7, right: 17 shifts). The shift positions are annotated with stars and their simulated magnitudes.

covariance in (2.2). For all methods, we set the maximum number of shifts to half the number of taxa in the tree.

We simulated datasets under OU models along 2 different phylogenies of flowering plants in the family Melastomataceae, one with 60 taxa and one with 215 taxa using the function rTraitCont in the R package ape (Paradis et al., 2004). The first tree (Figure 2.4) is the consensus phylogeny from Kriebel et al. (2015) pruned to a single accession per species. It was small enough for all methods to run reasonably fast, and was used to compare the methods' accuracies. The second tree was simply used to sample subtrees and compare the methods' running times as a function of tree size.

On the "small" tree we simulated traits under 4 different configurations: either no shift, or 3, 7 or 17 shifts as shown in Figure 2.4. We used $\alpha = 1$, corresponding to a moderate half-life (0.69) compared to the tree height, which was set to 1 by

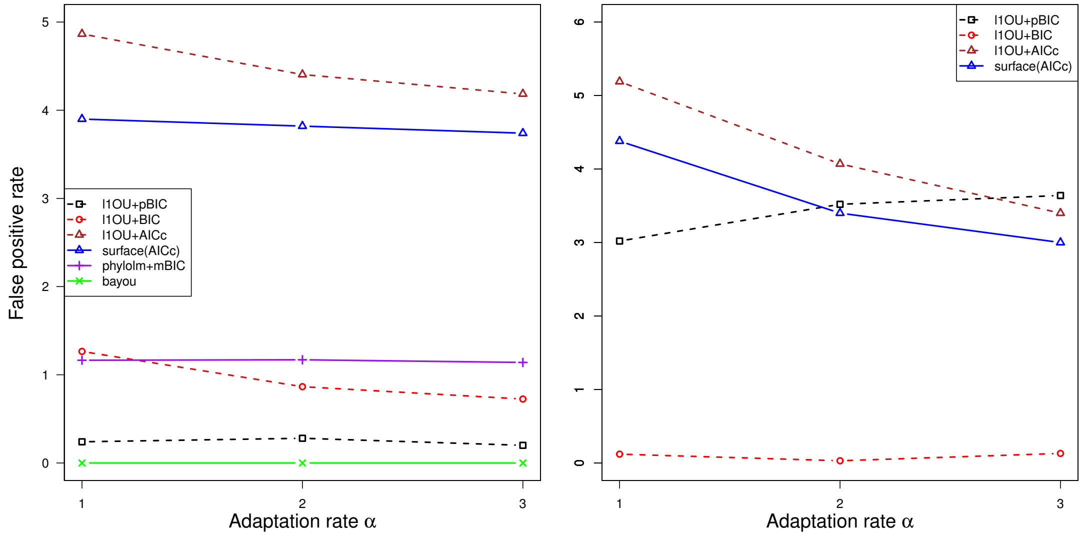


Figure 2.5: Number of false positives for different methods to detect shifts in the OU process. One trait (left) or four independent traits (right) were simulated under a homogeneous OU model with no shift.

rescaling all branch lengths. We set $\sigma^2 = 2$ to fix the stationary variance $\sigma^2/(2\alpha)$ at 1. In the absence of shifts, we varied α while keeping $\sigma^2/(2\alpha) = 1$. In the presence of shifts, we instead varied the shift magnitudes. They were first set to the values shown in Figure 2.4. They correspond to moderate magnitudes, just large enough to be detected individually (if their phylogenetic positions were known) with non-negligible power (Ho and Ané, 2014), because means at the tips differ by about 1 stationary standard deviation ($\beta_b(1 - e^{-\alpha a_b}) \approx \sigma/\sqrt{2\alpha} = 1$). These shift magnitudes were then all scaled by the same factor, varying from 1 to 4, to create easier scenarios. For each condition, we generated 200 replicate data sets with 1 trait each, and estimated the shift configuration using each method.

To compare the methods' accuracies, we first considered the scenario with no shifts and calculated the number of false positives, i.e. the average number of detected shifts, necessarily all false. Figure 2.5 (left) shows that ℓ1ou+AIC$_c$ and surface (both using AIC$_c$) have many more false positives than the other methods.

Next, we considered scenarios with 3, 7 or 17 true shifts and calculated the recall rate of each method (average proportion of branches with a true shift that were detected as having a shift) and precision (average proportion of detected shifts that were true, i.e. located on a branch with a true shift). Figure 2.6 shows that surface and ℓ1ou with AIC_c are liberal methods: both enjoy high recall rates (they find many of the true shifts) but tend to have low precision (they also find many false positives). Given our liberal threshold to call a shift in bayou ($\text{PP} \geqslant 0.1$), it is not surprising to find that bayou also has a tendency to be liberal, with high recall rates and low precision. However, both performance measures tended to be lower for bayou than for ℓ1ou with AIC_c. On the other extreme, ℓ1ou was conservative when coupled with pBIC, enjoying the highest precision (the detected shifts were mostly true) but a low recall rate (many true shifts were missed). When coupled with BIC and on a single trait, ℓ1ou provided an intermediate approach, which might provide a good balance to reach a reasonable precision with a reasonable recall rate. The phylolm stepwise method based on mBIC performed consistently more poorly than other methods. Its recall rate was among the lowest, but its precision was always comparable or lower than that or ℓ1ou with pBIC, for instance. Figure 2.6 (right) also shows that identifying 17 shifts on a 60-taxon tree is much more difficult than detecting 3 or 7 shifts. The performance of all methods went down significantly with 17 shifts. This is not surprising, because each shift was visible by an average of 3.5 extant species, compared to 8.6 when there were only 7 shifts. Detecting the exact position of each shift is likely to be much more difficult as the density of shifts increases within the tree.

To compare the methods' running time, we used a 215-taxon plant phylogeny expanded from Kriebel et al. (2015) and randomly subsampled between 32 and 215 taxa to obtain a smaller tree. For each tree size, we generated 2 replicate data sets with a single trait, using $\alpha = 1$, $\sigma^2 = 2$ and a true number of shifts that increased with the tree size (from 4 shifts on 32 taxa to 26 shifts on 215 taxa). The maximum number of estimated shifts was set for all methods to twice the true number of shifts. We kept a constant number of 400,000 generations in bayou, because it is unclear how this number should be set to obtain a comparable mixing convergence across tree sizes. However good mixing is likely to require more generations on large

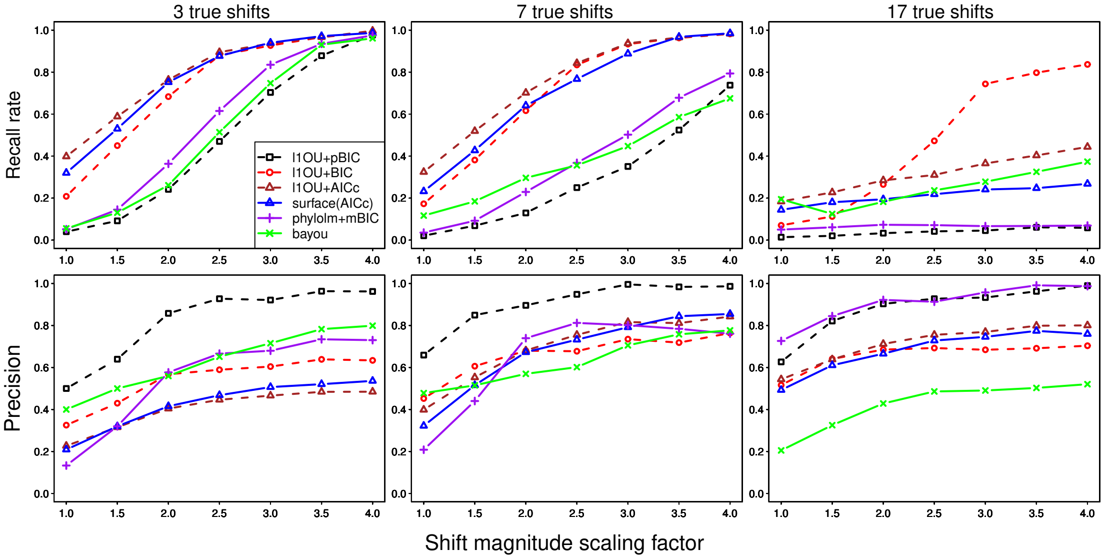


Figure 2.6: Recall rate (first row) and precision (second row) of various methods to detect the position of OU shifts on the tree with three, seven, and 17 true shifts (see Figure 2.4). The magnitudes of all shifts were increased by the same scaling factor.

trees with large numbers of edges to evaluate. Hence the running time for bayou is likely to be underestimated for large trees. All running times were obtained with a 2.7 GHz processor. Figure 2.7 displays the average elapsed time of each method, showing that previously proposed methods do not scale well to trees with a few hundred taxa. On the other hand, ℓ1ou is between one to two orders of magnitude faster than the other methods, with no loss of accuracy.

Shift Detection from Multiple Traits

We conducted 2 simulation experiments with multiple trait data. First, we explored the effect of conducting standard PCA to reduce the problem dimension, before detecting shifts on the first PC axis only. Second, we explored the performance of ℓ1ou and surface when applied to multiple traits.

We simulated data (100 replicates for each situation) under the same 60-taxon tree as before (fig 2.4) except that each data set contained $m = 20$ independent

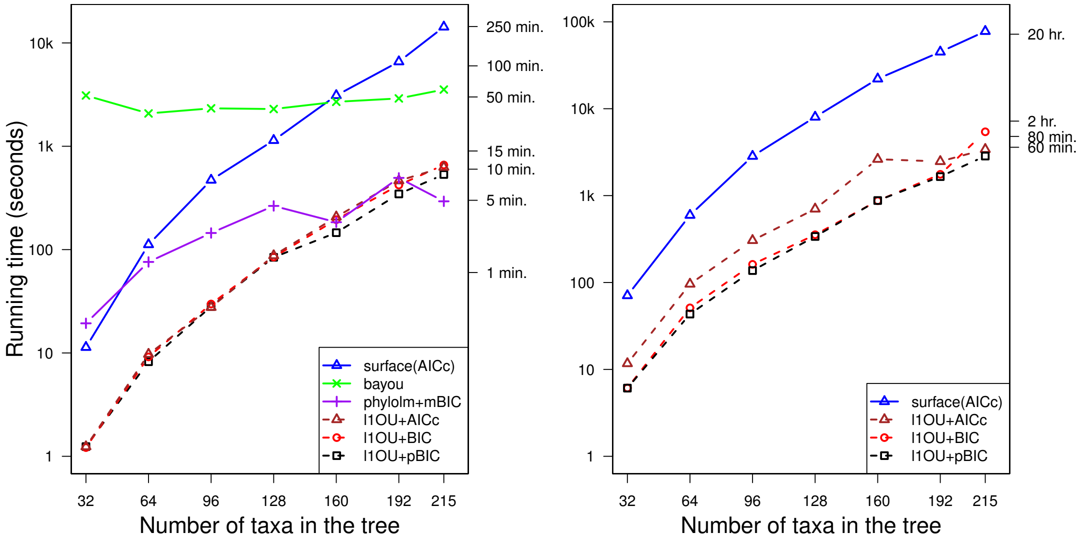


Figure 2.7: Average running time of different methods for OU shift detection versus the number of species in the tree, for data sets with a single trait (left) or four traits (right). Time is displayed on a log scale.

continuous traits simulated under OU model. For our first experiment, the true model had no shifts. We set $\alpha = 2$, corresponding to a moderate half-life 0.34, and $\sigma^2 = 4$ to fix the stationary variance at 1. Figure 2.8 shows the systematic error caused by using only the first PC, i.e. the axis with the largest variation in the data. As expected, some branches near the root are consistently detected as having a shift. Even though we used the most conservative method (ℓ1ou + pBIC) to analyze the first PC, at least one shift was detected near the root in 65% of the replicates, on one of the branches marked by a star. When using ℓ1ou + BIC or AIC$_c$, the occurrence of these false positives increased to 82% and 89%.

Second, we considered the same tree as before with 0, 3, 7 or 17 true shifts but with multiple traits (Figure 2.4). We used $\alpha = 1$ and $\sigma^2 = 2$ and generated $m = 4$ independent traits under the OU model. We chose 4 traits because this is representative of a number of applications, and because surface was too slow to handle 20 traits for many replicates (about one hour per replicate). When no shifts were simulated, we further varied α keeping the stationary variance $\sigma^2/(2\alpha) = 1$.

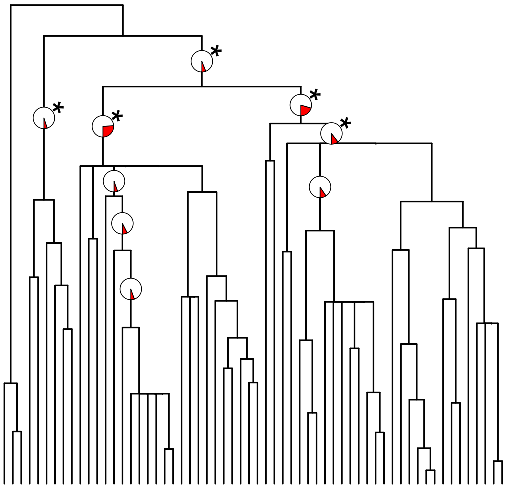


Figure 2.8: Shifts detected by analyzing the first standard principal component using ℓ1ou + pBIC. Data were generated under the BM model ($\alpha = 0$) with no shifts, $\sigma^2 = 4$ and 20 variables. Pie charts show the proportion of replicates for which a shift was detected on a given branch (shaded area), on branches for which this proportion was 5% or greater.

Figure 2.5 (right) shows that all methods except ℓ1ou+pBIC had a few false positives. In contrast to analyses with a single trait, ℓ1ou+pBIC appeared as most conservative. We then repeated the same simulations except that the 4 traits had residual correlation, either from correlated drift or from correlated selection. This caused an increase in the number of falsely detected shifts, for all methods (Figure A.9).

In simulations with shifts, the magnitudes shown in Figure 2.4 were multiplied by $+1$ or -1 randomly and independently for each trait. They were then all scaled by a common factor as before, varying from 1 to 4. Bayou and the phylolm stepwise method were not applied since they cannot handle multiple traits. As expected, using multiple independent traits improved both the recall rate and the precision of all methods, compared to using a single trait (Figure 2.9). Like before, surface and ℓ1ou + AIC$_c$ were very similar and were the most liberal methods and pBIC tended to be the most conservative With 4 traits ℓ1ou + BIC was also very conservative. However, there were situations when the most liberal methods kept detecting false

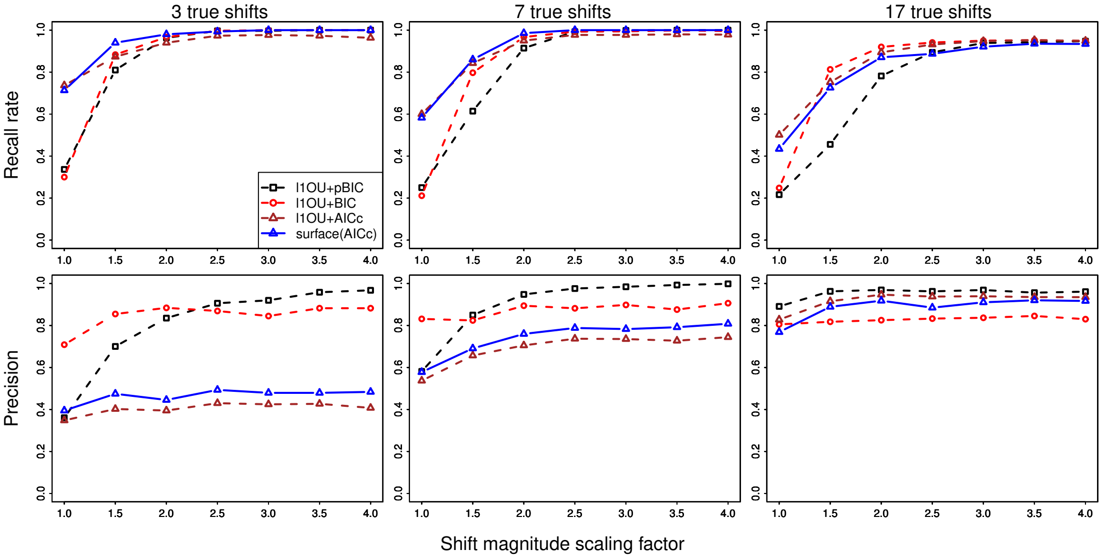


Figure 2.9: Recall rate (first row) and precision (second row) of multivariate methods to detect the position of OU shifts on the 60-taxon tree with three, seven, and 17 true shifts (see Figure 2.4), from 4 traits. The magnitudes of all shifts were increased by the same scaling factor.

shifts (precision capped around 50% with 3 true shifts and 3 false shifts detected) even when the signal-to-noise ratio increased (large shift magnitudes), while the most conservative method reached both a recall rate of 100% and a precision of 100%.

We also evaluated the accuracy of shift detection when phylogenetic PCA is first applied to reduce the dimension of the data, to detect shift positions by on each pPC axis separately. For each data set generated above, we applied pPCA (assuming a BM model as proposed by Revell, 2009) and applied various shift detection methods on the first axis. The multivariate version of ℓ1ou or surface on the original multivariate data had a better or comparable recall rate and precision than the same method applied to the single first pPC (Figure A.1).

2.8 Illustrations with Data on Anolis Lizards

Anolis lizards on the Caribbean islands have independently evolved a similar set of "ecomorphs", such that species of the same ecomorph category from different islands are similar morphologically (Losos et al., 1998). Mahler et al. (2013) studied similarities among islands by considering 11 traits including body size, limb and tail lengths, and adhesive toepad lamella number across 100 species. They applied pPCA and retained the first 4 axes, which together explained 93% of variation. Their data and tree are available in the supplementary material of Mahler et al. (2013). We applied surface and ℓ1ou + pBIC, BIC or AIC$_c$ to their 4 pPC traits, using the random root covariance in (2.2) and allowing for a maximum of 50 shifts. ℓ1ou + pBIC detected 12 shifts (in 13.8 minutes). Figure 2.10 shows that each of these shifts is supported by several of the 4 traits. Surface found 28 shifts (in 2 hours in 12 minutes), which include 11 of the 12 shifts detected here (Figure A.2). The one shift not detected by surface had the lowest bootstrap support (39%). All other shifts had support between 69% and 100%. The 28 shifts found by ℓ1ou + AIC$_c$ (in 13.2 minutes) included all 12 shifts found by ℓ1ou + pBIC and were very similar to those found by surface up to equivalent parsimonious configurations (Figure A.3). With this many shifts, the one configuration returned by ℓ1ou + AIC$_c$ (or by surface) is equivalent to many other configurations that define the same clustering of taxa. Therefore, this one configuration is masking a lot of uncertainty about the shift locations. Because pBIC is quite conservative, we can be more confident in its 12 shifts compared to the extra 16 shifts found by AIC$_c$ or by surface. On these data, ℓ1ou + BIC was most conservative and did not detect any shift. Figure A.4 shows the score profile plot of each method. For BIC, this profile shows a local optimum in BIC at 9 shifts, 7 of which were found by pBIC (Figure A.5).

Of the 12 shifts detected by ℓ1ou + pBIC, 4 occurred within Cuba (or as a dispersal to Cuba), 5 within Hispaniola, 2 within Jamaica and only 1 with Puerto Rico (or as a dispersal to Puerto Rico), based on a parsimonious geography reconstruction (Figure A.6). Overall, our results suggest that ecomorphological convergence is not as convincing as previously argued. First, over half of the shifts previously detected

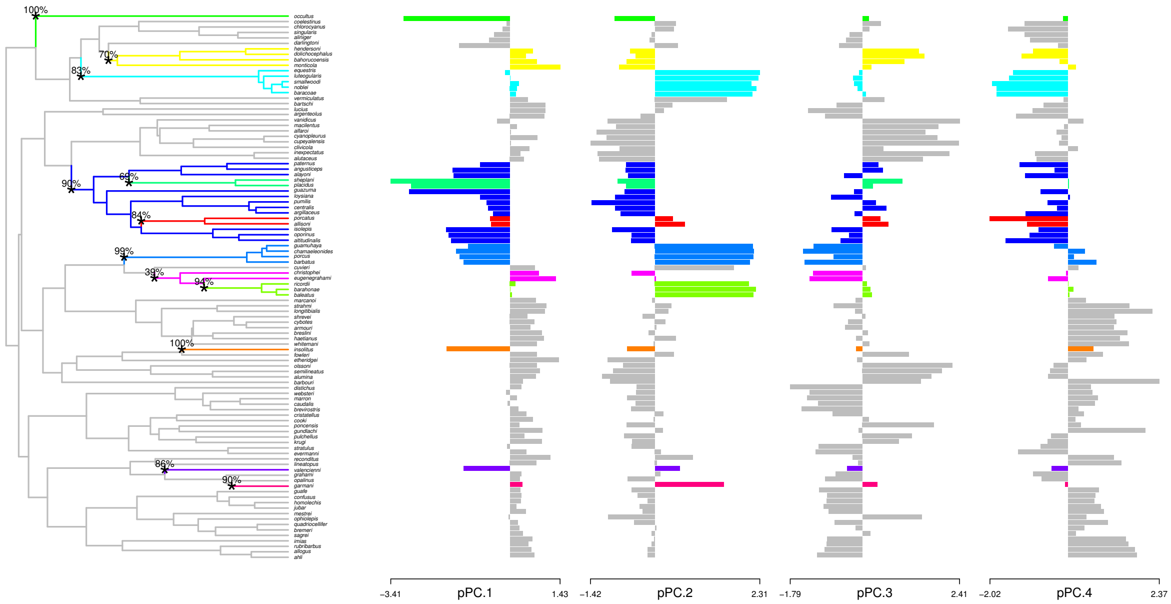


Figure 2.10: Shifts in Anolis lizard morphology. Re-analysis of the 4 traits from Mahler et al. (2013) with ℓ1ou and pBIC provided support for 12 evolutionary shifts in optimum morphology under an OU process. Left: shift configuration. Edges with a shift are annotated with a star and bootstrap support. Right: bar graphs showing the 4 traits combined for analysis.

are suspected to be unreliable. Second, 2 of the 4 islands only have 1 or 2 confirmed shifts, weakening the evidence for repeated convergence on separate islands.

When analyzing the first trait only (pPC1, which alone explained 40% of variation), fewer shifts were detected by all methods, showing the gain in detection power from combining multiple traits. Four shifts were detected with ℓ1ou + pBIC, all of which were also detected by ℓ1ou + AIC$_c$, which detected 16 shifts total. Using the generalized lasso (2.7) + AIC$_c$ on these 16 shifts, we detected a high level of convergent evolution with a total of 8 regimes only. In comparison, surface detected 12 shifts and 5 distinct optima, with some similarities but also marked differences (Figure A.7). These 2 convergent evolution models had very similar AIC$_c$ scores however (-86.37 and -86.40), highlighting great uncertainty about the exact phylogenetic placement of shifts and convergent evolution.

2.9 Discussion

In this work we adapted the lasso, now widely used for standard statistical model selection, to phylogenetic comparative data and the detection of shifts in the mean. The lasso penalizes parameters by their absolute values, which leads to sparse models with most parameters estimated at 0. The OU process that we used can model the response to a changing adaptation landscape, to which the lasso provides a parsimonious solution.

We also proposed a new phylogenetic criterion pBIC that explicitly accounts both for phylogenetic correlation, and for the large number of configurations with a given number of shifts k. This number of models grows extremely fast with k, leading to overfitting issues and high rates of falsely detected shifts with AIC. On the contrary, pBIC was shown to be conservative. Interestingly, the pBIC penalty for a k-shift model is not a simple function of the number of parameters, and/or of the number of configurations with k shifts (Massart, 2007). The penalty depends on the best shift configuration, and generalizes the notion of a shift's effective sample size (Ané, 2008). In particular, shifts leading to small clades are penalized less than shifts leading to large clades, especially if phylogenetic correlation is low, because their effective sample size is smaller. Our ℓ1ou method could be combined with any further improvements to pBIC. Also, pBIC can be generalized to other models with shifted means, making it applicable to models with jumps derived from the BM process for instance (see below).

