Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 17

Imputation Estimators Partially Correct for Model Misspecification

Vladimir N. Minin, *University of Washington*
John D. O'Brien, *University of Oxford*
Arseni Seregin, *University of Washington*

Recommended Citation:
Minin, Vladimir N.; O'Brien, John D.; and Seregin, Arseni (2011) "Imputation Estimators Partially Correct for Model Misspecification," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 17.
DOI: 10.2202/1544-6115.1650

Imputation Estimators Partially Correct for Model Misspecification

Vladimir N. Minin, John D. O'Brien, and Arseni Seregin

Abstract

Inference problems with incomplete observations often aim at estimating population properties of unobserved quantities. One simple way to accomplish this estimation is to impute the unobserved quantities of interest at the individual level and then take an empirical average of the imputed values. We show that this simple imputation estimator can provide partial protection against model misspecification. We illustrate imputation estimators' robustness to model specification on three examples: mixture model-based clustering, estimation of genotype frequencies in population genetics, and estimation of Markovian evolutionary distances. In the final example, using a representative model misspecification, we demonstrate that in non-degenerate cases, the imputation estimator dominates the plug-in estimate asymptotically. We conclude by outlining a Bayesian implementation of the imputation-based estimation.

KEYWORDS: exponential family, imputation, incomplete observations, model misspecification, robustness

Author Notes: VNM was partially supported by the UW Royalty Research Fund and by the National Scientific Foundation grant No. DMS-0856099. AS was partially supported by the National Scientific Foundation grant No. DMS-0804587. JDO was partially supported by the Wellcome Trust grant No. WT082930MA. We thank Adrian Raftery and his working group for their comments on model-based clustering.

1 Introduction

We are interested in robustness to model misspecification in problems with incomplete observations. Semiparametric approaches have enjoyed a lot of success in this area but these methods lack universality and so need to be fine-tuned for each problem at hand (Tsiatis, 2006; Little and An, 2004; Kang and Schafer, 2007; Chen et al., 2009). Consequently, when practitioners are faced with nonstandard problems with incomplete observations, they are often left to their own devices. As a first step to ameliorating this deficiency, we propose a general imputation-based estimation method that provides partial protection against model misspecification for incomplete data problems.

The idea of using imputation techniques to combat model misspecification is not new. Consider the standard missing data problem of estimating population mean μ given a sample $(r_1, g_1 r_1, \mathbf{w}_1), \ldots, (r_n, g_n r_n, \mathbf{w}_n)$, where g_i is a response variable, r_i is a response indicator taking value 1 if g_i is observed and 0 otherwise, and $\mathbf{w}_i$ is a vector of covariates. Assuming strong ignorability, meaning that g_i and r_i are independent given $\mathbf{w}_i$, we use only those individuals for which the response variable is available to fit a response model with $m_i = \mathrm{E}(g_i | w_i)$ to obtain $\hat{m}_i$ (Rosenbaum and Rubin, 1983). Intuitively, we can combine the empirical estimate of the mean of respondents with model-based predictions of missing g_is for non-respondents to arrive at $\hat{\mu} = (1/n)\sum_{i:r_i=1}^{n} g_i + (1/n)\sum_{i:r_i=0}^{n} \hat{m}_i$. This estimator, called an imputation estimator by Tsiatis and Davidian (2007), will be biased if the response model is misspecified. However, the bias vanishes as the number of non-respondents decreases to zero. Using conditioning on the observed data, we can rewrite Tsiatis and Davidian (2007)'s imputation estimator as $\hat{\mu} = (1/n)\sum_{i=1}^{n} \mathrm{E}(g_i \mid r_i, g_i r_i, \mathbf{w}_i)$. In a completely unrelated missing data setting, O'Brien et al. (2009) also use expectations of complete data conditional on the observed data to arrive at novel estimators of evolutionary distances. Although O'Brien et al. (2009) used imputation by conditional expectations explicitly, these authors did not recognize the full generality of their approach.

In this paper, we investigate the behavior of imputation estimators when they are applied to general problems with incomplete observations. After formulating the generalized imputation estimator, we consider three problems with incomplete observations. We start with a mixture model and demonstrate that imputation is useful for estimating densities of mixture components. Moreover, this imputation density estimation improves accuracy of mixture model-based clustering. Next, we turn to a statistical genetics problem of estimating genotype frequencies. To keep the genetic-specific intricacies to a minimum, we construct an artificial but representative example. In spite of the introduced simplification, our results are directly applicable to a topical problem of multilocus haplotype/genotype frequency estimation, where model misspecification occurs due to a failure to account for population structure (Allen and Satten, 2008; Kraft et al., 2005). In our last example, we consider imputation estimators of evolutionary distances between DNA sequences with partially observed continuous-time Markov chains introduced in O'Brien et al. (2009). We fill some theoretical gaps in their work. First, we identify situations where imputation estimators are not helpful. In doing so, we - for the first time to our knowledge - use the fact that so called group-based Markov models belong to the regular exponential family (Evans and Speed, 1993). Next, we compute almost sure limits of imputation and plug-in estimators for a particular model misspecification. Although we make several simplifying assumptions in this derivation, we believe that qualitatively our results are portable to more realistic applications considered by O'Brien et al. (2009). We conclude by outlining a Bayesian implementation of the imputation-based estimation.

2 Generalized imputation estimators

Assume that complete data $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ are independent and identically distributed with each $\mathbf{x}_i$ distributed according to a parametric family of sampling densities $p_T(\mathbf{x}; \theta_T)$ with parameters $\theta_T \in \Theta_T$. We observe each $\mathbf{x}_i$ through a transformed vector $\mathbf{y}_i = \mathbf{y}(\mathbf{x}_i)$. We further assume that the true sampling density $p_T(\mathbf{x}_1; \theta_T)$ is unknown to us and we have to erroneously postulate a misspecified model $p_F(\mathbf{x}_1; \theta_F)$, where $\theta_F \in \Theta_F$ with parameter spaces Θ_T and Θ_F of possibly different dimensions. Despite this model misspecification, we would like to estimate $\mu = \mathrm{E}_{\theta_T}[\mathbf{s}(\mathbf{x}_1)] = \int \mathbf{s}(\mathbf{x}_1) p_T(\mathbf{x}_1; \theta_T) \mathrm{d}\mathbf{x}_1$, where $\mathbf{s}$ is an arbitrary measurable function that maps complete data to an m-dimensional vector of summary statistics. Assuming that θ_F is identifiable from incomplete data $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$, one can simply maximize the likelihood of the observed data $\prod_{i=1}^n p_F(\mathbf{y}_i; \theta_F)$ to arrive at the maximum likelihood estimate $\hat{\theta}_F = \arg\max_{\theta_F \in \Theta_F} p_F(\mathbf{y}; \theta_F)$. Then, *ignoring model misspecification*, we use $\hat{\theta}_F$ to get the plug-in estimate of the complete-data summaries

$$\hat{\mu}_n^{pi} = \mathrm{E}_{\hat{\theta}_F}[\mathbf{s}(\mathbf{x}_1)] = \int \mathbf{s}(\mathbf{x}_1) p_F(\mathbf{x}_1; \hat{\theta}_F) \mathrm{d}\mathbf{x}_1. \tag{1}$$

This estimator is destined to be biased and asymptotically inconsistent in nearly all situations due to the model misspecification.

Consider an imputation estimator

$$\hat{\mu}_n^{im} = \frac{1}{n} \sum_{i=1}^n \mathrm{E}_{\hat{\theta}_F} [\mathbf{s}(\mathbf{x}_i) \,|\, \mathbf{y}_i] = \frac{1}{n} \sum_{i=1}^n \int \mathbf{s}(\mathbf{x}_i) p_F(\mathbf{x}_i \,|\, \mathbf{y}_i; \hat{\theta}_F) \mathrm{d}\mathbf{x}_i. \tag{2}$$

The motivation behind this new estimator is quite simple: in order to offer protection against model misspecification, we would like to use the empirical measure based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$. To accomplish this, we write $\mathrm{E}_{\theta_T}[\mathbf{s}(\mathbf{x}_1)] = \mathrm{E}_{\theta_T}\{\mathrm{E}_{\theta_T}[\mathbf{s}(\mathbf{x}_1) \,|\, \mathbf{y}_1]\} \approx \mathbb{P}_n \mathrm{E}_{\theta_T}[\mathbf{s}(\mathbf{x}_1) \,|\, \mathbf{y}_1]$ where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{y}_i)$ for any measurable function f. In the absence of a good alternative, we plug-in $\hat{\theta}_F$ for θ_T in the conditional expectations of $\mathbf{s}(\mathbf{x}_i)$ to arrive at our imputation estimator, $\hat{\mu}_n^{im}$.

If the family of distributions $\{p_F(\mathbf{y};\theta_F)\}$ satisfies usual regularity conditions we have $\hat{\theta}_F \overset{\text{a.s.}}{\to} \theta_0$. For example, if our model is not misspecified, i.e. $\Theta_F \equiv \Theta_T$, we would have $\theta_0 = \theta_T$. Consider the family of functions $\mathscr{F}$ which consists of conditional expectations: $\mathscr{F} = \{f(\mathbf{y}_1;\theta) = \mathrm{E}_\theta[\mathbf{s}(\mathbf{x}_1) \,|\, \mathbf{y}_1)], \theta \in \Theta_0\}$ for some bounded open neighborhood Θ_0 of the limiting value θ_0. If we assume that $\mathscr{F}$ has finite bracketing number $N_{[]}(\varepsilon, \mathscr{F}, L_1(P))$ for each $\varepsilon > 0$ and is pointwise continuous in θ, then one can show that $\mathbb{P}_n \mathrm{E}_{\hat{\theta}_F}[\mathbf{s}(\mathbf{x}_1) \,|\, \mathbf{y}_1] \overset{\text{a.s.}}{\to} \mathrm{E}_{\theta_T}\{\mathrm{E}_{\theta_0}[\mathbf{s}(\mathbf{x}_1) \,|\, \mathbf{y}_1]\}$ using standard empirical processes machinery (van der Vaart and Wellner, 2000). Assuming model misspecification almost inevitably leads to $\theta_0 \neq \theta_T$. Therefore, our imputation estimator has little chance of achieving asymptotic consistency. However, if the loss of information due to missing data is relatively small, our new estimator can be quite close to the true value both for finite sample sizes and asymptotically.

