Stochastic Models And Statistical Inference In Evolutionary Genetics: Using DNA Sequence Data To Learn About Population Divergence And Speciation

Rui Jorge Barrigana Ramos da Costa

Department of Statistical Science
University College London

This dissertation is submitted for the degree of
Doctor of Philosophy

May 2017

To my mother, my father and my brother.
To Maria and Gabriela.

Declaration

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Rui Jorge Barrigana Ramos da Costa
May 2017

Acknowledgements

I would like to express my sincere gratitude to my principal supervisor Dr. Hilde Wilkinson-Herbots for the continuous support of my PhD research, for her patience, motivation, and vast knowledge. I would also like to thank my subsidiary supervisor Prof. Ziheng Yang for helpful suggestions and valuable discussions.

I am indebted to Dr. Christian Hennig, Dr. Alexandros Beskos and Prof. Richard Chandler, for their support during my years as a Graduate Diploma, MSc, and PhD student in the UCL Department of Statistical Science, as well as to my fellow research students, for their friendship and generosity.

My PhD research was funded by an EPSRC research studentship, for which I am most grateful. I would also like to thank Dr. Karel Janko at the Czech Academy of Sciences for supplying the *Cobitis* DNA sequence data and for leading the research project that underlies section 3.3 of the present thesis. Thanks are also due to Dr. Yong Wang, Professor Jody Hey, Professor Nick Barton and Dr. Konrad Lohse for kindly providing the *Drosophila* DNA sequence data. The research on the IIM model benefited from the constructive comments of Professor Yun Song and Dr. Konrad Lohse, for which I am grateful as well.

Last, but not least, my deepest thanks to my family, especially my mother, Maria and Gabriela, for their love and generosity. They made this thesis possible.

Abstract

During speciation, the degree of clustering of a population in terms of genetic polymorphisms increases gradually until the exchange of genes between subpopulations is no longer possible. The isolation-with-migration (IM) model is used to estimate how long ago an ancestral population divided into two subpopulations, and to infer the level of gene flow between the subpopulations during genetic divergence. Its assumption of constant gene flow until the present is however particularly unrealistic in the context of two present-day species. In addition, traditional methods to fit the IM model are aimed at large numbers of DNA sequences from a small number of loci, and are computationally very expensive.

To overcome these limitations, this thesis begins by focusing on an extension of the IM model in which the initial period of gene flow is followed by a period of isolation: the so-called isolation-with-initial-migration (IIM) model. For an IIM model with potentially asymmetric gene flow and unequal subpopulation sizes, the distribution of the number of nucleotide differences between two homologous DNA sequences is derived. Based on this distribution, we develop a maximum-likelihood estimation method which is appropriate for data sets containing observations from many independent loci, and is both very efficient and able to deal with mutation rate heterogeneity. Using a data set of *Drosophila* sequences from approximately 30,000 loci, we show how alternative models, representing different evolutionary scenarios, can be distinguished by means of likelihood ratio tests. To enable inference on both historical and contemporary rates of gene flow between two closely related species, our estimation method is extended to a generalised IM (GIM) model, in which gene flow rates and population sizes can change at some point in the past.

Finally, we show how the theory of statistical inference under model misspecification can be used to improve the accuracy of interval estimation and comparison of speciation models; and we develop a simulation method to estimate the limiting distribution of the likelihood ratio statistic when the true parameter vector lies on the boundary of the parameter space.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Population genetics and the coalescent

Population genetics studies the dynamics of genotypic diversity in a population of DNA sequences, from a statistical perspective. Underlying any study in population genetics is a population genetic model, i.e., a set of statistical assumptions about the processes that affect the reproduction of DNA sequences and the creation of new alleles, namely mutation, natural selection, recombination, and mating system. Several stochastic processes, tracing the evolution of different aspects of genetic diversity, can usually be defined using the same population genetic model. The results of population genetics concern the statistical properties of these processes.

One of the main goals of classical population genetics is to make predictions regarding the evolution of allele frequencies, or of measures related to allele frequencies, such as the probability of sampling a pair of heterozygous DNA sequences at a given locus, or the probability of fixation of a given allele in a population. Most theoretical results are derived within the framework of increasingly complex Wright-Fisher population genetic models and of the continuous-time diffusion processes that approximate them (Wakeley, 2010).

In the 1970's, a new branch of population genetics called *coalescent theory* began to develop. It relied on essentially the same population genetic models, but studied a different class of stochastic processes under these models, namely the genealogy of a random sample of DNA sequences. In rough terms, the genealogy of the sample can be defined as a stochastic process that traces the ancestral lineages of the sample back into the past until their most recent common ancestor. The emergence of coalescent theory was to a large extent motivated by the knowledge of genotypic variation at the molecular level, i.e.,

by the emergence of DNA sequencing and DNA sequence data sets (Ewens, 2004). The analysis of these data sets using sequence evolution models from classical population genetics is extremely hard (Nordborg, 2007). However, it was realised that the distribution of the observable genetic differences in a sample depends to a large extent on the distribution of the topology and branch lengths of its genealogy, which in turn depends on the mechanisms of natural selection, mating structure and genetic recombination. In other words, it was realised that genetic polymorphisms can contain valuable information on the mechanisms responsible for genetic variation.

One of the two main aims of the present thesis is to enable the estimation of a set of population genetic models about the mating structure of a population and the historical evolution of this structure. This is achieved by making use of coalescent theory to derive the likelihood for data sets in which each observation consists of the number of nucleotide differences between two homologous DNA sequences, and which contain a large number of observations. The second main aim is to deliver a detailed analysis of some technical issues that should be raised when using this likelihood to make statistical inferences. The best way to clarify the object of our research is to present the population genetic model whose extensions we will be focusing on: the two-island isolation-with-migration (IM) model.

1.2 The isolation-with-migration (IM) model

1.2.1 Definition and genealogy

The IM model is a model for the reproduction of a haploid population of DNA sequences from a given genetic locus. Generations are discrete, which means that a new generation replaces the current one instantaneously after a fixed period of time. Suppose that $\tau_0 \in (0, \infty)$ and $N \in \mathbb{N}$, and that $[\cdot]$ denotes the integer part function. Until generation $[2N\tau_0]$ ago, the population had constant size $2N$ and evolved according to the Wright-Fisher model. In generation $[2N\tau_0]$ ago, it split into subpopulation 1 and subpopulation 2, with sizes $[2b_1N]$ and $[2b_2N]$ respectively, where $b_1, b_2 \in (0, \infty)$, and evolved according to a two-island Wright-Fisher model with potential gene flow until the present.

These reproduction assumptions, which define the IM model for the purposes of the present thesis, can be more thoroughly described. Before generation $[2N\tau_0]$ ago, each generation has $2N$ sequences which are chosen by simple random sample with replacement among the sequences in the previous generation.

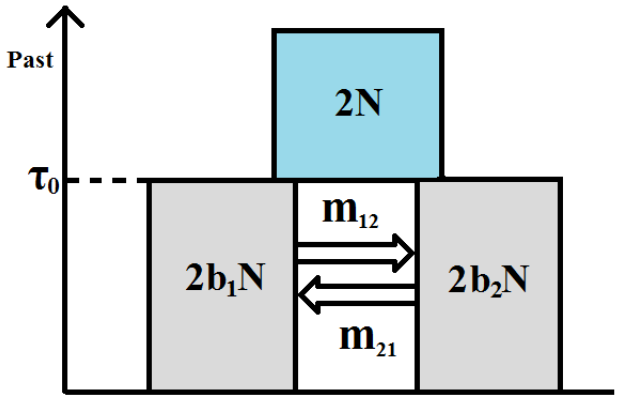

Fig. 1.1 A two-island IM model: the division of the ancestral population occurred in generation $[2N\tau_0]$ ago; each DNA sequence from subpopulation i migrates to subpopulation j independently with probability m_{ij}, for $i \in \{1, 2\}$, and $i \neq j$; the size of each population is given by the integer part of the value inside its corresponding box.

In other words, each sequence chooses independently one sequence from the previous generation to be a copy of, and this choice is carried out uniformly at random. In generation $[2N\tau_0]$, each sequence is still chosen uniformly at random from the $2N$ sequences of the previous generation: the difference is that there are now $[2b_1N] + [2b_2N]$ sequences distributed between two subpopulations. For $0 \leq k < [2N\tau_0]$, generation k ago is the result of the following two steps: first, each sequence from generation $k+1$ in subpopulation i ($i \in \{1, 2\}$) migrates independently to subpopulation j ($j \neq i$) with probability m_{ij}; then, generation k of subpopulation i is created by sampling uniformly at random, with replacement, $[2b_iN]$ sequences, so that reproduction undoes any change in population size caused by gene flow. A diagram of the two-island IM model is shown in fig. 1.1.

To derive the likelihood for a data set, we make use of a particular genealogical process, which we refer to as the *coalescent under the IM model*. To understand what the coalescent under the IM model is, it is useful to consider first the process $Z_N(t), t \in [0, \infty)$, where the time variable t is in units of $2N$ generations, and $Z_N(t)$ represents the state of the genealogy of a present-day sample, $[2Nt]$ generations ago, when the sample is drawn from a population which has evolved according to an IM model with parameter N. More specifically, the random variable $Z_N(t)$ gives the number of distinct ancestral lineages of the sample in each of the two subpopulations, for all t such that $0 \leq [2Nt] \leq [2N\tau_0]$, or in the ancestral population, for all t such that $[2Nt] > [2N\tau_0]$. According to the Wright-Fisher model of reproduction,

each new DNA sequence is generated by copying one single sequence from the previous generation, which means that recombination within loci is assumed to be impossible. Since two or more sequences can descend (i.e. be copies) from the same parent, the number of distinct ancestral lineages of the sample can only decrease or stay the same at each time step into the past, and the process is absorbed when the most recent common ancestor of the sample is reached.

Let the state space of the process, for all t such that $0 \leq [2Nt] \leq [2N\tau_0]$, be represented by $\mathcal{E} = \left\{(\alpha_k)_{k\in\{1,2\}} : \alpha_k \in \mathbb{N}, 1 \leq \alpha_1 + \alpha_2 \leq n\right\}$, where n is the size of the present-day sample and α_k is the number of lineages in subpopulation k. Moreover, let $\epsilon^{(i)}$ be a vector of length two with '1' in position i and '0' in the remaining position, and, for $i, j \in \{1, 2\}$ and $i \neq j$, assume that $M_i := \lim_{N\to\infty} 4N\, m_{ji}\, \frac{b_j}{b_i}$ exists and is finite. Suppose also that $u \in (0\,, \tau_0]$ and N is large enough to ensure that $0 < [2Nu]$. Because between generation $[2Nu]$ and generation 0 the population evolved according to a two-island Wright-Fisher model with gene flow, we know, from Notohara (1990), that

$$\lim_{N\to\infty} \mathrm{P}\left[Z_N(u) = \eta | Z_N(0) = \xi\right] = \left(e^{u\mathbf{Q_{mig}}}\right)_{\xi\eta} \quad ,$$

where $\mathbf{Q_{mig}} = \left(q_{\xi\eta}^{(mig)}\,;\, \xi, \eta \in \mathcal{E}\right)$ is a transition rate matrix with entries

$$q_{\xi\eta}^{(mig)} = \begin{cases} \alpha_i \dfrac{M_i}{2} & \text{if } \eta = \xi - \epsilon^{(i)} + \epsilon^{(j)} \ , \\ \dfrac{1}{b_i} \dfrac{\alpha_i\,(\alpha_i - 1)}{2} & \text{if } \eta = \xi - \epsilon^{(i)}, \\ -\displaystyle\sum_{i=1}^{2} \left[\frac{1}{b_i} \frac{\alpha_i\,(\alpha_i - 1)}{2} + \alpha_i \frac{M_i}{2} \right] & \text{if } \eta = \xi, \\ 0 & \text{otherwise,} \end{cases}$$

and where $\left(e^{u\mathbf{Q_{mig}}}\right)_{\xi\eta}$ denotes the (ξ, η) entry of the probability transition matrix $e^{u\mathbf{Q_{mig}}}$. In other words, in the limit of infinite population size, the process $Z_N(t), t \in [0, \tau_0]$, converges in distribution to a continuous-time Markov chain with transition rate matrix $\mathbf{Q_{mig}}$ (Kingman, 1982b, p. 31, for the type of convergence in question here).

Suppose now that $u \in (\tau_0, \infty)$ and we wish to compute the limit, as $N \to \infty$, of $\mathrm{P}\left[Z_N(u) = \eta | Z_N(0) = \xi\right]$, where $\xi \in \mathcal{E}$ and $\eta \in \{1, 2, ..., n\}$. For this purpose, we let $\mathcal{E}_i$, $i \in \{1, 2, ..., n\}$, denote a set composed of the elements of $\mathcal{E}$ which describe how many lineages, from a total of i lineages, are in each subpopulation,

i.e., $\mathcal{E}_i = \left\{(\alpha_k)_{k\in\{1,2\}} : \alpha_k \in \mathbb{N}, \alpha_1 + \alpha_2 = i\right\}$. Then

$$\begin{aligned}
&\lim_{N\to\infty} \mathrm{P}\left[Z_N(u) = \eta | Z_N(0) = \xi\right] \\
&= \lim_{N\to\infty} \sum_{\gamma\in\mathcal{E}} \mathrm{P}\left[Z_N(u) = \eta | Z_N\left(\tau_0\right) = \gamma, Z_N(0) = \xi\right] \mathrm{P}\left[Z_N\left(\tau_0\right) = \gamma | Z_N(0) = \xi\right] \\
&= \lim_{N\to\infty} \sum_{i=1}^{n} \sum_{\gamma\in\mathcal{E}_i} \mathrm{P}\left[Z_N(u) = \eta | Z_N\left(\tau_0\right) = \gamma, Z_N(0) = \xi\right] \mathrm{P}\left[Z_N\left(\tau_0\right) = \gamma | Z_N(0) = \xi\right] \quad ,
\end{aligned} \tag{1.1}$$

since $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_n$ is a partition of $\mathcal{E}$. To further simplify equation (1.1), suppose we know that there are i lineages ancestral to the sample at generation $[2N\tau_0]$. From the definition of the IM model, each of these lineages was chosen independently and uniformly at random from the $2N$ sequences of the previous generation. In fact, all lineages ancestral to the sample until generation $[2N\tau_0]$ (inclusive) were produced in this manner. This sampling scheme is the same regardless of the value of $Z_N(0)$, and of how the i lineages are distributed between the two subpopulations at τ_0. Hence for all $\gamma \in \mathcal{E}_i$,

$$\mathrm{P}\left[Z_N(u) = \eta | Z_N\left(\tau_0\right) = \gamma, Z_N(0) = \xi\right] = \mathrm{P}\left[Z_N(u) = \eta | Z_N\left(\tau_0\right) \in \mathcal{E}_i\right] \quad .$$