Bastide et al. (2015) recently considered the same problem and highlighted the same identifiability issues on shift configurations. They derived the exact number of non-equivalent (distinguishable) parsimonious configurations of k shifts, which could depend on the tree topology. This number, necessarily smaller than the number of ways to choose k edges, could be used to improve our pBIC derivation (affecting the term $2k\log(2n-3)$). To select k and for a single trait, Bastide et al. (2015) used a criterion penalty based on the number of distinguishable configurations, with guaranteed properties if α is known. For one trait, the maximum likelihood configuration with k shifts is found with Expectation-Maximization

(Dempster et al., 1977), which is probably more thorough but slower than our approach.

A major strength of our phylogenetic lasso method is its speed, being one or more orders of magnitudes faster than currently existing methods. Parallelization of our implementation could further reduce its running time. This is because the set of candidate models returned by the lasso can be evaluated for pBIC in parallel; this second step is the computational bottleneck, consuming much more time than the first lasso step. To achieve fast running times, we also implemented a linear-time algorithm (Stone, 2011) to obtain the square-root and inverse square-root of the covariance matrix, $\Sigma^{(\alpha)}$. This fast algorithm facilitates both the noise-whitening transformation for the phylogenetic lasso and the bootstrap procedure here, but it could have broader benefits for other applications. The matrices $\Sigma^{-1/2}$ and $\Sigma^{1/2}$ are not unique (many matrices satisfy $A'A = \Sigma$) and the matrices returned by the last algorithm are not symmetric, but they have an advantage of interpretability: each row corresponds to an edge in the tree, including a root edge. Therefore, they provide phylogenetically corrected residuals that map onto the phylogenetic tree. Their applications include model diagnostics and visualizations (Pennell et al., 2015) with possible interpretation as to the cause of potential model violations. Here, phylogenetically corrected residuals might be used to detect possible model violations that might correlate with shift configurations.

Our bootstrap procedure, which uses both $\Sigma^{-1/2}$ and $\Sigma^{1/2}$, is comparable to the fully parametric bootstrap method used by Ingram and Mahler (2013) for surface. Our method is partially non-parametric, however, in that we resample the phylogenetically corrected residuals instead of sampling from the OU process, to gain some robustness to potential violation of the OU model assumptions. The results from such bootstrap procedures should be interpreted with caution, however, because they can depend heavily on the shifts simulated under the bootstrap model. If this model only uses the shifts detected conservatively with pBIC, then any true shift that went undetected will necessarily receive low bootstrap support. On the lizard data for instance, adding an extra 2 shifts to the pBIC configuration increased the pBIC score by 4.99 only, but resulted in greatly increased bootstrap support for the newly added shifts (from close to 0% to 63% and 62%, see Figure A.8). It might also be advantageous to use the shifts detected with a more liberal criterion (AIC_c) in the bootstrap simulation model, but analyze the bootstrap data sets with a conservative criterion (pBIC). Hence these bootstrap values should be interpreted with caution and more work is needed to improve parametric bootstrap methods here, when model selection is involved.

For shifts located on neighboring edges, extra caution should be taken because of identifiability issues. For instance, the data contains no information on whether a shift is at the base on the ingroup clade versus the outgroup clades (i.e. on either edge connecting to the root). Even if the bootstrap support for a shift is 100% at the base of the ingroup clade, the user should keep in mind that there is still complete uncertainty about the exact placement of this shift on either side of the root, or its timing along either edge. Similarly, shifts detected on two sister clades should be interpreted with caution, even if each one receives 100% bootstrap support. The exact same data could be obtained with a shift on the edge ancestral to these 2 sister clades, and only 1 subsequent shift to one of the clades. Here again, the 100% bootstrap values ignore uncertainty due to a lack of identifiability. Bayesian methods can deal with this issue much more elegantly (Uyeda and Harmon, 2014), because one might place equal prior probabilities on all the non-distinguishable shift configurations. Posterior probabilities would reflect uncertainty between all these configurations, even uncertainty on the location of a shift along a given edge. Non-identifiable shift configurations might also have different posterior probabilities because their shared maximum likelihood might be achieved at different shift magnitudes, which are not necessarily equally likely a priori. Therefore, a Bayesian framework can distinguish between non-idenfiable shift configurations using biologically reasonable priors on shift magnitudes. Also, even though the posterior mean number of shifts depends on the prior number of shifts, Bayesian posterior distributions might quantify uncertainty over the various configurations with a fixed number of shifts better than bootstrap samples (see also visualization tools in Rabosky et al., 2014). This is because bootstrap samples are generated under a unique bootstrap simulation model, from the best estimated shift configuration only. More work could still be done to improve frequentist bootstrap procedures or other ways to quantify uncertainty, for the detection of phylogenetic shifts.

The lack of identifiability between different shift configurations is because the data truly bear on the clustering of taxa into groups. If there is evidence that two sister clades and their outgroup taxa make 3 different clusters each with its own adaptive optimum, then we might be able to estimate these 3 clusters with very high confidence. However there will still remain complete uncertainty (without fossil data) to know how many adaptive shifts occurred, at the base of which clade they occurred, and at what time. Therefore, the proposed method should be treated as an estimation of phylogenetically-consistent clusters, rather than exact shift positions.

In many applications, the underlying data are truly on a continuous scale but are discretized to facilitate analysis or to provide a taxonomic description. For instance, moss sporangium shape (Rose et al., 2015) might be described as either "round" or "linear", with some subjectivity involved when scoring intermediate species, or training needed to achieve consistent scoring between different observers. For the purpose of defining thresholds to categorize continuous measurements into discrete values, our method would provide an objective and phylogenetically-aware method. A liberal model selection criterion like AIC_c would be recommended, to detect sufficiently many categories and to prioritize the influence of the trait data over the species phylogenetic placement.

For the purpose of categorizing a continuous variable or for the study of adaptation, an interesting next step is to detect convergence, when different shifts lead to the same selective optimum value. For one variable, we used the generalized lasso to penalize differences between pairs of optima (Tibshirani and Taylor, 2011). However, more work is needed to adapt pBIC, to correctly integrate out the constrained shifts and to account for the number of convergent configurations with k shifts. For multiple traits, the ideas of the generalized and group lasso could be combined in an ℓ_1 penalty that favors convergent regimes shared by all traits. But further work is needed because there is no fast algorithm for this form of penalty yet.

Extending our method to account for residual correlation between traits would be desirable. Simulations showed that none of the available methods are robust to the presence of correlation among traits due to drift (Figure A.9), with a marked increased of falsely detected shifts. Models for correlated traits could also combine primary response variables with potential predictors into one multivariate variable, to model variation in the response explained by shifts as well as predictors (Hansen et al., 2008; Bartoszek et al., 2012). However, fitting phylogenetic multivariate OU models with arbitrary selection and drift covariance matrices is difficult computationally (e.g. Clavel et al., 2015) and new theory would be needed for these models, to select the appropriate number of shifts.

Another extension of our method would be to move away from the OU model with discontinuous jumps in the adaptive optimum but continuous trait evolution. For example, the OU model leads to the same trait distribution on present-day taxa as a BM punctuated by jumps causing discontinuity in the process (at an evolutionary time scale), provided that branch lengths in the tree are rescaled depending on α (Ho and Ané, 2014). If the OU model leads to unreasonably large shifts in optimal values, a BM model might provide jumps that are more reasonable biologically, even though the two models are statistically equivalent. This is likely to occur if phylogenetic correlation is high (low α, or slow adaptation), in which case the OU model needs an unreasonably large shift in the adaptive optimum to explain a moderate jump in the observed mean. While the OU model is statistically equivalent to a process with jumps, our lasso and pBIC in (2.5) both penalize the magnitude of shifts in the adaptive optima, rather than the magnitude of jumps in the observed means. Hence, our model and implementation would need to be adapted to BM evolution with jumps to penalize changes in observed means rather than in adaptive shifts, through an adaptation of $X^{(\alpha)}$ and of the phylogenetic covariance. Further work could also extend this BM model with jumps to allow for an unknown level of phylogenetic correlation, using an extra parameter like Pagel's λ (Lynch, 1991; Pagel, 1999) and a similar approach to vary λ (instead of α) across different runs of the lasso.

Finally, extending our method to account for measurement error should be easiest when multiple measurements are available per species, using the observed standard errors of species means as in Ives et al. (2007). Doing so could be most beneficial if two very closely sister species have quite different trait values, in the range of measurement error. A spurious shift to one of the two sister species might be needed to explain the trait difference if measurement error is ignored, with a possibly overestimated α (underestimated phylogenetic correlation).

2.10 Data Accessibility

The R package ℓ1ou is available open source at `https://github.com/khabbazian/l1ou`.

Chapter 3

Novel Sampling Design for Respondent-driven Sampling

Contribution

- I conducted all the simulations and result visualizations expect Figure 3.3.
- I implemented rdssim R package (https://github.com/khabbazian/rdssim).
- I developed all the theoretical results except Proposition 3.6.
- I wrote the first draft of the manuscript.

Abstract

Respondent-driven sampling (RDS) is a type of chain referral sampling popular for sampling hidden and/or marginalized populations. As such, even under the ideal sampling assumptions, the performance of RDS is restricted by the underlying social network: if the network is divided into weakly connected communities, then RDS is likely to oversample one of these communities. In order to diminish the "referral bottlenecks" between communities, we propose anti-cluster RDS (AC-RDS), an adjustment to the standard RDS implementation. Using a standard model in the RDS literature, namely, a Markov process on the social network that is indexed by a tree, we construct and study the Markov transition matrix for AC-RDS. We show that if the underlying network is generated from the Stochastic Blockmodel with equal block size, then the transition matrix for AC-RDS has a smaller spectral gap and consequently faster mixing properties than the standard random walk model for RDS. In addition, we show that AC-RDS reduces the covariance of the samples in the referral tree compared to the standard RDS and consequently leads to the smaller variance and design effect. We confirm the effectiveness of the new design using both the Add-Health networks and simulated networks.

Keywords: hard-to-reach population; social network; trees; Markov chains; upectral representation; anti-cluster RDS

3.1 Introduction

Public policy and public health programs depend on estimating characteristics of hard-to-reach or hidden populations (e.g. HIV prevalence among people who inject drugs). These hard-to-reach populations cannot be sampled with standard techniques because there is no way to construct a sampling frame. Heckathorn (1997, 2002) proposed respondent-driven sampling (RDS) as a variant of chain-referral methods, similar to snowball sampling (Goodman, 1961; Handcock and Gile, 2011), for collecting and analyzing data from hard-to-reach populations. Since then, RDS has been employed in over 460 studies spanning more than 69 countries (Malekinejad et al., 2008; White et al., 2015).

RDS encompasses a collection of methods to both sample a population and infer population characteristics (Salganik, 2012), referred to as RDS sampling and RDS inference. RDS sampling starts with a few "seed" participants chosen by a convenience sample of the target population. Then, the initial participants are given a few coupons to refer the second wave of respondents, the second wave refers the third wave, and so on. The participants receive a dual incentive to (i) take part in the study and (ii) successfully refer participants. The dual incentive helps RDS obtain many waves of sampling. With many waves of sampling, RDS has the potential to penetrate the broad target population and reduce its dependency on the initial convenience sample.

Since Heckathorn's original RDS paper, the statistical literature on RDS has created several estimators that seek to reduce the bias and estimate confidence intervals (Heckathorn, 2011). The most popular RDS estimators are generalized Horwitz-Thompson type estimators where the inclusion probabilities are derived from various models of the sampling procedure (Volz and Heckathorn, 2008; Gile, 2011; Gile and Handcock, 2011).

RDS performance has been evaluated through simulation studies (Goel and Salganik, 2010; Gile and Handcock, 2010), empirical studies (Wejnert, 2009; McCreesh et al., 2012), and theoretical analyses (Goel and Salganik, 2009). The main message of these studies is that (i) RDS can suffer from bias; (ii) in some cases, the current

RDS estimators do not reduce bias; and most importantly, (iii) the estimators have higher variance (and thus, design effect) than what was initially thought (Goel and Salganik, 2009, 2010; White et al., 2012). To help bridge the gap between theory and practice, Gile et al. (2015) suggests various diagnostics to examine the validity of the modeling assumptions.

Goel and Salganik (2009) and Verdery et al. (2015) analytically study the effects of the homophily or community structure on the variance of the estimator. Homophily, a common property of social networks, is the tendency of people to establish social ties with others who share common characteristics such as race, gender, and age. Strong homophily creates community structure in the social network. This in turn creates referral bottlenecks between different groups in the population; the RDS referral chain can struggle to cross these bottlenecks, failing to quickly explore the network. In such situations, RDS is sensitive to the initial convenience sample, leading to biased estimators. Moreover, the bottlenecks make successive samples dependent, leading to highly variable estimators. The results in Rohe (2015) show that if the strength of this bottleneck crosses a critical threshold, then the variance of the standard estimator decays slower than $1/n$, where n is the sample size.

To diminish referral bottlenecks, this dissertation proposes an adjustment to the current RDS implementation. Instead of asking participants to refer anyone from the target population, this dissertation proposes three basic types of "anti-cluster referral requests," which are described in Figure 3.1. These referral requests diminish referral bottlenecks by producing triples of participants that do not form a triangle in the social network.

As compared to alternative methods, anti-cluster requests are more successful in diminishing referral bottlenecks for three reasons. First, this approach preserves privacy. Many previous approaches have required participants to list all of their friends in the population. However, in sensitive populations, this is not allowed by institutional review boards. Second, anti-cluster requests do not require *a priori* knowledge about the nature of the bottleneck. For example, the most salient bottleneck could form on race, gender, neighborhood, or something else. If researchers knew which of these was most restricting the sampling process, then perhaps specific requests could be formed. However, in many populations, the bottlenecks are not known in advance. The final advantage is that the proposed adjustment is mathematically tractable; under certain assumptions, anti-cluster requests can form a reversible Markov chain.

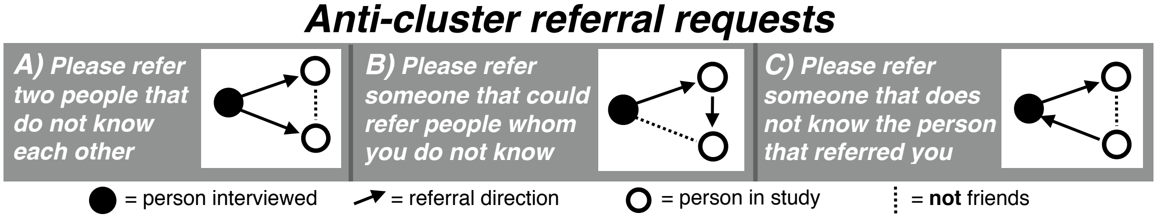

Anti-cluster referral requests
A) Please refer two people that do not know each other
B) Please refer someone that could refer people whom you do not know
C) Please refer someone that does not know the person that referred you
● = person interviewed
= referral direction
○ = person in study
= not friends

Figure 3.1: An illustration of the three anti-cluster referral requests considered. The referral requests for anti-cluster sampling are privacy preserving because they do not require participants to list all of their friends. Moreover, these requests do not require any knowledge about the community structures in the social network.

The remainder of the chapter is organized as follows. Section 3.2 describes designed RDS and presents our proposed design, anti-cluster RDS (AC-RDS). Section 3.3 sets the notation and provides the mathematical preliminaries. Section 3.4 gives our theoretical results, distinguishing between population sample graph results. Section 3.5 contains simulation experiments, which compare the performance of AC-RDS with standard RDS. We summarize the chapter and offer a discussion in Section 3.6. All of the proofs are provided in Appendix B.

3.2 Novel Sampling Designs

When preparing to sample a target population with RDS, some aspects can be controlled by researchers (e.g. how many referral coupons to give each participant) and others cannot. In particular, the social network is beyond the control of researchers. Community structures are an intrinsic part of social networks (Girvan and Newman, 2002) which, in RDS, lead to referral bottlenecks. To minimize these bottlenecks, RDS can be altered to make some referrals more or less likely. This is the essence of novel sampling designs for respondent-driven sampling.

In standard RDS, researchers ask each participant to refer their contacts in the target population, with no further instructions. In order to make statistical inferences, it is necessary to presume that participants refer a random set of their friends. The most common assumption is that each friend is equally likely to be referred. To test this assumption, suppose that the population of interest is divided into two communities, EAST and WEST. Furthermore, assume that people form most of their friendships within their own community. Under this simple model, referrals between communities are unlikely, creating a bottleneck.

As a thought experiment, suppose that these communities were known before performing the sample. The researchers could then request referrals from specific groups (e.g. flip a coin, if heads request WEST and if tails request EAST). This does not change the underlying social network, but it does change the probability of certain referrals. If participants followed this request, the referral bottleneck between EAST and WEST would be diminished. If 90% of a participant's friends belonged to the same community as the participant, then the standard approach would obtain a cross-community referral only 10% of the time. However, with the coin flip implementation, such a referral happens 50% of the time.

Mouw and Verdery (2012) propose an alternative technique which will be referred to as MW sampling. In MW sampling, researchers construct a sampling frame by asking RDS participants to name all of their friends in the target population. This list is combined with the friend lists from previous participants to form the sampling frame for selecting the next individual. Notice that this data can be used to construct partial information about the underlying social network. Based upon this partial social network information, MW sampling computes sampling probabilities for the individuals in the partial network who have not yet been sampled. The next individual who is sampled is then asked to list their friends in the target population. This process iterates on every additional sample. In computational experiments, Mouw and Verdery (2012) report a decrease in the design effect of this novel approach.

These two extensions of RDS (i.e. flipping a coin and MW) are both forms of Designed RDS; through novel implementations they adjust the probability of certain referrals, thereby diminishing the referral bottlenecks. Unfortunately, these two approaches have practical difficulties that prevent applications to hidden and/or marginalized populations. The coin flipping example requires prior information about the social network, which may unattainable given the hidden nature of the target population. The MW approach requires participants to reveal a friend list; for marginalized populations this is potentially unethical because it asks a participant to reveal sensitive information about an individual that has not provided consent. For example, if the target population is people who inject drugs, asking a participant to reveal friends in this population could be perceived as "snitching." Disclosing this type of information is often prevented by institutional review boards. Traditional forms of RDS do not require participants to reveal their friends to researchers (without consent from the friends).

Anti-cluster RDS is a type of Designed RDS that complements and builds upon both of these approaches. The implementation of anti-cluster RDS does not require *a priori* information on the communities in the social network, nor does it require that participants reveal sensitive information about individuals who have not consented. Anti-cluster sampling is designed to place larger referral probabilities on edges that belong to fewer triangles. There are at least two ways to consider why this strategy circumvents bottlenecks.

1. Many empirical networks share three properties. First, the number of edges is proportional to the number of nodes (i.e. the network is globally sparse). Second, friends of friends are likely to be friends (i.e. the network is locally dense). Third, shortest path lengths are small (i.e. the network has a small diameter); this is also known as the small-world phenomenon. Watts and Strogatz (1998) shows how a network can satisfy all three properties; take a deterministic graph that satisfies the first two features (e.g. a triangular tessellation), then select a few edges at random and randomly re-wire this edge to a randomly chosen node. Notice that these "random edges" are unlikely to be contained in a triangle. So, anti-cluster RDS is likely to make referrals along these edges, which connect to a node that is chosen uniformly

at random.

2. The Markov chain is a standard model for RDS. It presumes that people make referrals by selecting uniformly from the set of friends. A similar assumption could be made about anti-cluster referrals; the referral is drawn uniformly from the set of referrals that satisfy the anti-cluster request. If the Markov transition matrix for anti-cluster sampling can be shown to have a larger spectral gap than the Markov transition matrix for the simple random walk, then this suggests that anti-cluster sampling will obtain a more representative sample.

Here, we pursue the second approach.

3.3 Preliminaries

Framework

This work models the referral process as a Markov chain indexed by a tree (Benjamini and Peres, 1994). Markov chain indexed by a tree is a variant of branching Markov chains in which a fixed deterministic tree indicates branching. This model is a straightforward combination of the Markov models developed in the previous literature (e.g. Heckathorn (1997); Salganik and Heckathorn (2004); Volz and Heckathorn (2008) and Goel and Salganik (2009)) which allows multiple participation of an individual in the target population. This necessitates the following four mathematical pieces: an underlying social graph, a node feature which is measured on each sampled node (e.g. HIV status), a Markov transition matrix on this graph, and a referral tree to index the Markov process. Figure 3.2 gives a graphical depiction of this process.

The social network. RDS is based on the assumption that there are social ties among the individuals in the population and consequently there exists a social network that connects them. We denote the underling social network by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \ldots, N\}$ is a set containing the individuals in

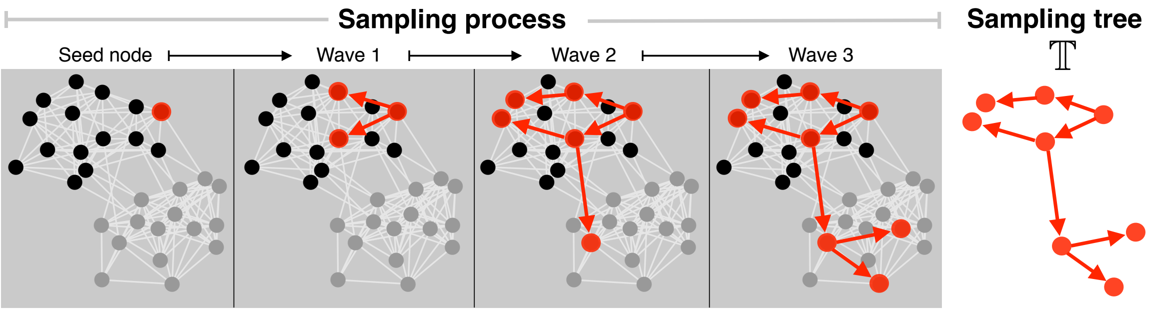


Figure 3.2: A graphical depiction of the referral process, which is modeled as a Markov chain indexed by a tree. This figure gives an example of a social network $\mathcal{G}$ and a referral tree $\mathbb{T}$.

the target population and $\mathcal{E} = \{(u, v) : \text{ u and v are friends}\}$ is a set containing the social ties. We reach a subset of $\mathcal{V}$ by starting from some "seed" node and then tracing edges in $\mathcal{E}$. Define

$$A(u, v) = \begin{cases} 1 & \text{if } (u, v) \in \mathcal{E}; \\ 0 & \text{o.w.} \end{cases} \tag{3.1}$$

and $\deg(u) = \sum_v A(u, v)$.

Node features. After reaching an individual $u \in \mathcal{V}$, we can measure their status $y(u)$, where $y : \mathcal{V} \to \mathbb{R}$ is some node feature. For instance, $y(u)$ could be a binary variable which is one if node u is HIV+ and zero otherwise. The aim of RDS is to estimate the population average of y over all nodes,

$$\mu = \frac{1}{N} \sum_{u \in \mathcal{V}} y(u).$$

Markov chain. By tracing the edges (social ties) in the network, RDS sampling procedure collects dependent samples from the nodes to estimate the population mean. If we assume that every participant recruits one individual and participants can be recruited multiple times then the sampling procedure can be modeled as a Markov chain (Gile and Handcock, 2010). These simplifying assumptions allow us to use the well-developed mathematical theory of Markov chains to gain insight into the behavior of the sampling procedure and understand the impact of the social network structure on the estimates.

Let $(X_i)_{i=1}^n$ be an irreducible Markov chain with the finite state space $\mathcal{V}$ of size N and transition matrix $P \in \mathbb{R}^{N \times N}$; for $u, v \in \mathcal{V}$ and for all $i \in 1, \dots, n-1$,

$$P(u,v) = \Pr(X_{i+1} = v | X_i = u).$$

Define P_A as the Markov transition matrix of the simple random walk,

$$P(u,v) = \frac{A(u,v)}{\deg(u)}.$$

The standard Markov model for RDS presumes that X_i is a simple random walk.

Novel designs. Designed RDS is a way of assigning differing weights to the edges. Define the $W : \mathcal{E} \to R_+$ as a weighting function on the edges $(u,v) \in \mathcal{E}$. If $(u,v) \in \mathcal{E}$ and $W(u,v) > 0$, then u can recruit v. For simplicity, define $W(u,v) = 0$ if $(u,v) \notin \mathcal{E}$. Then, W can be considered as a matrix. Define diagonal matrix T to contain the row sums, $T_{uu} = \sum_v W(u,v)$.