Assume that a misspecified complete-data sampling density belongs to the regular exponential family so that $p_F(\mathbf{x}_1;\theta_F) = a(\mathbf{x}_1)\exp\left[\theta_F^T \mathbf{t}(\mathbf{x}_1)\right]/b(\theta_F)$, where $\mathbf{t}(\mathbf{x}_1) = (t_1(\mathbf{x}_1), \ldots, t_r(\mathbf{x}_1))$ is an r-dimensional vector of minimal sufficient statistics and $\theta_F = (\theta_{F1}, \ldots, \theta_{Fr})$ is a natural parameter vector of the same dimension. Then, as noted by Sundberg (1974), the likelihood equations based on the observed data $\mathbf{y}$ can be written as $(1/n)\sum_{i=1}^{n} \mathrm{E}_{\theta_F}[\mathbf{t}(\mathbf{x}_i) \,|\, \mathbf{y}_i] = \mathrm{E}_{\theta_F}[\mathbf{t}(\mathbf{x}_1)]$. Therefore, if the complete-data summary $\mathbf{s}(\mathbf{x}_1)$ can be expressed as a linear transformation of the sufficient statistics $\mathbf{t}(\mathbf{x}_1)$, imposed by the falsely assumed regular exponential family model, then the plug-in estimator (1) and imputation estimator (2) *coincide exactly* regardless of the true sampling density of $\mathbf{x}_1$.

3 Mixture models and model-based clustering

Consider a mixture model with k components. Let $\mathbf{h} = (h_1, \ldots, h_n)$ be iid discrete random variables taking values in $\{1, \ldots, k\}$ with probabilities $\Pr(h_1 = j) = \alpha_j$, $\sum_{j=1}^{k} \alpha_j = 1$. Event $h_i = j$ indicates that the observed $\mathbf{y}_i$ is sampled from the density $p_{Fj}(\mathbf{y};\theta_{Fj})$. The complete-data sampling density becomes

$$p_F(h_i, \mathbf{y}_i; \theta_F) = \prod_{j=1}^{k} \left[\alpha_j p_{Fj}(\mathbf{y}_i;\theta_{Fj})\right]^{1_{\{h_i=j\}}}.$$

We obtain parameter estimates $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_k)$ and $\hat{\theta}_F = (\hat{\theta}_{F1}, \dots, \hat{\theta}_{Fk})$ by maximizing $\prod_{i=1}^n p_F(\mathbf{y}_i; \theta_F)$, where $p_F(\mathbf{y}_i; \theta_F) = \sum_{j=1}^k \alpha_j p_{Fj}(\mathbf{y}_i; \theta_{Fj})$. If we further assume regular exponential family sampling densities of mixture components sharing the same normalizing constant $a(\mathbf{y})$, $p_{Fj}(\mathbf{y}; \theta_F) = a(\mathbf{y}) \exp\left[\mathbf{t}_j(\mathbf{y})^T \theta_{Fj}\right] / b_j(\theta_{Fj})$, then the density of the ith completely observed sampling unit also belongs to the regular exponential family,

$$p_F(h_i, \mathbf{y}_i; \theta_F) = a(\mathbf{y}_i) \exp\left\{ \sum_{j=1}^k 1_{\{h_i=j\}} \mathbf{t}_j(\mathbf{y}_i)^T \theta_{Fj} + \sum_{j=1}^k 1_{\{h_i=j\}} \left[\ln \frac{\alpha_j}{b_j(\theta_{Fj})} \right] \right\}.$$

From our discussion of regular exponential family complete-data likelihoods, it is clear that plug-in and imputation estimators of mean complete-data summaries,

$$\mathrm{E}_{\theta_T}\left[1_{\{h_i=j\}} \mathbf{t}_j(\mathbf{y}_1)\right] \text{ and } \mathrm{E}_{\theta_T}\left[1_{\{h_i=j\}}\right], \tag{3}$$

will coincide exactly regardless of the true complete-data sampling model $p_T(\mathbf{y}_1, \mathbf{h}_1; \theta_T)$. In fact, plug-in and imputation estimators of the second mean complete-data summary, $\mathrm{E}_{\theta_T}\left[1_{\{h_i=j\}}\right]$, will coincide even if densities $p_{Fj}(\mathbf{y}_i; \theta_F)$ do not belong to the regular exponential family. To see this, note that the plug-in estimator in this context is $\hat{\alpha}_j^{pi} = \mathrm{E}_{\hat{\alpha}_j}\left[1_{\{h_i=j\}}\right] = \Pr(h_i = j) = \hat{\alpha}_j$. The estimated probability that observation i belongs to component j is

$$\hat{z}_{ij} = \mathrm{E}\left(1_{\{h_i=j\}} \,|\, \mathbf{y}_i\right) = \frac{\hat{\alpha} p_{Fj}(\mathbf{y}_i, \hat{\theta}_{Fj})}{\sum_{j=1}^k \hat{\alpha} p_{Fj}(\mathbf{y}_i, \hat{\theta}_{Fj})}.$$

The imputation estimate of the jth mixing proportion becomes $\hat{\alpha}_j^{im} = (1/n) \sum_{i=1}^n \mathrm{E}\left(1_{\{h_i=j\}} \,|\, \mathbf{y}_i\right) = (1/n) \sum_{i=1}^n \hat{z}_{ij}$. The likelihood equations for the mixture model can be rearranged to show that $\hat{\alpha}_j^{pi} = \hat{\alpha}_j^{im}$ (Redner and Walker, 1984). Notice that estimating all of the above complete-data expectations requires unambiguously identifying mixture component j, which we assume is possible by imposing constraints on mixture component parameters $\theta_{F1}, \dots, \theta_{Fk}$.

To make our discussion of mixture models more concrete, we simulate $n = 1000$ realizations from a mixture of two log-normal distributions with the log-scale means $\mu_1 = 1.5$ and $\mu_2 = 2.5$ and standard deviations $\sigma_1 = 0.2$ and $\sigma_2 = 0.25$ respectively. The mixing proportion, α, was set to 0.3, completing the set of true model parameters $\theta_T = (\mu_1, \mu_2, \sigma_1, \sigma_2, \alpha)$. Now, we assume a two-component normal mixture model with means ν_1, ν_2, possibly unequal standard deviations δ_1, δ_2, and a mixing proportion β. We estimate parameters $\theta_F = (\nu_1, \nu_2, \delta_1, \delta_2, \beta)$ of this misspecified model using maximum likelihood via the EM algorithm (Dempster et al., 1977; Fraley and Raftery, 2003). We show a histogram of simulated data with a normal mixture model fit in the left plot of Figure 1.

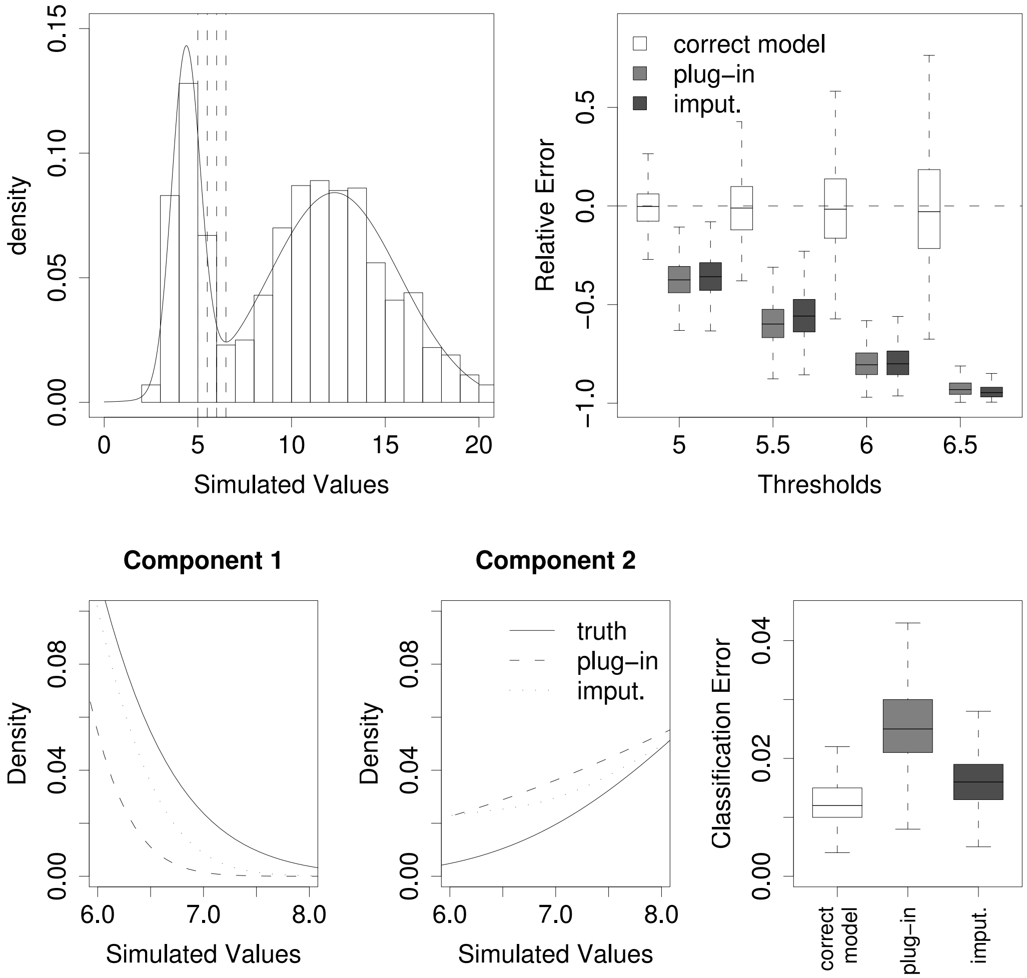

Figure 1: Mixture model example. The upper left plot shows a histogram of 1000 simulated realizations of the two-component log-normal model, described in the text. The solid line depicts the normal mixture density estimated from these simulated data. The dashed vertical lines indicate four values of threshold c, for which we estimate $\mu(c) = \mathrm{E}_{\theta_T}\left(1_{\{h_1=1\}}1_{\{y>c\}}\right)$. Results of conventional and robust estimation of these quantities are shown in the upper right plot of the figure. We repeat simulation and estimation 1000 times and plot box-plots of relative errors, $\frac{\hat{\mu}^{pi}(c)-\mu(c)}{\mu(c)}$ and $\frac{\hat{\mu}^{im}(c)-\mu(c)}{\mu(c)}$, for $c = 5.0, 5.5, 6.0, 6.5$. The bottom row shows results of mixture component density estimation and classification errors during model based clustering.