Since our process is in units of $2N$ generations, and the DNA sequences were generated according to the Wright-Fisher model until generation $[2N\tau_0]$ (inclusive), the limiting probability that a sample of i sequences at τ_0 has η ancestors at u follows from the results of Kingman (1982a,b). In particular, for any $i \in \{1, 2, ..., n\}$,

$$\lim_{N\to\infty} \mathrm{P}\left[Z_N(u) = \eta | Z_N(\tau_0) \in \mathcal{E}_i\right] = \left[e^{(u-\tau_0)\mathbf{Q_{anc}}}\right]_{i\eta} \quad ,$$

where the transition rate matrix $\mathbf{Q_{anc}} = \left(q_{i\eta}^{(anc)};\ i, \eta \in \{1, 2, ..., n\}\right)$ has entries

$$q_{i\eta}^{(anc)} = \begin{cases} -\frac{i(i-1)}{2} & \text{if } \eta = i \ , \\ \frac{i(i-1)}{2} & \text{if } \eta = i-1 \ , \\ 0 & \text{otherwise.} \end{cases}$$

Combining the results of Kingman (1982a,b) with those of Notohara (1990), equation (1.1) can be rewritten, for the case of $u > \tau_0$, as

$$\begin{aligned}
&\lim_{N\to\infty} \mathrm{P}\left[Z_N(u) = \eta | Z_N(0) = \xi\right] \\
&= \lim_{N\to\infty} \sum_{i=1}^{n} \mathrm{P}\left[Z_N(u) = \eta | Z_N\left(\tau_0\right) \in \mathcal{E}_i\right] \sum_{\gamma\in\mathcal{E}_i} \mathrm{P}\left[Z_N\left(\tau_0\right) = \gamma | Z_N(0) = \xi\right] \\
&= \sum_{i=1}^{n} \left[e^{(u-\tau_0)\mathbf{Q_{anc}}}\right]_{i\eta} \sum_{\gamma\in\mathcal{E}_i} \left[e^{\tau_0\mathbf{Q_{mig}}}\right]_{\xi\gamma}
\end{aligned}$$

The expression $\sum_{\gamma\in\mathcal{E}_i} \left[e^{\tau_0\mathbf{Q_{mig}}}\right]_{\xi\gamma}$ denotes the limiting probability that, at $t = \tau_0$, the sample has i ancestral lineages, given that, at $t = 0$, the process is in state $\xi \in \mathcal{E}$. If we define $\mathbf{V}$ as a matrix with a row for each $\xi \in \mathcal{E}$ and n columns, where the i^{th} column has 1's in the components corresponding to states in $\mathcal{E}_i$ and zeros otherwise, then $\sum_{\gamma\in\mathcal{E}_i} \left[e^{\tau_0\mathbf{Q_{mig}}}\right]_{\xi\gamma} = \left[e^{\tau_0\mathbf{Q_{mig}}}\ \mathbf{V}\right]_{\xi i}$, and

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{P}\left[Z_N(u) = \eta | Z_N(0) = \xi\right] &= \sum_{i=1}^{n} \left[e^{(u-\tau_0)\mathbf{Q_{anc}}}\right]_{i\eta} \left[e^{\tau_0\mathbf{Q_{mig}}}\ \mathbf{V}\right]_{\xi i} \\
&= \left[e^{\tau_0\mathbf{Q_{mig}}}\ \mathbf{V}\ e^{(u-\tau_0)\mathbf{Q_{anc}}}\right]_{\xi\eta} \quad .
\end{aligned}$$

Summarising, we may say that, as $N \to \infty$, the process $Z_N(t)$ converges in distribution to a continuous-time Markov process which is piecewise time-homogeneous. Its transition rate matrix is $\mathbf{Q_{mig}}$ for $0 \le t \le \tau_0$, and $\mathbf{Q_{anc}}$ for $t > \tau_0$. This limiting process is what we designate as the *coalescent under the IM model*.

1.2.2 Mutation model and inference

The vector of parameters of an IM model – say $\boldsymbol{\psi}$ – typically includes the rates of migration between subpopulations, the splitting time of the ancestral population, and the population sizes. Let $\mathbf{y}$ denote a polymorphism data set of n DNA sequences from a single locus; and let $\mathbf{U}$ be a random vector (or variable) whose value depends on the coalescent process of n sequences for that same locus, and whose distribution depends on $\boldsymbol{\psi}$. To estimate $\boldsymbol{\psi}$ by maximum-likelihood, a possible first step is to derive the distribution of $\mathbf{U}$, which we denote $p\left(\mathbf{u}; \boldsymbol{\psi}\right)$. Apart from this derivation, we also need a model of mutation: a set of assumptions from which the conditional distribution $p\left(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\theta}\right)$ can be derived, or at least estimated by simulation. Then, to obtain

the likelihood of the parameters given a data set for a single locus, the following integral, sometimes referred to as Felsenstein's decomposition (Felsenstein, 1988), is computed:

$$\mathrm{L}\left(\boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{y}\right) = p\left(\mathbf{y}; \boldsymbol{\psi}, \boldsymbol{\theta}\right) = \int p\left(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\theta}\right)\, p\left(\mathbf{u}; \boldsymbol{\psi}\right) \mathrm{d}\mathbf{u} \quad . \tag{1.2}$$

As follows from the previous sections, in our implementations $\mathbf{y}$ will be the number of nucleotide differences between two randomly sampled DNA sequences (also termed 'pairwise nucleotide differences'); hence $\mathbf{U}$ will be a function of the coalescent process (under an IM model) of two sequences only. More specifically, $\mathbf{U}$ is the time, in units of $2N$ generations, until their most recent common ancestor. The model of mutation we use to derive $p\left(\mathbf{y} \mid \mathbf{u}\,; \boldsymbol{\theta}\right)$ is the infinite-sites model of Watterson (1975), and consists of the following two assumptions. First, it is assumed that the number of mutations hitting any one lineage, during a single generation, follows a Poisson distribution with mean μ; or, equivalently, that during t units of $2N$ generations, the distribution of the number of these mutations follows a Poisson distribution with mean $\frac{\theta t}{2}$, where $\theta = 4N\mu$ is the so-called 'scaled mutation rate'. Second, it is assumed that each single nucleotide cannot be hit more than once by a mutation. As will become clear in the next chapters, the infinite-sites mutation model, along with the genealogical process described in the previous section, allow an (almost) fully analytical derivation of $p\left(\mathbf{y}; \boldsymbol{\psi}, \boldsymbol{\theta}\right)$.

When $\mathbf{y}$ is simply the number of pairwise nucleotide differences, the amount of information contained in $\mathrm{L}\left(\boldsymbol{\psi}, \boldsymbol{\theta}; \mathbf{y}\right)$ about $\boldsymbol{\psi}$ is quite limited. Meaningful estimates are typically based on the likelihood of many (at least hundreds) of observations, each one drawn from the population of DNA sequences at a different locus. If loci are chosen far apart from each other in the genome, or separated by recombination hotspots, their genealogical histories can be considered as independent realisations of the IM model. The likelihood for a data set comprising j observations hence becomes the product of j individual likelihoods.

Equation 1.2, or Felsenstein's decomposition, is applicable to a range of data types. For example, before the advent of whole-genome sequencing techniques, when a typical DNA data set would span at most a few independent loci, inference methods for the IM model sought to maximise the information used at each locus (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004, 2007; Hey, 2010): for these methods, $\mathbf{y}$ typically represents the full DNA sequence data of an alignment of possibly dozens of sequences at each locus; $\mathbf{u}$ gives information

on the branching structure (topology) of the genealogy, as well as the times between events of migration and coalescence; and $p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\theta})$ can be defined in terms of different models of nucleotide substitution. The integral in (1.2) is still computed, but not analytically, because an explicit expression for $p(\mathbf{u}; \boldsymbol{\psi})$ is virtually impossible to obtain (see also Wakeley, 2009, p. 266). Instead, it must be estimated by simulation, typically using Markov chain Monte Carlo methods.

1.2.3 Role in speciation studies

In a speciation process, the degree of clustering of a population in terms of genetic polymorphisms increases gradually until the exchange of genes between clusters (i.e. subpopulations) is no longer possible (Nosil, 2012). How significant must the physical and geographic barriers dividing a population be to allow the formation and differentiation of clusters/subpopulations is a question which was for long controversial in evolutionary biology (Mayr, 1997). Today, it is thought that speciation in the absence of significant physical barriers to gene flow, known as *sympatric speciation*, is possible as a result of disruptive selection (Futuyma, 2005). In this type of natural selection, an allele aa is favoured over a portion of the population's range, where a given resource is more abundant, while an allele AA is favoured over the remainder of the range, where another resource is more common. The opportunities for recombination events of Aa individuals are decreased, since heterozygous individuals are less fit than homozygous ones for any of the two resources, and mating events also tend to occur between individuals who share the same portion of the range. If, in addition, the locus under selection is situated in an area of reduced recombination, the genome region around it will diverge much faster than expected under random mating and free recombination, and may lead to a speciation event (Pinho and Hey, 2010; Hey, 2006).

Sympatric speciation due to disruptive selection is now considered the most credible explanation for some patterns of divergence observed in nature. A well-known putative example of incipient sympatric speciation is that of the *Rhagoletis pomonella*. This fly species has two different host plant races inhabiting roughly the same geographic area of the US. The split into one race whose larvae develop in native hawthorns and another race whose larvae develop in the domestic apple is most probably a very recent event, as the introduction of the domestic apple in the US did not occur before the 19th century (Futuyma, 2005, p. 395). Sympatric divergence has also been observed

in laboratory experiments, in particular with *Drosophila melanogaster*, when the disruptively selected character is one that causes assortative mating (Futuyma, 2005, p. 394).

As attested by the meta-analysis of research articles in Pinho and Hey (2010), the available methods to fit the IM model have been used extensively to look for signs of sympatric divergence, or *divergence with gene flow*, in DNA data. During this type of divergence, regions of the genome which are not linked to loci under disruptive selection are expected to experience higher rates of gene flow (Pinho and Hey, 2010). So when confronted with two sympatric subspecies, or two sympatric species which may have achieved reproductive isolation only recently, it is of interest to: a) assess whether there is evidence of a recent period of gene flow; and b) assess whether there is evidence that gene flow has affected all loci except a few that may have been under disruptive selection. For the first purpose, an IM model can be fitted to polymorphism data; for the second purpose, the variation of divergence across loci needs to be analysed.

1.2.4 Review of the relevant literature

The study of the coalescent under the reproduction models that make up each of the two stages of the IM model started a few decades ago. The main results for the coalescent in an isolated Wright-Fisher population (such as the ancestral population of the IM model) are actually the founding results of coalescent theory, and can be found in J. F. C. Kingman's seminal papers (Kingman, 1982a,b). The coalescent under the two-island model, with equal subpopulation sizes and symmetric migration, was first studied by Takahata (1988). The extension of the results of Takahata to the case of different population sizes and asymmetric migration rates was carried out in the aforementioned paper of Notohara (1990).

Before the first implementations of the IM model were developed, most inference methods used in speciation studies would either assume an n-island model with equilibrium migration (Nath and Griffiths, 1996; Beerli and Felsenstein, 1999; Bahlo and Griffiths, 2000), or a model of divergence in complete isolation (Nielsen, 1998; Nielsen et al., 1998; Wakeley and Hey, 1997). Methods such as those of Wakeley (1996a) and of Nielsen and Slatkin (2000) did allow the performance of statistical tests to distinguish between divergence with and without gene flow, but under restrictive assumptions about population sizes and other demographic variables (Nielsen and Wakeley, 2001).

In a 1996 paper, John Wakeley derived the expectation and the variance of the number of pairwise nucleotide differences for an IM model with symmetric migration between two equal-sized subpopulations (Wakeley, 1996b). For the same model, Rosenberg and Feldman (2002) examined how the distribution of the coalescence time of a sample of genes depends on the population divergence time. Excoffier (2004) obtained the distribution and the expectation of the number of pairwise nucleotide differences for an IM model with symmetric migration and an infinite number of subpopulations (i.e. where two lineages from different subpopulations cannot coalesce until they reach the ancestral population). More general results were derived by Wilkinson-Herbots (2008), for an IM model with a finite number of equal-sized descendant populations and symmetric migration.

The first estimation method for the IM model described in section 1.2.1 was developed by Nielsen and Wakeley (2001) and implemented in the computer program *MDIV*. Its development was prompted by the lack of an inference method that could yield joint estimates of the initial split of a population and the level of gene flow between the resulting subpopulations (Nielsen and Wakeley, 2001). The method relies on Markov chain Monte Carlo (MCMC) algorithms to compute either the likelihood or the posterior distribution of parameters, and is suitable for data sets consisting of an alignment of DNA sequences at a single non-recombining locus. The generalisation of the method of Nielsen and Wakeley (2001) to multiple unlinked loci with variable mutation rates was carried out in Hey and Nielsen (2004) and implemented in the computer program *IM*. An approximation to the full joint posterior density of the parameters was obtained by Hey and Nielsen (2007), using an MCMC algorithm to integrate over the space of genealogies and integrating out analytically other nuisance variables. The associated computer program was named *IMa*. In Hey (2010), this method was extended to an IM model with more than two subpopulations and a known phylogeny (computer program *IMa2*).

In the past decade, the availability of large data sets spanning the entire genome has increased considerably. However, the aforementioned MCMC-based implementations of the IM model are computationally expensive even for small numbers of loci, and their running times increase linearly with the number of loci (Wang and Hey, 2010). Fitting an IM model also provides a rather simplified picture of the divergence process, which for some research purposes is clearly insufficient (for example, if one wishes to know whether a process of sympatric speciation has been completed, or whether gene flow occurred due to secondary contact). In addition, Becquet and Przeworski (2009) and Strasburg

and Rieseberg (2010) showed that inference based on the programs *IM* and *IMa* can become unreliable if any of the assumptions made about population structure, recombination, or linkage is severely violated. For these reasons, there has been a significant increase in the demand for methods that not only scale well to genome-sized data, but are also able to estimate increasingly realistic models.