Through novel implementations, Designed RDS alters the edge weights. After weighting the edges, the Markov transition matrix becomes

$$P_W = T^{-1}W. \tag{3.2}$$

If Designed RDS increases an edge weight, it makes the edge more likely to be traversed.

We restrict the analysis to symmetric weighting matrices. Because of this restriction, P_W is reversible and has a stationary distribution $\pi : \mathcal{V} \to \mathbb{R}_+$ that is easily computable,

$$\pi(u) = \frac{T_{uu}}{\sum_v T_{vv}}.$$

Throughout, it will be presumed that X_0 is initialized with π. A more thorough treatment of Markov chains and their stationary distribution can be found in Levin et al. (2009).

Referral tree. In the Markov chain model, participant X_i refers participant X_{i+1}. This presumes that each participant refers exactly one individual. In practice, RDS participants usually refer between zero and three future participants. To allow for this heterogeneity, it is necessary to index the Markov process with a tree, not a chain. Let $\mathbb{T}$ denote a rooted tree with n nodes. See Figure 3.2 for a graphical depiction.

To simplify notation, $\sigma \in \mathbb{T}$ is used synonymously with σ belonging to the node set of $\mathbb{T}$. For any node $\sigma \in \mathbb{T}$ and $\sigma \neq \mathit{root}(\mathbb{T})$, denote $\mathit{parent}(\sigma) \in \mathbb{T}$ as the parent node of σ. The Markov process indexed by $\mathbb{T}$ is a set of random variables $\{X_\sigma \in V : \sigma \in \mathbb{T}\}$ such that $X_{\mathit{root}(\mathbb{T})}$ is initialized from π and

$$\Pr(X_\sigma = v | X_{\mathit{parent}(\sigma)} = u) = P(u, v), \text{ for } u, v \in \mathcal{V}.$$

The distribution of X_σ is completely determined by the state of $X_{\mathit{parent}(\sigma)}$. Up to this, everything is independent. Benjamini and Peres (1994) called this process a $(\mathbb{T}, P)$-*walk on* $\mathcal{G}$. In the social network $\mathcal{G}$, an edge represents friendship. In the referral tree, a directed edge (τ, σ) represents that random individual $X_\tau \in \mathcal{V}$ refers random individual $X_\sigma \in \mathcal{V}$ in the $(\mathbb{T}, P)$-*walk on* $\mathcal{G}$.

Statistical estimation. For any function on the nodes of the graph $y : V \to \mathbb{R}$, denote

$$\mu_{\pi,y} := E_\pi y := \sum_{u \in \mathcal{V}} y(u)\pi(u) \text{ and } \mu_y := Ey := \frac{1}{N} \sum_{u \in \mathcal{V}} y(u),$$

where $N := |\mathcal{V}|$ is the number of nodes in the social network. By assumption, $X_0 \sim \pi$. So, $X_\tau \sim \pi$ and the sample mean $1/n \sum_{\tau \in \mathbb{T}} y(X_\tau)$ estimates $\mu_{\pi,y}$, the population mean computed under the stationary distribution π. Thus, it is not a consistent estimate of the population mean, μ_y. In order to estimate μ_y, one can use inverse probability weighting (IPW), using the stationary distribution π. It can be shown that

$$\hat{\mu}_{IPW} = \frac{1}{n} \sum_{\tau \in \mathbb{T}} \frac{1}{N} \cdot \frac{y(X_\tau)}{\pi(X_\tau)}$$

is an unbiased and consistent estimator of μ_y. Typically, N is unknown. The Hajek estimator circumvents this problem while remaining asymptotically unbiased,

$$\frac{1}{\sum_{\tau \in \mathbb{T}} 1/\pi(X_\tau)} \sum_{\tau \in \mathbb{T}} \frac{y(X_\tau)}{\pi(X_\tau)}. \tag{3.3}$$

The typical "simple random walk" assumption in the RDS literature is that participants select uniformly from their contacts. This corresponds to $T_{uu} = \deg(u)$, making $\pi(u) \propto \deg(u)$ something that can be asked of participants. By these assumptions, Equation (3.3) reduces to the RDS II estimator (Heckathorn, 2007):

$$\hat{\mu}_y = \frac{1}{\sum_{\tau \in \mathbb{T}} 1/\deg(X_\tau)} \sum_{\tau \in \mathbb{T}} \frac{y(X_\tau)}{\deg(X_\tau)}.$$

The Variance and the Design Effect of RDS

Many empirical and social networks display community structures (Girvan and Newman, 2002). This can lead to referral bottlenecks in the Markov chain. These bottlenecks describe the fact that respondents are likely to refer people in their own community who have similar characteristics. This section specifies how bottlenecks makes successive samples dependent, increasing the variance of $\hat{\mu}_y$ and the design effect of RDS; the spectral properties of the Markov transition matrix reveal the strength of these bottlenecks and it controls the variance of estimators like $\hat{\mu}_{IPW}$. These results motivate the main results which show that anti-cluster sampling improves the relevant spectral properties of the Markov transition matrix under a certain class of Stochastic Blockmodels. As a result, anti-cluster sampling can decrease the variance of estimators like $\hat{\mu}_{IPW}$.

Let $\lambda_2(P_A)$ be the second largest eigenvalue of the Markov transition matrix for the simple random walk. The Cheeger bound demonstrates that the spectral properties of P_A can measure the strength of these communities. See Chung (1997) (Chapter 2) and Levin et al. (2009) page 215 for more details. This relationship between communities in G and the spectral properties of P_A is exploited in the literature on spectral clustering. In that literature, $\mathcal{G}$ is observed and the spectral clustering algorithm uses the leading eigenvectors of P_A to partition $\mathcal{V}$ into communities (Von Luxburg, 2007).

Intuitively, if there are strong communities in $\mathcal{G}$ and the node features y are relatively homogeneous within communities, then successive samples X_i and X_{i+t} will likely belong to the same community and have similar values $y(X_i)$ and $y(X_{i+t})$. This makes the samples highly dependent; the auto-covariance $\text{Cov}(y(X_i), y(X_{i+t}))$ will decay slowly. The next lemma decomposes the auto-covariance in the eigenbasis of the Markov transition matrix. This proposition shows that the auto-covariance decays like λ_2^t.

The following results apply to any reversible Markov matrix with $|\lambda_2| < 1$. In particular, they apply to both P_A (RDS) and P_W (AC-RDS). The assumption $|\lambda_2| < 1$ is equivalent to saying that the Markov chain is connected and aperiodic.

Proposition 3.1. *Let $X_i, i = 1, 2, \dots$ be a Markov chain with reversible transition matrix P. Suppose that X_1 is initialized with π, the stationary distribution of P. For $j = 1, 2, \dots, N$, let (f_j, λ_j) be the eigenpairs of P, ordered so that $|\lambda_i| \geqslant |\lambda_{i+1}|$. Because P is reversible, f_j and λ_j are real valued and the f_j are orthonormal with respect to the inner product $\langle f_\ell, f_j \rangle_\pi = \sum_{i \in \mathcal{V}} f_\ell(i) f_j(i) \pi(i)$. If $|\lambda_2| < 1$, then*

$$\text{Cov}(y(X_i), y(X_{i+t})) = \sum_{j=2}^{|\mathcal{V}|} \langle y, f_j \rangle_\pi^2 \lambda_j^t.$$

In previous research, Bassetti et al. (2006) and Verdery et al. (2015) used a similar expression to compute the variance.

Design effect. The design effect of RDS is a measure of the quality of the sampling mechanism. It is defined as

$$\text{DE}(\hat{\mu}) = \frac{\text{Var}_{RDS}(\hat{\mu})}{\text{Var}_\pi(\hat{\mu})},$$

where the subscript RDS denotes that the sample was collected with a $(\mathbb{T}, P)$-*walk on* $\mathcal{G}$ and the subscript π denotes that that samples were drawn independently from the stationary distribution.

The results in Rohe (2015) show that the design effect has a critical threshold that depends on λ_2 and m, the branching rate of the tree. If $\lambda_2 > m^{-1/2}$, then $\mathrm{DE}(\hat{\mu}_{IPW})$ grows with n. Moreover, under certain additional assumptions, design effect has a critical threshold,

$$\mathrm{DE}(\hat{\mu}_{IPW}) \asymp \begin{cases} c & \text{if } \lambda_2 \leqslant m^{-1/2} \\ n^{1-\alpha} & \text{if } \lambda_2 > m^{-1/2}, \end{cases} \tag{3.4}$$

where c is some constant, $\alpha = \log_m \lambda_2^{-2}$, and $\asymp$ is equality up to $(\log n)^2$ terms. This shows that if the referral bottleneck is too strong (i.e. $\lambda_2 > m^{-1/2}$), then the design effect grows polynomially with the sample size. Here, the standard error of $\hat{\mu}_{IPW}$ does not decay like $n^{-1/2}$, rather it converges at the slower rate of $n^{\log_m \lambda_2}$.

By redesigning the Markov transition matrix via novel referral requests, it is shown that λ_2 is partially malleable. We can diminish referral bottlenecks and λ_2.

Anti-Cluster Random Walk; Constructing the Weights W

This subsection describes a Markov model for AC-RDS. Section 3.4 then studies the spectral properties of the resulting AC-RDS Markov transition matrix. To describe the model we need the following notation. Let $\cdot$ denote element-wise matrix multiplication and let $J_{q\times q}$ denote a $q \times q$ matrix containing all ones. Finally, define the overbar operator for a $q \times q$ matrix B as $\bar{B} := J_{q\times q} - B$, so that $\bar{A} = J_{N\times N} - A$.

The assumption that RDS can be modeled as a simple random walk is a common working assumption in the RDS literature (Gile et al., 2015). It assumes that a participant selects a single referral by choosing a friend uniformly at random. To understand the theoretical properties of AC-RDS, we extend this idea of “uniformly at random." We model a participant’s response to the request “please refer two people that don’t know each other” in the following way: if i is friends with j, then the probability that i refers j is proportional to the number of friends of i that are not friends with j. This is equivalent to the participant making a list of all friend pairs (j, ℓ) for which the friends in a pair do not know each other $(A_{j,\ell} = 0)$. Then, the participant selects one pair from the list uniformly at random. Note that the

Markov transition matrix only allows a single referral. To maintain the Markov property, the participant is then instructed to select one person from the chosen pair with equal probability.

This model creates a Markov transition matrix which can be expressed with matrix notation. Under the model, if i has one coupon, then the probability that i refers j is proportional to the $(i, j)^{th}$ element of the matrix $(A\bar{A}) \cdot A$. To see this, note that the $(i, j)^{th}$ element of $A\bar{A}$ is the number of nodes ℓ that are friends with i but not friends with j, that is

$$[A\bar{A}]_{ij} = \sum_{\ell} A_{i\ell}(1 - A_{j\ell}).$$

Then, the element-wise multiplication ensures that i is friends with j, yielding the weight matrix $(A\bar{A}) \cdot A$.

Note that the weight matrix $(A\bar{A}) \cdot A$ is not symmetric and, thus, does not lead to a reversible Markov chain. However, we can use a second referral request to augment the first request to ensure reversibility. To this end, model the referral request "Please refer someone that knows many people that you do not know" as follows: if i is friends with j, then the probability that i refers j is proportional to the number of people that j knows that i does not know. In a similar fashion as above, this request produces the weight matrix $(\bar{A}A) \cdot A$.

To implement AC-RDS, choose between $(A\bar{A}) \cdot A$ and $(\bar{A}A) \cdot A$ with equal probability by flipping a coin. Consider the matrix $\tilde{W}$ given by

$$\tilde{W} = (A\bar{A} + \bar{A}A) \cdot A. \tag{3.5}$$

The $(i, j)^{th}$ element of $\tilde{W}$ is proportional to the probability that i refers j in the process described above. By design, $\tilde{W}$ is symmetric, making making $P_{\tilde{W}}$ a reversible Markov transition matrix.

These ideas for connecting implementation instructions for AC-RDS with the Markov model are summarized in Table 3.1. The next section studies the spectral properties of $P_{\tilde{W}}$ under a statistical model for $\mathcal{G}$.

Implementation instructions compared to the Markov model

Flip a coin	If heads,	If tails,
Implementation Instructions	Ask “Please refer two of your contacts in the population that don’t know each other.”	Ask “Please refer two of your contacts in the population that have many contacts you don’t know.”
Markov model, starting from node i	List all pairs of nodes (i,k) such that, $(i,j) \in \mathcal{E}$, $(i,k) \in \mathcal{E}$, and $(k,j) \notin \mathcal{E}$. Then choose a pair (j,k) uniformly and refer j or k uniformly at random.	List all pairs of nodes (j,k) such that $(i,j) \in \mathcal{E}$ and $(i,j) \notin \mathcal{E}$. Then choose a node pair (i,k) uniformly at random from the lit and refer j.

Table 3.1: The correspondence between AC-RDS implementation instructions and the Markov model for the referral process. Referral requests A and B from Figure 3.1 correspond to the left and right columns of this table. The first row of this table is the verbal request given to a participant. The second row of this table describes the Markov model for this request, as described in Section 3.3.

Finally, we note that the transition matrix $P_{\tilde{W}}$ does not use referral request C in Figure 3.1, “Please refer someone that does not know the person that referred you." Such a request cannot form a Markov chain on the nodes in the network because it depends on the previous participant. This non-Markovian behavior should not preclude the use of request C in practice; however, it does make establishing theoretical results for request C more difficult. Here, we focus on requests A and B and their Markov transition matrix $P_{\tilde{W}}$.

3.4 Theoretical Results

To study the spectral properties of $P_{\tilde{W}}$ under a statistical model for the underlying social network, we break the analysis into the “population results” and the “sampling results.” The “sampling” referred to in this section introduces an additional layer of randomness. In this section, “sampling” refers to the randomness in the generation of the underlying social network. To refer to the randomness of the Markov chain, this section will refer to “anti-cluster sampling,” “Markov sampling,” or “respondent-driven sampling.”

The “population results" in this section correspond to using the (weighted) adjacency matrix $\mathcal{A} = EA$, where the expectation is with respect to the statistical model for generating the network. Then, define

$$\tilde{\mathcal{W}} = (\mathcal{A}\bar{\mathcal{A}} + \bar{\mathcal{A}}\mathcal{A}) \cdot \mathcal{A}. \tag{3.6}$$

Define the Markov transition matrices $P_{\tilde{\mathcal{W}}}$ and $P_{\mathcal{A}}$ as in Equation (3.2). In these definitions, $P_{\mathcal{A}}$ corresponds to the population matrix for the standard simple random walk and $P_{\tilde{\mathcal{W}}}$ corresponds to the population matrix for AC-RDS.

The population results will show that under various statistical models for the underlying social network, the second eigenvalue of $P_{\tilde{\mathcal{W}}}$ is less than the second eigenvalue of $P_{\mathcal{A}}$. To extend these population results to a network which is sampled from the model, the sampling results use concentration of measure to show that A and $\tilde{W}$ are close to $\mathcal{A}$ and $\tilde{\mathcal{W}}$ under operator norm, respectively. Then, perturbation theorems show that the eigenvalues of $P_{\mathcal{A}}$ and $P_{\tilde{W}}$ are close to the eigenvalues of P_A and $P_{\tilde{\mathcal{W}}}$ respectively. Theorem 3.9 combines these results with Proposition 3.1 to show that AC-RDS reduces the covariance between Markov samples.

Population Graph Results

Anti-cluster sampling is motivated by the need to readily escape communities in a social network. The Stochastic Blockmodel (SBM) is a standard and popular model that parameterizes communities in the social network (Holland et al., 1983). For this reason, the analyses below use the SBM to study anti-cluster sampling.

Definition 3.2. *To sample a network from the* ***Stochastic Blockmodel****, assign each node* $i \in \{1, 2, \ldots, N\}$ *to a class* $z(i) \in \{1, 2, \ldots, K\}$*, where the* $z(i)$ *are independently generated from Multinomial*$(\pi_1, \ldots, \pi_K)$*. Conditionally on* z*, edges are independent and the probability of an edge between nodes* i *and* j *is* $B_{z(i)z(j)}$*, for some matrix* $B \in [0,1]^{K \times K}$.

The results below condition on the partition z. Conditional on this partition, $E[A|z]$ has a convenient block structure. Define the partition matrix $Z \in \{0,1\}^{N \times K}$

such that $Z_{ij} = 1$ if $z(i) = j$, otherwise $Z_{ij} = 0$. Define $\mathcal{A} = E[A|z]$ and note that

$$\mathcal{A} = ZBZ^T.$$

Let $\bar{\mathcal{A}} := J_{N\times N} - \mathcal{A}$. Define the population weighting matrix as in Equation (3.6). The following lemma shows that $\tilde{\mathcal{W}}$ retains the block structure of $\mathcal{A}$.

Lemma 3.3. *Define* $\bar{B} := J_{K\times K} - B$ *and* $\Pi \in \mathbb{R}^{K\times K}$ *as a diagonal matrix with* Π_{ii} *equal to the expected number of nodes in the* i*th block. Then,* $\tilde{\mathcal{W}} = (\mathcal{A}\bar{\mathcal{A}} + \bar{\mathcal{A}}\mathcal{A}) \cdot \mathcal{A}$ *can be expressed as*

$$\tilde{\mathcal{W}} = Z\left((B\Pi\bar{B} + \bar{B}\Pi B) \cdot B\right) Z^T.$$

The following lemma shows that under a certain class of Stochastic Blockmodels, anti-cluster sampling decreases the probability of an in-block referral.

Lemma 3.4. *For* $0 < r < p + r < 1$*, let* $B = pI + rJ_{K\times K}$*. If* $\Pi_{jj} r < \Pi_{ii}(p + r)$ *for all* $i \neq j$*, then for any two nodes* u *and* v *with* $z(u) = z(v)$*,*

$$P_{\tilde{\mathcal{W}}}(u, v) < P_{\mathcal{A}}(u, v).$$

Note that the assumption $\Pi_{jj} r < \Pi_{ii}(p + r)$ is implied by the first assumption, $0 < r < p + r < 1$, when every block has an equal population. The next proposition uses Lemma 3.4 to show that anti-cluster sampling reduces the second eigenvalue of the population Markov transition matrix.

Proposition 3.5 (Spectral gap of the population graph)**.** *Under the SBM with* K *blocks, let* $B = pI + rJ_{K\times K}$ *for* $0 < r < p + r < 1$*. If the* K *blocks have equal size, then*

$$0 < \lambda_2(P_{\tilde{\mathcal{W}}}) + \epsilon < \lambda_2(P_{\mathcal{A}}) < 1, \tag{3.7}$$

where $\epsilon > 0$ *depends on* $K, p,$ *and* r*, but is independent of the number of nodes in the graph* N*. Specifically,* $\lambda_2(P_{\mathcal{A}}) = 1/(R + 1)$*, where* $R = Kr/p$*. In the asymptotic setting where* K

grows and r *shrinks, while* p *and* R *stay fixed,*

$$\lambda_2(P_{\tilde{\mathcal{W}}}) \to \frac{1}{cR+1}, \quad \textit{with} \quad c = \frac{R+1}{R+1-p}. \tag{3.8}$$

For any single node, note that R is roughly the expected number of out-of-block edges divided by the expected number of in-block edges. To see this, multiply the numerator and denominator of Kr/p by the block population N/K. As such, it is approximately the odds that a random walker will change blocks. When R is large, the Markov chain mixes quickly and $\lambda_2(P_{\mathcal{A}})$ is small to reflect that.

AC-RDS is most useful in social networks with tight communities, where the walk is slow to mix; this corresponds to a larger value of p and a smaller value of R. In this setting, c in Equation (3.8) is large, thus making $\lambda_2(P_{\tilde{\mathcal{W}}})$ much smaller than $\lambda_2(P_{\mathcal{A}})$. In particular, if p is close to one, then $c \approx 1 + R^{-1}$ becomes very large for small values of R. Notice that the second part of Proposition 3.5 makes no assumption on N, the number of nodes in the network.

The next proposition shows that anti-cluster sampling continues to perform well, even when the community structure is exceedingly strong and standard approaches will fail to mix well. Here, the reduction of λ_2 from anti-cluster sampling is dramatic.

Proposition 3.6. *Under the SBM with* 2 *blocks of equal sizes, let* $\epsilon > 0$ *and suppose that* $B_{ii} = (1-\epsilon)$ *and* $B_{ij} = \epsilon$ *for* $i \neq j$. *Then,*

$$\lim_{\epsilon \searrow 0} \lambda_2(P_{\mathcal{A}}) = 1$$

and

$$\lim_{\epsilon \searrow 0} \lambda_2(P_{\tilde{\mathcal{W}}}) = 1/3.$$

For any Markov transition matrix P, $\lambda_2(P) \leqslant 1$. The graph is disconnected if and only if $\lambda_2 = 1$; this is the most extreme form of a bottleneck. In the above proposition, if $\epsilon = 0$, then the sampled graph will contain two disconnected cliques, one for each block. Under this regime, both P_A and $P_{\tilde{W}}$ will have second eigenvalues equal to one. However, if ϵ converges to zero from above, then the above proposition shows that $\lambda_2(P_{\tilde{\mathcal{W}}})$ approaches $1/3$, while $\lambda_2(P_{\mathcal{A}})$ approaches 1.

Propositions 3.5 and 3.6 suppose balanced block sizes (i.e. an equal number of nodes). To study unbalanced cases, the necessary algebra quickly becomes uninterpretable. We explore the role of unbalanced block sizes with numerical experiments in Section 3.5.

Sample Graph Results

Theorem 3.7 gives conditions which ensure that the population eigenvalues (i.e. $\lambda_\ell(P_{\tilde{\mathcal{W}}})$) are close to the sampled eigenvalues (i.e $\lambda_\ell(P_{\tilde{W}})$). As such, the population results in the previous section appropriately represent the behavior of Markov sampling (both AC-RDS and RDS) on a network sampled from the Stochastic Blockmodel.

Theorem 3.7 (Concentration of the anti-cluster random walk)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a random graph with independent edges and $\mathcal{A} = \mathrm{E}A$ be the expected adjacency matrix. Let $\mathcal{D}_i := \sum_k \mathcal{A}_{ik}$, $F_{ij} := \sum_k \mathcal{A}_{ik}(1 - \mathcal{A}_{kj})$, and $G_{ij} := \sum_k (1 - \mathcal{A}_{ik})\mathcal{A}_{kj}$. Define $F_{\min} = \min_{i,j=1,\cdots,|\mathcal{V}|} F_{ij}$. If $F_{\min} = \omega(\ln N)$ and there exits a constant c_1 such that $F_{ij} + G_{ij} \geqslant c_1 \mathcal{D}_i$ for all $i, j = 1, \cdots, |\mathcal{V}|$, then with probability at least $1 - \epsilon$,*

$$\left\| T^{-\frac{1}{2}} \tilde{W} T^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}} \tilde{\mathcal{W}} \mathcal{T}^{-\frac{1}{2}} \right\|^2 \leqslant \frac{c_2 \ln \frac{10N}{\epsilon}}{F_{\min}},$$

where c_2 is a constant, $\|\cdot\|$ denotes the operator norm, T is a diagonal matrix with the row sums of $\tilde{W}$ down its diagonal, and similarly for $\mathcal{T}$ and $\tilde{\mathcal{W}}$. Moreover,

$$|\lambda_\ell(P_{\tilde{W}}) - \lambda_\ell(P_{\tilde{\mathcal{W}}})|^2 = \mathcal{O}\left(\frac{\ln \frac{10N}{\epsilon}}{F_{\min}} \right), \quad \textit{for all } \ell \in 2, \ldots, N.$$

A similar result for $|\lambda_\ell(P_A) - \lambda_\ell(P_{\mathcal{A}})|$ was shown in Chung and Radcliffe (2011).

Remark 3.8. *F_{ij} shows the number of friends of node i that are not in the friend list of node j. So $F_{\min} = \omega(\ln N)$ makes sure that the number of individuals that a node can* *refer under AC-RDS adjustment grow with a rate faster than* $\ln \mathrm{N}$. *Roughly speaking, it is similar to the sparsity condition required for the concentration results of any random graph with independent edges. Since $\mathcal{A}$ is a symmetric matrix,* $F_{ij} = G_{ji}$ *and consequently*

$$\min_{i,j=1,\cdots,|\mathcal{V}|} F_{ij} = \min_{i,j=1,\cdots,|\mathcal{V}|} G_{ij}.$$

In the first condition of the theorem, c_1 *makes sure that the ratio* $\frac{\mathcal{D}_i}{F_{ij}+G_{ij}}$ *stays bounded as it is needed to prove the results. These sampling results are sufficiently general to apply to all of the models studied in the previous section.*

Theorem 3.9 presents the asymptotic behavior of AC-RDS in reducing the correlation among samples collected from a random graph under a Stochastic Blockmodel. The theorem is an aggregation of all the previous results. The result is asymptotic in the size of the population, not in the size of the sample.