To avoid the label switching problem, we define mixture component labels by the inequality $\nu_1 < \nu_2$. Equation (3) says that if we try to estimate $\mathrm{E}_{\theta_T}\left(1_{\{h_1=1\}}y_1\right)$, $\mathrm{E}_{\theta_T}\left(1_{\{h_1=1\}}y_1^2\right)$ or $\mathrm{E}_{\theta_T}\left(1_{\{h_1=1\}}\right)$, it does not matter whether we use the plug-in or imputation approach. Instead, we choose to estimate the proportion of samples from the first mixture component that fall to the right of some threshold c, $\mu(c) = \mathrm{E}_{\theta_T}\left(1_{\{h_1=1\}}1_{\{y_1>c\}}\right)$. The plug-in estimate of this quantity is

$$\hat{\mu}^{pi}(c) = \mathrm{E}_{\hat{\theta}_F}\left(1_{\{h_1=1\}}1_{\{y_1>c\}}\right) = \left[1 - \Phi\left(\frac{c - \hat{\nu}_1}{\hat{\delta}_1}\right)\right]\hat{\beta}, \tag{4}$$

where Φ is the standard normal cdf. Our imputation estimator becomes

$$\hat{\mu}^{im}(c) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}_F}\left(1_{\{h_i=1\}}1_{\{y_i>c\}}\,|\,y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ij}1_{\{y_i>c\}}.$$

Since tails of mixture components can be estimated via imputation, it should be possible to devise an imputation estimator of mixture components' densities. Indeed, if we use a nonparametric kernel density estimator, where each observed point i is weighted by z_{ij}, we arrive at an imputation estimate of the jth component density. This is potentially useful, because more accurate estimation of component densities may lead to more accurate model-based clustering (Fraley and Raftery, 2002).

The right plot of Figure 1 demonstrates results of estimating $\mu(c)$ for threshold values $c = 5.0, 5.5, 6.0, 6.5$, depicted in the left plot of the figure by the dashed vertical lines. We consider these values of c, because they fall into the region where sampled points cannot be easily assigned to either of the two mixture components. We simulate 1000 test data sets using already described settings. For each of the simulated data set, we compute plug-in estimates of $\mu(c)$ using the fitted correct log-normal and the misspecified normal model and the imputation under the misspecified normal model. We show box plots of the corresponding relative errors in the upper right plot of Figure 1. Although, the performance of the plug-in and imputation estimators under model misspecification is disappointingly similar, imputation density estimates, plotted in the bottom row, look more promising. We used plug-in density estimates under the correct and misspecified model and imputation density estimates to assign simulated points to two clusters. We then computed clustering classification error using R package MCLUST (Fraley and Raftery, 2003). As shown in the lower bottom plot, clustering accuracy improves significantly under imputation estimates of mixture component densities and approaches the accuracy of clustering under the correct mixture model.

DOI: 10.2202/1544-6115.1650

4 Estimating genotype frequencies

Here, we turn to a classical problem in statistical genetics: estimating allele and genotype frequencies from incomplete observations (Ceppelini et al., 1955). Suppose that we measure some observable characteristic, called a phenotype, in n individuals and record them in a vector $\mathbf{y} = (y_1, \ldots, y_n)$, where each y_i takes one of M possible values in $\mathscr{C} = \{c_1, \ldots, c_M\}$. We further assume that each individual i has an unobserved genotype $\mathbf{x}_i = (x_{i1}, x_{i2})$, defined as an unordered pair of gene variants, called alleles, on two paired chromosomes of this individual. Suppose there are R possible alleles, $\mathscr{G} = (g_1, \ldots, g_R)$. Genotypes are assumed to determine observed phenotypes via a deterministic function $h : \mathscr{G} \times \mathscr{G} \rightarrow \mathscr{C}$ such that $h(g_k, g_l) = h(g_l, g_k)$. Making certain population genetics assumptions allows us to assume that unobserved genotypes are iid with

$$p_T((g_k, g_l); \mathbf{p}, f) = \begin{cases} p_k^2(1-f) + f p_k & \text{if } k = l \\ 2 p_k p_l (1-f) & \text{if } k \neq l, \end{cases} \tag{5}$$

where $\mathbf{p} = (p_1, \ldots, p_R)$ are population allele frequencies and f is called an inbreeding coefficient. We erroneously assume that $f = 0$, reducing the model for genotype probabilities to the celebrated Hardy-Weinberg equilibrium (Hardy, 1908; Weinberg, 1908). The falsely misspecified complete-data likelihood for datum 1 becomes

$$p_F(\mathbf{x}_1; \mathbf{p}) = \prod_{k>l} (2 p_k p_l)^{1_{\{\mathbf{x}_1 = (g_k, g_l)\}}} \prod_{k=1}^{R} (p_i)^{2 \times 1_{\{\mathbf{x}_1 = (g_k, g_k)\}}} \propto \prod_{k=1}^{R} p_k^{t_k},$$

where $t_k = 2 \times 1_{\{\mathbf{x}_1 = (g_k, g_k)\}} + \sum_{l=1}^{R} 1_{\{\mathbf{x}_1 = (g_k, g_l)\}}$. The misspecified observed-data likelihood for datum 1 is $p_F(y_1; \mathbf{p}) = \sum_{\mathbf{x}_1 : h(\mathbf{x}_1) = y_1} p_F(\mathbf{x}_1; \mathbf{p})$.

Since the complete-data likelihood is in the regular exponential family with sufficient statistics $(t_1, \ldots, t_R)$, the plug-in and imputation estimates of $\mathrm{E}(\sum_{i=1}^{R} c_i t_i)$ will coincide exactly. Suppose our objective is to estimate genotype frequencies $\mu_{kl} = \mathrm{E}\left(1_{\{\mathbf{x}_1 = (g_k, g_l)\}}\right) = \Pr\left(\mathbf{x}_1 = (g_k, g_l)\right)$. The complete-data summary $1_{\{\mathbf{x}_1 = (g_k, g_l)\}}$ can not be expressed as a linear combination of the sufficient statistics, so plug-in and imputation estimation will not necessarily produce identical results. After obtaining maximum likelihood estimates of allele frequencies, $\hat{\mathbf{p}}$, the plug-in approach yields

$$\hat{\mu}_{kl}^{pi} = \hat{p}_i^2 1_{\{k=l\}} + 2 \hat{p}_k \hat{p}_l 1_{\{k \neq l\}}.$$

The imputation estimator becomes

$$\hat{\mu}_{kl}^{im} = \frac{1}{n} \sum_{i=1}^{n} \Pr\left(\mathbf{x}_i = (g_k, g_l) \,|\, y_i = c_j\right) 1_{\{y_i = c_j\}} = \frac{n_j \hat{p}_k^2 1_{\{k=l\}}}{n p_F(c_j; \hat{\mathbf{p}})} + \frac{n_j 2 \hat{p}_k \hat{p}_l 1_{\{k \neq l\}}}{n p_F(c_j; \hat{\mathbf{p}})},$$

where $h(g_k, g_l) = c_j$ and $n_j = \sum_{i=1}^{n} 1_{\{y_i = c_j\}}$.

Table 1: Mappings of complete to observed data during genotype frequencies estimation. Ambiguous phenotypes are highlighted in bold.

(g_k,g_l)	(A,A)	(A,B)	(A,C)	(A,D)	(B,B)	(B,C)	(B,D)	(C,C)	(C,D)	(D,D)
$h_1(g_k,g_l)$	aa	ab	ac	ad	bb	bdc	bdc	cc	cd	dd
$h_2(g_k,g_l)$	aa	ab	ac	ad	bd	bdc	bdc	cc	cd	bd

Consider a particular case of the above model with four alleles: $\mathscr{G} = \{A,B,C,D\}$. Table 1 defines two mappings from genotypes to phenotypes, $h_1 : \mathscr{G} \times \mathscr{G} \to \mathscr{C}_1$ and $h_2 : \mathscr{G} \times \mathscr{G} \to \mathscr{C}_2$, where $\mathscr{C}_1 = \{aa,ab,ac,ad,bb,bdc,cc,cd,dd\}$ and $\mathscr{C}_2 = \{aa,ab,ac,ad,bd,bdc,cc,cd\}$. Notice that $\mathscr{C}_1$ has 9 phenotypes and $\mathscr{C}_2$ has 8 phenotypes. Therefore, the loss of information due to missing data is larger under mapping h_2 than under h_1. We simulate 1000 observed phenotypes under both mappings using complete-data model (5) with $p_A = 0.3$, $p_B = 0.2$, $p_C = 0.2$, $p_D = 0.3$ and $f = 0, 0.125, 0.25, 0.375, 0.5$. For each of these 10 simulated data sets, we estimate allele frequencies $\hat{p}_A$, $\hat{p}_B$, $\hat{p}_C$, and $\hat{p}_D$ using the EM algorithm and assuming that $f = 0$.

For phenotypes that unambiguously correspond to exactly one genotype, the empirical phenotype frequency can be used to estimate the corresponding genotype frequency. Therefore, it only makes sense to compare plug-in and imputation estimation for genotypes that correspond to ambiguously defined phenotypes. For example, under both h_1 and h_2 genotypes (B,C) and (B,D) correspond to the phenotype bcd. Suppose our goal is to estimate these genotype frequencies: $\mu_{BC} = \Pr(\mathbf{x}_1 = (B,C))$ and $\mu_{BD} = \Pr(\mathbf{x}_1 = (B,D))$. Plug-in estimates of these population-level quantities are

$$\hat{\mu}_{BC}^{pi} = 2\hat{p}_B\hat{p}_C \quad \text{and} \quad \hat{\mu}_{BD}^{pi} = 2\hat{p}_B\hat{p}_D.$$

Imputation estimates are obtained as

$$\hat{\mu}_{BC}^{im} = \frac{n_{bcd}}{n}\frac{\hat{p}_B\hat{p}_C}{\hat{p}_B\hat{p}_C + \hat{p}_B\hat{p}_D} \quad \text{and} \quad \hat{\mu}_{BD}^{im} = \frac{n_{bcd}}{n}\frac{\hat{p}_B\hat{p}_D}{\hat{p}_B\hat{p}_C + \hat{p}_B\hat{p}_D},$$

where $n_{bcd} = \sum_{i=1}^{n} 1_{\{y_i = bcd\}}$ and $n = 1000$. Figure 2 shows box plots of relative errors of plug-in and imputation estimators, obtained by repeating the above simulation and estimation steps 1000 times. In the case of 9 phenotypes, corresponding to h_1 mapping, the imputation estimation offers remarkable protection against model misspecification. Decreasing the number of observed phenotypes from 9 to 8 results in the imputation estimators outperforming the plug-in one only for $f = 0.125$ and $f = 0.25$. For the rest of inbreeding coefficient values, plug-in estimation produces better estimates of μ_{BC}, while imputation estimation offers better estimates of μ_{BD}. However, overall imputation relative errors are still smaller than plug-in errors.