To improve efficiency and scalability, one possible strategy is to work with summary statistics rather than full data patterns. The MCMC-based program *MIMAR* of Becquet and Przeworski (2007, 2009) uses the four summary statistics studied by Wakeley and Hey (1997) to fit the IM model and drops the assumption of no intralocus recombination. Gutenkunst et al. (2009) introduced a method based on the joint sample frequency spectrum (JSFS) that is able to fit a range of demographic models incorporating multiple populations, periods of migration and admixture, splits and joins of populations and changes in population sizes. Based on the same type of data, the more recent implementation of Kamm et al. (2016) can already deal with a large number of individuals and populations, but does not yet include gene flow.

Genome-scale data sets, even when stemming from just a few individuals, tend to be more informative than data sets consisting of many individuals but covering only a relatively short genomic region. In fact, as the sample size for a single locus increases, the probability that an extra sequence adds a deep (i.e. informative) branch to the coalescent tree quickly becomes negligible (see for example Hein et al., 2005, p. 28-29). Data sets of a small number of individuals per locus are also more suitable for likelihood-based inference: if at each locus the observation consists only of a few sequences, the coalescent process of these sequences is relatively simple and can more easily be used to derive the likelihood for the locus concerned.

Among the methods designed for whole-genome sequence data of only a few individuals are those of Mailund et al. (2012), Schiffels and Durbin (2014) and Steinrücken et al. (2015). The fact that they are designed for full polymorphism data makes these methods computationally more expensive than JSFS-based methods. However, they rely on the coalescent with recombination modelled as a hidden Markov process, i.e., they are able to capture the linkage information present in the data. Presently, complex models of demographic history can already be fitted using this approach (see, for example, Steinrücken et al., 2015).

Arguably the only implementations that can be considered *fast* are those based on *blockwise likelihood* methods. These implementations are aimed at

a small number of sampled individuals, and use the information in each of a large number of relatively short and well separated loci: because recombination within loci is disregarded, it is considerably easier to derive explicitly the likelihood for each locus; and because linkage between loci is assumed to be negligible, the likelihood for a data set is just the product of the likelihoods for the individual loci.

Blockwise likelihood methods for the standard two-deme IM model have been developed, for example, by Wilkinson-Herbots (2008) and Wang and Hey (2010), for pairs of DNA sequences at a large number of independent loci, and by Lohse et al. (2011) and Andersen et al. (2014) for larger numbers of sequences at each locus. Lohse et al. (2011) also developed a more general Laplace transform method to calculate blockwise likelihoods for a range of demographic scenarios, which was further extended and efficiently automated in Lohse et al. (2016). Zhu and Yang (2012) developed an implementation, based on triplets of sequences, of an IM model with three species with known phylogeny and symmetric migration between two of them.

Some authors have focused on blockwise likelihood methods for models of divergence which drop the assumption of constant gene flow until the present, and which are therefore more realistic in the context of speciation. In particular, Innan and Watanabe (2006) considered a model in which the level of gene flow between two subpopulations gradually decreases until they become completely isolated from each other. Their calculation of the likelihood given the number of nucleotide differences between pairs of sequences relies on the numerical computation of the coalescence time density at different points in time, which can be computationally expensive. IM models in which gene flow is allowed to cease at some point in the past – hereafter referred to as isolation-with-initial-migration (IIM) models – have also been considered by, for example, Teshima and Tajima (2002), Becquet and Przeworski (2009), Wilkinson-Herbots (2012), Mailund et al. (2012) and Lohse et al. (2015).

1.2.5 Our implementations

The whole of chapter 2, with the exception of section 2.2, is included in Costa and Wilkinson-Herbots (2017). In that chapter, we extend the work of Wilkinson-Herbots (2012), who derived explicit formulae for the distribution of the coalescence time of a pair of sequences under the IIM model, as well as the distribution of the number of nucleotide differences between them. The analytical results of Wilkinson-Herbots (2012) enable a very fast computation of

the likelihood given a data set consisting of observations on pairs of sequences at a large number of independent loci (Wilkinson-Herbots, 2015; Lohse et al., 2015; Janko et al., 2016). However, for mathematical reasons, this work adopted two biologically unrealistic assumptions which may affect the reliability of estimates: symmetric migration and equal subpopulation sizes during the migration period (Figure 1.2, top left-hand plot). We study a more general IIM model which allows for unequal subpopulation sizes during gene flow, as well as during the isolation stage, and drops the assumption of symmetric gene flow during the migration period (Figure 1.2, top right-hand plot).

The core of chapter 2 is dedicated to the description of an efficient method to compute the likelihood given a set of observations on the number of nucleotide differences between pairs of sequences, when each pair comes from a different locus and we assume free recombination between loci and no recombination within loci. Unlike Wilkinson-Herbots (2012), who exploited a result by Griffiths (1981) concerning the equilibrium symmetric island model, we rely on the eigendecomposition of the coalescent matrix. Since we manage to obtain (almost) explicit expression for the likelihood, our method is very fast, and efficient enough to easily deal with asymmetric bidirectional gene flow, unequal population sizes, mutation rate heterogeneity and large numbers of mutations. We also illustrate how to use it to fit the IIM model to real data: the data set of Drosophila sequences from Wang and Hey (2010), containing over 30,000 observations (i.e. loci), is used for this purpose. Finally, we demonstrate, using the same data set, how different models representing different evolutionary scenarios can be compared using likelihood ratio tests. More specifically, we compare three main scenarios: a) divergence without gene flow; b) divergence with potentially asymmetric gene flow until the present; and c) divergence with potentially asymmetric gene flow until some time in the past, and in isolation since then.

The standard two-deme IM model, as well as the IIM model, are nested in what we will call the 'generalised IM' (GIM) model. This last model can be described as a two-island IM model in which migration rates and population sizes are allowed to change at some point in the past (Figure 1.2, bottom plot), and includes, as a special case, a scenario of secondary contact after a period of isolation. In chapter 3, we extend our IIM results to the GIM model, enabling inference on both historical and contemporary rates of gene flow between two closely related species. Once more, we illustrate the implementation of our method using a real data set, this time of several species of Cobitis fish.

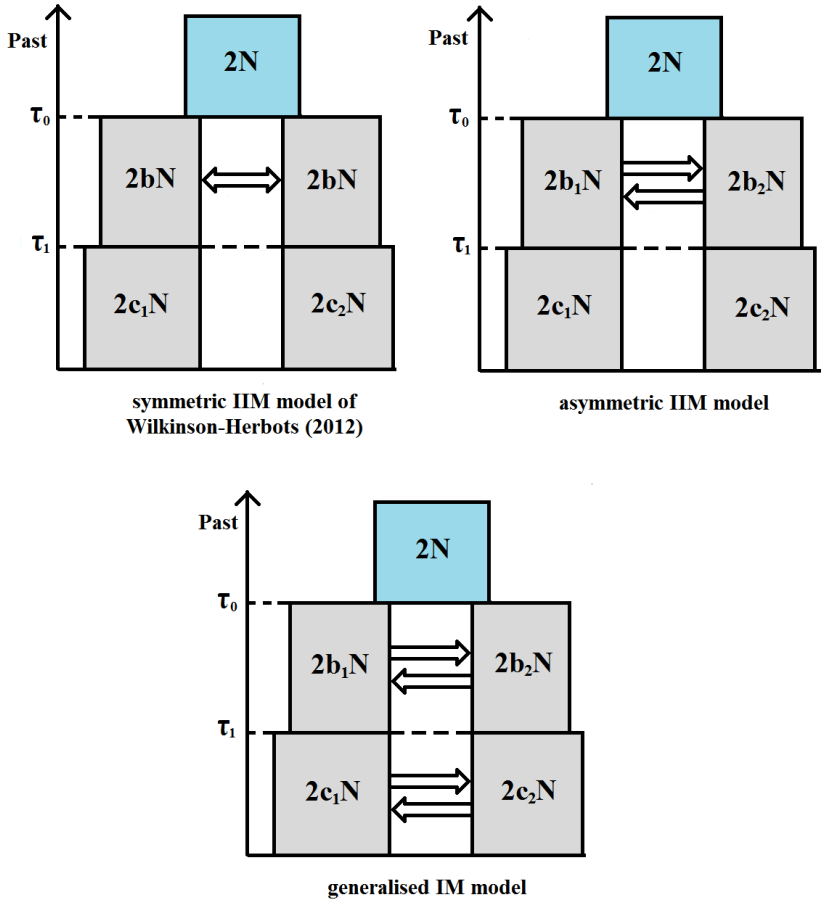

Fig. 1.2 Three extensions of the IM model: the isolation-with-initial-migration (IIM) model of Wilkinson-Herbots (2012), with symmetric gene flow and symmetric subpopulation sizes during gene flow (top left-hand side); an IIM model which drops both symmetry assumptions of Wilkinson-Herbots (2012) (top right-hand side); and an IM model in which migration rates and population sizes are allowed to change at time τ_1 (bottom). The size of each population is given by (the integer part of) the value inside its corresponding box.

An essential output of the present thesis is the R code implementing the theory of chapters 2 and 3, which we believe will be a useful resource for evolutionary geneticists. Costa and Wilkinson-Herbots (2017) includes an R script with detailed instructions that can be used to: fit isolation, IM, and IIM models; simulate observations from these models; and compute Wald-type confidence intervals. This R program does not require high-performance computing resources, as fitting a full IIM model to a data set containing tens of thousands of observations can be carried out in a couple of minutes, using a personal computer. Its latest version is available at https://github.com/ruibarrigana/GIM, and incorporates the following additional features: a wider range of models to fit and simulate from, including many models nested in the full GIM model; the capability of fitting any model using a single R function; and the capability of computing confidence intervals based on the profile likelihood.

1.3 Estimation and model comparison

A substantial part of the present thesis is concerned with deriving the distribution of the number of nucleotide differences between pairs of DNA sequences. For this purpose, we use probability theory, in particular coalescent theory. In chapter 4, however, we rely mostly on general statistics theory, linear algebra, and basic convex geometry to address some inference issues that we were faced with while fitting different extensions of the IM model to real data and comparing how well they fit.

Section 4.1 is essentially devoted to the application of the theory in White (1996) and Jesus and Chandler (2011) to our inference problems. Its main goal is to assess the impact of dropping the assumption of correct model specification: how it changes the distribution of point estimators and of the likelihood ratio statistic; how confidence intervals and p-values should be computed in the light of these changes; and how it affects the very meaning of estimation.

In section 4.2, we focus on the asymptotic distribution of the likelihood ratio statistic under the following (irregular) setting. When comparing how well two models fit some polymorphism data, using the likelihood ratio statistic, it sometimes happens that, under the null hypothesis, the true vector of parameters lies on the boundary of the parameter space, rather than on its interior. Fundamental theoretical results regarding this research topic are given in the papers of Chernoff (1954) and Self and Liang (1987), as well as explicit derivations for relatively simple cases (for example, a vector of p parameters,

where the 2 parameters of interest are assumed to lie on the boundary and the remaining $p-2$ parameters are interior points). Kopylev and Sinha (2011) derive the limiting distribution for a few other specific situations (for example, one parameter of interest and two nuisance parameters on the boundary). Liang and Self (1996) and Chen and Liang (2010) adapt the theory of Chernoff (1954) and Self and Liang (1987) to a pseudolikelihood setting, but their explicit results for the limiting distribution of the likelihood ratio statistic also concern very particular cases.

To avoid the case-by-case approach of these papers, we started by developing a program in R which estimated, by simulation, the mixing coefficients of the asymptotic mixture distribution of the likelihood ratio statistic for the case of only two parameters on the boundary (the parameters of interest); later, we succeded in extending our program to an arbitrary number of parameters of interest on the boundary of the parameter space, and an arbitrary number of nuisance parameters on its interior. The R program is available at https://github.com/ruibarrigana/boundary. Section 4.2 evolved mainly from our wish to demonstrate the correctness of our estimation method. For this purpose, we relied on basic linear algebra and convex geometry, and also benefited from the work of Silvapulle and Sen (2011), in particular their theorem 3.4.2 (our equation 4.8).

Chapter 2

The asymmetric IIM model

2.1 Definition

Recall our definition of an IM model: a population of DNA sequences in which reproduction followed a Wright-Fisher model until $[2N\tau_0]$ generations ago, and, since then until the present, followed a Wright-Fisher two-island model with potential gene flow. If we take an IM model and add the assumption that, $[2N\tau_1]$ generations ago, for $0 < \tau_1 < \tau_0$, gene flow ceased, we obtain an isolation-with-initial-migration (IIM) model. Figure 2.1 illustrates the fullest IIM model dealt with in this thesis. The subpopulation on the left-hand side

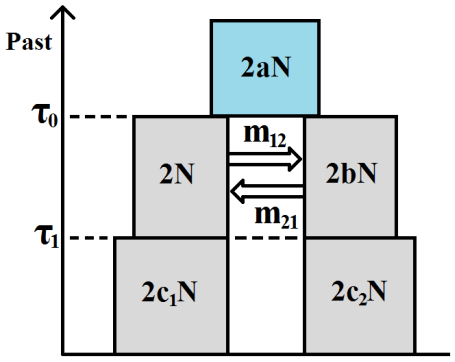

Fig. 2.1 The IIM model. The size of each population is given by the integer part of the value inside its corresponding box. The parameters m_{12} and m_{21} represent the probabilities of migration of each sequence between τ_0 and τ_1. Between τ_1 and the present, there is no gene flow between the subpopulations.

of the diagram is defined as 'subpopulation 1', and the one on the right-hand side as 'subpopulation 2'. The population sizes, in units of DNA sequences, are given by the integer part of the values inside the boxes of the diagram. Note the

slightly different parameterisation of the population sizes: subpopulation 1 now has $[2N]$ sequences during the migration stage, and the ancestral population has $[2aN]$ sequences. In each generation of the migration stage, each sequence in subpopulation i migrates to subpopulation j independently with probability m_{ij}, for $i, j \in \{1, 2\}$ and $i \neq j$.

As before, we are interested in the process $Z_N(t), t \in [0, \infty)$, running backward in time, where t is in units of $2N$ generations, and which gives the number of lineages ancestral to the sample in generation $[2Nt]$ ago: in each subpopulation, for all t such that $[2Nt] \leq [2N\tau_0]$, or in the ancestral population, for all t such that $[2Nt] > [2N\tau_0]$. We will work with present-day samples of two sequences only, so we can simplify the notation for the state space of $Z_N(t)$ as follows. During the isolation stage and the migration stage, the process can only be in state 1 – both lineages in subpopulation 1 –, state 2 – both lineages in subpopulation 2 –, state 3 – one lineage in each subpopulation –, or state 4 – in which lineages have coalesced. In the ancestral population, the lineages have either coalesced already – state 4 –, or have not – state 0. Only states 1, 2 and 3 can be initial states, according to whether we sample two sequences from subpopulation 1, two sequences from subpopulation 2, or one sequence from each subpopulation.