Theorem 3.9 (Dependency reduction property of AC-RDS)**.** *Let* $\mathcal{G}$ *be a random graph with* N *nodes sampled from a Stochastic Blockmodel with* $B = pI_{K\times K} + rJ_{K\times K}$ *for* $0 < r < p + r < c < 1$ *and an equal number of nodes in each of the* K *blocks. Let* $(X_i)_{i=1}^n$ *and* $(X_i^{ac})_{i=1}^n$ *be two Markov chains with transition matrix* P_A *and* $P_{\tilde{W}}$ *respectively.*

The parameters p, r *and* K *can change with* N. *If* $\ln(N)/(p\frac{N}{K} + rN) \to 0$, *then asymptotically almost surely, for all* i, $i + t \in \{1, \ldots, n\}$, *and* $t \neq 0$,

$$\mathrm{Cov}(y(X_i^{ac}), y(X_{i+t}^{ac})) < \mathrm{Cov}(y(X_i), y(X_{i+t})),$$

where $y : V \to \mathbb{R}$ *is any bounded node feature.*

Remark 3.10. *The quantity* $p\frac{N}{K} + rN$ *is* $\mathcal{D}_{\min}$, *the minimum expected degree. The condition* $\ln(N)/(p\frac{N}{K} + rN) \to 0$ *is needed to use Theorem 3.7. Note that* $F_{ij} + G_{ij} > 2cD_{\min}$ *for all* $i, j = 1, \cdots, |\mathcal{V}|$.

3.5 Simulation Study

We conduct three sets of simulations to compare the performance of AC-RDS with standard RDS. The first set investigates the impact of unequal block sizes on the results of Propositions 3.5 and 3.6. The second set investigates the impact of community structures and homophily using the stochastic block model. In the third set, we consider an empirical social network with unknown community structure. Finally, we consider the difference of simulating the sample with and without replacement from the underlying network.

The Role of Unequal Block Sizes

Propositions 3.5 and 3.6 assume that the blocks contain an equal number of nodes. Here we explore the role of unequal block sizes on those results. As a measure of unbalance, we use the ratio of the largest block size to the smallest block size. The results of the study are displayed in Figure 3.3. The horizontal axis in both panels gives this ratio of unbalance; when this value is large (farther to the right), the blocks are exceedingly unbalanced. The vertical axis controls the expected number of in-block versus out-of-block edges with a parameter ϵ. In the left panel, ϵ plays the dual role as in Proposition 3.6. In the right panel, ϵ does not control the in-block probabilities (i.e. the diagonal of B); here, the diagonal of B is set to .8 across all experiments.

The spectral gap is given by $1 - \lambda_2$, we are interested in exploring the ratio

$$\text{ratio of spectral gaps} = \frac{1 - \lambda_2(P_{\tilde{\mathcal{W}}})}{1 - \lambda_2(P_{\mathcal{A}})}. \tag{3.9}$$

For a range of unbalances and values of ϵ, Figure 3.3 plots the ratio of spectral gaps. In all simulations, this value is greater than one, indicating that anti-cluster sampling decreases λ_2 relative to the random walk model of RDS. For example, the contour at 5.3 represents the class of models such that anti-cluster sampling increases the spectral gap by over five-fold.

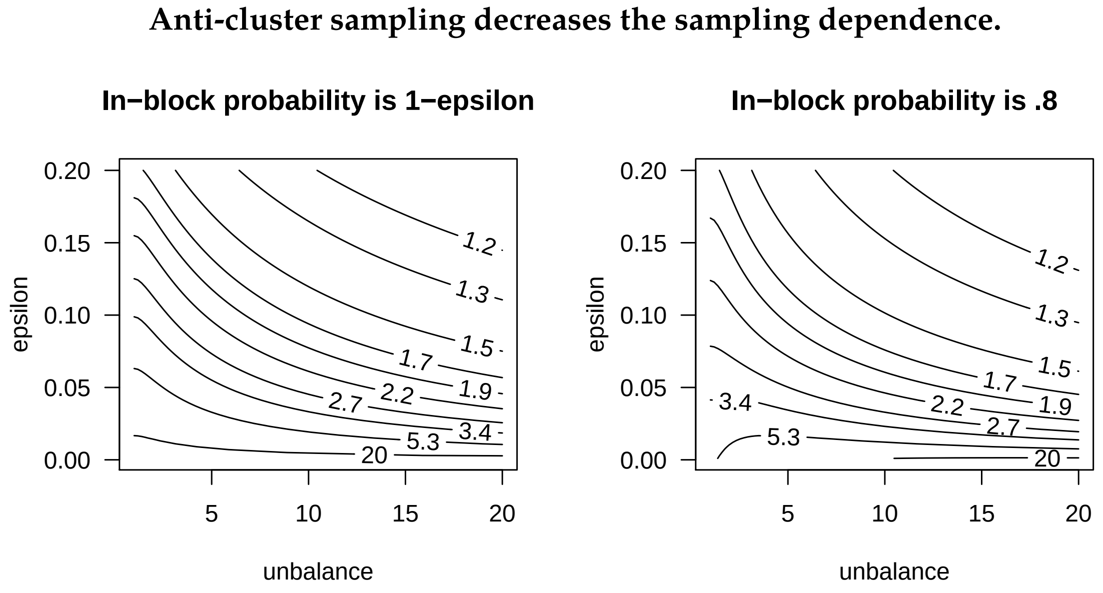


Figure 3.3: Results for the simulation study described in Section 3.5, which examines the impact of unequal block sizes on the results of Propositions 3.5 and 3.6. As a measure of unbalance (the x-axis), we use the ratio of the largest block size to the smallest block size. For a range of SBM parameterizations (as described in the text), these two panels display the ratio of spectral gaps as given in (3.9). All values are greater than one, indicating that anti-cluster sampling will increase the spectral gap, thus decreasing the dependence between adjacent samples in the RDS. The benefits of anti-cluster sampling are especially prominent when ϵ is small; this corresponds to a model setting in which there are drastically fewer edges between blocks.

Random Networks

Here we investigate the impact of community structures and homophily using the stochastic block model. We use a SBM with 2000 nodes and 50 communities of equal size to generate the underlying social network. To illustrate the impact of community structures, we vary the ratio of expected in-block and out-of-block node degree from 1/2 to 4. We fix the in-block probabilities to 0.9 and change the out-of-block probabilities from 0.036 to 0.0045. This ratio also controls the probability of generating an out-of-community referral. For example, with the ratio equal to one, the probability of an out-of-community referral is 1/2.

We simulate Markovian referral trees in which each participant refers exactly three members with replacement. The three referrals are samples from the neighbors of the participants. RDS uses uniform samples, whereas AC-RDS uses non-uniform samples based on the weights described in 3.5. To show the effect of the communities, we choose the binary node feature to be based on the community membership. The value is set to zero if the node belongs to communities 1 through 25, otherwise, the value is to set one. For both designs, we use the RDS II estimator to estimate the community proportion, where the inclusion probabilities are the stationary distribution of the simple random walk.

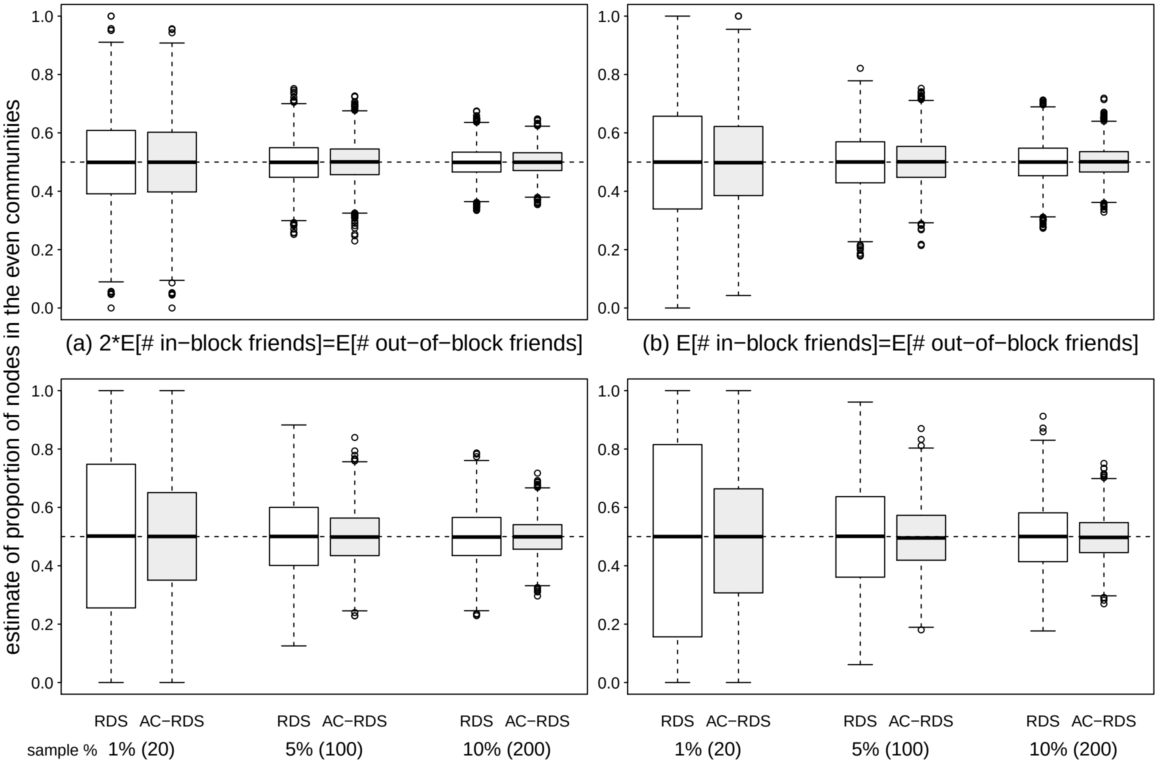
The datasets are simulated in the following way. First we generate a realization of an SBM and compute the stationary distribution of the simple random walk. We simulate the referral procedure of RDS and AC-RDS starting from a uniformly selected node and continuing until a certain number of samples are collected, either 1%, 5%, or 10% of the total nodes. We compute the RDS II estimates of the feature from samples collected by both procedures.

This study is based on 5000 simulated datasets. Figure 3.4 displays boxplots for the 5000 RDS II estimates of the proportion in different settings. Comparing RDS to AC-RDS, we see that AC-RDS collects more representative samples. Additionally, as we increase the degree of homophily, the performance of AC-RDS suffers less. In (a) and (b), the chance that participants make referrals outside of their community is relatively high, 2/3 and 1/2, respectively. In these cases, both designs perform similarly. However, in (c) and (d), where there is a smaller chance of cross-community referral, AC-RDS collects more representative samples by encouraging participants to leave their communities more often. This is exactly the intended outcome of AC-RDS. In fact, at the population level, this is the result proven in Lemma 3.4.

Add-Health Networks

This set of simulations is based on friendship networks from the National Longitudinal Survey of Adolescent Health dataset (available at http://www.cpc.unc.edu/

The estimates with samples from SBM collected by AC-RDS outperforms RDS.



(a) 2*E[# in−block friends]=E[# out−of−block friends]

(b) E[# in−block friends]=E[# out−of−block friends]

(c) E[# in−block friends]=2*E[# out−of−block friends]

(d) E[# in−block friends]=4*E[# out−of−block friends]

Figure 3.4: Simulation results for the random network study. Here we consider RDS sampling of random graphs drawn from a SBM with balanced communities. Under different settings, the figure compares RDS II estimates of the community proportion from samples collected by standard RDS and AC-RDS. The boxplots display the estimated proportions for the 5000 simulated datasets. The true value is 0.5.

addhealth), which we refer to as the Add-Health Study. In these networks, students are represented by nodes and an edge between two students indicates any type of friendship. To collect friendship data, the students were asked to list up to five friends of each gender, and whether they had interaction within a certain period of time. Here, we treat the friendship networks as undirected networks. That is, an edge connecting two students means that either student, not necessarily both, reported a friendship. For this study, we use the four largest networks in the dataset. Table 3.2 contains summary information for the largest connected component of these four networks. We use gender as the binary node feature and focus on estimating the proportion of males in the population.

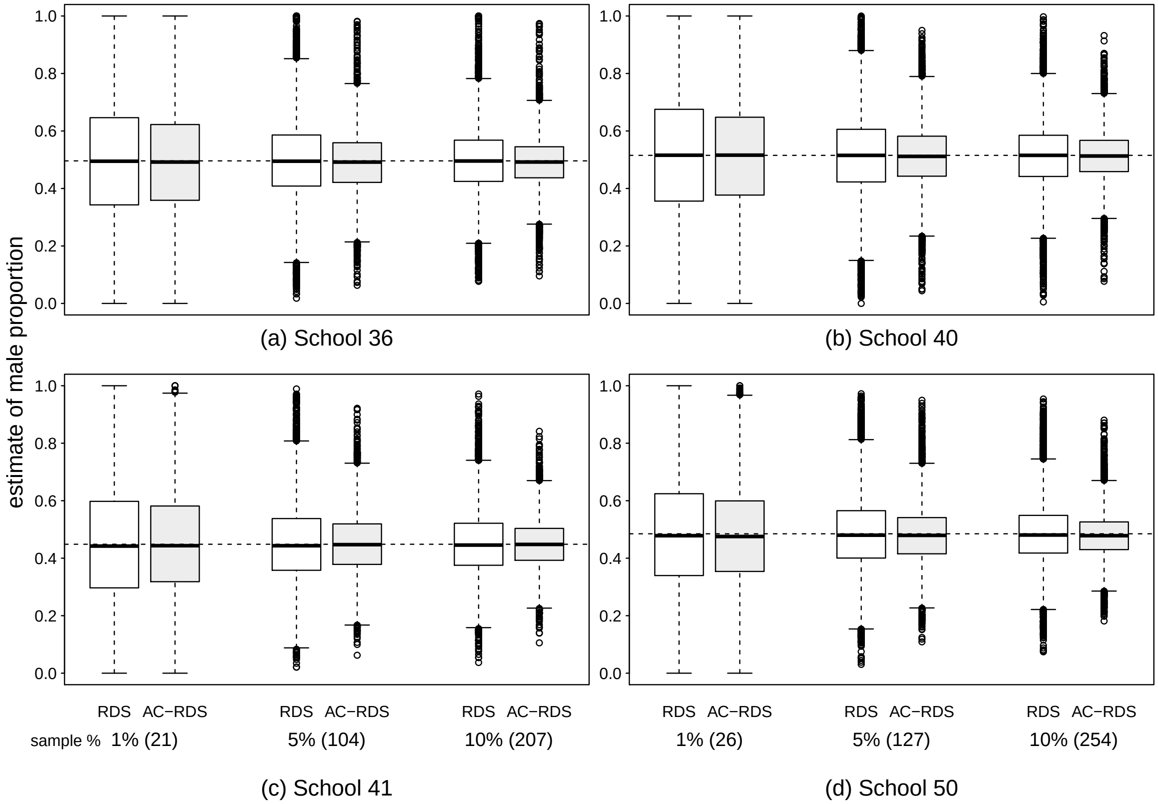
School id	# Nodes	# Edges	CC	covariance	covariance^{ac}
School 36	2152	7986	0.178	0.0260	0.0056
School 40	1996	8522	0.144	0.0265	0.0030
School 41	2064	8646	0.139	0.0243	0.0042
School 50	2539	10455	0.141	0.0276	0.0069

Table 3.2: Network characteristics for the four largest friendship networks in the Add-Health study. This table provides characteristics for the largest connected component of each network. An edge between student nodes indicates that either student reported a friendship. The clustering coefficient (CC) is the ratio of the number of triangles and connected triplets. The last two columns represent the covariance of the samples collected under RDS and AC-RDS, respectively.

We simulate the referral procedure of RDS and AC-RDS starting from a uniformly selected node and continuing until a certain number of samples are collected, either 1%, 5%, or 10% of the total nodes. In these simulations, each participant refers exactly three members with replacement. We compute the RDS II estimate of the male proportion using the node degree for the weights.

This study is based on $10,000$ simulated datasets. Figure 3.5 displays boxplots for the $10,000$ RDS II estimates of the male proportion under different settings.

The estimates with samples from Add-Health friendship networks collected by AC-RDS outperforms RDS.



(a) School 36

(b) School 40

(c) School 41

(d) School 50

Figure 3.5: Simulation results for the Add-Health study. Here we consider RDS sampling of Add-Health friendship networks. Under different settings, the figure compares RDS II estimates of the male proportion from samples collected by standard RDS and AC-RDS. The boxplots display the estimated proportions for the $10,000$ simulated datasets.

Comparing Sampling With and Without Replacement

We consider the impact on AC-RDS when simulating the sample with and without replacement from the underlying network. In the Random Networks simulation model, there is only a small difference between the two sampling settings. This is likely because the network is dense. In smaller networks, one expects there to be a greater difference between with and without replacement sampling. In fact, in the Add-Health simulation model, under a without replacement setting and a referral rate of one or two, the trees die quickly and often do not collect enough samples to attain 1% of the total nodes.

3.6 Discussion

By employing the social network to drive referrals, many RDS studies have successfully attracted large sets of participants from a variety of marginalized and hard-to-reach populations.

In the respondent-driven sampling, bottlenecks create dependencies between the samples; successive samples are much more likely to belong to the same community. Because of these dependencies, bottlenecks increase the variability of the resulting estimators. While researchers cannot alter the social network to diminish bottlenecks, researchers can use novel implementations of RDS to implicitly encourage participants to refer friends in different communities. In comparison to other such techniques in the literature, AC-RDS does not require participants to reveal sensitive information, nor does it require *a priori* knowledge on what forms the bottlenecks (e.g. race, gender, neighborhood, some combination of these factors, or some entirely different factors).

We call this approach anti-cluster RDS. This terminology stems from two distinct, but related, definitions of "clustering" in networks. First, the classical use of "clustering" in social networks is the clustering coefficient, a summary statistic of a network which describes the propensity of nodes to form triangles. This idea of "clustering" is a local measure. The second form of "clustering" is more global and is often used synonymously with community structure; the idea is that “clusters” of individuals form communities. Both of these types of clusters emerge due to homophily, the tendency of individuals to become friends with people who are similar. As such, homophily produces a local-global duality in “clustering.” AC-RDS requests are built upon local structures in the network (which of your friends are friends) and immediately access the global network patterns, which could be unknown to the researchers and/or participants.

Section 3.4 studies theoretical properties of AC-RDS. We first argue that AC-RDS can be approximated by a reversible Markovian process. Propositions 3.5 and 3.6 show that AC-RDS can decrease λ_2, the eigenvalues of the Markov transition matrix, on the population graph. Theorem 3.7 shows that these gains from Propositions 3.5 and 3.6 will continue to hold if the graph is sampled with independent edges. In addition, Theorem 3.9 shows that AC-RDS reduces the covariance of the samples in the referral tree under the Stochastic Blockmodel with equal block size.

Chapter 4

Future Work

Phylogenetic data analysis: The model selection methods that detect occurrence of evolutionary shifts lack rigorous theoretical studies to guarantee closeness of selected models to the optimum. The computational complexity of the problem even under the standard criteria is unknown. In addition, it is necessary to have a procedure to quantify the uncertainty in the estimated model. The procedure essentially needs a well-defined metric to measure the distance between various models and form the possible shift configuration space. It seems possible to adapt the metric space introduced by Billera, Holmes and Vogtmann (BHV) to accomplish that goal.

Tree-structured models for summarizing data: Hierarchical clustering (HC) type techniques output the same mathematical objects as phylogenetic trees, a variate of latent tree models. Dendrograms, the output of HC methods, provides a summary and insight into the geometry and global structure of the data. In contrast to flat clusters, dendrograms can reveal multiscale structure in datasets that the standard methods fail to capture. Furthermore, they are popular methods to display microarray data and study the co-expressions of the genes' time series under various conditions. Despite the popularity, there is a gap between the theory and practice of the HC methods. For example, their stability and consistency are not well studied. These properties can be studied under the mathematical framework of a stochastic process on an unobserved tree. Furthermore, various computational methods that are developed for phylogenetic trees such as metric spaces of the trees (e.g. BHV distance) and ultrametric spaces can be used to study dendrograms.

Topologically inspired methods: Trees (as in graph theory) are special one-dimensional topological structures. Their higher dimensional generalizations are called "simplicial complexes" which have a fully developed theory and ubiquitous applications in mathematical models of spaces and point clouds. There are powerful algorithms that already use the geometry of simplicial complexes in data science. In addition to the rich geometric dependencies in modeling data as higher-dimensional topological structures, it seems plausible to adapt the theory of operators and their spectrum that are quite useful for study of familiar statistical properties such as clustering, hierarchical clustering, and nonlinear dependencies among data features.

Appendix A

Appendix for Chapter 2

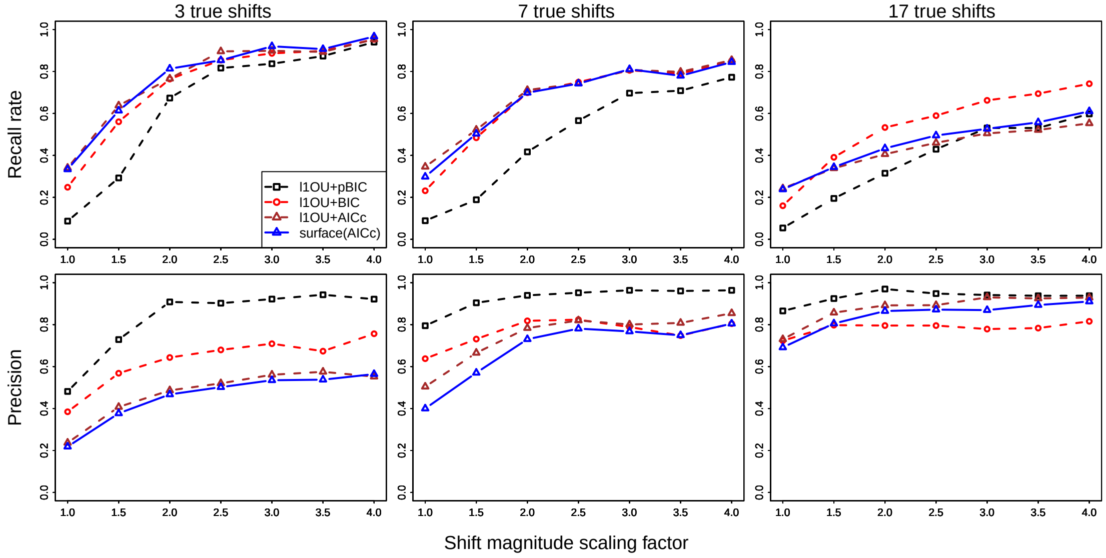


Figure A.1: Recall rate (first row) and precision (second row) of first pPC from 4 traits. The magnitudes of all shifts were increased by the same scaling factor.