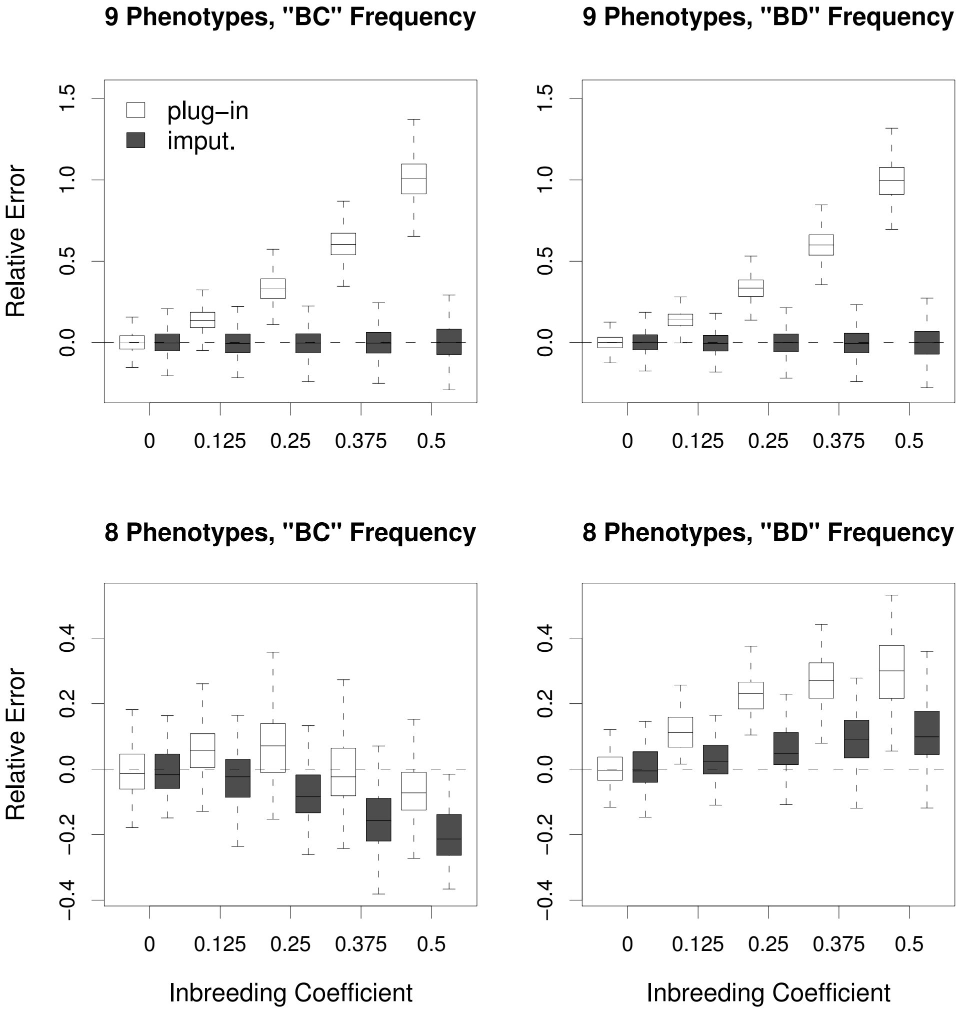

Figure 2: Genotype frequency estimation. We plot box plots of relative errors, of plug-in and imputation estimates of genotype frequencies (μ_{BC} and μ_{BD}) for two incomplete data mappings, with 9 and 8 observed phenotypes. Each pair of white and grey box plots corresponds to an inbreeding coefficient that ranges from 0 to 0.5.

5 Labeled evolutionary distances

Imputation estimation was proposed by O'Brien et al. (2009) in the context of estimation of evolutionary distances between molecular sequences, a standard problem in molecular evolutionary biology (Gu and Li, 1998; Yang, 2006). Consider a $2 \times n$ matrix $\mathbf{y} = \{y_{ij}\}$, where each y_{ij} takes values in the $\mathscr{S} = \{1, \ldots, s\}$. We assume that all columns in $\mathbf{y}$ are independently generated by the same reversible and irreducible continuous-time Markov chain (CTMC) $\{X_t\}$, defined on the finite state-space $\mathscr{S}$ by infinitesimal generator $\Lambda(\theta_T)$. This Markov process models the evolution of DNA sequences so that the state space $\mathscr{S}$ usually consists of 4 nucleotide bases, however, a couple of alternative state-spaces are also often used. Each column $\mathbf{y}_i$ in $\mathbf{y}$ is produced by first drawing y_{1i} from the stationary distribution of $\{X_t\}$, $\pi(\theta_T) = (\pi_1(\theta_T), \ldots \pi_s(\theta_T))$, running the chain for an unknown time t and setting $y_{2i} = X_t$. For each realization i, we observe only the starting and ending states of the Markov chain on the time interval $[0,t]$. Here, model misspecification usually manifests itself through an incorrect parameterization of the infinitesimal generator, $\Lambda(\theta_F)$. The misspecified likelihood of the observed data is $p_F(\mathbf{y};\theta_F) = \prod_{i=1}^{n} \pi_{y_{1i}}(\theta_T) p_{y_{1i}y_{2i}}(\theta_F,t)$, where $\mathbf{P}(\theta_F,t) = e^{\Lambda(\theta_F)t} = \{p_{ij}(\theta_F,t)\}$ and $p_{ij}(\theta_F,t) = \Pr(X_t = j \,|\, X_0 = i)$ are finite-time transition probabilities of $\{X_t\}$. Notice that transition probabilities depend on Λ and t only through their product. Therefore, we require the identifiability constraint $t = 1$.

In this example, complete data consist of the full Markov chain trajectory $\mathbf{x}_i = \{X_{ri} : 0 < r < t\}$. A complete-data summary of scientific interest is $s(\mathbf{x}_1) = N_{\mathscr{L}}$, the number of transitions of X_t during the time interval $[0,1]$, labeled by the set of ordered state pairs $\mathscr{L}$. In the absence of complete Markov trajectories, we are interested in the mean number of labeled transitions of the stationary Markov chain, available analytically via

$$\mu = \mathrm{E}_{\theta_T}[s(\mathbf{x}_1)] = \mathrm{E}_{\theta_T}(N_{\mathscr{L}}) = \pi(\theta_T)^T \Lambda_{\mathscr{L}}(\theta_T)\mathbf{1}, \tag{6}$$

where $\mathbf{1}$ is an s-dimensional column vectors of 1s and $\Lambda_{\mathscr{L}} = \{\lambda_{uv} \times 1_{\{(u,v)\in\mathscr{L}\}}\}$. In molecular evolution, this expected number of labeled Markov transitions translates into mean number of labeled substitutions, allowing evolutionary biologists to measure molecular sequence similarity in a flexible manner (O'Brien et al., 2009).

The plug-in approach for estimating μ proceeds by first fitting a possibly misspecified Markov model, $\Lambda(\theta_F)$ and then using the resulting parameter estimates to compute complete-data summary expectations. More specifically, we obtain $\hat{\theta}_F = \arg\max_{\theta_F} p(\mathbf{y};\theta_F)$ and obtaining plug-in and imputation estimators

$$\hat{\mu}^{pi} = \pi(\hat{\theta}_F)^T \Lambda(\hat{\theta}_F)\mathbf{1} \qquad \text{and} \qquad \hat{\mu}^{im} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}_F}\left[N_1^{\mathscr{L}} \,|\, X_0 = y_{1i}, X_1 = y_{2i}\right].$$

O'Brien et al. (2009) execute two extensive simulation studies that demonstrate that the imputation estimator offers remarkable protection against misspecification of a Markovian substitution model.

5.1 Complete-data likelihood

After falsely assuming a misspecified model parameterization $\Lambda(\theta_F)$ (and $\pi(\theta_F)$ as a result) we condition on the initial Markov chain states and write the misspecified conditional complete-data likelihood

$$p_F(X_{[0,1]};\theta_F) \propto \left[\prod_{u\neq v} \lambda_{uv}^{n_{uv}}(\theta_F)\right] \times e^{\sum_{u=1}^{s} T_u \lambda_{uu}(\theta_F)}, \tag{7}$$

where n_{uv} is the number of times X_t instantaneously jumped from u to v and T_u is the total time X_t spent in state u during the time interval $[0,1]$, both summed over all n realizations of the Markov chain (Guttorp, 1995). The complete-data likelihood belongs to the curved exponential family with sufficient statistics $\mathbf{n} = \{n_{uv}\}_{u\neq v}$ and $\mathbf{T} = (T_1,\ldots,T_s)$.

Nearly all Markov infinitesimal generators used in molecular evolutionary biology fall into the set $\mathscr{A} = \{\Lambda = \{\lambda_{uv}\} : \lambda_{uv} = \pi_v \alpha_{uv} \text{ for } u \neq v\}$, where $\alpha = \{\alpha_{uv}\}$ is a symmetric matrix. Such parameterization ensures reversibility of the Markov chain, a common assumption in the field of molecular evolution (Yang, 2006).