As $N \to \infty$, the genealogical process converges in distribution to a succession of three continuous-time Markov chains, one for each stage of the IIM model (Kingman, 1982a,b; Notohara, 1990). We refer to this stochastic process in continuous time as the *coalescent under the IIM model*. During the isolation stage, the convergence is to a Markov chain defined by the generator matrix

$$\mathbf{Q}_{\mathbf{iso}}^{(i)} = \begin{array}{c} \\ {\scriptstyle (i)} \\ {\scriptstyle (4)} \end{array} \begin{array}{c} \begin{array}{cc} {\scriptstyle (i)} & {\scriptstyle (4)} \end{array} \\ \begin{bmatrix} -\frac{1}{c_i} & \frac{1}{c_i} \\ 0 & 0 \end{bmatrix} \end{array}, \tag{2.1}$$

with $i \in \{1, 2\}$ being the initial state (Kingman, 1982a,b). If 3 is the initial state, the two lineages cannot coalesce before τ_1. During the ancestral stage, the genealogical process converges to a Markov chain with generator matrix

$$\mathbf{Q}_{\mathbf{anc}} = \begin{array}{c} \\ {\scriptstyle (0)} \\ {\scriptstyle (4)} \end{array} \begin{array}{c} \begin{array}{cc} {\scriptstyle (0)} & {\scriptstyle (4)} \end{array} \\ \begin{bmatrix} -\frac{1}{a} & \frac{1}{a} \\ 0 & 0 \end{bmatrix} \end{array} \tag{2.2}$$

(Kingman, 1982a,b). In between, during the migration stage, the convergence is to a Markov chain with generator matrix

$$\mathbf{Q_{mig}} = \begin{array}{c} \\ (1) \\ (3) \\ (2) \\ (4) \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (3) & (2) & (4) \end{array} \\ \begin{bmatrix} -(1+M_1) & M_1 & 0 & 1 \\ \frac{M_2}{2} & -\left(\frac{M_1+M_2}{2}\right) & \frac{M_1}{2} & 0 \\ 0 & M_2 & -(1/b+M_2) & 1/b \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{2.3}$$

(Notohara, 1990). In this matrix, $M_1/2 := \lim_{N\to\infty} 2Nm_{21}b$ represents the rate of backward migration (in continuous time) of a single sequence when in subpopulation 1; the corresponding rate of migration of a single sequence in subpopulation 2 is represented by $M_2/2 := \lim_{N\to\infty} 2Nm_{12}/b$. The rates of coalescence for two lineages in subpopulation 1 or 2 are 1 and $1/b$ respectively. Note that state 3 corresponds to the second row and column, and state 2 to the third row and column. This swap was dictated by mathematical convenience: the matrix $\mathbf{Q_{mig}}$ should be as symmetric as possible because this facilitates a proof in the next section.

2.2 Motivation

The IIM model just defined is the IIM model of Wilkinson-Herbots (2012) without the assumptions of symmetric migration rates and of equal population sizes during divergence. There are essentially two reasons to drop these unrealistic assumptions. First, speciation under asymmetric gene flow is an object of study in its own right. For example, Servedio (2000) studied the impact of asymmetric gene flow and unequal subpopulation sizes on the mechanisms of reinforcement, i.e., on the development of mating preferences that decrease the production of unfit hybrids. Telschow et al. (2006) showed how gene flow asymmetries can be caused by sex ratio imbalances, rather than by differences in subpopulation sizes and densities.

A second reason is the lack of robustness of the estimators of the symmetric IIM model. Whether subpopulation sizes and migration rates are asymmetric or not, and how asymmetric they are, may not be relevant to answer a specific research question. However, according to a short robustness study we carried out (figures 2.2 and 2.3), there is reason to believe that substantial violations

of the assumptions of symmetry and equal population sizes can translate into substantial estimator biases.

Figure 2.2 shows boxplots of estimates of the symmetric migration rate (M) obtained by fitting the IIM model of Wilkinson-Herbots (2012). Figure 2.3 shows the same but for estimates of the parameter $V = \theta(\tau_0 - \tau_1)$, which is defined as twice the expected number of mutations hitting a single lineage during the migration period, and thus gives a measure of the duration of gene flow on a mutational scale. Each boxplot depicts the distribution of 100 estimates, and each estimate was obtained from a different data set. A single data set contained 40,000 independent observations on the number of pairwise nucleotide differences: 10,000 for pairs of sequences drawn from subpopulation 1; 10,000 for pairs of sequences drawn from subpopulation 2; and 20,000 for pairs in which there is one sequence from each subpopulation. The data sets were generated by simulation from an *asymmetric* IIM model in which $M_2 = 0.3$, $\tau_0 = 3$, $\tau_1 = 1.5$, $a = 1.5$, $c_1 = 2.3$, $c_2 = 1.7$ and $\theta = 2$, but using different values of b and M_1. For each of three different values of b (1, 1.25 and 2), we generated four batches of 100 data sets each, with each batch corresponding to a different value of M_1 (0.3, 0.15, 0.05 or 0). In figure 2.2, a red horizontal line indicates the average migration rate $\frac{M_1+M_2}{2}$ used in the simulation of each batch of 100 data sets. In figure 2.3 the red horizontal lines indicate the value of V used to simulate all data sets.

We would expect that if $\hat{M}$ is to give a reasonably accurate picture of the overall level of gene flow between the subpopulations, then it should be close to the average migration rate in the two directions. The boxplots in figure 2.2 suggest that the performance of $\hat{M}$ is strongly affected by the asymmetry of population sizes, but is fairly robust to the asymmetry in migration rates. Similarly, the distribution of $\hat{V}$ shifts away from the true value of V, as the true value of b increases, but is fairly insensitive to decreases in M_2.

2.3 Coalescence time distribution

2.3.1 Models with bidirectional gene flow

Suppose that the genealogical process of two present-day sequences is the coalescent under the IIM model with initial state $i \in \{1, 2, 3\}$. Let $T^{(i)}$ denote the time until their most recent common ancestor (or their coalescence time) and let $S^{(i)}$ denote the number of nucleotide differences between them. To find $f_T^{(i)}$, the density of the coalescence time $T^{(i)}$ given that there is gene flow in both

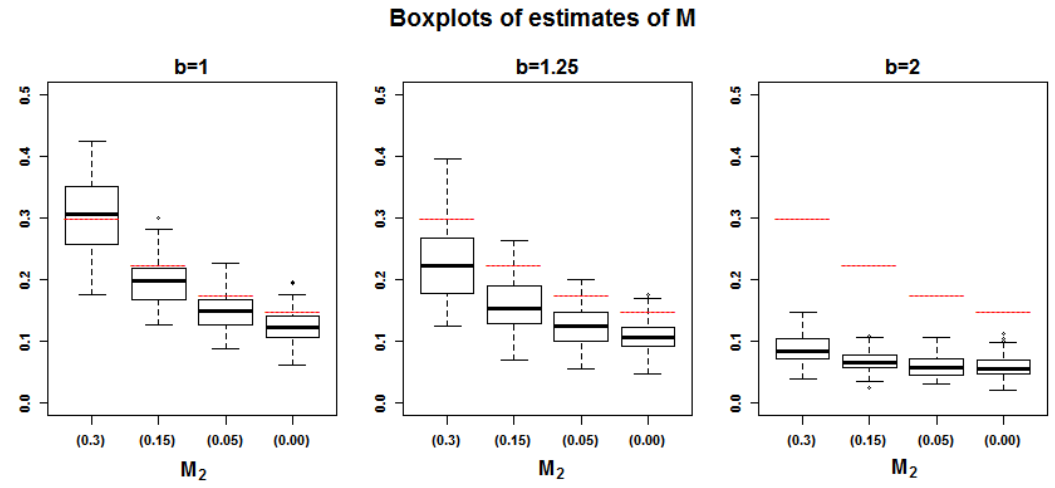

Fig. 2.2 Boxplots of estimates of M under asymmetric migration and unequal population sizes during divergence. The true values of M_2 are given on the x axis. The parameter M_1 is fixed at 0.3 throughout. The parameter b, given on the top of each graph, represents the relative size of subpopulation 2 with respect to subpopulation 1 (see Figure 2.1). The red horizontal lines indicate the true average migration rate.

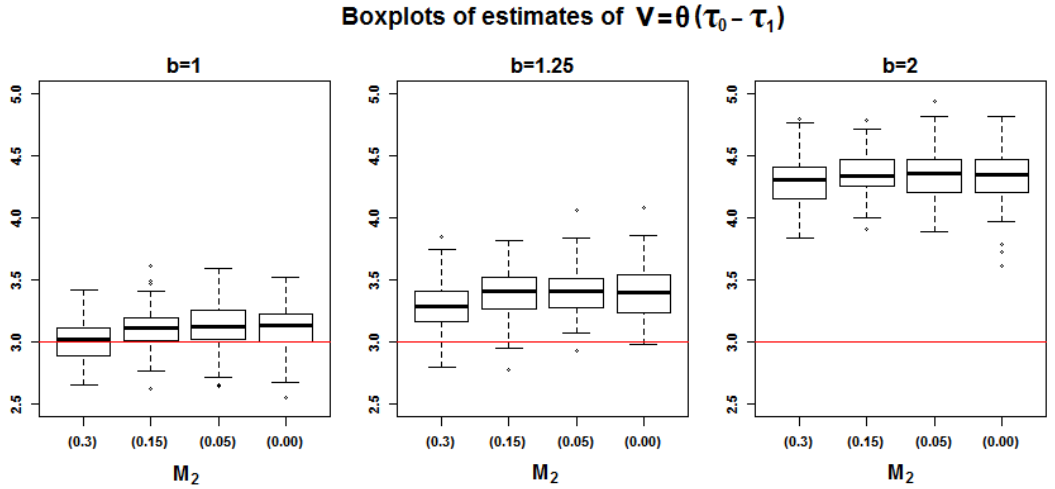

Fig. 2.3 Boxplots of estimates of $V = \theta(\tau_0 - \tau_1)$ under asymmetric migration and unequal population sizes during divergence. The true values of M_2 are given on the x axis. The parameter M_1 is fixed at 0.3 throughout. The parameter b, given on the top of each graph, represents the relative size of subpopulation 2 with respect to subpopulation 1 (see Figure 2.1). The red horizontal lines indicate the true value of V.

directions, we consider separately the three Markov chains mentioned above. We let $T_{iso}^{(i)}$ ($i \in \{1,2\}$), $T_{mig}^{(i)}$ ($i \in \{1,2,3\}$) and $T_{anc}^{(0)}$ denote the times until absorption of the time-homogeneous Markov chains defined by the generator matrices $\mathbf{Q_{iso}^{(i)}}$, $\mathbf{Q_{mig}}$ and $\mathbf{Q_{anc}}$ respectively. And we let the corresponding *pdf*'s (or *cdf*'s) be denoted by $f_{iso}^{(i)}$, $f_{mig}^{(i)}$ and $f_{anc}^{(0)}$ (or $F_{iso}^{(i)}$, $F_{mig}^{(i)}$ and $F_{anc}^{(0)}$). Then $f_T^{(i)}$ can be expressed in terms of the distribution functions just mentioned:

$$f_T^{(i)}(t) = \begin{cases} f_{iso}^{(i)}(t) & \text{for } 0 \le t \le \tau_1, \\ \left[1 - F_{iso}^{(i)}(\tau_1)\right] f_{mig}^{(i)}(t - \tau_1) & \text{for } \tau_1 < t \le \tau_0, \\ \left[1 - F_{iso}^{(i)}(\tau_1)\right] \left[1 - F_{mig}^{(i)}(\tau_0 - \tau_1)\right] f_{anc}^{(0)}(t - \tau_0) & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

for $i \in \{1,2\}$. If 3 is the initial state,

$$f_T^{(3)}(t) = \begin{cases} f_{mig}^{(3)}(t - \tau_1) & \text{for } \tau_1 < t \le \tau_0, \\ \left[1 - F_{mig}^{(3)}(\tau_0 - \tau_1)\right] f_{anc}^{(0)}(t - \tau_0) & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The important conclusion to draw from these considerations is that to find the distribution of the coalescence time under the IIM model, we only need to find the distributions of the absorption times under the simpler processes just defined.

A Markov process defined by the matrix of transition rates $\mathbf{Q_{anc}}$, and starting in state 0, is simply Kingman's coalescent (Kingman, 1982a,b). For such a process, the distribution of the coalescence time is exponential, with rate equal to the inverse of the relative population size:

$$f_{anc}^{(0)}(t) = \frac{1}{a} e^{-\frac{1}{a}t}, \qquad 0 \le t < \infty. \tag{2.6}$$

A Markov process defined by $\mathbf{Q_{iso}^{(i)}}$, $i \in \{1,2\}$, is again Kingman's coalescent, so

$$f_{iso}^{(i)}(t) = \frac{1}{c_i} e^{-\frac{1}{c_i}t}, \qquad 0 \le t < \infty. \tag{2.7}$$

Finally, with respect to the 'structured' coalescent process defined by the matrix $\mathbf{Q_{mig}}$, we prove below that, for $i \in \{1,2,3\}$,

$$f_{mig}^{(i)}(t) = -\sum_{j=1}^{3} V_{ij}^{-1} V_{j4} \lambda_j e^{-\lambda_j t}, \tag{2.8}$$

where V_{ij} is the (i,j) entry of the (non-singular) matrix $\mathbf{V}$, whose rows are the left eigenvectors of $\mathbf{Q_{mig}}$. The (i,j) entry of the matrix $\mathbf{V^{-1}}$ is denoted by V_{ij}^{-1}. The λ_j ($j \in \{1,2,3\}$) are the absolute values of those eigenvalues of $\mathbf{Q_{mig}}$ which are strictly negative (the remaining one is zero). Since the λ_j are real and strictly positive, the density function of $T_{mig}^{(i)}$ is a linear combination of exponential densities.

Proof:

This proof has three parts. Part (i) proves the result under two assumptions: a) $\mathbf{Q_{mig}}$ has three strictly negative eigenvalues and one zero eigenvalue, all of them real; and b) $\mathbf{Q_{mig}}$ is diagonalisable. Part (ii) proves assumption a). Part (iii) proves assumption b). To simplify the notation, we denote $\mathbf{Q_{mig}}$ by $\mathbf{Q}$ throughout the proof.