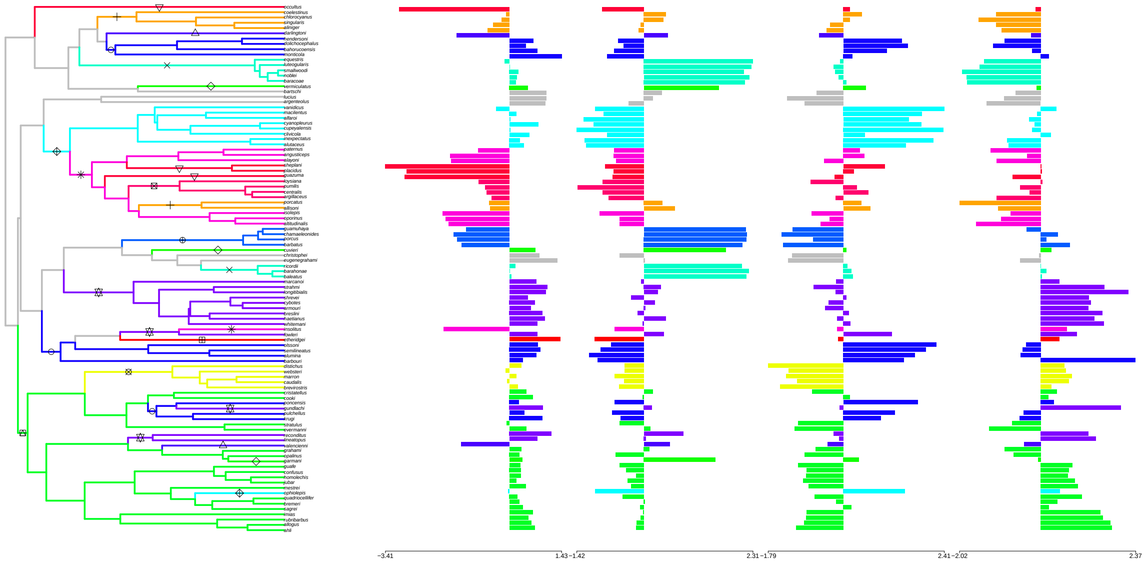


Figure A.2: Left: phylogenetic placement of 28 shifts in the selection optimum in Anolis lizards, as detected by surface (running time of the forward phase: 2 hours and 12 minutes). Each shift is annotated with a symbol on the corresponding edge. The same symbol is used for shifts inferred to have converged to the same selection optimum. Right: bar graphs showing the 4 traits combined for analysis. The tree and data are from Mahler et al. (2013), available in their supplementary material at `www.sciencemag.org/content/suppl/2013/07/18/341.6143.292.DC1/`.

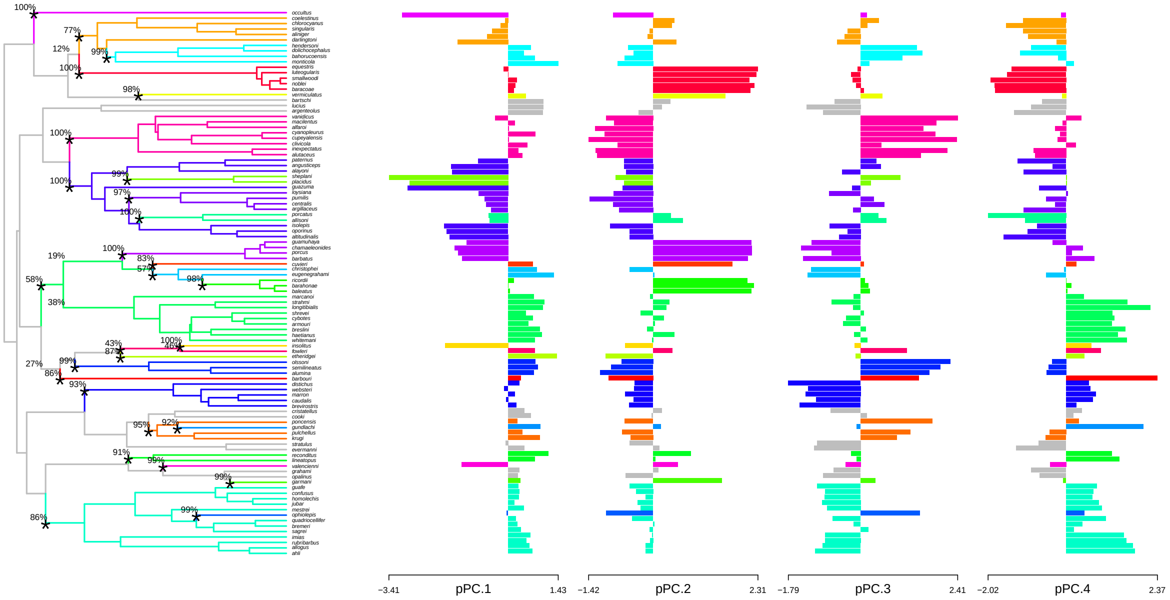


Figure A.3: Left: phylogenetic placement of 28 shifts in the selection optimum in Anolis lizards, as detected by ℓ1ou+AIC$_c$ (running time: 13.2 minutes). Each shift is annotated with an asterisk and its bootstrap support. Right: bar graphs showing the 4 traits combined for analysis, as in Figure A.2.

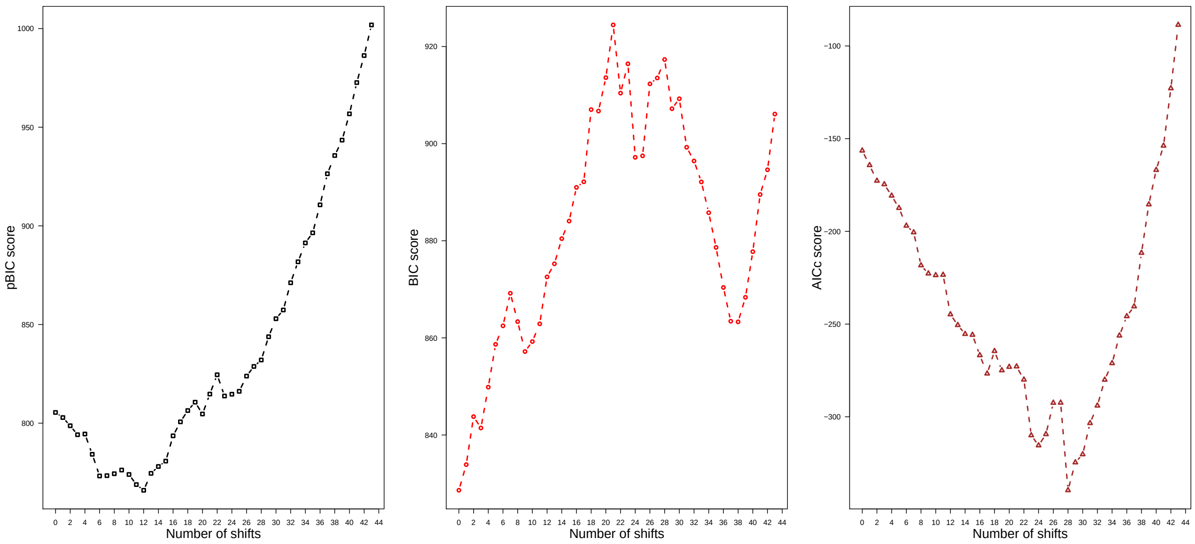


Figure A.4: Profile plot of the pBIC (left), BIC (middle) and AIC$_c$ (right) scores among all candidate models evaluated during the ℓ1ou approach, for the Anolis lizards data (all 4 traits).

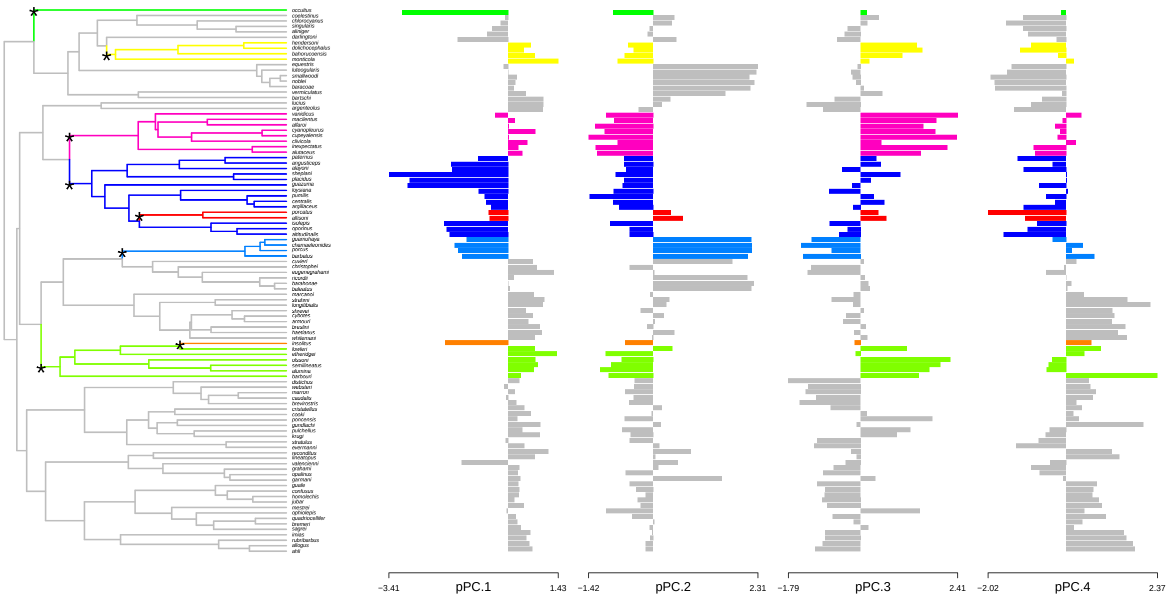


Figure A.5: Left: phylogenetic placement of 9 shifts in the selection optimum in Anolis lizards, corresponding to a local optimum at 9 found by ℓ1ou+BIC (see Figure A.4). Right: bar graphs showing the 4 traits combined for analysis, as in Figure A.2.

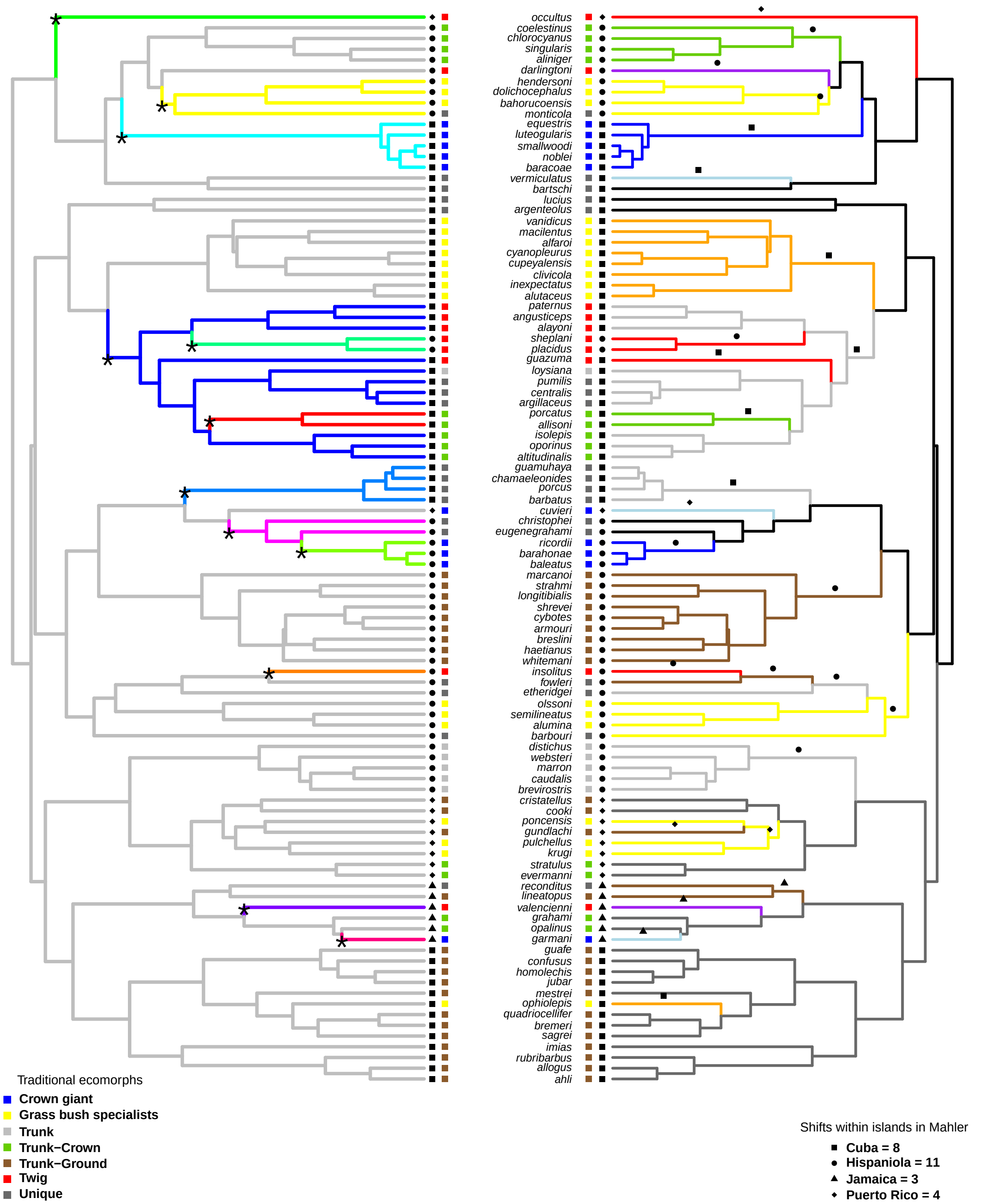


Figure A.6: Configurations selected by ℓ1ou+ pBIC (left, 12 shifts) and by surface (right, 28 shifts), with the traditional ecomorph designations present in Mahler et al. (2013) (colored squares at the tips) and the geographic location of each species (black shapes at the tips). Right: edge colors correspond to those in Mahler et al. (2013). Shapes above edges with shifts correspond to the island where the shift likely occurred.

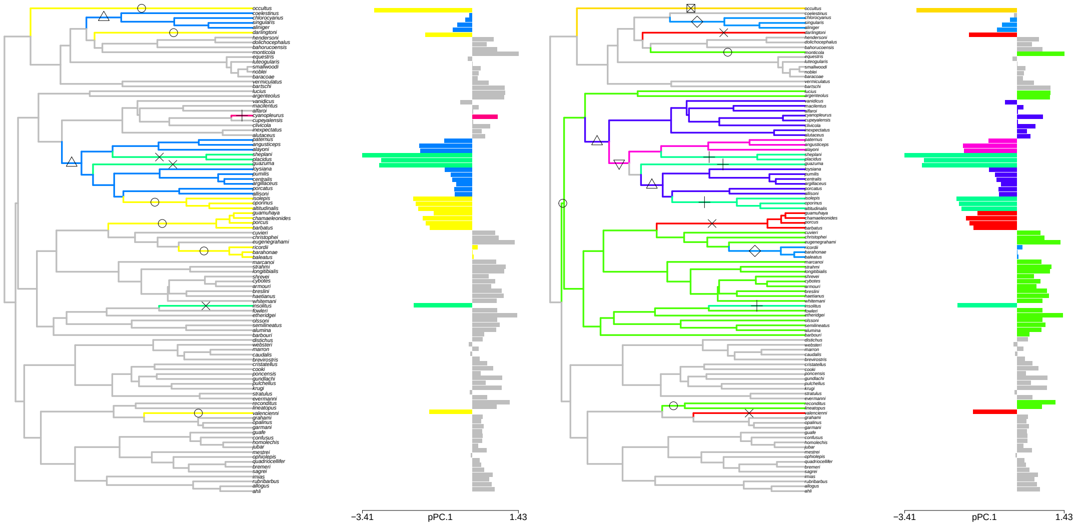


Figure A.7: Shifts in Anolis lizard morphology from the first pPC only, using surface (left: 12 shifts, 5 unique optima, 3 having convergence: yellow, blue and green) or using ℓ1ou + AIC$_c$ followed by detection of convergent evolution (right: 16 shifts, 8 unique optima, 5 having convergence: light blue, dark blue, red, light green and pale green). The two configurations have marked differences, but similar AIC$_c$ scores (−86.37 and −86.68). Each shift is annotated with a symbol on the corresponding edge. The same symbol is used for convergent shifts.

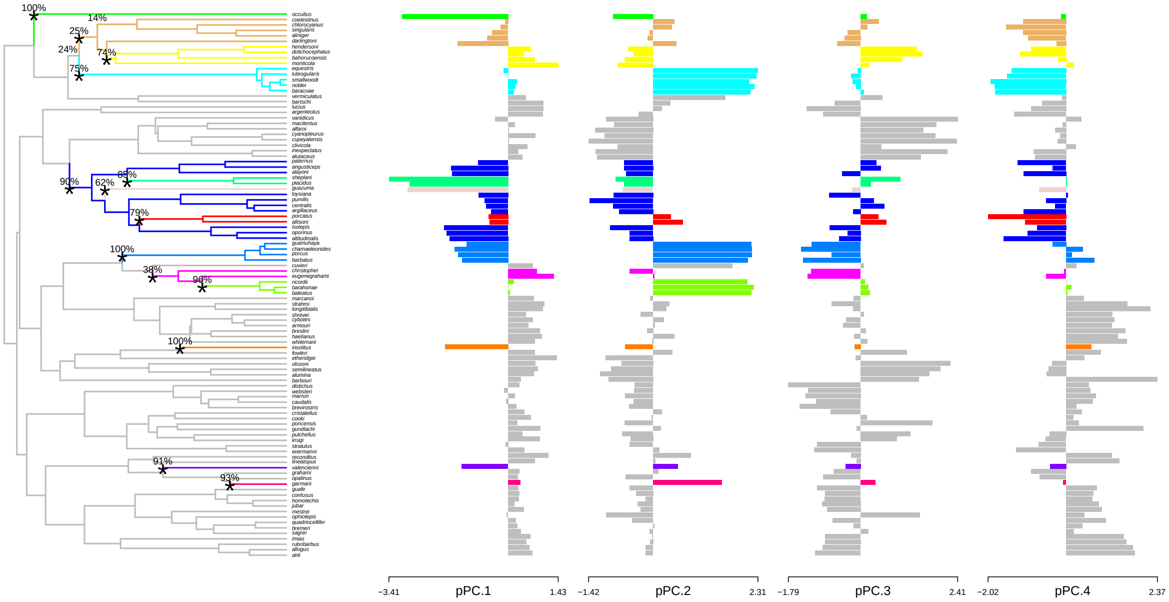


Figure A.8: Sensitivity of bootstrap support to the configuration used to simulate bootstrap replicates. Left: Scenario with 14 evolutionary shifts: the 12 found by ℓ1ou + pBIC plus 2 extra shifts. Right: bar graphs showing the 4 traits combined for analysis (from Mahler et al., 2013). This 14-shift configuration has a pBIC score of 770.97, close to that of the optimal 12-shift configuration, 765.98. When the bootstrapping model uses all 14 evolutionary shifts, bootstrap support (shown above edges) is strong for the extra 2 shifts, highlighting the sensitivity of bootstrap values to the configuration used to simulate bootstrap data sets. One of the extra shifts (leading to a single species, in beige) receives 62% support, compared to no support when the bootstrapping model uses only 12 shifts (Figure 9). The other extra shift (leading to 5 tips, in light brown) receives support that is split between 3 edges: the original edge (25%), its parent edge (24%), and one of its daughter edge (14%), for a total support of 63%. Positioning the shift on the parent edge leads to the same clustering of taxa, i.e. to an equivalent non-identifiable shift configuration. The daughter edge separates only 1 of the 5 species in the cluster defined by the shift.

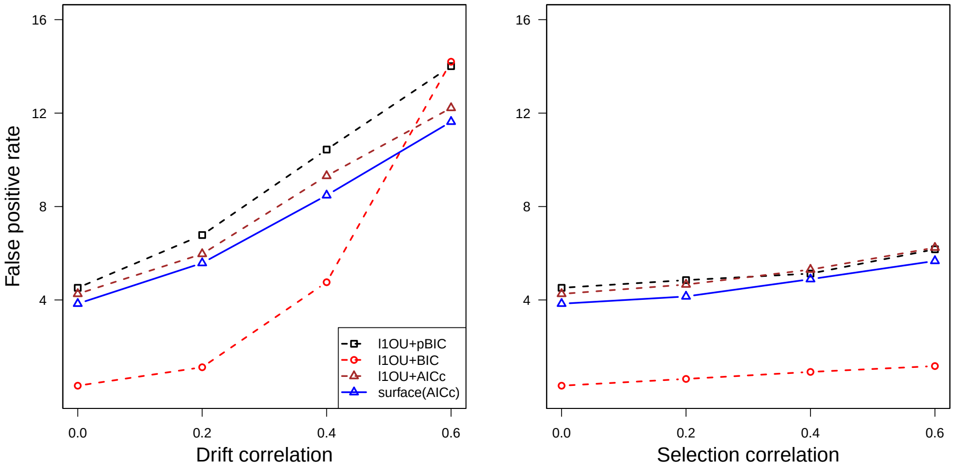


Figure A.9: Number of false positives (false shifts). Four traits were simulated with no shifts as in Figure 4 (right) but with residual correlation, under a multivariate OU process with either correlated drift (left) or correlated selection (right), in R using mvSLOUCH v1.2.1 (Bartoszek et al., 2012). The drift covariance had diagonal elements of $\sigma^2 = 2$ and the selection matrix had diagonal elements of $\alpha = 1$. To simulate correlation due to drift, the off-diagonal elements of the drift covariance were all set to $2r$ (left). To simulate correlation due to selection, the off-diagonal elements of the selection matrix were all set to r (right). Correlation due to drift had a large impact with an increased number of falsely detected shifts, for all of the methods currently available to handle multiple traits. More work is needed to develop shift detection methods that account for residual correlation among traits.

Appendix B

Appendix for Chapter 3

This appendix provides the proofs contained in the Chapter 3. We begin by presenting some preliminary lemmas. We then provide the proofs for the results given in Sections 3.3, 3.4, and 3.4.

B.1 Preliminary Lemmas

This section contains lemmas which are used to prove our main results. First we state two standard results, given here for convenience.

Lemma B.1. *Let* A *be a symmetric matrix and* D *a diagonal matrix. Then*

$$\|DA\| = \|D^{\frac{1}{2}}AD^{\frac{1}{2}}\|$$

.

Lemma B.2 (Bernstein's inequality)**.** *Let* $X_1, \cdots, X_N$ *be independent random variables and* $|X_i - EX_i| \leqslant S$ *for* $i = 1, \cdots, N$. *Let* $\sigma^2 := \sum_{i=1}^{N} E[X_i - EX_i]^2$. *Then for all* $t \geqslant 0$,

$$\Pr\left(\left|\sum_{i=1}^{N} X_i - EX_i\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{\frac{1}{2}t^2}{\sigma^2 + \frac{1}{3}St}\right).$$

We use the following result from Rohe et al. (2011) in the proof of Proposition 3.5.

Lemma B.3. *[Rohe et al. (2011)] Under the Stochastic Blockmodel, if* $B = pI + rJ$ *and there are an equal number of nodes in each block, then*

$$\lambda_i(P_{\mathcal{A}}) = \begin{cases} 1 & i = 1 \\ (Kr/p + 1)^{-1} & i = 2, \ldots, K \\ 0 & o.w. \end{cases}$$

For completeness we include the proof here.

Proof. The matrix $B \in \mathbb{R}^{k \times k}$ is the sum of two matrices,

$$B = pI + rJ_k\mathbf{1}_k^T,$$

where $I_k \in \mathbb{R}^{k \times k}$ is the identity matrix, $\mathbf{1}_k \in \mathbb{R}^k$ is a vector of ones, $r \in (0,1)$ and $p \in (0, 1-r)$. Let $Z \in \{0,1\}^{N \times K}$ be such that $Z^T\mathbf{1}_N = s\mathbf{1}_K$ for some $s \in \mathbb{R}$. This guarantees that all K blocks have equal size s. The Stochastic Blockmodel has the population adjacency matrix, $\mathcal{A} = ZBZ^T$. Moreover, $P_{\mathcal{A}} = ZB_LZ^T$ for

$$B_L = \frac{1}{Nr + sp}\left(pI_K + r\mathbf{1}_K\mathbf{1}_K^T\right).$$

The eigenvalues are found by construction.

- The constant vector $\mathbf{1}_N$ is an eigenvector with eigenvalue 1;

$$\begin{aligned} ZB_LZ^T\mathbf{1}_N &= \frac{s}{Nr+sp}Z\left(pI_K + r\mathbf{1}_K\mathbf{1}_K^T\right)1_K \\ &= \frac{s}{Nr+sp}Z(p+Kr)\mathbf{1}_K + \frac{s(p+Kr)}{Nr+sp}\mathbf{1}_N = \mathbf{1}_N, \end{aligned}$$

  where the last line follows because $N = sK$.