5.2 Group-based models

Notice that the likelihood (7) simplifies significantly if we assume a reversible model with equal diagonal entries of Λ:

$$p_F(X_{[0,1]};\alpha) \propto \prod_{u<v} \alpha_{uv}^{n_{uv}+n_{vu}}, \tag{8}$$

because $\sum_{u=1}^{s} T_u = 1$ is the length of the observational time interval. It turns out that in molecular evolution, only so called group-based models satisfy this property (Evans and Speed, 1993). Group-based Markov evolutionary models can be defined as continuous-time random walks on Abelian groups. If we define an Abelian group on a Markov chain state space $\mathscr{S}$ with algebraic operation "+", then entries of the corresponding group-based CTMC generator Λ must satisfy $\lambda_{uv} = g(u-v)$ for some function $g : \mathscr{S} \to [0,\infty)$. For example, the most general group-based model on the state space of DNA bases $\{A,G,C,T\}$ is a Kimura three-parameter model with

$$\Lambda^{\text{K3P}}(\alpha,\beta,\gamma) = \begin{pmatrix} - & \alpha & \beta & \gamma \\ \alpha & - & \gamma & \beta \\ \beta & \gamma & - & \alpha \\ \gamma & \beta & \alpha & - \end{pmatrix}, \tag{9}$$

corresponding to the Klein group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ (Evans and Speed, 1993).

Group-based models, constructed with algebraic symmetry in mind, find extensive use in statistical phylogenetics (Sturmfels and Sullivant, 2005; Steel et al., 1998). For us, these models are appealing because they turn the completed-data CTMC likelihood into the regular exponential family form. If we break all possible DNA substitutions into three classes and define their corresponding counts $N_{AG,CT} = n_{AG} + n_{GA} + n_{CT} + n_{TC}$, $N_{AC,GT} = n_{AC} + n_{CA} + n_{GT} + n_{TG}$, and $N_{AT,GC} = n_{AT} + n_{TA} + n_{GC} + n_{CG}$, then these counts form the sufficient statistics for the Kimura three-parameter model. From our discussion of the regular exponential family it follows that plug-in and imputation estimates of $\text{E}_{\alpha,\beta,\gamma}\left(c_1 N_{AG,CT} + c_2 N_{AC,GT} + c_3 N_{AT,GC}\right)$ will coincide exactly regardless of the true sampling model and of the choice of constants c_1, c_2, and c_3. This fact was not noticed by O'Brien et al. (2009), because the authors did not consider group-based models explicitly in their work.

5.3 A closer look at observed data likelihood equations

Instead of invoking properties of the regular exponential family, one can find more general conditions under which imputation and plug-in estimates of labeled evolutionary distances coincide, as demonstrated by the theorem below.

Theorem 1. *Let* $\mathbf{y} = \{y_{ij}\}, i = 1,2, j = 1,\dots,n$, *be a pairwise sequence alignment generated by a CTMC with an unknown infinitesimal generator* $\Lambda(\theta_T)$ *as described at the beginning of this section. We take* $\Lambda(\theta_F)$ *to be a misspecified model and* $\hat{\theta}_F = (\hat{\theta}_{F1},\dots,\hat{\theta}_{Fr})$ *to be the corresponding maximum likelihood estimator obtained from the observed data* $\mathbf{y}$. *If*

$$\Lambda_{\mathscr{L}}(\hat{\theta}_F) - \mathbf{I} \times \pi^T(\hat{\theta}_F)\Lambda_{\mathscr{L}}(\hat{\theta}_F)\mathbf{1} \in \left\langle \frac{\partial \Lambda(\theta_F)}{\partial \theta_{F1}}\Big|_{\theta_F = \hat{\theta}_F}, \dots, \frac{\partial \Lambda(\theta_F)}{\partial \theta_{Fd}}\Big|_{\theta_F = \hat{\theta}_F} \right\rangle, \tag{10}$$

where $\mathscr{L} \subset S^2 \setminus \{(i,i) : i \in S\}$ *is a set of ordered Markov state pairs and* $\Lambda_{\mathscr{L}} = \{\lambda_{uv} \times 1_{\{(u,v)\in\mathscr{L}\}}\}$, *then*

$$E_{\hat{\theta}_F}(N_{\mathscr{L}}) = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{\theta}_F}(N_{\mathscr{L}} \mid X_0 = y_{1i}, X_1 = y_{2i}),$$

DOI: 10.2202/1544-6115.1650

where $N_{\mathscr{L}}$ is the unobserved number of Markov chain transitions labeled by the set $\mathscr{L}$.

To illustrate the above theorem, consider a Kimura two-parameter model $\Lambda^{\mathrm{K2P}}(\alpha,\beta) = \Lambda^{\mathrm{K3P}}(\alpha,\beta,\beta)$, obtained by setting $\gamma = \beta$ in matrix (9) (Kimura, 1980). For both of these models, the stationary distribution is $\pi^T = (0.25, 0.25, 0.25, 0.25)$. Let

$$\mathscr{L}_1 = \{(A,G),(G,A),(C,T),(T,C)\} \text{ and} \tag{11}$$

$$\mathscr{L}_2 = \{(A,C),(C,A),(A,T),(T,A),(C,G),(G,C),(T,G),(G,T)\} \tag{12}$$

be two substitutional classes of interest. The partial derivatives of the Kimura two-parameter generator,

$$\frac{\partial}{\partial\alpha}\Lambda^{\mathrm{K2P}}(\alpha,\beta) = \frac{1}{\alpha}\left[\Lambda_{\mathscr{L}_1} - \mathbf{I}\times\pi^t\Lambda_{\mathscr{L}_1}\mathbf{1}\right] \text{ and}$$
$$\frac{\partial}{\partial\beta}\Lambda^{\mathrm{K2P}}(\alpha,\beta) = \frac{1}{\beta}\left[\Lambda_{\mathscr{L}_1} - \mathbf{I}\times\pi^t\Lambda_{\mathscr{L}_2}\mathbf{1}\right],$$

satisfy condition (10). Therefore, Theorem 1 says that plug-in and imputation estimators of $\mathrm{E}_{\alpha,\beta}\left(N_{\mathscr{L}_1}\right)$ and $\mathrm{E}_{\alpha,\beta}\left(N_{\mathscr{L}_2}\right)$ coincide exactly. Of course this example reiterates the fact that complete-data likelihood of the Kimura two-parameter model belongs to the regular exponential family with sufficient statistics $N_{\mathscr{L}_1}$ and $N_{\mathscr{L}_2}$.

5.4 Misspecified Kimura model: asymptotic behavior

Studying asymptotic properties of our imputation estimator is challenging in general even for the specific problem of the evolutionary distance estimation. Therefore, we turn to an elementary example to obtain some basic asymptotic results. First, we introduce the simplest group-based model on the nucleotide state space, known as the Jukes-Cantor model. The infinitesimal generator of this Markov chain is obtained by setting $\alpha = \beta$ in the Kimura two-parameter model, $\Lambda^{\mathrm{JC}}(\gamma) = \Lambda^{\mathrm{K2P}}(\gamma,\gamma)$ (Jukes and Cantor, 1969).

Theorem 2. *Assume that observed sequence data* $\mathbf{y}$ *was generated from the Kimura two-parameter model with generator $\Lambda^{K2P}(\alpha,\beta)$. Let $\hat{\gamma}$ be the maximum likelihood*

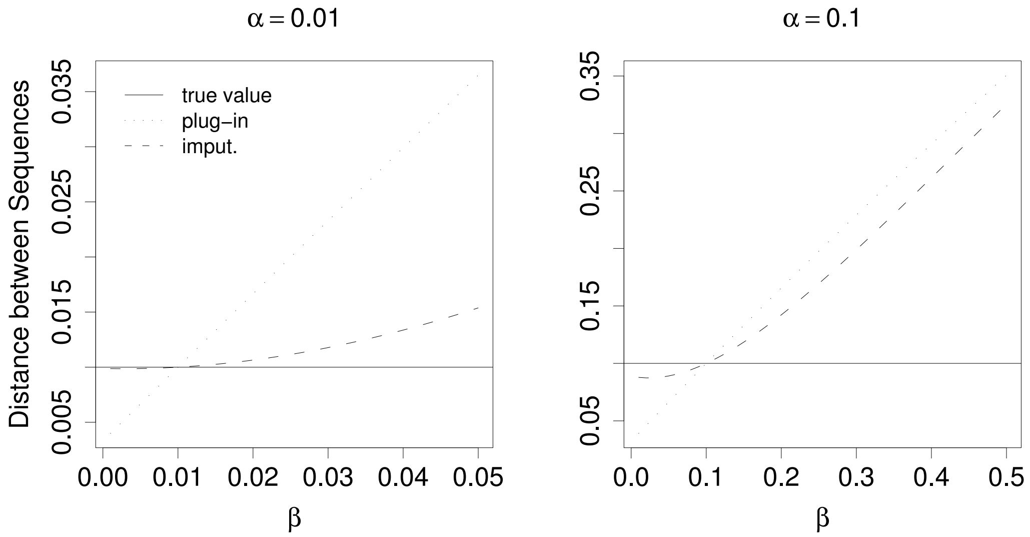

Figure 3: Estimator limits in the Kimura model. In both panels of the figure we plot the true value of the mean number of labeled substitutions $\mu = \mathrm{E}(N_{\mathscr{L}})$ (solid line), the a.s. limits of the plug-in (dotted line) and imputation (dashed line) estimators.

estimate, obtained by fitting a Jukes-Cantor model with generator $\Lambda^{JC}(\gamma)$ to **y**. *Then as the number of columns in* **y**, *n, approaches infinity,*

$$E_{\hat{\gamma}}\left(N_{\mathscr{L}_1}\right) \overset{a.s.}{\to} \beta - \frac{1}{4}\ln \quad \frac{1+2e^{2(\beta-\alpha)}}{3}\Bigg)$$

$$\frac{1}{n}\sum_{i=1}^{n} E_{\hat{\gamma}}\left(N_{\mathscr{L}_1} | X_0 = y_{1i}, X_1 = y_{2i}\right) \overset{a.s.}{\to} \left[\beta - \frac{1}{4}\ln \quad \frac{1+2e^{2(\beta-\alpha)}}{3}\Bigg)\right] \times \left[1 + \frac{4(e^{-4\beta}+2e^{-2(\alpha+\beta)})(e^{-4\beta}-e^{-2(\alpha+\beta)})}{3(3-e^{-4\beta}-e^{-2(\alpha+\beta)})}\right],$$

where $\mathscr{L}_1$ is defined by equation (11).

Corollary 1. *Under the conditions of Theorem 2, define*

$$\mu = E_{\alpha,\beta}\left(N_{\mathscr{L}_1}\right), \mu_\infty^{pi} = \lim_{n\to\infty} E_{\hat{\gamma}}\left(N_{\mathscr{L}_1}\right), \mu_\infty^{im} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E_{\hat{\gamma}}\left(N_{\mathscr{L}_1} | X_0 = y_{1i}, X_1 = y_{2i}\right).$$

Then $|\mu_\infty^{im} - \mu| < |\mu_\infty^{pi} - \mu|$ when $\alpha \neq \beta$. In other words, the imputation estimator asymptotically is always better than the plug-in one.