(i)

Consider the continuous-time Markov chain defined by the matrix $\mathbf{Q}$. Let $P_{ij}(t)$, the (i,j) entry of the matrix $\mathbf{P}(t)$, be the probability that the process is in state j at time t into the past, given that the process starts in state i. $\mathbf{P}(t)$ can be calculated by solving the following initial value problem:

$$\begin{aligned} \mathbf{P}'(t) &= \mathbf{P}(t)\mathbf{Q} \quad ; \\ \mathbf{P}(0) &= \mathbf{I_4} \quad , \end{aligned}$$

where $\mathbf{I_4}$ is the four by four identity matrix. Under the assumptions that $\mathbf{Q}$ is diagonalisable and that its eigenvalues are real, the solution to this initial value problem is given by:

$$\begin{aligned} \mathbf{P}(t) &= \mathbf{P}(0)e^{\mathbf{Q}t} \\ &= \mathbf{V^{-1}}e^{\mathbf{B}t}\mathbf{V} \quad , \end{aligned}$$

where $\mathbf{B}$ denotes the diagonal matrix containing the real eigenvalues β_j, $j \in \{1,2,3,4\}$, of $\mathbf{Q}$, and $\mathbf{V}$ is the matrix of left eigenvectors of $\mathbf{Q}$. Note that $P_{i4}(t)$ is the probability that the process has reached coalescence by time t, if it started in state i. In other words, it is the *cdf* of $T_{mig}^{(i)}$:

$$P_{i4}(t) = F_{mig}^{(i)}(t) = \mathbf{v_{i.}^{-1}}e^{\mathbf{B}t}\mathbf{v_{.4}} \quad ,$$

where $\mathbf{v_{i.}^{-1}}$ is the i^{th} row vector of $\mathbf{V^{-1}}$, and $\mathbf{v_{.4}}$ the 4^{th} column vector of $\mathbf{V}$. Differentiating, we get the *pdf*:

$$\begin{aligned} f_{mig}^{(i)}(t) &= \mathbf{v_{i.}^{-1}} \mathbf{B} e^{\mathbf{B}t} \mathbf{v_{.4}} \\ &= \sum_{j=1}^{4} V_{ij}^{-1} V_{j4} \beta_j e^{\beta_j t} \quad . \end{aligned}$$

If we denote the eigenvalue equal to zero by β_4, and the remaining eigenvalues are strictly negative, this *pdf* can be written as a linear combination of exponential densities:

$$f_{mig}^{(i)}(t) = -\sum_{j=1}^{3} V_{ij}^{-1} V_{j4} \lambda_j e^{-\lambda_j t} \quad , \tag{2.9}$$

where $\lambda_j = |\beta_j|$ for $j \in \{1, 2, 3\}$.

(ii)

As $\mathbf{Q}$ is given by equation (2.3), its characteristic polynomial, $\mathcal{P}_{\mathbf{Q}}(\beta)$, is of the form $\beta \times \mathcal{P}_{\mathbf{Q^{(r)}}}(\beta)$, where $\mathbf{Q^{(r)}}$ is the three by three upper-left submatrix of $\mathbf{Q}$, that is:

$$\mathbf{Q^{(r)}} = \begin{bmatrix} -(1+M_1) & M_1 & 0 \\ M_2/2 & -(M_1+M_2)/2 & M_1/2 \\ 0 & M_2 & -(1/b+M_2) \end{bmatrix}.$$

Thus the eigenvalues of $\mathbf{Q}$ are the solutions to $\beta \times \mathcal{P}_{\mathbf{Q^{(r)}}}(\beta) = 0$. Consequently, one of them is zero (β_4, say) and the remaining three eigenvalues are also eigenvalues of $\mathbf{Q^{(r)}}$.

Now consider the similarity transformation

$$\mathbf{S} = \mathbf{D}\mathbf{Q^{(r)}}\mathbf{D^{-1}} = \begin{bmatrix} -(1+M_1) & \sqrt{\frac{M_1 M_2}{2}} & 0 \\ \sqrt{\frac{M_1 M_2}{2}} & -\frac{M_1+M_2}{2} & \sqrt{\frac{M_1 M_2}{2}} \\ 0 & \sqrt{\frac{M_1 M_2}{2}} & -(\frac{1}{b}+M_2) \end{bmatrix},$$

$$\text{where } \mathbf{D} = \begin{bmatrix} \sqrt{\frac{M_2}{2M_1}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{\frac{M_1}{2M_2}} \end{bmatrix} \quad .$$

Because $\mathbf{S}$ is a symmetric matrix, its eigenvalues are real. Therefore, all the eigenvalues of $\mathbf{Q^{(r)}}$ are real (a similarity transformation does not change the eigenvalues). $\mathbf{S}$ is also a negative definite matrix, since its first, second and

third upper-left determinants are respectively negative, positive, and negative. Hence its eigenvalues are all strictly negative, and so are the eigenvalues of $\mathbf{Q^{(r)}}$. Hence $\mathbf{Q}$ has one zero eigenvalue (β_4) and three real, strictly negative eigenvalues (β_1, β_2 and β_3).

(iii)

Being a symmetric matrix, $\mathbf{S}$ has three independent eigenvectors. A similarity transformation preserves the number of independent eigenvectors, so $\mathbf{Q^{(r)}}$ has three independent eigenvectors as well. We denote by $\mathbf{V^{(r)}}$ the matrix whose rows are the left eigenvectors of $\mathbf{Q^{(r)}}$.

By definition, any left eigenvector $\mathbf{v_{j.}}$ of $\mathbf{Q}$ satisfies the system of equations $\mathbf{x}(\mathbf{Q} - \mathbf{I}\beta_j) = \mathbf{0}$, where $\mathbf{x} = [x_1\ x_2\ x_3\ x_4]$. The first three linear equations of this system are identical to $\mathbf{x^{(r)}}(\mathbf{Q^{(r)}} - \mathbf{I}\beta_j) = \mathbf{0}$, for $j \in \{1, 2, 3\}$ and $\mathbf{x^{(r)}} = [x_1\ x_2\ x_3]$, which is solved by $\mathbf{x^{(r)}} = \mathbf{v_{j.}^{(r)}}$. So this implies that, for $\beta_j \in \{\beta_1, \beta_2, \beta_3\}$, any row vector $\mathbf{x}$ in $\mathbb{R}^4$ that has $\mathbf{v_{j.}^{(r)}}$ as its first three elements will solve the first three equations of the system, whatever the value of x_4. If $x_4 = (V_{j1}^{(r)} + \frac{1}{b}V_{j3}^{(r)})/\beta_j$, that vector will be an eigenvector of $\mathbf{Q}$, because it also solves the fourth equation of the system:

$$\begin{bmatrix} \text{——}\mathbf{v_{j.}^{(r)}}\text{——} & \frac{V_{j1}^{(r)} + \frac{1}{b}V_{j3}^{(r)}}{\beta_j} \end{bmatrix} \begin{bmatrix} -(1+M_1) - \beta_j & M_1 & 0 & 1 \\ \frac{M_2}{2} & -\frac{(M_1+M_2)}{2} - \beta_j & \frac{M_1}{2} & 0 \\ 0 & M_2 & -(\frac{1}{b} + M_2) - \beta_j & \frac{1}{b} \\ 0 & 0 & 0 & -\beta_j \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix},$$

for $\beta_j \in \{\beta_1, \beta_2, \beta_3\}$. For the case of $\beta_j = \beta_4 = 0$, a row eigenvector is $[0\ 0\ 0\ 1]$. Collecting these row eigenvectors in a single matrix, we get $\mathbf{V}$. So,

$$\mathbf{V} = \begin{bmatrix} \text{——} & \mathbf{v_{1.}^{(r)}} & \text{——} & \frac{(V_{11}^{(r)} + \frac{1}{b}V_{13}^{(r)})}{\beta_1} \\ \text{——} & \mathbf{v_{2.}^{(r)}} & \text{——} & \frac{(V_{21}^{(r)} + \frac{1}{b}V_{23}^{(r)})}{\beta_2} \\ \text{——} & \mathbf{v_{3.}^{(r)}} & \text{——} & \frac{(V_{31}^{(r)} + \frac{1}{b}V_{33}^{(r)})}{\beta_3} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

If the matrix $\mathbf{V}$ can be shown to be invertible, then $\mathbf{Q}$ is diagonalisable. This will be the case if the system $\mathbf{xV} = \mathbf{0}$ can only be solved by $\mathbf{x} = [0\ 0\ 0\ 0]$.

Now since the three by three upper-left submatrix of $\mathbf{V}$, $\mathbf{V^{(r)}}$, is full-ranked, $x_1 = x_2 = x_3 = 0$ is a necessary condition for $\mathbf{xV} = \mathbf{0}$. But then $x_4 = 0$, from the last equation of the system. Thus we have shown that $\mathbf{Q}$ is diagonalisable. □

Substituting the *pdf*'s from equations (2.6), (2.7) and (2.8) into the equations (2.4) and (2.5), and denoting by $\mathbf{A}$ the three by three matrix with entries $A_{ij} = -V_{ij}^{-1}V_{j4}$, we obtain

$$f_T^{(i)}(t) = \begin{cases} \frac{1}{c_i}e^{-\frac{1}{c_i}t} & \text{for } 0 \le t \le \tau_1, \\ e^{-\frac{1}{c_i}\tau_1}\sum_{j=1}^{3} A_{ij}\lambda_j e^{-\lambda_j(t-\tau_1)} & \text{for } \tau_1 < t \le \tau_0, \\ e^{-\frac{1}{c_i}\tau_1}\sum_{j=1}^{3} A_{ij}e^{-\lambda_j(\tau_0-\tau_1)}\frac{1}{a}e^{-\frac{1}{a}(t-\tau_0)} & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise,} \end{cases} \tag{2.10}$$

for $i \in \{1, 2\}$, and

$$f_T^{(3)}(t) = \begin{cases} \sum_{j=1}^{3} A_{3j}\lambda_j e^{-\lambda_j(t-\tau_1)} & \text{for } \tau_1 < t \le \tau_0, \\ \sum_{j=1}^{3} A_{3j}e^{-\lambda_j(\tau_0-\tau_1)}\frac{1}{a}e^{-\frac{1}{a}(t-\tau_0)} & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

If $M_1 = M_2$ and $b = 1$ (i.e., in the case of symmetric gene flow and equal subpopulation sizes during the gene flow period), results (2.10) and (2.11) above simplify to the corresponding results in Wilkinson-Herbots (2012) – in this case, the coefficient A_{i3} in the linear combination is zero for $i \in \{1, 2, 3\}$.

2.3.2 Models with unidirectional gene flow and without gene flow

If either M_1 or M_2 is equal to zero, or if both are equal to zero, the above derivation of $f_{mig}^{(i)}$ is no longer applicable, as the similarity transformation in part (ii) of the proof is no longer defined (see the denominators in some entries of the matrix $\mathbf{D}$). In this section, we derive $f_{mig}^{(i)}$, the density of the absorption time of the Markov chain defined by the matrix $\mathbf{Q_{mig}}$ given in equation (2.3), starting from state i, when one or both the migration rates are zero. Again, this is all we need to fill in equations (2.4) and (2.5) and obtain the distribution of the coalescence time of a pair of DNA sequences under the

IIM model. Having gene flow in just one direction considerably simplifies the coalescent. For this reason, we resort to moment generating functions, instead of eigendecomposition, and derive fully explicit *pdf*'s.

Migration from subpopulation 2 to subpopulation 1 backward in time ($M_1 = 0$, $M_2 > 0$)

Let $T_{mig}^{(i)}$ again be defined as the absorption time of the Markov chain generated by $\mathbf{Q_{mig}}$, now with $M_1 = 0$ and $M_2 > 0$, given that the initial state is $i \in \{1, 2, 3\}$. We condition on the state of the coalescent after the first transition to obtain the following system of equations for the *mgf* of $T_{mig}^{(i)}$, where s denotes a dummy variable:

$$\mathrm{E}\left[\exp\left(-sT_{mig}^{(1)}\right)\right] = \left(\tfrac{1}{1+s}\right)$$

$$\mathrm{E}\left[\exp\left(-sT_{mig}^{(2)}\right)\right] = \left(\tfrac{M_2}{1/b+M_2+s}\right)\mathrm{E}\left[\exp\left(-sT_{mig}^{(3)}\right)\right] + \left(\tfrac{1/b}{1/b+M_2+s}\right)$$

$$\mathrm{E}\left[\exp\left(-sT_{mig}^{(3)}\right)\right] = \left(\tfrac{M_2}{M_2+2s}\right)\mathrm{E}\left[\exp\left(-sT_{mig}^{(1)}\right)\right] \quad .$$

(see also more general equations in Wilkinson-Herbots, 1998, and Lohse et al., 2011). Solving this system of equations and applying a partial fraction decomposition (analogous to the work done in Griffiths, 1981, and Nath and Griffiths, 1993, for the case of symmetric migration and equal population sizes), the distributions of $T_{mig}^{(1)}$, $T_{mig}^{(2)}$ and $T_{mig}^{(3)}$ can be expressed as linear combinations of exponential distributions:

$$\mathrm{E}\left[\exp\left(-sT_{mig}^{(1)}\right)\right] = \left(\frac{1}{1+s}\right) \quad ,$$

$$\begin{aligned}\mathrm{E}\left[\exp\left(-sT_{mig}^{(2)}\right)\right] &= \left(\tfrac{M_2}{1/b+M_2+s}\right)\left(\tfrac{M_2}{M_2+2s}\right)\left(\tfrac{1}{1+s}\right) + \left(\tfrac{1/b}{1/b+M_2+s}\right) \\ &= \left(\tfrac{bM_2^2}{(M_2-2)(1-b+bM_2)}\right)\left(\tfrac{1}{1+s}\right) + \left(\tfrac{4bM_2}{(2-M_2)(2+bM_2)}\right)\left(\tfrac{M_2}{M_2+2s}\right) \\ &+ \left(\tfrac{1/b}{1/b+M_2} + \tfrac{b^2M_2^2}{(2+bM_2)(1-b+bM_2)(1/b+M_2)}\right)\left(\tfrac{1/b+M_2}{1/b+M_2+s}\right) \quad ,\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left[\exp\left(-sT_{mig}^{(3)}\right)\right] &= \left(\tfrac{M_2}{M_2+2s}\right)\left(\tfrac{1}{1+s}\right) \\ &= \left(\tfrac{M_2}{M_2-2}\right)\left(\tfrac{1}{1+s}\right) + \left(\tfrac{2}{2-M_2}\right)\left(\tfrac{M_2}{M_2+2s}\right) \quad .\end{aligned}$$

Thus we obtain the following *pdf*'s:

$$f_{mig}^{(1)}(t) = e^{-t} \quad ,$$

$$f_{mig}^{(2)}(t) = \left(\tfrac{bM_2^2}{(M_2-2)(1-b+bM_2)}\right) e^{-t} + \left(\tfrac{4bM_2}{(2-M_2)(2+bM_2)}\right) \tfrac{M_2}{2} e^{-\frac{M_2}{2}t} + \left(\tfrac{1}{1+bM_2} + \tfrac{b^2M_2^2}{(2+bM_2)(1-b+bM_2)(1/b+M_2)}\right) \left(\tfrac{1}{b} + M_2\right) e^{-(1/b+M_2)t} \quad ,$$

$$f_{mig}^{(3)}(t) = \left(\tfrac{M_2}{M_2-2}\right) e^{-t} + \left(\tfrac{2}{2-M_2}\right) \tfrac{M_2}{2} e^{-\frac{M_2}{2}t} \quad ,$$

for $t > 0$.