- Let $b_2, \ldots, b_K \in \mathbb{R}^K$ be a set of orthogonal vectors which are also orthogonal to $\mathbf{1}_K$. For any i, Zb_i is an eigenvector with eigenvalue $(Kr/p+1)^{-1}$,

$$ZB_LZ^T(Zb_i) = ZB_LsI_{K\times K}b_i = \frac{s}{Nr+sp}Z\left(pI_K + r\mathbf{1}_K\mathbf{1}_K^T\right)b_i = \frac{ps}{Nr+sp}(Zb_i).$$

  Because Zb_i and Zb_j are orthogonal for $i \neq j$, the multiplicity of the eigenvalue $(Kr/p+1)^{-1}$ is at least $K-1$.

Because $\mathrm{rank}(P_{\mathcal{A}}) \leqslant \mathrm{min}(\mathrm{rank}(Z), \mathrm{rank}(B_L), \mathrm{rank}(Z^T)) \leqslant K$, there are at most K nonzero eigenvalues. The results follow. □

The following result is used for the computation of the eigenvalues in the proof of Proposition 3.6

Lemma B.4. *Let* P *be a block constant Markov transition matrix, with blocks of identical sizes. Let* P *contain the block values*

$$\mathsf{P} = \left(\begin{array}{c|c} p & r \\ \hline r & p \end{array} \right),$$

then

$$\lambda_2(\mathsf{P}) = \frac{p - r}{p + r}.$$

Proof. This follows from Lemma B.3 using $K = 2$. □

Lemma B.5 (Operator norm of non-negative irreducible matrices)**.** *Let* $A \in \mathsf{R}^{N \times N}$ *be a non-negative, irreducible matrix. Let* $r_i(A) := \sum_{j=1}^{N} A_{ij}$*. Then,*

$$\|A\| \leqslant \max_i r_i(A).$$

Proof. By Perron-Frobenius theorem, A has a real leading eigenvalue and if $Ay \leqslant \mu y$ then $\lambda_1(A) \leqslant \mu$, where $y \in \mathbb{R}^N$, $\mu \in \mathbb{R}$, $y \geqslant 0$, and $\mu \geqslant 0$. Take $y = \mathbf{1}$ and $\mu = \max_i r_i(A)$. Therefore,

$$\|A\| = \lambda_1(A) \leqslant \max_i r_i(A).$$

□

Lemma B.6. *For any* $W \in \mathsf{R}^{N \times N}$*, define diagonal matrix* T *to contain the row sums down the diagonal,* $\mathsf{T}_{uu} = \sum_v W(u, v)$*. If* $\mathsf{T}_{uu} > 0$ *for all* u*, then the eigenvalues of* $\mathsf{P}_W = \mathsf{T}^{-1}W$ *are equal to the eigenvalues of* $\mathsf{L}_W = \mathsf{T}^{-1/2}W\mathsf{T}^{-1/2}$.

Proof. Let x, λ be an eigenpair of L_W,

$$\mathsf{T}^{-1/2}W\mathsf{T}^{-1/2}x = \lambda x \implies \mathsf{T}^{-1/2}\left(\mathsf{T}^{-1/2}W\left(\mathsf{T}^{-1/2}x\right)\right) = \lambda\left(\mathsf{T}^{-1/2}x\right),$$

where the left hand side is $\mathsf{P}_W(\mathsf{T}^{-1/2}x)$. This implies that $\mathsf{T}^{-1/2}x, \lambda$ is an eigenpair of P_W. □

B.2 Design Effect and Variance

Here we provide the proof of Proposition 3.1 from Section 3.3.

Proof of Proposition 3.1. Lemma 12.2 in Levin et al. (2009) shows that (i) f_j and λ_j are real valued and (ii) the f_j are orthonormal with respect to $\langle f_\ell, f_j \rangle_\pi$. Because $\lambda_2 < 1$, f_1 is the constant vector. To decompose the covariance,

$$\begin{aligned} \text{Cov}\,(y(X_i), y(X_{i+t})) &= \text{E}\,[(y(X_i) - \text{E}[y(X_i)])(y(X_{i+t}) - \text{E}[y(X_{i+t})])] \\ &= \text{E}\,[y(X_i)y(X_{i+t})] - \text{E}^2[y(X_1)] \\ &= \text{E}\,[y(X_1)y(X_{1+t})] - \text{E}^2[y(X_1)]. \end{aligned}$$

The first term

$$\begin{aligned} \text{E}[y(X_1)y(X_{1+t})] &= \sum_{u,v \in \mathcal{V}} y(u)y(v)\,\Pr(X_1 = u, X_{1+t} = v) \\ &= \sum_{u,v \in \mathcal{V}} y(u)y(v)\pi_u P^t(u,v) \\ &= \sum_{u,v \in \mathcal{V}} y(u)y(v)\pi_u \pi_v \sum_{j=1}^{|\mathcal{V}|} f_j(u) f_j(v) \lambda_j^t \\ &= \sum_{u,v \in \mathcal{V}} y(u)y(v)\pi_u \pi_v \{1 + \sum_{j=2}^{|\mathcal{V}|} f_j(u) f_j(v) \lambda_j^t\} \\ &= \sum_{u,v \in \mathcal{V}} y(u)y(v)\pi_u \pi_v + \sum_{j=2}^{|\mathcal{V}|} \lambda_j^t \sum_{u,v \in \mathcal{V}} y(u)y(v)\pi_u \pi_v f_j(u) f_j(v) \\ &= \text{E}^2[y(X_1)] + \sum_{j=2}^{|\mathcal{V}|} \langle y, f_j \rangle_\pi^2 \lambda_j^t. \end{aligned}$$

Hence,

$$\mathrm{Cov}\left(y(X_i), y(X_{i+t})\right) = \sum_{j=2}^{|\mathcal{V}|} \langle y, f_j \rangle_\pi^2 \lambda_j^t.$$

□

B.3 Population Graph Results

Here we provide the proofs of Lemmas 3.3, 3.4 and Propositions 3.5, 3.6 which are the results given in Section 3.4.

Proof of Lemma 3.3. From the definition of Z and $\bar{\mathcal{A}}$ it follows that $Z^T Z = \Pi$ and $\bar{\mathcal{A}} = J_{n\times n} - ZBZ^T = Z\bar{B}Z^T$. Then,

$$\mathcal{A}\bar{\mathcal{A}} = ZBZ^T Z\bar{B}Z^T = ZB\Pi\bar{B}Z^T$$

and similarly,

$$\bar{\mathcal{A}}\mathcal{A} = Z\bar{B}\Pi BZ^T.$$

Hence,

$$\begin{aligned}(\mathcal{A}\bar{\mathcal{A}} + \bar{\mathcal{A}}\mathcal{A}) \cdot \mathcal{A} &= \left(Z(B\Pi\bar{B} + \bar{B}\Pi B)Z^T\right) \cdot (ZBZ^T) \\ &= Z\left((B\Pi\bar{B} + \bar{B}\Pi B) \cdot B\right) Z^T.\end{aligned}$$

□

Proof of Lemma 3.4. We first show that

$$\frac{[(B\Pi\bar{B}) \cdot B]_{ij}}{[(B\Pi\bar{B}) \cdot B]_{ii}} > \frac{B_{ij}}{B_{ii}} = \frac{r}{p+r}. \tag{B.1}$$

We have

$$[(B\Pi\bar{B})\cdot B]_{ij} = r(\Pi_{ii}(p+r)(1-r) + \Pi_{jj}r(1-p-r) + \sum_{l\neq i, l\neq j}\Pi_{ll}r(1-r))$$
$$[(B\Pi\bar{B})\cdot B]_{ii} = (p+r)(\Pi_{ii}(p+r)(1-p-r) + \sum_{l\neq i}\Pi_{ll}r(1-r)).$$

We re-write (B.1) as follows:

$$\frac{r(\Pi_{ii}(p+r)(1-r) + \Pi_{jj}r(1-p-r) + \sum_{\substack{l\neq i\\ l\neq j}}\Pi_{ll}r(1-r))}{(p+r)(\Pi_{ii}(p+r)(1-p-r) + \sum_{l\neq i}\Pi_{ll}r(1-r))} > \frac{r}{p+r} \tag{B.2}$$
$$p(\Pi_{ii}(p+r) - \Pi_{jj}r) > 0, \tag{B.3}$$

where (B.2) to (B.3) follows from some algebraic manipulation and (B.3) is always true because of the lemma assumptions. In addition, by going through the same procedure, it can be shown that

$$\frac{[(B\Pi\bar{B} + \bar{B}\Pi B)\cdot B]_{ij}}{[(B\Pi\bar{B} + \bar{B}\Pi B)\cdot B]_{ii}} > \frac{B_{ij}}{B_{ii}}.$$

In terms of the expected adjacency matrices the above statement is equivalent to the followings: Suppose that nodes i and l belong to the same block and j belongs to a different block, then

$$\frac{\tilde{\mathcal{W}}_{ij}}{\tilde{\mathcal{W}}_{il}} > \frac{\mathcal{A}_{ij}}{\mathcal{A}_{il}}. \tag{B.4}$$

Now, we show $P_{\tilde{\mathcal{W}}}(u,v) < P_{\mathcal{A}}(u,v)$ when u and v belong to the same block. We have

$$\sum_{w\in\mathcal{V}} P_{\tilde{\mathcal{W}}}(u,w) = \sum_{w\in\mathcal{V}} P_{\mathcal{A}}(u,w) = 1$$
$$\sum_{w\in\mathcal{V}} [\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uw} = \sum_{w\in\mathcal{V}} [\mathcal{D}^{-1}\mathcal{A}]_{uw}.$$

Assume u and v belong to block $\mathcal{C}$ of size $|\mathcal{C}|$. Factor out the transition probability between u and v. Then,

$$[\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uv}\left(|\mathcal{C}|+\sum_{w\notin\mathcal{C}}\frac{[\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uw}}{[\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uv}}\right)=[\mathcal{D}^{-1}\mathcal{A}]_{uv}\left(|\mathcal{C}|+\sum_{w\notin\mathcal{C}}\frac{[\mathcal{D}^{-1}\mathcal{A}]_{uw}}{[\mathcal{D}^{-1}\mathcal{A}]_{uv}}\right),$$

and since the summations are along the rows,

$$[\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uv}\left(|\mathcal{C}|+\sum_{w\notin\mathcal{C}}\frac{\tilde{\mathcal{W}}_{uw}}{\tilde{\mathcal{W}}_{uv}}\right)=[\mathcal{D}^{-1}\mathcal{A}]_{uv}\left(|\mathcal{C}|+\sum_{w\notin\mathcal{C}}\frac{\mathcal{A}_{uw}}{\mathcal{A}_{uv}}\right).$$

Therefore, based on inequality (B.4),

$$[\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uv}<[\mathcal{D}^{-1}\mathcal{A}]_{uv}.$$

Suppose the case where $\Pi_{ii}=\Pi_{jj}$ for all i and j, similarly, for $w\notin\mathcal{C}$

$$[\mathcal{T}^{-1}\tilde{\mathcal{W}}]_{uw}>[\mathcal{D}^{-1}\mathcal{A}]_{uw}. \tag{B.5}$$

□

Proof of Proposition 3.5. The first part of this proof focuses on the inequality $\lambda_2(P_{\tilde{\mathcal{W}}})<\lambda_2(P_{\mathcal{A}})$.

Let $\mathcal{B}^{RW}:=\mathcal{D}^{-1}_{k\times k}B$, and $\mathcal{B}^{AC}:=\mathcal{T}^{-1}_{k\times k}[(B\Pi\bar{B}+\bar{B}\Pi B)\cdot B]$. Since $\Pi_{ii}=\Pi_{jj}$, then $\mathcal{B}^{RW}$ and $\mathcal{B}^{AC}$ are symmetric matrices and have equal row sum. Hence,

$$\begin{aligned}\lambda_2(P_{\mathcal{A}})&=\lambda_2(\mathcal{D}^{-1}\mathcal{A})=\lambda_2(\mathcal{B}^{RW}),\\ \lambda_2(P_{\tilde{\mathcal{W}}})&=\lambda_2(\mathcal{T}^{-1}\tilde{\mathcal{W}})=\lambda_2(\mathcal{B}^{AC}).\end{aligned}$$

Let $f:\{1,2,\cdots,k\}\to\mathbb{R}$ and r be the row sum of $\mathcal{B}^{RW}$ and $\mathcal{B}^{AC}$. Then $I-\frac{1}{r}\mathcal{B}^{AC}$

and $I - \frac{1}{r}\mathcal{B}^{RW}$ are Laplacian matrices. Therefore,

$$\begin{aligned}\lambda_2(I - \frac{1}{r}\mathcal{B}^{AC}) &= \inf_{\substack{f:\sum_u f(u)=0 \\ f:\sum_u f^2(u)=1}} \frac{1}{2r} \sum_{u,v\; u\neq v} \mathcal{B}^{AC}_{uv}(f(v) - f(u))^2 \\ &> \inf_{\substack{f:\sum_u f(u)=0 \\ f:\sum_u f^2(u)=1}} \frac{1}{2r} \sum_{u,v\; u\neq v} \mathcal{B}^{RW}_{uv}(f(v) - f(u))^2 = \lambda_2(I - \frac{1}{r}\mathcal{B}^{RW}),\end{aligned}$$

where the inequality follows from inequality (B.5) and the fact that $\mathcal{B}^{AC}_{uv} > \mathcal{B}^{RW}_{uv}$ for $u \neq v$. So we conclude that

$$\lambda_2(\mathcal{B}^{AC}) < \lambda_2(\mathcal{B}^{RW})$$

and, therefore proves the inequality in (3.7),

$$\lambda_2(P_{\tilde{\mathcal{W}}}) < \lambda_2(P_{\mathcal{A}}).$$

The fact that $\lambda_2(P_{\mathcal{A}}) = 1/(R+1)$ follows immediately from Lemma B.3.

The rest of the proof is dedicated to Equation (3.8). From Lemma 3.3, $\tilde{\mathcal{W}} = Z\tilde{B}Z^T$ for $\tilde{B} = (B\Pi\bar{B} + \bar{B}\Pi B) \cdot B$. Define $r' = 1 - r$ and note that $\Pi = N/KI$ and so it can be temporarily ignored as a constant. First,

$$B\bar{B} = (r'J - pI)(rJ + pI) = (r'rK + r'p - pr)J - p^2I.$$

Then, define $u = (r'rK + r'p - pr)$.

$$(B\bar{B}) \cdot B = (uJ - p^2I) \cdot (rJ + pI) = p(u - rp - p^2)I + urJ$$

Reincorporating the constants from $\Pi = N/KI$ and a 2 to account for $\bar{B}B$, it follows that $\tilde{B} = \tilde{p}I + \tilde{r}J$, for

$$\tilde{p} = 2p(N/K)(u - rp - p^2) \quad \text{and} \quad \tilde{r} = 2(N/K)ur.$$

Note that $\tilde{r}$ and $\tilde{p}$ depend on the block populations N/K and thus the number of nodes in the graph N. However, this term cancels out in the ratio $\tilde{r}/\tilde{p}$. So, neither $\lambda_2(P_{\tilde{\mathcal{W}}})$ nor $\lambda_2(P_{\mathcal{A}})$ depend on N. As such,

$$\lambda_2(P_{\tilde{\mathcal{W}}}) + \epsilon < \lambda_2(P_{\mathcal{A}})$$

for some $\epsilon > 0$ that is independent of N.

As K grows and r shrinks, $u \to p(R+1)$ and

$$\tilde{p} \to 2p(N/K)(p(R+1) - p^2) \quad \text{and} \quad \tilde{r} \to 2rp(N/K)(R+1).$$

Using Lemma B.3 on $\tilde{B}$,

$$\lambda_2(P_{\tilde{\mathcal{W}}}) = \frac{1}{K(\tilde{r}/\tilde{p}) + 1}.$$

Then,

$$\frac{K\tilde{r}}{\tilde{p}} \to \frac{Krp(R+1)}{p(p(R+1) - p^2)} = \frac{Kr(R+1)}{p(R+1-p)} = R\frac{R+1}{R+1-p},$$

which concludes the proof. □

Proof of Proposition 3.6. Both $P_{\mathcal{A}}$ and $P_{\tilde{\mathcal{W}}}$ satisfy the conditions of Lemma B.4. It is only necessary to compute the probabilities. For $P_{\mathcal{A}}$, $p = 1 - \epsilon$ and $r = \epsilon$. So,

$$\lambda_2(P_{\mathcal{A}}) = \frac{1 - 2\epsilon}{1} \to 1.$$

To compute $\lambda_2(P_{\tilde{\mathcal{W}}}$, notice that it is only necessary to determine p and r up to proportionality. Under the assumed model, $\bar{B}_{11} = \epsilon$, $\bar{B}_{12} = 1 - \epsilon$, and $\Pi \propto I$. Moreover, the matrix $(B\Pi\bar{B} + \bar{B}\Pi B) \cdot B$ contains the elements $p = 2(1-\epsilon)^2$ and $r = (1-\epsilon)^2 + \epsilon^2$ for $P_{\tilde{\mathcal{W}}}$. By Lemma B.4.

$$\lambda_2(P_{\tilde{\mathcal{W}}}) = \frac{2(1-\epsilon)^2 - (1-\epsilon)^2 + \epsilon^2}{2(1-\epsilon)^2 + (1-\epsilon)^2 + \epsilon^2} = \frac{(1-\epsilon)^2 + \epsilon^2}{3(1-\epsilon)^2 + \epsilon^2} \to 1/3.$$

□

B.4 Sampled Graph Results

Here we provide the proofs of Theorems 3.7 and 3.9 from Section 3.4.

Proof of Theorem 3.7. By Lemma B.6, and Wyle's inequality,

$$|\lambda_\ell(\mathrm{P}_{\tilde{W}}) - \lambda_\ell(\mathrm{P}_{\tilde{\mathcal{W}}})| = |\lambda_\ell(\mathrm{T}^{-\frac{1}{2}}\tilde{W}\mathrm{T}^{-\frac{1}{2}}) - \lambda_\ell(\mathcal{T}^{-\frac{1}{2}}\tilde{\mathcal{W}}\mathcal{T}^{-\frac{1}{2}})| \leqslant \|\mathrm{T}^{-\frac{1}{2}}\tilde{W}\mathrm{T}^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}}\tilde{\mathcal{W}}\mathcal{T}^{-\frac{1}{2}}\|.$$

The rest of the proof studies the righthand side of this inequality.

In order to reduce the required notation, let $\mathrm{N} := |\mathcal{V}|$,

$$\begin{aligned}
\tilde{\mathrm{A}} &:= (A\bar{\mathrm{A}} + \bar{\mathrm{A}}A),\\
\tilde{\mathcal{A}} &:= (\mathcal{A}\bar{\mathcal{A}} + \bar{\mathcal{A}}\mathcal{A}),\\
\tilde{W} &:= (A\bar{\mathrm{A}} + \bar{\mathrm{A}}A) \cdot A = \tilde{\mathrm{A}} \cdot A,\\
\tilde{\mathcal{W}} &:= (\mathcal{A}\bar{\mathcal{A}} + \bar{\mathcal{A}}\mathcal{A}) \cdot \mathcal{A} = \tilde{\mathcal{A}} \cdot \mathcal{A}.
\end{aligned}$$

By the triangle inequality,

$$\|\mathrm{T}^{-\frac{1}{2}}\tilde{W}\mathrm{T}^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}}\tilde{\mathcal{W}}\mathcal{T}^{-\frac{1}{2}}\| \leqslant \|\mathcal{T}^{-\frac{1}{2}}(\tilde{W} - \tilde{\mathcal{W}})\mathcal{T}^{-\frac{1}{2}}\| + \|\mathrm{T}^{-\frac{1}{2}}\tilde{W}\mathrm{T}^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}}\tilde{W}\mathcal{T}^{-\frac{1}{2}}\|.$$

Also,

$$\begin{aligned}
\|\mathcal{T}^{-\frac{1}{2}}(\tilde{W} - \tilde{\mathcal{W}})\mathcal{T}^{-\frac{1}{2}}\| &= \|\mathcal{T}^{-\frac{1}{2}}(\tilde{\mathrm{A}} \cdot \mathrm{A} - \tilde{\mathcal{A}} \cdot \mathcal{A})\mathcal{T}^{-\frac{1}{2}}\|\\
&\leqslant \|\mathcal{T}^{-\frac{1}{2}}((\tilde{\mathrm{A}} - \tilde{\mathcal{A}}) \cdot \mathrm{A})\mathcal{T}^{-\frac{1}{2}}\| + \|\mathcal{T}^{-\frac{1}{2}}((A - \mathcal{A}) \cdot \tilde{\mathcal{A}})\mathcal{T}^{-\frac{1}{2}}\|.
\end{aligned}$$

The proof is divided into four parts. Part 1 bounds $\|\mathcal{T}^{-\frac{1}{2}}((\tilde{\mathrm{A}} - \tilde{\mathcal{A}}) \cdot \mathrm{A})\mathcal{T}^{-\frac{1}{2}}\|$. Part 2 bounds $\|\mathcal{T}^{-\frac{1}{2}}((A - \mathcal{A}) \cdot \tilde{\mathcal{A}})\mathcal{T}^{-\frac{1}{2}}\|$. Part 3 bounds $\|\mathrm{T}^{-\frac{1}{2}}\tilde{W}\mathrm{T}^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}}\tilde{W}\mathcal{T}^{-\frac{1}{2}}\|$. Part 4 combines these bounds

Part 1- Matrix $\mathcal{T}$ is diagonal and matrices $\tilde{A}$ and $\tilde{\mathcal{A}}$ are symmetric. Now, we apply Lemma B.1.

$$\begin{aligned}\|\mathcal{T}^{-\frac{1}{2}}((\tilde{A}-\tilde{\mathcal{A}})\cdot A)\mathcal{T}^{-\frac{1}{2}}\| &= \|\mathcal{T}^{-1}(\tilde{A}-\tilde{\mathcal{A}})\cdot A\| \\ &\leqslant \|\mathcal{T}^{-1}(A\bar{A}-\mathcal{A}\bar{\mathcal{A}})\cdot A\| + \|\mathcal{T}^{-1}(\bar{A}A-\bar{\mathcal{A}}\mathcal{A})\cdot A\|.\end{aligned}$$

At this step, we show an upper-bound for the first term and the result naturally carries on to the second term.

$$\|\mathcal{T}^{-1}(A\bar{A}-\mathcal{A}\bar{\mathcal{A}})\cdot A\| \leqslant \|\mathcal{T}^{-1}|A\bar{A}-\mathcal{A}\bar{\mathcal{A}}|\cdot A\|, \tag{B.6}$$

where $|\cdot|$ is the element-wise absolute value operator and the inequality follows from the fact that for any matrix M, $\|M\| \leqslant \||M|\|$ (e.g. Mathias, 1990, Theorem 2.5). Now, we bound the row sum of $|A\bar{A}-\mathcal{A}\bar{\mathcal{A}}|\cdot A$ by a concentration inequality and then we use Lemma B.5 to bound the operator norm.