To illustrate the above theorem and its corollary we plot the true value (μ) and a.s. limits of the plug-in (μ_∞^{pi}) and imputation (μ_∞^{im}) estimators as function of β in Figure 3. We fix $\alpha = 0.01$ in the left panel and $\alpha = 0.1$ in the right panel. Roughly speaking, the left panel shows the behavior of the estimators when the overall substitution rate is low, while the right panel corresponds to a high substitution rate scenario. The lower the substitution rate, the better our imputation estimator behaves asymptotically. This property of the imputation estimation is expected, because low substitution rate translates into smaller loss of information due to missing data, which in turn makes the imputation estimation more powerful. We have already seen this behavior of the imputation estimator in the previous examples.

6 Bayesian implementation

6.1 General recipe

Although all examples so far were analyzed from the maximum likelihood perspective, one can easily perform imputation-based estimation in a Bayesian framework. To accomplish this, we first need to assign a prior distribution $p(\theta_F)$ to the parameters of our misspecified model $p_F(\mathbf{y};\theta_F)$. We assume that it is possible to obtain either the posterior distribution $p_F(\theta_F \mid \mathbf{y})$ or the augmented posterior $p_F(\theta_F, \mathbf{x} \mid \mathbf{y})$, possibly approximating these distributions via Markov chain Monte Carlo (MCMC) (Tanner and Wong, 1987). Using these posterior distributions, we define plug-in and two imputation predictive distributions

$$p\left(\mathrm{E}_{\theta_F}[\mathbf{s}(\mathbf{x}_1)] \mid \mathbf{y}\right), \quad p\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(\mathbf{x}_i) \,\Big|\, \mathbf{y}\right), \quad \text{and} \quad p\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\theta_F}[\mathbf{s}(\mathbf{x}_i) \mid \mathbf{y}_i] \,\Big|\, \mathbf{y}\right).$$

As before, we hope that the latter two will provide us some protection against model misspecification. These last two predictive distributions have the same mean, but conditioning reduces the variance of the third distribution. This is similar to Rao-Blackwellization in Monte Carlo sampling (Casella and Robert, 1996), but since we are working under the assumption of model misspecification, smaller variance is not necessarily a desirable property of a predictive distribution.

6.2 Bayesian estimation of genotype frequencies

To illustrate the Bayesian implementation of our procedure, we revisit the genotype frequency estimation example. We generate 10 phenotype samples using the two genotype-to-phenotype mappings defined in Table 1 and setting the inbreeding coefficient f and true allele frequencies to the values we used in the original example. We place Dirichlet$(1,1,1,1)$ prior on allele frequencies and approximate the posterior distribution of complete data (genotype counts) and allele frequencies via Gibbs sampling.

Recall that our goal is to estimate genotype frequency $\mu_{k,l} = \Pr(\mathbf{x}_1 = (g_k, g_l))$. For $(g_k, g_l) = (B, C)$ and $(g_k, g_l) = (B, D)$, we report posterior distributions of

$$2p_k p_l, \frac{1}{n}\sum_{i=1}^{n} 1_{\{\mathbf{x}_i = (g_k, g_l)\}} = \frac{m_{kl}}{n}, \text{ and } \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\mathbf{x}_i = (g_k, g_l) \,|\, y_i) = \frac{n_j 2 p_k p_l}{n(2p_B p_C + 2p_B p_D)},$$

where $m_{kl} = \sum_{i=1}^{n} 1_{\{\mathbf{x}_i = (g_k, g_l)\}}$ and $n_j = \sum_{i=1}^{n} 1_{\{y_i = BCD\}}$. We report box plots of these posterior distributions in Figure 4. These box plots are not directly comparable to results in Figure 2, because our Bayesian analysis is based only on ten data sets, while the maximum likelihood analysis was done on 10,000 simulated data sets, one thousand for each value of f and for each genotype-to-phenotype mapping. To make these analyses comparable, one can study frequentist properties of Bayesian plug-in and imputation estimators based, for example, on the posterior median-based estimators of allele frequencies and genotype counts. Our Bayesian results are nonetheless consistent with the maximum likelihood analysis: imputation estimators outperform the plug-in estimate in the case of nine phenotypes, none of the estimators have a uniform advantage across all values of the inbreeding coefficient in the eight phenotype case.

We end our discussion of the Bayesian implementation of our imputation estimation by pointing out that this inferential framework is already being used in evolutionary biology, albeit somewhat informally (Zhai et al., 2007; Minin and Suchard, 2008b). These methods extend the idea of imputation evolutionary distance estimation to multiple sequences.

7 Discussion

We generalize the notion of imputation estimators and demonstrate that such estimators can be useful in a variety of incomplete data problems under model misspecification. We use simulations as our main tool in the first two examples and provide some simple asymptotic results in our last example. So far, our experience suggests that imputation estimators perform very well under mild model misspecification and when the loss of information due to missing data is reasonably small. Intuitively, it is clear that imputation estimators should be more successful as the amount of missing data decreases, because in the absence of missing information

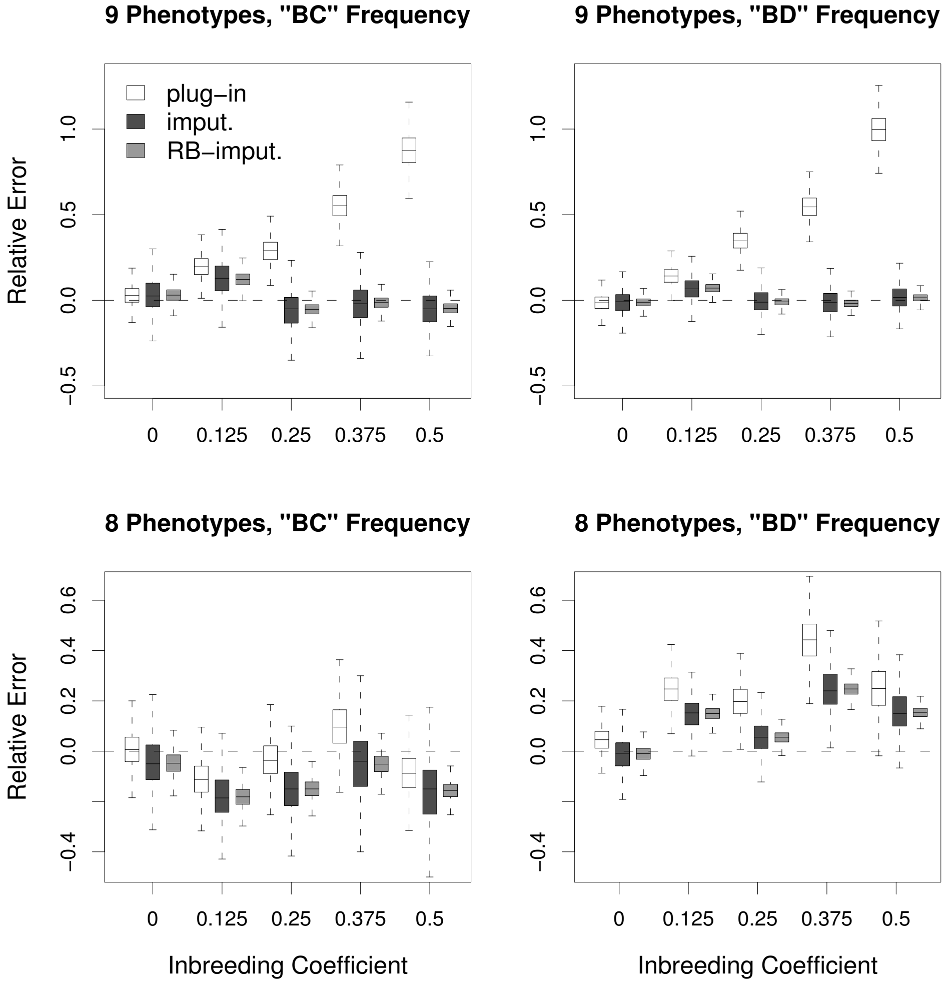

Figure 4: Genotype frequency estimation. We plot box plots of relative errors of plug-in, imputation, and Rao-Blackwellized imputation estimates of genotype frequencies (μ_{BC} and μ_{BD}) for two incomplete data mappings, with 9 and 8 observed phenotypes. Each trio of white, dark grey, and light grey box plots corresponds to an inbreeding coefficient that ranges from 0 to 0.5.

these estimators turn into sample means, which are model-free and consistent estimates of appropriate population-level quantities. However, to make this intuition useful, we need to connect formally efficiency of imputation estimators with the amount of missing data and degree of model misspecification. We hope to be able to make these connections in our future work.

Studying sampling properties of imputation estimators proved to be difficult in general, especially since in practice the true sampling density of the observed data is unknown. In fact, in all our examples, we do not discuss how to compute the variance of the maximum likelihood-based imputation estimators. We recommend to use nonparametric bootstrap to explore sampling properties of imputation estimators. However, one should interpret bootstrap results with care, because imputation estimators remain biased even asymptotically. Similar care needs to be applied to the interpretation of predictive distributions in the Bayesian context.

Although we have not emphasized this throughout the paper, imputation estimators are usually easy to compute, which makes them particularly useful when a compromise between model complexity and computational efficiency results in an intentionally misspecified model. In our examples of model misspecification, we considered Gaussian mixture components, Hardy-Weinberg genotype frequencies, and parametric Markov models of DNA substitution. All these highly popular models owe a large portion of their success to their computational tractability. We argue that imputation estimators can take these and many other simple and computationally efficient models one step further outside of their usual domain of application.

Determining the most general conditions under which imputation estimators are guaranteed to improve upon plug-in estimators remains an open problem even when the sample size approaches infinity. In our experience, in the worst case scenario imputation and plug-in estimators essentially coincide, so the risk of replacing a plug-in estimator with its imputation counterpart is minimal. On the other hand, a discrepancy between the imputation and plug-in estimators indicates model misspecification. Sundberg (1974) used this observation to construct goodness-of-fit test statistics based on the differences between imputation and plug-in estimators. Sundberg (1974)'s constructions, underutilized in general, can be especially useful in the field of statistical phylogenetics, where goodness-of-fit tests are nonexistent. As we previously argued, in some cases one intentionally misspecifies the model to achieve computational tractability. We believe that imputation estimation is a promising remedy for such intentional model violations. We hope that our general imputation estimation framework and worked out examples in this paper will help researchers recognize missing data problems in which imputation estimation leads to robustness to model misspecification.