The *pdf* of the coalescence time of a pair of DNA sequences under an IIM model with $M_1 = 0$ and $M_2 > 0$ can thus be expressed by equations (2.10) and (2.11) above, but now with

$$\boldsymbol{\lambda} = \begin{bmatrix} 1 & \dfrac{M_2}{2} & \dfrac{1}{b} + M_2 \end{bmatrix} \quad ,$$

and

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{bM_2^2}{(M_2-2)(1-b+bM_2)} & \frac{4bM_2}{(2-M_2)(2+bM_2)} & \frac{1}{1+bM_2} + \frac{b^2M_2^2}{(2+bM_2)(1-b+bM_2)(1/b+M_2)} \\ \frac{M_2}{M_2-2} & \frac{2}{2-M_2} & 0 \end{bmatrix} .$$

Migration from subpopulation 1 to subpopulation 2 backward in time ($M_1 > 0$, $M_2 = 0$)

In the opposite case of unidirectional migration, and using the same derivation procedure, we find that:

$$f_{mig}^{(1)}(t) = \left(\tfrac{b^2M_1^2}{(bM_1-2)(b-1+bM_1)}\right) \tfrac{1}{b} e^{-\frac{1}{b}t} + \left(\tfrac{4M_1}{(2-bM_1)(2+M_1)}\right) \tfrac{M_1}{2} e^{-\frac{M_1}{2}t} + \left(\tfrac{1}{(1+M_1)} + \tfrac{M_1^2}{(2+M_1)(b-1+bM_1)(1+M_1)}\right) (1+M_1)\, e^{-(1+M_1)t}$$

$$f_{mig}^{(2)}(t) = \tfrac{1}{b} e^{-\frac{1}{b}t}$$

$$f_{mig}^{(3)}(t) = \left(\tfrac{bM_1}{bM_1-2}\right) \tfrac{1}{b} e^{-\frac{1}{b}t} + \left(\tfrac{2}{2-bM_1}\right) \tfrac{M_1}{2} e^{-\frac{M_1}{2}t} \quad .$$

As a result, the *pdf* of the coalescence time of a pair of sequences under the IIM model, $f_T^{(i)}(t)$, is again given by equations (2.10) and (2.11), now with

$$\boldsymbol{\lambda} = \begin{bmatrix} \frac{1}{b} & \frac{M_1}{2} & 1 + M_1 \end{bmatrix}$$

and

$$\mathbf{A} = \begin{bmatrix} \frac{b^2 M_1^2}{(bM_1-2)(b-1+bM_1)} & \frac{4M_1}{(2-bM_1)(2+M_1)} & \frac{1}{1+M_1} + \frac{M_1^2}{(2+M_1)(b-1+bM_1)(1+M_1)} \\ 1 & 0 & 0 \\ \frac{bM_1}{bM_1-2} & \frac{2}{2-bM_1} & 0 \end{bmatrix}.$$

Distribution of the time until coalescence under an IIM model with $M_1 = M_2 = 0$

In this case, the IIM model reduces to a complete isolation model where both descendant populations may change size at time τ_1 into the past. The distribution of the absorption time $T_{mig}^{(i)}$ corresponding to $\mathbf{Q_{mig}}$ will now be either exponential, if both sampled sequences are from the same subpopulation (i.e. for $i \in \{1, 2\}$), or coalescence will not be possible at all until the ancestral population is reached, if we take a sequence from each subpopulation (i.e. if $i = 3$). It follows that the *pdf* of the coalescence time of a pair of sequences in the IIM model is given by equations (2.10) and (2.11) where

$$\boldsymbol{\lambda} = \begin{bmatrix} 1 & \frac{1}{b} & 0 \end{bmatrix}$$

and $\mathbf{A}$ is the 3×3 identity matrix.

2.4 The likelihood for a multilocus data set

2.4.1 Distribution of the number of pairwise nucleotide differences

Let $S^{(i)}$ denote the number of nucleotide differences in a random sample of two sequences from a given locus, when the ancestral process of these sequences follows the coalescent under the IIM model and the initial state is state i ($i \in \{1, 2, 3\}$). Recall the infinite-sites assumption and assume that the distribution of the number of mutations hitting one sequence in a single generation is Poisson with mean μ. As before, time is measured in units of $2N$

generations and we use the coalescent approximation. Given the coalescence time $T^{(i)}$ of two sequences, $S^{(i)}$ follows a Poisson distribution with mean $\theta T^{(i)}$, where $\theta = 4N\mu$ denotes the scaled mutation rate. Since the *pdf* of $T^{(i)}$, $f_T^{(i)}$, is known, the likelihood $L^{(i)}$ of an observation from a single locus corresponding to the initial state i can be derived by integrating out $T^{(i)}$:

$$L^{(i)}(\boldsymbol{\gamma}, \theta; s) = P\left(S^{(i)} = s; \boldsymbol{\gamma}, \theta\right) = \int_0^\infty P\left(S^{(i)} = s | T^{(i)} = t\right) f_T^{(i)}(t) \mathrm{d}t,$$

where $\boldsymbol{\gamma}$ is the vector of parameters of the coalescent under the IIM model, that is, $\boldsymbol{\gamma} = (a, b, c_1, c_2, \tau_1, \tau_0, M_1, M_2)$. There is no need to compute this integral numerically: because $f_T^{(i)}$ has been expressed in terms of a piecewise linear combination of exponential or shifted exponential densities, we can use standard results for a Poisson process superimposed onto an exponential or shifted exponential distribution.

Equations (18) and (29) of Wilkinson-Herbots (2012) use this superimposition to derive the distribution of S under a mathematically much simpler IIM model with symmetric migration and equal subpopulation sizes during the period of migration. These equations can now be adapted to obtain the *pmf* of S under each of the migration scenarios dealt with in this thesis. The changes accommodate the fact that the density of the coalescence time during the migration stage of the model is now given by a different linear combination of exponential densities, where the coefficients in the linear combination, as well as the parameters of the exponential densities, are no longer the same. The *pmf* of S has the following general form: for $s \in \{0, 1, 2, 3, ...\}$,

$$\begin{aligned} P(S^{(i)} = s) \quad = \quad & \tfrac{(c_i\theta)^s}{(1+c_i\theta)^{s+1}} \left(1 - e^{-\tau_1\left(\frac{1}{c_i}+\theta\right)} \sum_{l=0}^{s} \frac{\left(\frac{1}{c_i}+\theta\right)^l \tau_1^l}{l!} \right) \\ & + e^{-\frac{1}{c_i}\tau_1} \sum_{j=1}^{3} A_{ij} \frac{\lambda_j \theta^s}{(\lambda_j+\theta)^{s+1}} \left(e^{-\theta\tau_1} \sum_{l=0}^{s} \frac{(\lambda_j+\theta)^l \tau_1^l}{l!} \right. \\ & \left. - e^{-\lambda_j(\tau_0-\tau_1)-\theta\tau_0} \sum_{l=0}^{s} \frac{(\lambda_j+\theta)^l \tau_0^l}{l!} \right) \\ & + \tfrac{e^{-\frac{1}{c_i}\tau_1-\theta\tau_0}(a\theta)^s}{(1+a\theta)^{s+1}} \left(\sum_{l=0}^{s} \frac{\left(\frac{1}{a}+\theta\right)^l \tau_0^l}{l!} \right) \sum_{j=1}^{3} A_{ij} e^{-\lambda_j(\tau_0-\tau_1)} \quad , \end{aligned} \tag{2.12}$$

if $i \in \{1, 2\}$, and

$$\begin{aligned} P(S^{(3)} = s) &= \sum_{j=1}^{3} A_{3j} \frac{\lambda_j \theta^s}{(\lambda_j + \theta)^{s+1}} \left(e^{-\theta\tau_1} \sum_{l=0}^{s} \frac{(\lambda_j + \theta)^l \tau_1^l}{l!} \right. \\ &\quad \left. - e^{-\lambda_j(\tau_0 - \tau_1) - \theta\tau_0} \sum_{l=0}^{s} \frac{(\lambda_j + \theta)^l \tau_0^l}{l!} \right) \\ &\quad + \frac{e^{-\theta\tau_0}(a\theta)^s}{(1+a\theta)^{s+1}} \left(\sum_{l=0}^{s} \frac{\left(\frac{1}{a} + \theta\right)^l \tau_0^l}{l!} \right) \sum_{j=1}^{3} A_{3j} e^{-\lambda_j(\tau_0 - \tau_1)} \quad . \end{aligned} \tag{2.13}$$

As defined in section 2.3.1, under bidirectional migration, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ is the vector of the absolute values of the strictly negative eigenvalues of $\mathbf{Q_{mig}}$ and $A_{ij} = -V_{ij}^{-1} V_{j4}$. When migration occurs in one direction only, or when there is no gene flow, the matrix $\mathbf{A}$ and the vector $\boldsymbol{\lambda}$ are defined as in section 2.3.2. In the special case of $M_1 = M_2$ and $b = 1$, equations (2.12) and (2.13) reduce to the results of Wilkinson-Herbots (2012).

2.4.2 Multiple loci

To jointly estimate all the parameters of the IIM model, our method requires a large set of observations on each of the three initial states (i.e., on pairs of sequences from subpopulation 1, from subpopulation 2, and from both subpopulations). To compute the likelihood of such a data set, we use the assumption that observations are independent, so we should have no more than one observation or pair of sequences per locus and there should be free recombination between loci, i.e., loci should be sufficiently far apart, or at least separated by recombination hotspots.

Let each locus for the initial state i be assigned a label $j_i \in \{1_i, 2_i, 3_i, ..., J_i\}$, where J_i is the total number of loci associated with initial state i. Denote by $\theta_{j_i} = 4N\mu_{j_i}$ the scaled mutation rate at locus j_i, where μ_{j_i} is the mutation rate per sequence per generation at that locus. Let θ denote the average scaled mutation rate over all loci and denote by $r_{j_i} = \frac{\theta_{j_i}}{\theta}$ the relative mutation rate of locus j_i. Then $\theta_{j_i} = r_{j_i}\theta$. If the relative mutation rates are known, we can represent the likelihood of the observation at locus j_i simply by $L(\boldsymbol{\gamma}, \theta; s_{j_i})$. By independence, the likelihood of the data set is then given by

$$L(\boldsymbol{\gamma}, \theta; \mathbf{s}) = \prod_{i=1}^{3} \prod_{j_i=1}^{J_i} L(\boldsymbol{\gamma}, \theta; s_{j_i}) \quad . \tag{2.14}$$

In our likelihood method, the r_{j_i} are treated as known constants. In practice, however, the relative mutation rates at the different loci are usually estimated using outgroup sequences (Yang, 2002; Wang and Hey, 2010).

2.5 Results on simulated data

We generated three batches of data sets by simulation, each batch having 100 data sets. Each data set consists of thousands of independent observations, where each observation represents the number of nucleotide differences between two DNA sequences belonging to the same locus, when the genealogy of these sequences follows an IIM model. Each data set of batches 1, 2 and 3 contains 8000, 40,000 and 800,000 observations respectively. In each data set, half of the observations correspond to initial state 3, 1/4 to initial state 1, and 1/4 to initial state 2.

All data sets were generated using the following parameter values: $a = 0.75$, $\theta = 2$, $b = 1.25$, $c_1 = 1.5$, $c_2 = 2$, $\tau_0 = 2$ and $\tau_1 = 1$, $M_1 = 0.5$ and $M_2 = 0.75$. Each observation in a data set refers to a different genetic locus j, and hence was generated using a different scaled mutation rate θ_j for that locus. For batch 1, we first fixed the average mutation rate over all sites to be $\theta = 2$. Then, a vector of 8000 relative size scalars r_j was randomly generated using a Gamma(15,15) distribution. The scaled mutation rate at locus j was then defined to be $\theta_j = r_j\theta$, where r_j denotes the relative mutation rate at locus j, that is, the relative size of θ_j with respect to the average mutation rate, θ. All data sets in batch 1 were generated using the same vector of relative mutation rates. The generation of the mutation rates θ_j used in batches 2 and 3 was carried out following the same procedure.

When fitting the IIM model to data sets generated in this manner, the relative mutation rates r_j are included as known constants in the log-likelihood function to be maximised. So, as far as mutation rates are concerned, only the average over all loci is estimated (i.e. the parameter θ). To increase the robustness and performance of the fitting procedure (see also Wilkinson-Herbots, 2015, and the references therein), we found the maximum-likelihood estimates for a reparameterised model with parameters θ, $\theta_a = \theta a$, $\theta_b = \theta b$, $\theta_{c_1} = \theta c_1$, $\theta_{c_2} = \theta c_2$, $V = \theta\,(\tau_0 - \tau_1)$, $T_1 = \theta\tau_1$, M_1 and M_2.