Denote the sum of the i-th row by $r_i(\cdot)$. We have

$$r_i\left(\mathcal{T}^{-1}|A\bar{A}-\mathcal{A}\bar{\mathcal{A}}|\cdot A\right) = \frac{1}{\mathcal{T}_{ii}}\sum_j A_{ij}\left|\sum_k A_{ik}\bar{A}_{kj}-\mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}\right|. \tag{B.7}$$

Define $F_{ij} = \sum_k \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}$ and $G_{ij} = \sum_k \bar{\mathcal{A}}_{ik}\mathcal{A}_{kj}$. For fixed i and j, random variables $A_{ik}\bar{A}_{kj}$ are independent with the expected value $E[A_{ik}\bar{A}_{kj}] = \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}$ and the variance

$$\sigma_{ij}^2 = \sum_k E(A_{ik}\bar{A}_{kj}-\mathcal{A}_{ik}\bar{\mathcal{A}}_{kj})^2 \leqslant \sum_k E(A_{ik}\bar{A}_{kj})^2 + (\mathcal{A}_{ik}\bar{\mathcal{A}}_{kj})^2 \leqslant 2F_{ij}.$$

Let $\Delta_{F_{ij}} := \sqrt{10F_{ij}\ln\frac{2N^2}{\delta}}$. By the Bernstein's inequality (stated in Lemma B.2 for convenience) and union bound,

$$\Pr\left(\left|\sum_k A_{ik}\bar{A}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}\right| \geqslant \Delta_{F_{ij}}\right) \leqslant 2\exp\left(-\frac{\frac{1}{2}\Delta_{F_{ij}}^2}{\sigma_i^2 + \frac{1}{3}S\Delta_{F_{ij}}}\right) \tag{B.8}$$

$$= 2\exp\left(-\frac{5F_{ij}\ln\frac{2N^2}{\delta}}{4F_{ij} + \frac{1}{3}S\Delta_{F_{ij}}}\right)$$

$$\leqslant \frac{\delta}{N^2},$$

where the last inequality follows from the assumption that $F_{\min} \gg \ln N$. So with high probability,

$$\sum_j A_{ij}\left|\sum_k A_{ik}\bar{A}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}\right| \leqslant \sum_j A_{ij}\Delta_{F_{ij}}.$$

Now we bound $|\sum_j (A_{ij} - \mathcal{A}_{ij})\Delta_{F_{ij}}|$ by using the Bernstein's inequality again. $E[A_{ij}\Delta_{F_{ij}}] = \mathcal{A}_{ij}\Delta_{F_{ij}}$ and

$$\sum_j E[A_{ij}\Delta_{F_{ij}} - \mathcal{A}_{ij}\Delta_{F_{ij}}]^2 \leqslant 2\sum_j \mathcal{A}_{ij}\Delta_{F_{ij}}^2.$$

By Bernstein's inequality, we obtain that with high probability

$$|\sum_j (A_{ij} - \mathcal{A}_{ij})\Delta_{F_{ij}}| \leqslant \sqrt{2\sum_j \mathcal{A}_{ij}\Delta_{F_{ij}}^2},$$

and consequently

$$\sum_j A_{ij}\left|\sum_k A_{ik}\bar{A}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}\right| \leqslant \sum_j \mathcal{A}_{ij}\Delta_{F_{ij}} + \sqrt{2\sum_j \mathcal{A}_{ij}\Delta^2_{F_{ij}}} \tag{B.9}$$

$$\leqslant 2\sum_j \mathcal{A}_{ij}\sqrt{10F_{ij}\ln\frac{2N^2}{\delta}}$$

$$\leqslant 10\sum_j \mathcal{A}_{ij}\sqrt{F_{ij}\ln\frac{N}{\delta}}.$$

Furthermore,

$$\mathcal{T}_{ii} = \sum_j \tilde{\mathcal{W}}_{ij} = \sum_j \mathcal{A}_{ij}\sum_k \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj} + \bar{\mathcal{A}}_{ik}\mathcal{A}_{kj} = \sum_j \mathcal{A}_{ij}(F_{ij}+G_{ij}). \tag{B.10}$$

From (B.7), (B.9) and (B.10),

$$r_i\left(\mathcal{T}^{-1}|A\bar{A} - \mathcal{A}\bar{\mathcal{A}}|\cdot A\right) \leqslant \frac{10\sum_j \mathcal{A}_{ij}\sqrt{F_{ij}\ln\frac{N}{\delta}}}{\sum_j \mathcal{A}_{ij}F_{ij}} \leqslant \frac{10\sum_j \mathcal{A}_{ij}F_{ij}\sqrt{\frac{\ln\frac{N}{\delta}}{F_{ij}}}}{\sum_j \mathcal{A}_{ij}F_{ij}} \leqslant 10\sqrt{\frac{\ln\frac{N}{\delta}}{F_{\min}}}. \tag{B.11}$$

Following the same steps, we obtain

$$r_i\left(\mathcal{T}^{-1}|\bar{A}A - \bar{\mathcal{A}}\mathcal{A}|\cdot A\right) \leqslant \frac{10\sum_j \mathcal{A}_{ij}\sqrt{G_{ij}\ln\frac{N}{\delta}}}{\sum_j \mathcal{A}_{ij}G_{ij}} \leqslant 10\sqrt{\frac{\ln\frac{N}{\delta}}{G_{\min}}}. \tag{B.12}$$

Therefore,

$$\|\mathcal{T}^{-\frac{1}{2}}((\tilde{A} - \tilde{\mathcal{A}})\cdot A)\mathcal{T}^{-\frac{1}{2}}\| \leqslant \frac{10\ln^{\frac{1}{2}}\frac{N}{\delta}}{\min\{F^{\frac{1}{2}}_{\min}, G^{\frac{1}{2}}_{\min}\}}. \tag{B.13}$$

Part 2-

$$\|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot\tilde{\mathcal{A}})\mathcal{T}^{-\frac{1}{2}}\| \leqslant \|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot(\mathcal{A}\bar{\mathcal{A}}))\mathcal{T}^{-\frac{1}{2}}\| + \|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot(\bar{\mathcal{A}}\mathcal{A}))\mathcal{T}^{-\frac{1}{2}}\|. \tag{B.14}$$

It is sufficient to show an upper-bound for the first term and the result would be true for the second term. Let J be the square matrix of order N with all entries one.

$$\begin{aligned}\|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot\mathcal{A}\bar{\mathcal{A}})\mathcal{T}^{-\frac{1}{2}}\| &= \|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot\mathcal{A}(J-\mathcal{A}))\mathcal{T}^{-\frac{1}{2}}\| \\ &= \|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot(\mathcal{A}J)-(A-\mathcal{A})\cdot\mathcal{A}\mathcal{A})\mathcal{T}^{-\frac{1}{2}}\| \\ &\leqslant \|\mathcal{T}^{-\frac{1}{2}}\mathcal{D}(A-\mathcal{A})\mathcal{T}^{-\frac{1}{2}}\| + \|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot\mathcal{A}\mathcal{A})\mathcal{T}^{-\frac{1}{2}}\| \\ &= \|\mathcal{T}^{-\frac{1}{2}}\mathcal{D}^{\frac{1}{2}}(A-\mathcal{A})\mathcal{D}^{\frac{1}{2}}\mathcal{T}^{-\frac{1}{2}}\| + \|\mathcal{T}^{-\frac{1}{2}}((A-\mathcal{A})\cdot\mathcal{A}\mathcal{A})\mathcal{T}^{-\frac{1}{2}}\|.\end{aligned} \tag{B.15}$$

For $i, j = 1, \cdots, N$ define $A^{ij} \in \{0,1\}^{N\times N}$ to be the matrices with 1 at positions ij and ji, and 0 everywhere else. We have,

$$\mathcal{T}^{-\frac{1}{2}}\mathcal{D}^{\frac{1}{2}}(A-\mathcal{A})\mathcal{D}^{\frac{1}{2}}\mathcal{T}^{-\frac{1}{2}} = \sum_{i=1}^{N}\sum_{j>i}^{N}\sqrt{\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}(A_{ij}-\mathcal{A}_{ij})A^{ij}.$$

The right hand side is a sum of independent symmetric matrices. So, we can apply Theorem 5 of Chung and Radcliffe (2011) to bound it. Let

$$M := \max_{ij=1,\cdots,N}\left\|\sqrt{\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}(A_{ij}-\mathcal{A}_{ij})A^{ij}\right\| \leqslant \max_{ij=1,\cdots,N}\sqrt{\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}, \tag{B.16}$$

and

$$\begin{aligned}\nu^2 &:= \left\|\sum_{i=1}^{N}\sum_{j>i}^{N}\operatorname{Var}\left(\sqrt{\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}(A_{ij}-\mathcal{A}_{ij})A^{ij}\right)\right\| \\ &= \left\|\sum_{i=1}^{N}\sum_{j>i}^{N}\left[\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}(\mathcal{A}_{ij}-\mathcal{A}_{ij}^2)A^{ii}\right]\right\| \\ &\leqslant \max_{i=1,\cdots,N}\left(\sum_{j=1}^{N}\left[\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}(\mathcal{A}_{ij}-\mathcal{A}_{ij}^2)\right]\right) \\ &\leqslant \max_{i=1,\cdots,N}\left(\sum_{j=1}^{N}\left[\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}\mathcal{A}_{ij}\right]\right) \leqslant \max_{ij=1,\cdots,N}\frac{\mathcal{D}_{ii}^2\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}.\end{aligned} \tag{B.17}$$

Take

$$\Delta := \max_{ij=1,\cdots,N} 2\sqrt{\frac{\mathcal{D}_{ii}^2\mathcal{D}_{jj}\ln(2N/\delta)}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}. \tag{B.18}$$

Note that

$$\begin{aligned}\Delta &= \max_{ij}\sqrt{\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}\sqrt{\frac{\mathcal{D}_{ii}^2\mathcal{D}_{jj}\ln(2N/\delta)}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}\\ &= \max_{ij}\frac{\mathcal{D}_{ii}^2\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}\sqrt{\frac{\ln(2N/\delta)}{\mathcal{D}_{ii}}} \leqslant \nu^2\sqrt{\frac{\ln(2N/\delta)}{\mathcal{D}_{\min}}}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Pr\left(\left\|\sum_{i=1}^{N}\sum_{j>i}^{N}\sqrt{\frac{\mathcal{D}_{ii}\mathcal{D}_{jj}}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}(A_{ij}-\mathcal{A}_{ij})A^{ij}\right\| \geqslant \Delta\right) &\leqslant 2N\exp\left(-\frac{\Delta^2}{2\nu^2+2M\Delta/3}\right) \\ &\leqslant \delta.\end{aligned} \tag{B.19}$$

For the second term of (B.15), we obtain

$$\mathcal{T}^{-\frac{1}{2}}\left((A-\mathcal{A})\cdot\mathcal{A}\mathcal{A}\right)\mathcal{T}^{-\frac{1}{2}} = \sum_{i=1}^{N}\sum_{j>i}^{N}\sqrt{\frac{1}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}(A_{ij}-\mathcal{A}_{ij})(\sum_{k=1}^{N}\mathcal{A}_{ik}\mathcal{A}_{kj})A^{ij}.$$

Because $|\sum_k \mathcal{A}_{ik}\mathcal{A}_{kj}| \leqslant \sqrt{\mathcal{D}_{ii}\mathcal{D}_{jj}}$, we obtain the same bound as (B.19).

$$\begin{aligned}\Pr\left(\left\|\sum_{i=1}^{N}\sum_{j>i}^{N}\sqrt{\frac{1}{\mathcal{T}_{ii}\mathcal{T}_{jj}}}(A_{ij}-\mathcal{A}_{ij})(\sum_{k=1}^{N}\mathcal{A}_{ik}\mathcal{A}_{kj})A^{ij}\right\| \geqslant \Delta\right) &\leqslant 2N\exp\left(-\frac{\Delta^2}{2\nu^2+2M\Delta/3}\right)\\ &\leqslant \delta.\end{aligned}$$

In addition,

$$\begin{aligned}\mathcal{T}_{ii} &= \sum_j \mathcal{A}_{ij} \sum_k \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj} + \bar{\mathcal{A}}_{ik}\mathcal{A}_{kj} \\ &= \sum_j \mathcal{A}_{ij}(F_{ij} + G_{ij}) \\ &\geqslant \sum_j \mathcal{A}_{ij} c_1 \mathcal{D}_{ii} \geqslant c_1 \mathcal{D}_{ii}^2,\end{aligned}$$

where the ineqaulity follows from the assumption that $F_{ij} + G_{ij} > c_1\mathcal{D}_{ii}$ for all i and $j \in 1, \cdots, N$.

So

$$\|\mathcal{T}^{-\frac{1}{2}}((A - \mathcal{A}) \cdot \mathcal{A}\bar{\mathcal{A}})\mathcal{T}^{-\frac{1}{2}}\| \leqslant 4\sqrt{\frac{\ln \frac{N}{\delta}}{c_1 \mathcal{D}_{\min}}}.$$

It follows from (B.14) that

$$\|\mathcal{T}^{-\frac{1}{2}}((A - \mathcal{A}) \cdot \tilde{\mathcal{A}})\mathcal{T}^{-\frac{1}{2}}\| \leqslant 8\sqrt{\frac{\ln \frac{N}{\delta}}{c_1 \mathcal{D}_{\min}}} \tag{B.20}$$

Part 3- First we bound $|T_{ii} - \mathcal{T}_{ii}|$ and then $\|\mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}} - I\|$.

$$\begin{aligned}|T_{ii} - \mathcal{T}_{ii}| &= |r_i(\tilde{A} \cdot A) - r_i(\tilde{\mathcal{A}} \cdot \mathcal{A})| \\ &\leqslant |r_i((A\bar{A}) \cdot A) - r_i((\mathcal{A}\bar{\mathcal{A}}) \cdot \mathcal{A})| + |r_i((\bar{A}A) \cdot A) - r_i((\bar{\mathcal{A}}\mathcal{A}) \cdot \mathcal{A})|\end{aligned}$$

The first term

$$\begin{aligned}|r_i((A\bar{A}) \cdot A) - r_i((\mathcal{A}\bar{\mathcal{A}}) \cdot \mathcal{A})| &= |\sum_j A_{ij} \sum_k A_{ik}\bar{A}_{kj} - \sum_j \mathcal{A}_{ij} \sum_k \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}| \\ &\leqslant \sum_j A_{ij} |\sum_k A_{ik}\bar{A}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}| \\ &+ |\sum_j (A_{ij} - \mathcal{A}_{ij}) \sum_k \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}|\end{aligned} \tag{B.21}$$

To bound the first term of Inequality (B.21), we use (B.8) and (B.9). So with probability at least $1-\delta$,

$$\sum_j A_{ij}|\sum_k A_{ik}\bar{A}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}| \leqslant 10\sum_j \mathcal{A}_{ij}\sqrt{F_{ij}\ln\frac{N}{\delta}}. \tag{B.22}$$

The second term

$$\begin{aligned}
|\sum_j (A_{ij}-\mathcal{A}_{ij})\sum_k \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}| &= |\sum_k\sum_j (A_{ij}-\mathcal{A}_{ij})\mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}| \\
&\leqslant \sum_k \mathcal{A}_{ik}|\sum_j A_{ij}\bar{\mathcal{A}}_{kj} - \mathcal{A}_{ij}\bar{\mathcal{A}}_{kj}| \\
&= \sum_k \mathcal{A}_{ik}|\sum_j A_{ij}\bar{\mathcal{A}}_{jk} - \mathcal{A}_{ij}\bar{\mathcal{A}}_{jk}| \\
&= \sum_j \mathcal{A}_{ij}|\sum_k A_{ik}\bar{\mathcal{A}}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}|.
\end{aligned}$$

$E[A_{ik}\bar{\mathcal{A}}_{kj}] = \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}$. In addition, we can obtain the same upper bound for variance to use . So with probability at least $1-\delta$,

$$\sum_j \mathcal{A}_{ij}|\sum_k A_{ik}\bar{\mathcal{A}}_{kj} - \mathcal{A}_{ik}\bar{\mathcal{A}}_{kj}| \leqslant 10\sum_j \mathcal{A}_{ij}\sqrt{F_{ij}\ln\frac{N}{\delta}}. \tag{B.23}$$

From (B.22) and (B.23), we have

$$|r_i((A\bar{A})\cdot A) - r_i((\mathcal{A}\bar{\mathcal{A}})\cdot\mathcal{A})| \leqslant 20\sum_j \mathcal{A}_{ij}\sqrt{F_{ij}\ln\frac{N}{\delta}}. \tag{B.24}$$

Following the same steps,

$$|r_i((\bar{A}A)\cdot A) - r_i((\bar{\mathcal{A}}\mathcal{A})\cdot\mathcal{A})| \leqslant 20\sum_j \mathcal{A}_{ij}\sqrt{G_{ij}\ln\frac{N}{\delta}}. \tag{B.25}$$

Therefore,

$$|T_{ii} - \mathcal{T}_{ii}| \leqslant 40 \sum_{j} \mathcal{A}_{ij}(\sqrt{F_{ij} \ln \frac{N}{\delta}} + \sqrt{G_{ij} \ln \frac{N}{\delta}}).$$

$$\begin{aligned}
\|\mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}} - I\| &\leqslant \max_{i=1,\cdots,N} \left| \sqrt{\frac{T_{ii}}{\mathcal{T}_{ii}}} - 1 \right| \\
&\leqslant \max_{i=1,\cdots,N} \left| \frac{T_{ii}}{\mathcal{T}_{ii}} - 1 \right| \\
&\leqslant \max_{i=1,\cdots,N} \frac{40 \sum_j \mathcal{A}_{ij}(\sqrt{F_{ij} \ln \frac{N}{\delta}} + \sqrt{G_{ij} \ln \frac{N}{\delta}})}{\sum_j \mathcal{A}_{ij}(F_{ij} + G_{ij})} \leqslant \frac{40 \ln^{\frac{1}{2}} \frac{N}{\delta}}{\min\{G_{\min}^{\frac{1}{2}}, F_{\min}^{\frac{1}{2}}\}}.
\end{aligned}$$

Furthermore,

$$\|\mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}}\| \leqslant 1 + \frac{40 \ln^{\frac{1}{2}} \frac{N}{\delta}}{\min\{G_{\min}^{\frac{1}{2}}, F_{\min}^{\frac{1}{2}}\}} < 2, \tag{B.26}$$

where the last inequality follows from the Theorem's assumptions.

Let the Laplacian matrix $L^{ac} := T^{-\frac{1}{2}} \tilde{W} T^{-\frac{1}{2}}$. So

$$\begin{aligned}
\|T^{-\frac{1}{2}} \tilde{W} T^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}} \tilde{W} \mathcal{T}^{-\frac{1}{2}}\| &= \|T^{-\frac{1}{2}} \tilde{W} T^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}} T^{-\frac{1}{2}} \tilde{W} T^{-\frac{1}{2}} T^{+\frac{1}{2}} \mathcal{T}^{-\frac{1}{2}}\| \\
&= \|I - L^{ac} - \mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}} \{I - L^{ac}\} T^{+\frac{1}{2}} T^{-\frac{1}{2}}\| \\
&= \|\{\mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}} - I\}\{I - L^{ac}\} T^{+\frac{1}{2}} \mathcal{T}^{-\frac{1}{2}} + \{I - L^{ac}\}\{I - T^{+\frac{1}{2}} \mathcal{T}^{-\frac{1}{2}}\}\| \\
&\leqslant \|\mathcal{T}^{-\frac{1}{2}} T^{+\frac{1}{2}} - I\| \cdot \|T^{+\frac{1}{2}} \mathcal{T}^{-\frac{1}{2}}\| + \|I - T^{+\frac{1}{2}} \mathcal{T}^{-\frac{1}{2}}\|,
\end{aligned}$$

where the inequality follows from the fact that $\|I - L^{ac}\| \leqslant 1$. Now

$$\|T^{-\frac{1}{2}} \tilde{W} T^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}} \tilde{W} \mathcal{T}^{-\frac{1}{2}}\| \leqslant \frac{120 \ln^{\frac{1}{2}} \frac{N}{\delta}}{\min\{G_{\min}^{\frac{1}{2}}, F_{\min}^{\frac{1}{2}}\}}$$

Part 4- Let $\epsilon := 10\delta$. Hence, based on **Part 1**, **2**, and **3** results,

$$\|T^{-\frac{1}{2}}\tilde{W}T^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}}\tilde{\mathcal{W}}\mathcal{T}^{-\frac{1}{2}}\| \leqslant \frac{10\ln^{\frac{1}{2}}\frac{N}{\delta}}{\min\{G_{\min}^{\frac{1}{2}}, F_{\min}^{\frac{1}{2}}\}} + \frac{8\ln^{\frac{1}{2}}\frac{N}{\delta}}{c_1^{\frac{1}{2}}\mathcal{D}_{\min}^{\frac{1}{2}}} + \frac{120\ln^{\frac{1}{2}}\frac{N}{\delta}}{\min\{G_{\min}^{\frac{1}{2}}, F_{\min}^{\frac{1}{2}}\}}.$$

Note that $G_{\min} = F_{\min}$ and $D_{\min} \geqslant F_{\min}$. So

$$\|T^{-\frac{1}{2}}\tilde{W}T^{-\frac{1}{2}} - \mathcal{T}^{-\frac{1}{2}}\tilde{\mathcal{W}}\mathcal{T}^{-\frac{1}{2}}\| \leqslant \frac{138\ln^{\frac{1}{2}}\frac{10N}{\epsilon}}{\min\{c_1^{\frac{1}{2}}\mathcal{D}_{\min}^{\frac{1}{2}}, F_{\min}^{\frac{1}{2}}\}} \leqslant \frac{138\ln^{\frac{1}{2}}\frac{10N}{\epsilon}}{c_1^{\frac{1}{2}}F_{\min}^{\frac{1}{2}}},$$

with probability at least $1 - \epsilon$. □

Proof of Theorem 3.9. Define f_j and f_j^{ac} as the jth eigenvectors of P_A and $P_{\tilde{W}}$ with respect to the inner product $\langle\cdot,\cdot\rangle_{\pi}$ and $\langle\cdot,\cdot\rangle_{\pi^{ac}}$ respectively. From Proposition 3.1, it is enough to show that

$$\sum_{j=2}^{|\mathcal{V}|}\langle y, f_j^{ac}\rangle_{\pi^{ac}}^2\lambda_j(P_{\tilde{W}})^t < \sum_{j=2}^{|\mathcal{V}|}\langle y, f_j\rangle_{\pi}^2\lambda_j(P_A)^t.$$

Step 1 of the proof shows that the above holds true in the population, i.e. comparing Markov chains on $P_{\tilde{\mathcal{W}}}$ and $P_{\mathcal{A}}$. Step 2 of the proof shows that the sample versions converge to the population versions.

Part 1- From Lemma B.3, for $i = 2, \ldots, K$,

$$\lambda_i(P_{\mathcal{A}}) = \lambda_2(P_{\mathcal{A}}) \quad \text{and} \quad \lambda_i(P_{\tilde{\mathcal{W}}}) = \lambda_2(P_{\tilde{\mathcal{W}}}). \tag{B.27}$$

Moreover, for $i > K, \lambda_i(P_{\mathcal{A}}) = \lambda_i(P_{\tilde{\mathcal{W}}}) = 0$. Under the theorem conditions, $P_{\tilde{\mathcal{W}}}$ and $P_{\mathcal{A}}$ have the same stationary distribution; refer to this as $\bar{\pi}$ (in fact, this distribution is uniform on the nodes). Define $\bar{f}_j$ and $\bar{f}_j^{ac}$ as the jth eigenvectors of $P_{\mathcal{A}}$ and $P_{\tilde{\mathcal{W}}}$ with respect to $\langle\cdot,\cdot\rangle_{\bar{\pi}}$ respectively.