Appendix

Proof of Theorem 1. Defining $m_{kl} = \sum_{i=1}^{n} 1_{\{y_{1i}=k, y_{2i}=l\}}$, the misspecified complete-data log-likelihood takes the following form:

$$l(\mathbf{y}, \theta_F) = \sum_{k \in S} \sum_{l \in S} m_{kl} \ln p_{kl}(\theta_F, 1), \tag{A-1}$$

where $p_{kl}(\theta_F, 1)$ is the probability of $X_1 = l$ conditional on starting $X_0 = k$. Recall that $\mathbf{P}(\theta_F, 1) = e^{\Lambda(\theta_F)} = \{p_{ij}(\theta_F, 1)\}$. Differentiating (A-1) with respect model parameters, we arrive at the likelihood equations

$$\sum_{k\in E}\sum_{l\in E} \frac{m_{kl}}{p_{kl}(\theta_F,1)} \frac{\partial p_{kl}(\theta_F,1)}{\partial \theta_{Fj}} = 0, j = 1, \ldots, r. \tag{A-2}$$

From the backward Kolmogorov equation $\frac{\mathrm{d}\mathbf{P}(\theta_F,t)}{\mathrm{d}t} = \Lambda(\theta_F)\mathbf{P}(\theta_F,t)$ with initial condition $\mathbf{P}(\theta_F, 0) = \mathbf{I}$, we derive the following integral expression for the partial derivatives of transition probabilities:

$$\frac{\partial}{\partial\theta_{Fj}} \mathrm{P}(\theta_F, 1) = \int_0^1 e^{\Lambda(\theta_F)\tau} \frac{\partial}{\partial\theta_{Fj}} \Lambda(\theta_F) e^{\Lambda(\theta_F)(1-\tau)} \mathrm{d}\tau.$$

Next, we write the imputation estimator in terms of m_{kl},

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}_F}\left(N_{\mathscr{L}} | X_0 = y_{1i}, X_1 = y_{2i}\right) = \frac{1}{n}\sum_{k\in S}\sum_{l\in S} \frac{m_{kl}}{p_{kl}(\hat{\theta}_F,1)} \mathrm{E}_{\hat{\theta}_F}\left(N_{\mathscr{L}} 1_{\{X_1=l\}} | X_0 = k\right),$$

where

$$\mathrm{E}_{\hat{\theta}_F}\left(N_{\mathscr{L}} 1_{\{X_1=l\}} | X_0 = k\right) = \left\{ \int_0^1 e^{\Lambda(\hat{\theta}_F)\tau} \Lambda_{\mathscr{L}}(\hat{\theta}_F) e^{\Lambda(\hat{\theta}_F)(1-\tau)} \mathrm{d}\tau \right\}_{kl}. \tag{A-3}$$

Derivation of the formula (A-3) can be found in (Ball and Milne, 2005) or (Minin and Suchard, 2008a). Condition (10) says that there exist real constants $c_1, \ldots, c_r$ such that

$$\Lambda_{\mathscr{L}}(\hat{\theta}_F) - \mathbf{I} \times \pi^T(\hat{\theta}_F)\Lambda_{\mathscr{L}}(\hat{\theta}_F)\mathbf{1} = \sum_{i=1}^{r} c_i \frac{\partial \Lambda(\theta_F)}{\partial \theta_{Fi}}\Big|_{\theta_F = \hat{\theta}_F}.$$

Therefore, the difference between the plug-in and imputation estimators becomes

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}_F}\left(N_{\mathscr{L}} | X_0 = y_{1i}, X_1 = y_{2i}\right) - \mathrm{E}_{\hat{\theta}_F}\left(N_{\mathscr{L}}\right) =$$

$$\frac{1}{n}\sum_{k\in S}\sum_{l\in S} \frac{m_{kl}}{p_{kl}(\hat{\theta}_F,1)} \left\{ \int_0^1 e^{\Lambda(\hat{\theta}_F)} [\Lambda_{\mathscr{L}}(\hat{\theta}_F) - \mathbf{I} \times \pi^T(\hat{\theta}_F)\Lambda_{\mathscr{L}}(\hat{\theta}_F)\mathbf{1}] e^{\Lambda(\hat{\theta}_F)(1-\tau)} \mathrm{d}\tau \right\}_{kl} =$$

$$\frac{1}{n}\sum_{i=1}^{r} c_i \sum_{k\in S}\sum_{l\in S} \frac{m_{kl}}{p_{kl}(\hat{\theta}_F,1)} \left\{ \int_0^1 e^{\Lambda(\hat{\theta}_F)} \frac{\partial \Lambda(\theta_F)}{\partial \theta_{Fi}}\Big|_{\theta_F = \hat{\theta}_F} e^{\Lambda(\hat{\theta}_F)(1-\tau)} \mathrm{d}\tau \right\}_{kl} =$$

$$\frac{1}{n}\sum_{i=1}^{r} c_i \sum_{k\in S}\sum_{l\in S} \frac{m_{kl}}{p_{kl}(\hat{\theta}_F,1)} \frac{\partial p_{kl}(\theta_F,1)}{\partial \theta_{Fi}}\Big|_{\theta_F = \hat{\theta}_F} = 0,$$

because $\hat{\theta}_F$ satisfies likelihood equations (A-2). □

Proof of Theorem 2. As before, let $m_{kl} = \sum_{i=1}^{n} 1_{\{y_{1i}=k, y_{2i}=l\}}$. Using these site counts, define

$$m_{\mathscr{L}_1} = \sum_{(k,l)\in\mathscr{L}_1} m_{kl}, \quad m_{\mathscr{L}_2} = \sum_{(k,l)\in\mathscr{L}_2} m_{kl}, \quad m_D = \sum_{k=l} m_{kl},$$
$$f_{\mathscr{L}_1} = \frac{m_{\mathscr{L}_1}}{n}, \qquad f_{\mathscr{L}_2} = \frac{m_{\mathscr{L}_2}}{n}, \qquad f_D = \frac{m_D}{n},$$

where $\mathscr{L}_2$ is defined by equation (12). Transition probabilities of the Kimura two-parameter model are obtained as

$$p_{kl}(\alpha,\beta,t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{if } (k,l) \in \mathscr{L}_1, \\ \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \text{if } (k,l) \in \mathscr{L}_2, \\ \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{if } k = l. \end{cases} \tag{A-4}$$

Since the stationary distribution of the Kimura two-parameter model is uniform, $(m_{\mathscr{L}_1}, m_{\mathscr{L}_2}, m_D) \sim \text{Multinomial}(p_{\mathscr{L}_1}, p_{\mathscr{L}_2}, p_D)$, where

$$p_{\mathscr{L}_1} = \sum_{(k,l)\in\mathscr{L}_1} \frac{1}{4} p_{kl}(\alpha,\beta,1) = \frac{1}{4} + \frac{1}{4}e^{-4\beta} - \frac{1}{2}e^{-2(\alpha+\beta)},$$
$$p_{\mathscr{L}_2} = \sum_{(k,l)\in\mathscr{L}_2} \frac{1}{4} p_{kl}(\alpha,\beta,1) = \frac{1}{2} - \frac{1}{2}e^{-4\beta},$$
$$p_D = \sum_{k=l} \frac{1}{4} p_{kl}(\alpha,\beta,1) = \frac{1}{4} + \frac{1}{4}e^{-4\beta} + \frac{1}{2}e^{-2(\alpha+\beta)}.$$

Therefore,

$$f_{\mathscr{L}_1} \overset{\text{a.s.}}{\to} p_{\mathscr{L}_1}, f_{\mathscr{L}_2} \overset{\text{a.s.}}{\to} p_{\mathscr{L}_2}, \text{ and } f_D \overset{\text{a.s.}}{\to} p_D \tag{A-5}$$

by the strong law of large numbers. We will need these a.s. limits when we express both plug-in and imputation estimators in terms of $f_{\mathscr{L}_1}$, $f_{\mathscr{L}_2}$, and f_D.

The maximum likelihood estimator of γ, $\hat{\gamma} = -\frac{1}{4}\ln\left[1 - \frac{4}{3}(1 - f_D)\right]$, exists only if $1 - f_D < 3/4$. Since we know that $1 - f_D \overset{\text{a.s.}}{\to} \frac{3}{4} - \frac{1}{4}e^{-4\beta} - \frac{1}{2}e^{-2(\alpha+\beta)} < \frac{3}{4}$, we can safely assume that $\hat{\gamma}$ is well defined for large enough n. The plug-in estimator

$$\mathrm{E}_{\hat{\gamma}}(N_{\mathscr{L}_1}) = \hat{\gamma} \overset{\text{a.s.}}{\to} \frac{1}{4}\ln\left[1 - \frac{4}{3}\left.\frac{3 - e^{-4\beta} - 2e^{-2(\alpha+\beta)}}{4}\right)\right] = \beta - \frac{1}{4}\ln\left.\frac{1 + 2e^{2(\beta-\alpha)}}{3}\right).$$

To derive the limit of the imputation estimator we start with

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{\gamma}}\left(N_{\mathscr{L}_1} \,|\, X_0 = y_{1i}, X_1 = y_{2i}\right) = \sum_{k\in S}\sum_{l\in S} \frac{f_{kl}}{p_{kl}(\hat{\gamma},1)} \mathrm{E}_{\hat{\gamma}}\left(N_{\mathscr{L}_1} 1_{\{X_1=l\}} \,|\, X_0 = k\right). \tag{A-6}$$

Setting $\alpha = \beta$ in (A-4), we obtain transition probabilities for the Jukes-Cantor model:

$$p_{kl}(\gamma,t) = \left(\frac{1}{4} - \frac{1}{4}e^{-4\gamma}\right) 1_{\{k\neq l\}} + \left(\frac{1}{4} + \frac{3}{4}e^{-4\gamma}\right) 1_{\{k=l\}}. \tag{A-7}$$