The boxplots of the maximum-likelihood estimates obtained for the three batches of simulated data are shown in Figures 2.4 and 2.5. For each parameter, the boxplots on the left, centre, and right-hand side refer to batches 1, 2 and 3 respectively. From the boxplots of time and population size parameters,

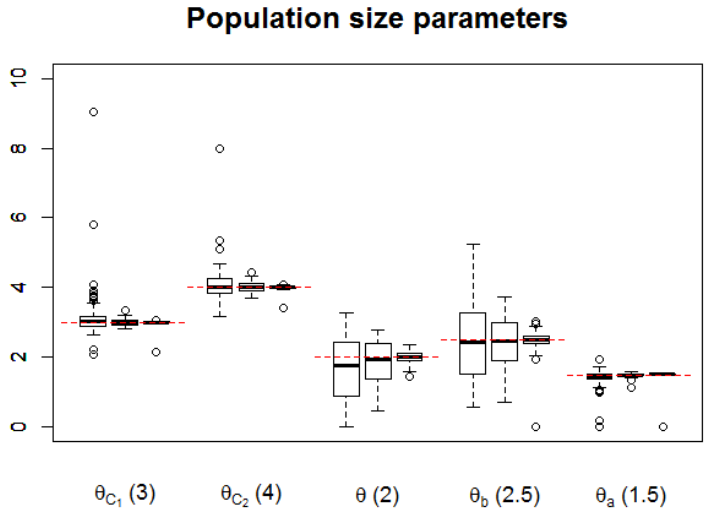

Fig. 2.4 Estimates of population size parameters for simulated data. For each parameter, the estimates shown on the left, centre and right-hand side boxplots are based on sample sizes of 8000, 40 000 and 800 000 loci respectively. The values stated in parentheses are the true parameter values used to generate the data. Horizontal dashed lines indicate the true parameter values for each group of boxplots.

it is seen that the estimates are centred around the true parameter values. Estimates for the migration rates are skewed to the right for batches 1 and 2, possibly because the true parameter values for these rates are closer to the boundary (zero) than the ones for population sizes and splitting times. For all types of parameters, increasing the sample size will decrease the variance of the maximum-likelihood estimator, as would be expected from using the correct expressions for the likelihood. In the case of the migration rate parameters, increasing the sample size eliminates most of the skewness.

The three q-q plots in Figure 2.6 show the maximum-likelihood estimates of θ_{c_1} (a size parameter) obtained from simulated data, plotted against the theoretical quantiles of the standard normal distribution. Figures 2.7 and 2.8 show the corresponding plots for T_1 (a time parameter) and M_1 (a migration parameter). In each figure, the left-hand side, centre and right-hand side q-q plots are based on simulation batches 1, 2 and 3 respectively. It is clear from figures 2.6 to 2.8 that the distributions of the maximum-likelihood estimates of θ_{c_1}, T_1 and M_1 become increasingly Gaussian as the number of observations grows. This is also true for the estimates of the remaining parameters (results not shown). We note also that the distributions of the time and population size estimates have already a reasonably Gaussian shape for a sample size of

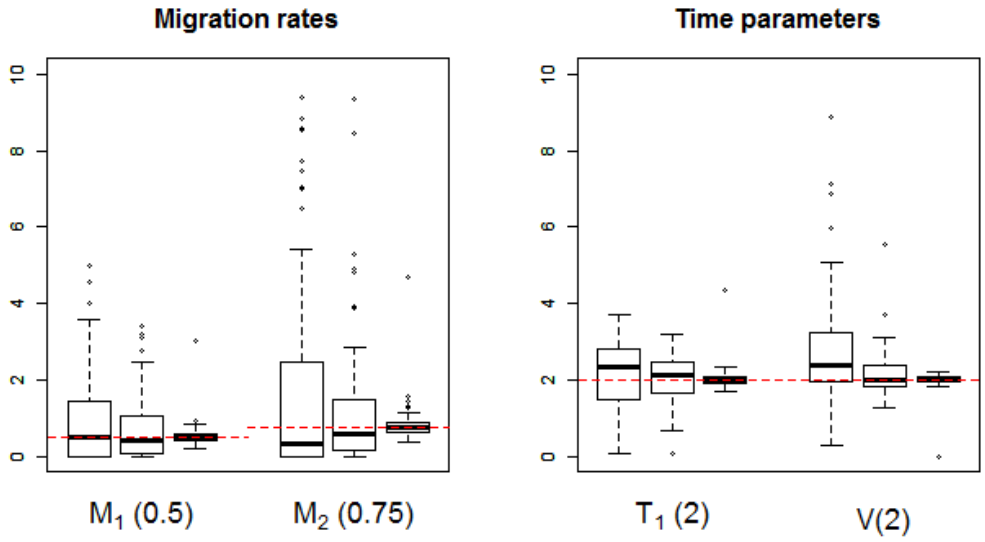

Fig. 2.5 Estimates of migration rates and time parameters for simulated data. For each parameter, the estimates shown on the left, centre and right-hand side boxplots are based on sample sizes of 8000, 40 000 and 800 000 loci respectively. The values stated in parentheses are the true parameter values used to generate the data. Horizontal dashed lines indicate the true parameter values for each group of boxplots.

8000 loci. Again, this is true for the estimates of the remaining time and size parameters as well. The lack of approximate normality of the migration rate estimates for smaller sample sizes suggests care should be taken when making inferences about these parameters – see section 5.1.

2.6 The data from Wang and Hey (2010)

2.6.1 Maximum-likelihood estimation

To illustrate our method, we apply it to a real, multilocus data set from two closely related species of *Drosophila*: *D. simulans* and *D. melanogaster*. The DNA sequence data of Wang and Hey (2010) consist of two subsets: a large subset, which we will call the 'Wang subset', containing 30247 blocks of intergenic sequence, and a smaller subset, which we will refer to as the 'Hutter subset', consisting of 378 blocks of intergenic sequence. Loci in the Wang subset were sampled by Wang and Hey (2010) from a genome alignment of four inbred lines, two from *D. simulans*, and one from each of *D. melanogaster* and *D. yakuba*. To take into account the assumption of no recombination within

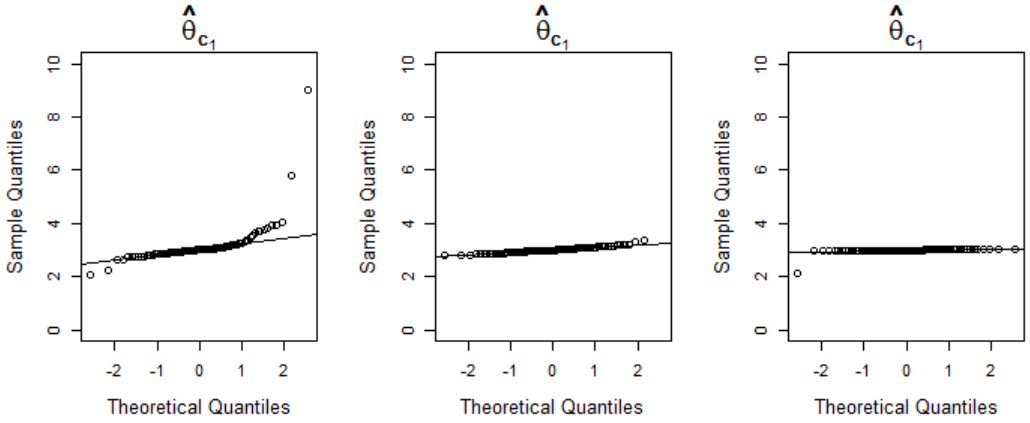

Fig. 2.6 Q-Q plots of maximum-likelihood estimates of the parameter θ_{c_1} obtained from simulated data, against the theoretical quantiles of the standard normal distribution. The estimates shown in the left-hand side, centre and right-hand side q-q plots are based on sample sizes of 8000, 40 000 and 800 000 loci respectively. In the central q-q plot, one outlier with a value above 10 is not shown.

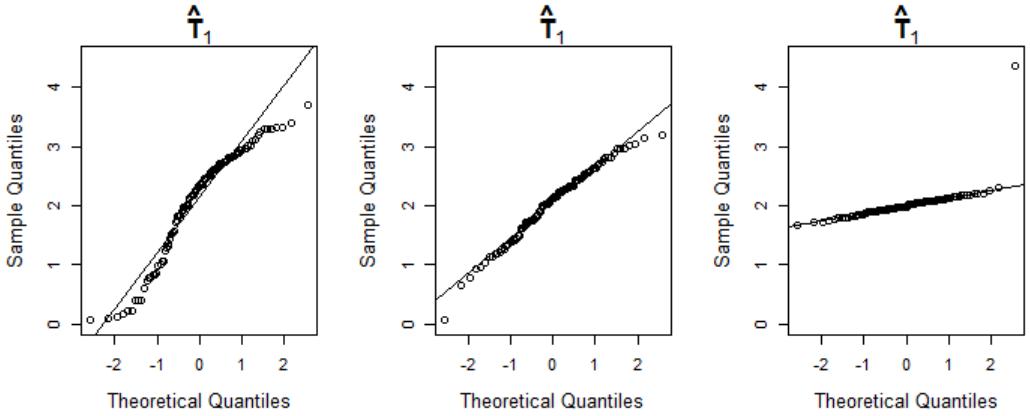

Fig. 2.7 Q-Q plots of maximum-likelihood estimates of the parameter T_1 obtained from simulated data, against the theoretical quantiles of the standard normal distribution. The estimates shown in the left-hand side, centre and right-hand side q-q plots are based on sample sizes of 8000, 40 000 and 800 000 loci respectively.

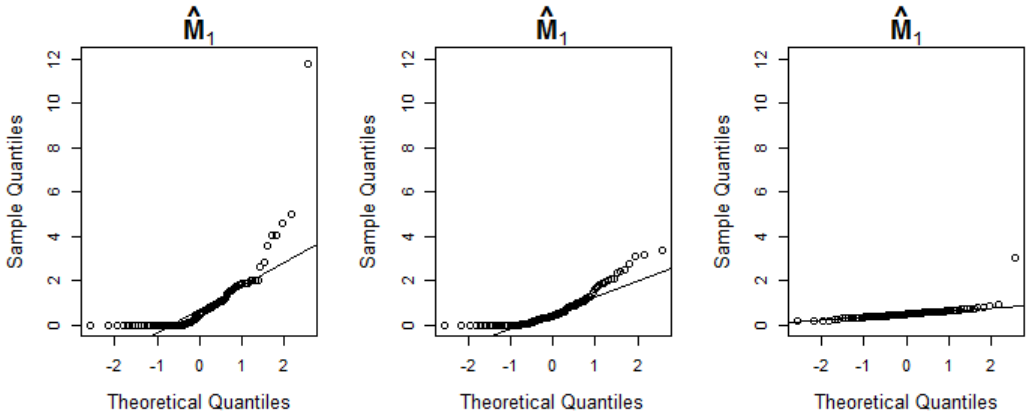

Fig. 2.8 Q-Q plots of maximum-likelihood estimates of the parameter M_1 obtained from simulated data, against the theoretical quantiles of the standard normal distribution. The estimates shown in the left-hand side, centre and right-hand side q-q plots are based on sample sizes of 8000, 40 000 and 800 000 loci respectively.

loci and free recombination between loci, and based on the findings of Hey and Nielsen (2004) regarding the density of apparent recombination events in *Drosophila*, Wang and Hey (2010) chose a locus length of approximately 500 bp and a space of at least 2000 bp between loci. To build the Hutter subset, they drew 378 pairs of *D. melanogaster* sequences from the data set of Hutter et al. (2007), which consists of 378 blocks of sequence sampled from 24 inbred lines of *D. melanogaster*, with an average locus length of 536 bp and an average distance of about 52 kb between consecutive loci. They then joined each of these sequence pairs with their respective *D. yakuba* orthologs from the *simulans-melanogaster-yakuba* genome alignment. Our models are fitted to the *D. melanogaster* and *D. simulans* sequences from both subsets. The *D. yakuba* sequences are only used as outgroup sequences, to estimate the relative mutation rates at the different loci and to calibrate time.

Since our method uses only one pair of sequences at each of a large number of independent loci, and requires observations for all initial states, the following procedure was adopted to select a suitable set of data. According to the genome assembly they stem from, sequences in the Wang subset were given one of three possible tags: 'Dsim1', 'Dsim2' or 'Dmel'. To each of the 30247 loci in the Wang subset we assigned a letter: loci with positions 1, 4, 7,... in the genome alignment were assigned the letter A; loci with positions 2, 5, 8,... were assigned the letter B; and loci with positions 3, 6, 9,..., the letter C. A data set was then built by selecting observations corresponding to initial states 1 and 3

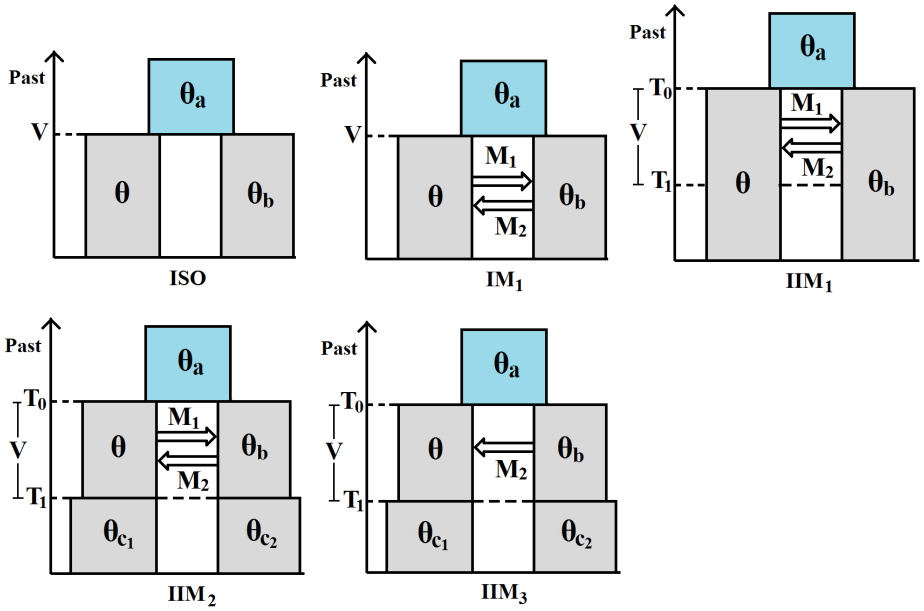

Fig. 2.9 Models fitted to the data of Wang and Hey (2010): $\theta_a = \theta a$, $\theta_b = \theta b$, $\theta_{c_1} = \theta c_1$, $\theta_{c_2} = \theta c_2$, $V = T_0 - T_1 = \theta(\tau_0 - \tau_1)$ and $T_1 = \theta\tau_1$.

from the Wang subset (we used the Dsim1-Dsim2 sequences from loci A, the Dmel-Dsim1 sequences from loci B, and the Dmel-Dsim2 sequences from loci C), whilst observations corresponding to initial state 2 were obtained from the Hutter subset by comparing the two *D. melanogaster* sequences available at each locus.