So,

$$\sum_{j=2}^{|\mathcal{V}|}\langle y, \bar{f}_j\rangle_{\bar{\pi}}^2\lambda_j(P_{\mathcal{A}})^t = \lambda_2(P_{\mathcal{A}})^t\sum_{j=2}^{K}\langle y, \bar{f}_j\rangle_{\bar{\pi}}^2.$$

Similarly,

$$\sum_{j=2}^{|\mathcal{V}|} \langle y, \bar{f}_j^{ac} \rangle_{\bar{\pi}}^2 \lambda_j(P_{\tilde{\mathcal{W}}})^t = \lambda_2(P_{\tilde{\mathcal{W}}})^t \sum_{j=2}^{K} \langle y, \bar{f}_j^{ac} \rangle_{\bar{\pi}}^2.$$

Proposition 3.5 shows that $\lambda_2(P_{\tilde{\mathcal{W}}})^t + \epsilon < \lambda_2(P_{\mathcal{A}})^t$, where ϵ does not change asymptotically as $|\mathcal{V}|$ grows. So, part 1 of the proof will be finished after showing that $\sum_{j=2}^{K} \langle y, \bar{f}_j^{ac} \rangle_{\bar{\pi}}^2 = \sum_{j=2}^{K} \langle y, \bar{f}_j \rangle_{\bar{\pi}}^2$. To compare these terms, note that the construction of the eigenvalues in the proof of Lemma B.3 shows that the span of sets $\{\bar{f}_j^{ac} \cdot \bar{\pi}^{\frac{1}{2}} : j = 1, \dots, K\}$ and $\{\bar{f}_j \cdot \bar{\pi}^{\frac{1}{2}} : j = 1, \dots, K\}$ are identical. Therefore, Parseval's Identity implies,

$$\sum_{j=1}^{K} \langle y, \bar{f}_j^{ac} \rangle_{\bar{\pi}}^2 = \sum_{j=1}^{K} \langle y, \bar{f}_j \rangle_{\bar{\pi}}^2.$$

Because these are the top eigenvectors of Markov transition matrices, $\bar{f}_1^{ac} = \bar{f}_1 = \mathbf{1}$. Thus,

$$\sum_{j=2}^{K} \langle y, \bar{f}_j^{ac} \rangle_{\bar{\pi}}^2 = \sum_{j=2}^{K} \langle y, \bar{f}_j \rangle_{\bar{\pi}}^2.$$

This first part of the proof shows that

$$\sum_{j=2}^{|\mathcal{V}|} \langle y, \bar{f}_j^{ac} \rangle_{\bar{\pi}}^2 \lambda_j(P_{\tilde{\mathcal{W}}})^t + \epsilon < \sum_{j=2}^{|\mathcal{V}|} \langle y, \bar{f}_j \rangle_{\bar{\pi}}^2 \lambda_j(P_{\mathcal{A}})^t.$$

Part 2- To ease notation, let $\lambda_j := \lambda_j(P_A)$ and $\bar{\lambda}_j := \lambda_j(P_{\mathcal{A}})$. This part of the proof shows

$$\left| \sum_{j=2}^{|\mathcal{V}|} \langle y, \bar{f}_j \rangle_{\bar{\pi}}^2 \bar{\lambda}_j^t - \langle y, f_j \rangle_{\pi}^2 \lambda_j^t \right| \to 0, \text{ as the size of the graph } |\mathcal{V}| = N \text{ increases.}$$

The proof for the anti-cluster random walk follows the same steps.

$$\left|\sum_{j=2}^{|\mathcal{V}|}\langle y,\bar{f}_j\rangle_{\bar{\pi}}^2\bar{\lambda}_j^t-\langle y,f_j\rangle_{\pi}^2\lambda_j^t\right| = \left|\sum_{j=2}^{|\mathcal{V}|}\langle y,\bar{f}_j\rangle_{\bar{\pi}}^2\bar{\lambda}_j^t-\langle y,f_j\rangle_{\pi}^2(\bar{\lambda}_j^t+(\lambda_j^t-\bar{\lambda}_j^t))\right| \tag{B.28}$$

$$\leqslant \left|\sum_{j=2}^{K}\bar{\lambda}_j^t\left(\langle y,\bar{f}_j\rangle_{\bar{\pi}}^2-\langle y,f_j\rangle_{\pi}^2\right)\right| + \left|\sum_{j=2}^{|\mathcal{V}|}\langle y,f_j\rangle_{\pi}^2\left|\lambda_j^t-\bar{\lambda}_j^t\right|\right| \tag{B.29}$$

$$\leqslant \bar{\lambda}_2^t\cdot\left|\sum_{j=2}^{K}\langle y,f_j\rangle_{\pi}^2-\langle y,\bar{f}_j\rangle_{\bar{\pi}}^2\right| + \max_j\left|\lambda_j^t-\bar{\lambda}_j^t\right|\cdot\langle y,y\rangle_{\pi}^2. \tag{B.30}$$

Note that since y is a bounded function, $\langle y,y\rangle_{\pi}^2$ is bounded. Hence, Theorem 1 with $\epsilon = 1/N^2$ and the Borel-Cantelli Theorem imply that the second term congress to zero almost surely as N increases.

To show that the first term in (B.30) converges to zero, we must study the convergence of the eigenspace. Let $\cdot$ denote element wise multiplication. Let $\pi^{\frac{1}{2}}$ be vector with the elements $\sqrt{\pi_i}$. Let $diag(\pi^{\frac{1}{2}})$ be a diagonal matrix with $\pi^{\frac{1}{2}}$ down the diagonal. For some constant c,

$$\mathcal{D}^{-\frac{1}{2}}\mathcal{A}\mathcal{D}^{-\frac{1}{2}}(\bar{f}_j\cdot\bar{\pi}^{\frac{1}{2}}) = \mathcal{D}^{-\frac{1}{2}}\mathcal{A}\mathcal{D}^{-\frac{1}{2}}diag(\bar{\pi}^{\frac{1}{2}})\bar{f}_j = \mathcal{D}^{-\frac{1}{2}}\mathcal{A}cI\bar{f}_j = c\bar{\lambda}_j f_j.$$

Together with the fact that $\langle\bar{f}_j\cdot\bar{\pi}^{\frac{1}{2}},\bar{f}_i\cdot\bar{\pi}^{\frac{1}{2}}\rangle\in\{0,1\}$ is equal to one if and only if $i=j$, this shows that $\bar{f}_j\cdot\bar{\pi}^{\frac{1}{2}}$ forms an orthonormal basis of the eigenspace of $\mathcal{D}^{-\frac{1}{2}}\mathcal{A}\mathcal{D}^{-\frac{1}{2}}$. Similarly for $f_j\cdot\pi^{\frac{1}{2}}$ and $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

Let $\bar{V}\in\mathbb{R}^{N\times(K-1)}$ and $V\in\mathbb{R}^{N\times(K-1)}$ be matrices with columns defined by $\bar{V}_{\cdot j} := \bar{f}_{j+1}\cdot\bar{\pi}^{\frac{1}{2}}$ and $V_{\cdot j} := f_{j+1}\cdot\pi^{\frac{1}{2}}$ for $j=1,\cdots,K-1$. Note that the columns of V and $\bar{V}$ are orthonormal. Furthermore, define the corresponding orthogonal projection matrices $\bar{Q} = \bar{V}\bar{V}^{\mathsf{T}} \in \mathbb{R}^{N \times N}$ and $Q = VV^{\mathsf{T}} \in \mathbb{R}^{N \times N}$.

$$\left| \sum_{j=2}^{K} \langle y, f_j \rangle_{\pi}^2 - \langle y, \bar{f}_j \rangle_{\bar{\pi}}^2 \right| = \left| \sum_{j=2}^{K} \langle y \cdot \pi^{\frac{1}{2}}, f_j \cdot \pi^{\frac{1}{2}} \rangle^2 - \langle y \cdot \bar{\pi}^{\frac{1}{2}}, \bar{f}_j \cdot \bar{\pi}^{\frac{1}{2}} \rangle^2 \right| \tag{B.31}$$

$$= \left| \left\| V^{\mathsf{T}} \left(y \cdot \pi^{\frac{1}{2}} \right) \right\|^2 - \left\| \bar{V}^{\mathsf{T}} \left(y \cdot \bar{\pi}^{\frac{1}{2}} \right) \right\|^2 \right| \tag{B.32}$$

$$= \left| \left\| Q \left(y \cdot \pi^{\frac{1}{2}} \right) \right\|^2 - \left\| \bar{Q} \left(y \cdot \bar{\pi}^{\frac{1}{2}} \right) \right\|^2 \right| \tag{B.33}$$

$$\leqslant \left\| Q \left(y \cdot \pi^{\frac{1}{2}} \right) - \bar{Q} \left(y \cdot \bar{\pi}^{\frac{1}{2}} \right) \right\|^2 \tag{B.34}$$

$$\leqslant \left\| (Q - \bar{Q}) \left(y \cdot \pi^{\frac{1}{2}} \right) \right\|^2 + \left\| \bar{Q} \left(y \cdot \pi^{\frac{1}{2}} - y \cdot \bar{\pi}^{\frac{1}{2}} \right) \right\|^2 \tag{B.35}$$

$$\leqslant \left\| Q - \bar{Q} \right\|^2 \cdot \langle y, y \rangle_{\pi} + \left\| \bar{Q} \left(y \cdot \left(\pi^{\frac{1}{2}} - \bar{\pi}^{\frac{1}{2}} \right) \right) \right\|^2 \tag{B.36}$$

First we show convergence of the first term. From Davis-Kahan Sin theorem (e.g. Yu et al., 2015, Theorem 1), it follows that

$$\frac{\left\| D^{-\frac{1}{2}} A D^{-\frac{1}{2}} - \mathcal{D}^{-\frac{1}{2}} \mathcal{A} \mathcal{D}^{-\frac{1}{2}} \right\|}{\delta} \geqslant \left\| \sin \Theta \left(V, \bar{V} \right) \right\| = \left\| Q - \bar{Q} \right\|, \tag{B.37}$$

where the equality follows from (Stewart, 1990, Theorem 5.5 pp. 43) and

$$\delta = \min \left\{ |\bar{\lambda}_{K+1} - \lambda_K|, |\bar{\lambda}_1 - \lambda_2| \right\}.$$

$\bar{\lambda}_{K+1} = 0$ and $\bar{\lambda}_1 = 1$. So $\delta = \min\{\lambda_K, 1 - \lambda_2\}$. Furthermore, $\lambda_K > |\bar{\lambda}_K - |\lambda_K - \bar{\lambda}_K||$ and $\lambda_2 > |\bar{\lambda}_2 - |\lambda_2 - \bar{\lambda}_2||$. Recall $\bar{\lambda}_2 = \bar{\lambda}_K$. Then, Theorem 3.7 implies $|\lambda_j - \bar{\lambda}_j| \to 0$ a.s., which is less than $\bar{\lambda}_2$. So, $\delta > \frac{1}{2}\bar{\lambda}_2$. Theorem 3.7 also implies that the numerator on the left hand side of Equation (B.37) coverages to zero a.s..

Now, we focus on the second term.

$$\left\| \bar{Q}^{\mathsf{T}} \left(y \cdot \left(\pi^{\frac{1}{2}} - \bar{\pi}^{\frac{1}{2}} \right) \right) \right\|^2 \leqslant \left\| y \cdot \left(\pi^{\frac{1}{2}} - \bar{\pi}^{\frac{1}{2}} \right) \right\|^2 \leqslant \|y\|_{\infty}^2 \cdot \left\| \pi^{\frac{1}{2}} - \bar{\pi}^{\frac{1}{2}} \right\|^2. \tag{B.38}$$

Based on the theorem assumptions $\|y\|_{\infty}$ is bounded. So we just need to show the convergence of the second term. Note that $\left\|\pi^{\frac{1}{2}}\right\| = \left\|\bar{\pi}^{\frac{1}{2}}\right\| = 1$. So

$$\left\|\pi^{\frac{1}{2}} - \bar{\pi}^{\frac{1}{2}}\right\| = 2\sin\frac{\Theta(\pi^{\frac{1}{2}}, \bar{\pi}^{\frac{1}{2}})}{2}.$$

Recall that $\pi^{\frac{1}{2}}$ and $\bar{\pi}^{\frac{1}{2}}$ are leading eigenvectors of the sample and population Laplacian matrices respectively. Then, it follows from Davis-Kahan Sin theorem and concentration of eigenvalues of the Laplacian matrices that $\left|\sin\Theta\left(\pi^{\frac{1}{2}}, \bar{\pi}^{\frac{1}{2}}\right)\right| \to 0$. Therefore, we conclude that $\left\|\pi^{\frac{1}{2}} - \bar{\pi}^{\frac{1}{2}}\right\| \to 0$ a.s..

□

references

Adamczak, Radosław, and Piotr Miłoś. 2015. CLT for Ornstein-Uhlenbeck branching particle system. *Electron. J. Probab.* 20(42):1–35.

Ané, Cécile. 2008. Analysis of comparative data with hierarchical autocorrelation. *Annals of Applied Statistics* 2(3):1078–1102.

Ané, Cécile, Lam Si Tung Ho, and Sebastien Roch. 2015. Phase transition on the convergence rate of parameter estimation under an Ornstein-Uhlenbeck diffusion on a tree. `http://arxiv.org/abs/1406.1568`, accessed 2015-12-24.

Bartoszek, K., and S. Sagitov. 2015. Phylogenetic confidence intervals for the optimal trait value. *Journal of Applied Probability* 52(4).

Bartoszek, Krzysztof, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F. Hansen. 2012. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology* 314:204–215.

Bassetti, Federico, Persi Diaconis, et al. 2006. Examples comparing importance sampling and the Metropolis algorithm. *Illinois Journal of Mathematics* 50(1-4): 67–91.

Bastide, Paul, Mahendra Mariadassou, and Stéphane Robin. 2015. Detection of adaptive shifts on phylogenies using shifted stochastic processes on a tree. `http://arxiv.org/abs/1508.00225`, accessed 2015-12-24.

Beaulieu, Jeremy M., Dwueng-Chwuan Jhwueng, Carl Boettiger, and Brian C. O'Meara. 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* 66:2369–2383.

Benjamini, Itai, and Yuval Peres. 1994. Markov chains indexed by trees. *The Annals of Probability* 219–243.

Bininda-Emonds, Olaf, Marcel Cardillo, Kate E. Jones, Ross D. E. MacPhee, Robin M. D. Beck, Richard Grenyer, Samantha A. Price, Rutger A. Vos, John L. Gittleman, and Andy Purvis. 2007. The delayed rise of present-day mammals. *Nature* 446(7135):507–512.

Bühlmann, Peter, and Sara Van De Geer. 2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Butler, Marguerite A., and Aaron A. King. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist* 164(6):683–695.

Chung, Fan, and Mary Radcliffe. 2011. On the spectra of general random graphs. *Electronic Journal of Combinatorics* 18(1):215–229.

Chung, Fan RK. 1997. *Spectral graph theory*, vol. 92. American Mathematical Soc.

Clavel, Julien, Gilles Escarguel, and Gildas Merceron. 2015. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution* 6(11):1311–1319.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 39(1):1–38.

Eastman, Jonathan M, Michael E Alfaro, Paul Joyce, Andrew L Hipp, and Luke J Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65(12):3578–3589.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least angle regression. *The Annals of statistics* 32(2):407–499.

Eldar, Yonina C, and Gitta Kutyniok. 2012. *Compressed sensing: theory and applications*. Cambridge University Press.

Felsenstein, Joseph. 1985. Phylogenies and the comparative method. *The American Naturalist* 125(1):1–15.

Freckleton, Robert P, and Paul H Harvey. 2006. Detecting non-Brownian trait evolution in adaptive radiations. *PLoS Biol* 4(11):e373.

Gile, Krista J. 2011. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association* 106(493).

Gile, Krista J, and Mark S Handcock. 2010. Respondent-driven sampling: An assessment of current methodology. *Sociological methodology* 40(1):285–327.

———. 2011. Network model-assisted inference from respondent-driven sampling data. *arXiv preprint arXiv:1108.0298*.

Gile, Krista J, Lisa G Johnston, and Matthew J Salganik. 2015. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Girvan, Michelle, and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99(12):7821–7826.

Goel, Sharad, and Matthew J Salganik. 2009. Respondent-driven sampling as markov chain monte carlo. *Statistics in medicine* 28(17):2202–2229.

———. 2010. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences* 107(15):6743–6747.

Goodman, Leo A. 1961. Snowball sampling. *The annals of mathematical statistics* 148–170.

Handcock, Mark S, and Krista J Gile. 2011. Comment: On the concept of snowball sampling. *Sociological Methodology* 41(1):367–371.

Hansen, Thomas F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51(5):1341–1351.

Hansen, Thomas F., J. Pienaar, and Steven Hecht Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* 62(8):1965–1977.

Harmon, Luke J., Jonathan B. Losos, T. Jonathan Davies, Rosemary G. Gillespie, John L. Gittleman, W. Bryan Jennings, Kenneth H. Kozak, Mark A. McPeek, Franck Moreno-Roark, Thomas J. Near, Andy Purvis, Robert E. Ricklefs, Dolph Schluter, James A. Schulte II, Ole Seehausen, Brian L. Sidlauskas, Omar Torres-Carvajal, Jason T. Weir, and Arne Ø. Mooers. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64(8):2385–2396.

Heckathorn, Douglas D. 1997. Respondent-driven sampling: A new approach to the study of hidden populations. *Social problems* 174–199.

———. 2002. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social problems* 49(1):11–34.

———. 2007. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology* 37(1): 151–207.

———. 2011. Comment: Snowball versus respondent-driven sampling. *Sociological methodology* 41(1):355–366.

Ho, Lam Si Tung, and Cécile Ané. 2013. Asymptotic theory with hierarchical autocorrelation: Ornstein-Uhlenbeck tree models. *Annals of Statistics* 41(2):957–981.

Ho, Lam Si Tung, and Cécile Ané. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution* 5(11): 1133–1146.

Ho, Lam Si Tung, and Cécile Ané. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* 63(3):397–408.

Holland, P.W., K.B. Laskey, and S. Leinhardt. 1983. Stochastic blockmodels: First steps. *Social Networks* 5(2):109–137.

Hopkins, Melanie J., and Scott Lidgard. 2012. Evolutionary mode routinely varies among morphological traits within fossil species lineages. *Proceedings of the National Academy of Sciences* 109(50):20520–20525.

Hunt, Gene, Michael A. Bell, and Matthew P. Travis. 2008. Evolution toward a new adaptive optimum: Phenotypic evolution in a fossil stickleback lineage. *Evolution* 62:700–710.

Ingram, Travis, and Yoshiaki Kai. 2014. The geography of morphological convergence in the radiations of Pacific *Sebastes* rockfishes. *The American Naturalist* 184(5): E115–E131.

Ingram, Travis, and D Luke Mahler. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. *Methods in Ecology and Evolution* 4(5):416–425.

Ives, Anthony R., Peter E. Midford, and Theodore Garland, Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology* 56:252–270.

Khabbazian, M., B. Hanlon, Z. Russek, and K. Rohe. 2016a. Novel sampling design for respondent-driven sampling. *ArXiv e-prints, 7, submitted to Journal of the American Statistical Association*. `1606.00387`.

Khabbazian, Mohammad, Ricardo Kriebel, Karl Rohe, and Cécile Ané. 2016b. Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution* 7(7):811–824.

Kriebel, Ricardo, Fabián A. Michelangeli, and Lawrence M. Kelly. 2015. Discovery of unusual anatomical and continuous characters in the evolutionary history of *Conostegia* (Miconieae: Melastomataceae). *Molecular Phylogenetics and Evolution* 82, Part A:289 – 313.

Levin, David Asher, Yuval Peres, and Elizabeth Lee Wilmer. 2009. *Markov chains and mixing times*. American Mathematical Soc.

Losos, Jonathan B. 2009. *Lizards in an evolutionary tree: ecology and adaptive radiation of anoles*, vol. 10. Univ of California Press.

Losos, Jonathan B, Todd R Jackman, Allan Larson, Kevin de Queiroz, and Lourdes Rodrıguez-Schettino. 1998. Contingency and determinism in replicated adaptive radiations of island lizards. *Science* 279(5359):2115–2118.

Lynch, Michael. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45(5):1065–1080.

Mahler, D. Luke, and Travis Ingram. 2014. *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice*, chap. Phylogenetic comparative methods for studying clade-wide convergence, 425–450. Springer-Verlag Berlin Heidelberg.

Mahler, D Luke, Travis Ingram, Liam J Revell, and Jonathan B Losos. 2013. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science* 341(6143):292–295.

Malekinejad, Mohsen, Lisa Grazina Johnston, Carl Kendall, Ligia Regina Franco Sansigolo Kerr, Marina Raven Rifkin, and George W Rutherford. 2008.

Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior* 12(1):105–130.

Massart, Pascal. 2007. *Concentration inequalities and model selection*, vol. 1896 of *Ecole d'Eté de Probabilités de Saint-Flour*. Springer-Verlag Berlin Heidelberg.

Mathias, Roy. 1990. The spectral norm of a nonnegative matrix. *Linear Algebra and its Applications* 139:269–284.

McCreesh, Nicky, Simon Frost, Janet Seeley, Joseph Katongole, Matilda Ndagire Tarsh, Richard Ndunguse, Fatima Jichi, Natasha L Lunel, Dermot Maher, Lisa G Johnston, et al. 2012. Evaluation of respondent-driven sampling. *Epidemiology (Cambridge, Mass.)* 23(1):138.

Mouw, Ted, and Ashton M. Verdery. 2012. Network sampling with memory: A proposal for more efficient sampling from social networks. *Sociological Methodology* 42(1):206–256. `http://smx.sagepub.com/content/42/1/206.full.pdf+html`.

O'Meara, Brian C, Cécile Ané, Michael J Sanderson, and Peter C Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60(5):922–933.

Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401(6756):877–884.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Pennell, Matthew W., Richard G. FitzJohn, William K. Cornwell, and Luke J. Harmon. 2015. Model adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist* 186(2):E33–E50.

Rabosky, Daniel L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PloS ONE* 9(2):e89543.

Rabosky, Daniel L., Michael Grundler, Carlos Anderson, Pascal Title, Jeff J. Shi, Joseph W. Brown, Huateng Huang, and Joanna G. Larson. 2014. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution* 5(7):701–707.

Revell, Liam J. 2009. Size-correction and principal components for interspecific comparative studies. *Evolution* 63(12):3258–3268.

Rohe, Karl. 2015. Network driven sampling; a critical threshold for design effects. *arXiv preprint arXiv:1505.05461*.

Rohe, Karl, Sourav Chatterjee, and Bin Yu. 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* 1878–1915.

Rose, Jeffrey, Ricardo Kriebel, and Ken Sytsma. 2015. Shape analysis of moss sporophytes (Bryophyta) provide insights into land plant evolution. *In review*.

Salganik, Matthew J. 2012. Commentary: Respondent-driven sampling in the real world. *Epidemiology* 23(1):148–150.

Salganik, Matthew J, and Douglas D Heckathorn. 2004. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology* 34(1):193–240.

Scales, Jeffrey A, Aaron A King, and Marguerite A. Butler. 2009. Running for your life or running for your dinner: What drives fiberâŁłtype evolution in lizard locomotor muscles? *The American Naturalist* 173(5):543–553.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464.

Stack, J Conrad, Luke J Harmon, and Brian O'Meara. 2011. RBrownie: an R package for testing hypotheses about rates of evolutionary change. *Methods in Ecology and Evolution* 2(6):660–662.

Stewart, Gilbert W. 1990. Matrix perturbation theory.

Stone, Eric A. 2011. Why the phylogenetic regression appears robust to tree misspecification. *Systematic Biology* 60(3):245–260.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 267–288.

Tibshirani, Ryan J., and Jonathan Taylor. 2011. The solution path of the generalized lasso. *Ann. Statist.* 39(3):1335–1371.

Uyeda, Josef C, Daniel S Caetano, and Matthew W Pennell. 2015. Comparative analysis of principal components can be misleading. *Systematic Biology* 64(4): 677–689.

Uyeda, Josef C, and Luke J Harmon. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology* 63(6):902–918.

Van De Geer, Sara A, Peter Bühlmann, et al. 2009. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* 3:1360–1392.

Verdery, Ashton M., Ted Mouw, Shawn Bauldry, and Peter J. Mucha. 2015. Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE* 10(12):e0145296.

Volz, Erik, and Douglas D Heckathorn. 2008. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics* 24(1):79.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4):395–416.

Watts, Duncan J, and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *nature* 393(6684):440–442.

Wejnert, Cyprian. 2009. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological methodology* 39(1):73–116.

White, Richard G, Avi J Hakim, Matthew J Salganik, Michael W Spiller, Lisa G Johnston, Ligia Kerr, Carl Kendall, Amy Drake, David Wilson, Kate Orroth, et al. 2015. Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies:"STROBE-RDS" statement. *Journal of clinical epidemiology* 68(12):1463–1471.

White, Richard G, Amy Lansky, Sharad Goel, David Wilson, Wolfgang Hladik, Avi Hakim, and Simon DW Frost. 2012. Respondent driven sampling–where we are and where should we be going? *Sexually transmitted infections* 88(6):397–399.

Wickett, Norman J., Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, Michael S. Barker, J. Gordon Burleigh, Matthew A. Gitzendanner, Brad R. Ruhfel, Eric Wafula, Joshua P. Der, Sean W. Graham, Sarah Mathews, Michael Melkonian, Douglas E. Soltis, Pamela S. Soltis, Nicholas W. Miles, Carl J. Rothfels, Lisa Pokorny, A. Jonathan Shaw, Lisa DeGironimo, Dennis W. Stevenson, Barbara Surek, Juan Carlos Villarreal, BÃ©atrice Roure, HervÃ© Philippe, Claude W. dePamphilis, Tao Chen, Michael K. Deyholos, Regina S. Baucom, Toni M. Kutchan, Megan M. Augustin, Jun Wang, Yong Zhang, Zhijian Tian, Zhixiang Yan, Xiaolei Wu, Xiao Sun, Gane Ka-Shu Wong, and James Leebens-Mack. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* 111(45):E4859–E4868.

Yu, Y, T Wang, and RJ Samworth. 2015. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* 102(2):315–323.

Yuan, Ming, and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Zhang, Nancy R, and David O Siegmund. 2007. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63(1):22–32.

Zhao, Peng, and Bin Yu. 2006. On model selection consistency of lasso. *J. Mach. Learn. Res.* 7:2541–2563.