To get the functional form $\mathrm{E}_{\hat{\gamma}}\left(N_{\mathscr{L}_1} 1_{\{X_1=l\}} \,|\, X_0 = k\right)$, we first notice that Λ^{JC} and $\Lambda^{\mathrm{JC}}_{\mathscr{L}_1}$ commute, leading to

$$\int\limits_0^1 e^{\Lambda^{\mathrm{JC}}(\gamma)\tau} \Lambda^{\mathrm{JC}}_{\mathscr{L}_1}(\gamma) e^{\Lambda^{\mathrm{JC}(\gamma)}(1-\tau)} \mathrm{d}\tau = \Lambda^{\mathrm{JC}}_{\mathscr{L}_1}(\gamma) e^{\Lambda^{\mathrm{JC}}(\gamma)} \int\limits_0^1 \mathrm{d}\tau = \Lambda^{\mathrm{JC}}_{\mathscr{L}_1}(\gamma) e^{\Lambda^{\mathrm{JC}}(\gamma)}.$$

Hence,

$$\mathrm{E}_{\hat{\gamma}}\left(N_{\mathscr{L}_1} 1_{\{X_1=l\}} \,|\, X_0 = k\right) = \hat{\gamma}\left(\frac{1}{4} + \frac{3}{4}e^{-4\hat{\gamma}}\right) 1_{\{(k,l)\in\mathscr{L}_1\}} + \hat{\gamma}\left(\frac{1}{4} - \frac{1}{4}e^{-4\hat{\gamma}}\right) 1_{\{(k,l)\notin\mathscr{L}_1\}}. \tag{A-8}$$

Plugging (A-7) and (A-8) to (A-6), we arrive at

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{\gamma}}\left(N_{\mathscr{L}_1} \,|\, X_0 = y_{1i}, X_1 = y_{2i}\right) &= \hat{\gamma}\left[1 + 4e^{-4\hat{\gamma}}\left(\frac{f_{\mathscr{L}_1}}{1-e^{-4\hat{\gamma}}} - \frac{f_D}{1+3e^{-4\hat{\gamma}}}\right)\right] \\
= -\frac{1}{4}\ln\left[1 - \frac{4}{3}(1-f_D)\right] &\times \left[1 + \left(1 - \frac{4}{3}(1-f_D)\right)\left(\frac{3f_{\mathscr{L}_1}}{1-f_D} - 1\right)\right].
\end{aligned}$$

Plugging in limits (A-5) in the above formula produces the desired result. □

Proof of Corollary 1. Defining $A = (e^{-4\beta} + 2e^{-2(\alpha+\beta)})/3$, we write the limiting difference of the imputation and plug-in estimates as

$$\mu^{im}_{\infty} - \mu^{pi}_{\infty} = -\frac{A\ln A}{3(1-A)} e^{-4\beta}(1 - e^{-2(\alpha-\beta)}). \tag{A-9}$$

Since $0 < A \leq 1$, $\mu^{im}_{\infty} - \mu^{pi}_{\infty}$ and $\alpha - \beta$ always have the same sign. Moreover, using $e^x \geq 1 + x$, we can show that

$$0 < -\frac{A\ln A}{1-A} = \frac{\ln(1/A)}{1/A - 1} < 1, \tag{A-10}$$

when $\alpha \neq \beta$. Recall that $\mu = \pi^T \Lambda^{K2P}(\alpha,\beta)\mathbf{1} = \alpha$. leading to

$$\mu^{pi}_{\infty} - \mu = \beta - \frac{1}{4}\ln\left.\frac{1+2e^{2(\beta-\alpha)}}{3}\right) - \alpha = -\frac{1}{4}\ln\left.\frac{e^{4(\alpha-\beta)} + 2e^{2(\alpha-\beta)}}{3}\right).$$

Hence, $\mu^{pi}_{\infty} - \mu$ and $\alpha - \beta$ always have opposite signs.

Case 1: $\alpha > \beta$. We have $0 < e^{-4\beta}(1-e^{-2(\alpha-\beta)}) < 2(\alpha-\beta)$, which together with A-10 imply $0 < \mu_\infty^{im} - \mu_\infty^{pi} < \frac{2}{3}(\alpha-\beta)$. Next, we use concavity of logarithm and arrive at $\mu_\infty^{pi} - \mu < (\alpha-\beta)/3$. Combining these last two inequalities, we have $\mu_\infty^{im} - \mu < 0$. Therefore,

$$\mu_\infty^{pi} - \mu = \mu_\infty^{pi} - \mu_\infty^{im} + \mu_\infty^{im} - \mu < \mu_\infty^{im} - \mu < 0,$$

which proves the desired inequality.

Case 2: $\alpha < \beta$. Recall that $\mu_\infty^{pi} > \mu$. Plugging in $0 > e^{-4\beta}\left(1-e^{-2(\alpha-\beta)}\right) = e^{-2(\alpha+\beta)}\left(e^{2(\alpha-\beta)}-1\right) > e^{2(\alpha-\beta)}-1$ to (A-9) we arrive at $\frac{1}{3}\left(e^{2(\alpha-\beta)}-1\right) < \mu_\infty^{im} - \mu_\infty^{p} < 0$. So

$$\mu_\infty^{im} - \mu = \mu_\infty^{im} - \mu_\infty^{pi} + \mu_\infty^{pi} - \mu > \frac{1}{3}\left(e^{2(\alpha-\beta)}-1\right) - \frac{1}{4}\ln \quad \frac{e^{4(\alpha-\beta)}+2e^{2(\alpha-\beta)}}{3}\Bigg).$$

Defining the function on the right-hand side of the above inequality as $w(\alpha-\beta)$, we show that $w(0) = 0$ and

$$w'(\delta) = \frac{2}{3}e^{2\delta} - \frac{3\left(e^{2\delta}+1\right)}{e^{2\delta}+2} = \frac{\left(e^{2\delta}-1\right)\left(2e^{2\delta}+3\right)}{3\left(e^{2\delta}+2\right)} < 0$$

for $\delta < 0$. Therefore, we have $\mu_\infty^{im} - \mu > w(\alpha-\beta) > 0$ and

$$\mu_\infty^{pi} - \mu = \mu_\infty^{pi} - \mu_\infty^{im} + \mu_\infty^{im} - \mu > \mu_\infty^{im} - \mu > 0.$$

□

References

Allen, A. and G. Satten (2008): "Robust estimation and testing of haplotype effects in case-control studies," Genetic Epidemiology, 32, 29–40.

Ball, F. and R. Milne (2005): "Simple derivations of properties of counting processes associated with Markov renewal processes," Journal of Applied Probability, 42, 1031–1043.

Casella, G. and C. Robert (1996): "Rao-Blackwellisation of sampling schemes," Biometrika, 83, 81–94.

Ceppelini, R., M. Siniscalco, and C. Smith (1955): "The estimation of gene frequencies in a random mating population," Annals of Human Genetics, 20, 97–115.

Chen, Y.-H., N. Chatterjee, and R. Carroll (2009): "Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies," Journal of the American Statistical Association, 104, 220–233.

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

Evans, S. and T. Speed (1993): "Invariants of some probability models used in phylogenetic inference," The Annals of Statistics, 21, 355–377.

Fraley, C. and A. Raftery (2002): "Model-based clustering, discriminant analysis, and density estimation," Journal of the American Statistical Association, 97, 611–631.

Fraley, C. and A. Raftery (2003): "Enhanced software for model-based clustering, density estimation, and discriminant analysis: Mclust." Journal of Classification, 20, 263–286.

Gu, X. and W. Li (1998): "Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution," Proceedings of the National Academy of Sciences, USA, 95, 5899–5905.

Guttorp, P. (1995): Stochastic Modeling of Scientific Data, Suffolk, Great Britain: Chapman & Hall.

Hardy, G. (1908): "Mendelian proportions in a mixed population," Science, 28, 49–50.

Jukes, T. and C. Cantor (1969): Evolution of protein molecules, New York: Academic Press, 21–32.

Kang, J. and J. Schafer (2007): "Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data," Statistical Science, 22, 523–539.

Kimura, M. (1980): "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," Journal of Molecular Evolution, 16, 111–120.

Kraft, P., D. Cox, R. Paynter, D. Hunter, and I. D. Vivo (2005): "Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques," Genetic Epidemiology, 28, 261–272.

Little, R. and H. An (2004): "Robust likelihood-based analysis of multivariate data with missing values," Statistica Sinica, 14, 949–968.

Minin, V. and M. Suchard (2008a): "Counting labeled transitions in continuous-time Markov models of evolution," Journal of Mathematical Biology, 56, 391–412.

Minin, V. and M. Suchard (2008b): “Fast, accurate and simulation-free stochastic mapping of discrete traits,” Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 3985–3995.

O’Brien, J., V. Minin, and M. Suchard (2009): “Learning to count: Robust estimates for labeled distances between molecular sequences,” Molecular Biology and Evolution, 26, 801–814.

Redner, R. and H. Walker (1984): “Mixture densities, maximum likelihood and the EM algorithm,” SIAM Review, 26, 195–239.

Rosenbaum, P. and D. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” Biometrika, 70, 41–55.

Steel, M., M. Hendy, and D. Penny (1998): “Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results,” Discrete Applied Mathematics, 88, 367–396.

Sturmfels, B. and S. Sullivant (2005): “Toric ideals of phylogenetic invariants,” Journal of Computational Biology, 12, 204–228.

Sundberg, R. (1974): “Maximum likelihood theory for incomplete data from an exponential family,” Scandinavian Journal of Statistics, 1, 49–58.

Tanner, M. and W. Wong (1987): “The calculation of posterior distributions by data augmentation,” Journal of the American Statistical Association, 82, 528–540.

Tsiatis, A. (2006): Semiparametric Theory and Missing Data, New York: Springer.

Tsiatis, A. and M. Davidian (2007): “Comment: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data,” Statistical Science, 22, 569–573.

van der Vaart, A. and J. Wellner (2000): Weak convergence and empirical processes, New York: Springer-Verlag, corrected second printing edition.

Weinberg, W. (1908): “Uber̈ den nachweis der vererbung beim menschen,” Jahreshefte des Vereins für vaterländische Naturkunde in Wüttemberg, 64, 368–382.

Yang, Z. (2006): Computational Molecular Evolution, USA: Oxford University Press.

Zhai, W., M. Slatkin, and R. Nielsen (2007): “Exploring variation in the d_N/d_S ratio among sites and lineages using mutational mappings: applications to the influenza virus,” Journal of Molecular Evolution, 65, 340–348.