To estimate the relative mutation rates r_{j_i}, we use the *ad hoc* approach proposed by Yang (2002), which was also used in Wang and Hey (2010) and Lohse et al. (2011). Estimates are computed by means of the following estimator:

$$\hat{r}_{j_i} = \frac{J\,\bar{k}_{j_i}}{\sum_{m=1}^{3}\sum_{n=1}^{J_m}\bar{k}_{n_m}} \quad , \tag{2.15}$$

where J is the total number of loci, and $\bar{k}_{j_i}$ is the average of the numbers of nucleotide differences observed in pairs of one ingroup sequence and one outgroup sequence, at locus j_i.

Table 2.1 contains the maximum-likelihood estimates for the models shown in Figure 2.9. Note that the parameters of time and population size have been reparameterised as in section 2.5, and recall that M_1 and M_2 are the scaled migration rates backward in time. In the diagrams, the left and right subpopulations represent *D. simulans* and *D. melanogaster* respectively.

Table 2.1 Results for the data of Wang and Hey (2010): maximum-likelihood estimates and values of the maximised log-likelihood, for the models shown in Figure 2.9.

Model	$\boldsymbol{\theta_a}$	$\boldsymbol{\theta}$	$\boldsymbol{\theta_b}$	$\boldsymbol{\theta_{c_1}}$	$\boldsymbol{\theta_{c_2}}$	$\mathbf{T_1}$	$\mathbf{V}$	$\mathbf{M_1}$	$\mathbf{M_2}$	$\log \mathbf{L}(\boldsymbol{\psi})$
ISO	4.757	5.628	2.665	-	-	-	13.705	-	-	-90879.14
IM_1	3.974	5.641	2.493	-	-	-	14.965	0.000	0.053	-90276.00
IIM_1	3.191	5.581	2.589	-	-	6.931	9.928	0.000	0.528	-90069.44
IIM_2	3.273	3.357	1.929	6.623	2.647	6.930	9.778	0.000	0.223	-89899.22
IIM_3	3.273	3.357	1.929	6.623	2.647	6.930	9.778	-	0.223	-89899.22

2.6.2 Model selection

In this section, we use a series of likelihood ratio tests for nested models to determine which of the models listed in Table 2.1 fits the data of Wang and Hey (2010) best. The use of such tests in the present situation is not entirely straightforward. We wish to apply a standard large-sample theoretical result which states that, as the number of observations increases, the distribution of the likelihood ratio statistic given by

$$D = -2 \log \lambda (\mathbf{s}) \quad ,$$

where

$$\lambda (\mathbf{s}) = \frac{\sup\limits_{\boldsymbol{\psi} \in \Omega_0} L (\boldsymbol{\psi}; \mathbf{s})}{\sup\limits_{\boldsymbol{\psi} \in \Omega} L (\boldsymbol{\psi}; \mathbf{s})} \tag{2.16}$$

approaches a χ^2 distribution. In equation (2.16), Ω denotes the parameter space and Ω_0 represents the parameter space according to the null hypothesis (H_0). The number of degrees of freedom of the limiting distribution is given by the difference between the dimensions of the two spaces. A list of sufficient regularity conditions for this result can be found, for example, in Casella and Berger (2001, p. 516). One of them is clearly not met in the present case: in the pairwise comparison of some of our models, every point of Ω_0 is a boundary point of Ω. In other words, if H_0 is true, the vector of true parameters $\boldsymbol{\psi}^* \in \Omega_0$, whichever it might be, is on the boundary of Ω. This irregularity is present, for example, when $M_1 = M_2 = 0$ according to H_0 and $[M_1 \quad M_2]^\intercal \in \left\{ [0, \infty)^2 \setminus \{0\}^2 \right\}$ according to H_1. The problem of parameters on the boundary has been the subject of papers such as Self and Liang (1987) and Kopylev and Sinha (2011). The limiting distribution of the likelihood ratio statistic under this irregularity has been derived in these papers, but only for very specific cases. In most of these cases, the use of the naive χ^2_r distribution, with r being the number of additional free parameters according to H_1, turns

out to be conservative, because the correct null distribution is a mixture of χ^2_ν distributions with $\nu \leq r$. Our analysis of the data of Wang and Hey (2010) involves two likelihood ratio tests with parameters on the boundary (ISO vs. IM_1, and IM_1 vs. IIM_1), so we need to check that the naive χ^2_r distribution is also conservative in these cases. This was verified in a short simulation study which we now describe.

We generated 100 data sets from the ISO model, each one consisting of 40,000 observations, and fitted both the ISO model (H_0) and the IM_1 model (H_1) to obtain a sample of 100 realisations of the likelihood ratio statistic. A q-q plot (Figure 2.10, left boxplot) shows that the estimated quantiles of the null distribution are smaller than the corresponding theoretical quantiles of the χ^2 distribution with 2 degrees of freedom (the difference between the dimensions of Ω_0 and Ω in this particular case). In other words, the naive χ^2 distribution should be conservative. Using χ^2_2 instead of the correct null distribution, at a significance level of 5%, the null hypothesis (i.e. the ISO model) was falsely rejected in only 1 out of the 100 simulations performed.

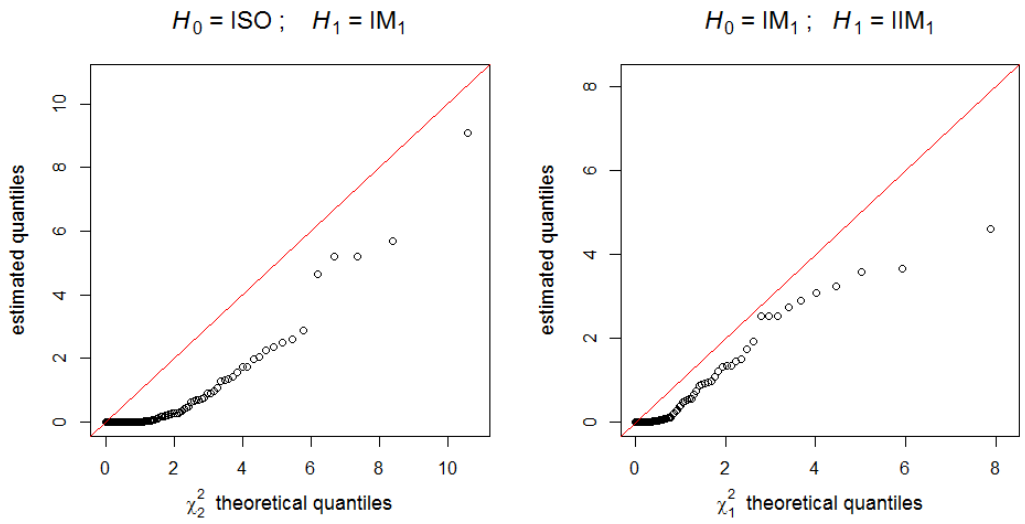

Fig. 2.10 Q-Q plots of the estimated quantiles of the likelihood ratio statistic null distribution against the χ^2 distribution theoretical quantiles. Left plot: H_0 = ISO model, H_1 = IM_1 model. Right plot: H_0 = IM_1 model, H_1 = IIM_1 model.

A similar simulation was carried out with respect to another pair of nested models: the IM_1 model (now as H_0), in which $\tau_1 = 0$, and the IIM_1 model (H_1), in which $\tau_1 \geq 0$. Again the naive χ^2 distribution (this time with only one degree of freedom) was found to be conservative (Figure 2.10, right boxplot). And once more, only in one out of the 100 simulations performed is the null

hypothesis (the IM_1 model) falsely rejected at the 5% significance level, if χ^2_1 is used instead of the correct null distribution.

To select the model that best fitted the data of Wang and Hey (2010), we performed the sequence of pairwise comparisons shown in Table 2.2. For any sensible significance level, this sequence of comparisons leads to the choice of IIM_2 as the best fitting model. In fact, assuming the naive χ^2 as the null distribution, a significance level as low as 1.2×10^{-74} is enough to reject H_0 in each of the three tests. However, since $\hat{M}_1 = 0$ for this model (see Table 2.1), a final (backward) comparison is in order: that between IIM_2 and IIM_3 (which corresponds to fixing M_1 at zero in IIM_2). The nested model in this comparison has one parameter less and, as can be seen in Table 2.1, has the same likelihood. So, in the end, we should prefer IIM_3 to IIM_2.

Table 2.2 Forward selection of the best model for the data of Wang and Hey (2010).

$\mathbf{H_0}$	$\mathbf{H_1}$	$-2\log\lambda(\mathbf{S})$	**P-value**
ISO	IM_1	1206.293	1.140E-262
IM_1	IIM_1	413.12	7.673E-92
IIM_1	IIM_2	340.44	1.187E-74

2.6.3 Confidence intervals for the selected model

Wald-type confidence intervals are straightforward to calculate whenever the vector of estimates is neither on the boundary of the model's parameter space, nor too close to it. In that case, it is reasonable to assume that the vector of *true* parameters does not lie on the boundary either, which justifies the application of standard large-sample results: the vector of maximum-likelihood estimators is consistent, and its distribution approaches a multivariate Gaussian distribution as the sample size grows (see, for example, Pawitan, 2001, p. 258). The confidence intervals can then be calculated using the inverse of the observed Fisher information. In the case of the data of Wang and Hey (2010), the vector of estimates of the selected model (IIM_3) is an interior point of the parameter space. Assuming that the vector of true parameters is also away from the boundary, we computed the Wald 95% confidence intervals shown in Table 2.3. In agreement with our assumption, we note that none of the confidence intervals includes zero.

Table 2.3 Results for the data of Wang and Hey (2010): point estimates and confidence intervals under the model IIM_3.

Parameter	**Estimate**	**95% Confidence intervals**	
		Wald	**Profile likelihood**
θ_a	3.273	(3.101, 3.445)	(3.100, 3.444)
θ	3.357	(3.139, 3.575)	(3.097, 3.578)
θ_b	1.929	(0.079, 3.779)	(0.672, 5.010)
θ_{c_1}	6.623	(6.407, 6.839)	(6.415, 6.843)
θ_{c_2}	2.647	(2.304, 2.990)	(2.331, 3.021)
T_1	6.930	(6.540, 7.320)	(6.542, 7.319)
V	9.778	(9.457, 10.099)	(9.456, 10.098)
M_2	0.223	(0.190, 0.256)	(0.186, 0.259)

For large sample sizes, and for true parameter values not too close to the boundary of the parameter space, the Wald intervals are both accurate and easy to compute. To check how well the Wald intervals for the IIM_3 model fare against the more accurate (see Pawitan, 2001, pp. 47-48), but also computationally more expensive, profile likelihood intervals, we included these in Table 2.3. The two methods yield very similar confidence intervals for all parameters except θ_b. The cause of this discrepancy should lie in the fact that we only had pairs of *D. melanogaster* sequences available from a few hundred loci (θ_b is the size of the *D. melanogaster* subpopulation during the migration stage).

2.6.4 Conversion of estimates

The conversion of point estimates and confidence intervals to more conventional units is based on the estimates of Powell (1997) of the duration of one generation ($g = 0.1$ years) and the speciation time between *D. yakuba* and the common ancestor of *D. simulans* and *D. melanogaster* (10 million years) – see also Wang and Hey (2010) and Lohse et al. (2011). Using these values, we estimated μ, the mutation rate per locus per generation, averaged over all loci, to be $\hat{\mu} = 2.31 \times 10^{-7}$.

In Tables 2.4, 2.5 and 2.6, we show the converted estimates for the best-fitting model IIM_3. The effective population size estimates, in units of diploid individuals, are all based on estimators of the form $\hat{N} = \frac{1}{4\hat{\mu}} \times \hat{\theta}$. For example, the estimate of the ancestral population effective size N_a is given by $\frac{1}{4\hat{\mu}} \times \hat{\theta}_a$. The estimates in years of the time since the onset of speciation and of the time since the end of gene flow are given by $\hat{t}_0 = \frac{g}{2\hat{\mu}} \times (\hat{T}_1 + \hat{V})$ and $\hat{t}_1 = \frac{g}{2\hat{\mu}} \times \hat{T}_1$

respectively. With respect to gene flow, we use $\hat{m}_{12} = \hat{\mu} \times \frac{\hat{M}_2 \hat{b}}{\hat{\theta}}$ as the estimator of the expected fraction of subpopulation 1 which, in each generation, migrates to subpopulation 2, forward in time, and $\hat{s}_{12} = \frac{\hat{M}_2 \hat{b}}{2}$ as the estimator of the expected number of migrant sequences from subpopulation 1 to subpopulation 2 in each generation, also forward in time.

Table 2.4 Effective population size estimates for the data of Wang and Hey (2010) under the model IIM_3 (values in millions of diploid individuals).

Population	**Population size**	**95% Confidence intervals**	
		Wald	**Profile likelihood**
Ancestral population (N_a)	3.549	(3.362, 3.736)	(3.362, 3.735)
D. simulans, migration stage (N)	3.640	(3.404, 3.877)	(3.359, 3.880)
D. melanogaster, migration stage (N_b)	2.092	(0.085, 4.099)	(0.729, 5.433)
D. simulans , isolation stage (N_{c_1})	7.182	(6.949, 7.415)	(6.957, 7.421)
D. melanogaster, isolation stage (N_{c_2})	2.871	(2.498, 3.243)	(2.528, 3.276)

If g and $\hat{\mu}$ are treated as constants, then each of the estimators just given can be expressed as a constant times a product – or a ratio – of the estimators of non-converted parameters. For example, we have that

$$\hat{m}_{12} = \hat{\mu} \times \frac{\hat{M}_2 \hat{b}}{\hat{\theta}} = \text{constant} \times \frac{\hat{M}_2 \hat{b}}{\hat{\theta}} \quad ,$$

and

$$\hat{N}_a = \frac{\hat{\theta}_a}{4\hat{\mu}} = \text{constant} \times \hat{\theta}_a \quad .$$

Hence if we denote the vector of estimators of the converted parameters by $\hat{\boldsymbol{\psi}}_{\boldsymbol{c}}$, then $\hat{\boldsymbol{\psi}}_{\boldsymbol{c}} = \mathbf{W}\hat{\boldsymbol{\psi}}$, where $\mathbf{W}$ is a diagonal matrix and $\hat{\boldsymbol{\psi}} = [\hat{M}_2\hat{b}/\hat{\theta} \quad \hat{\theta}_a \quad \ldots]^\intercal$. Because $\hat{\boldsymbol{\psi}}$ is a maximum-likelihood estimator (of a reparameterised model), its distribution, for a large sample size, is approximately multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}^{-1}$, where $\mathbf{I}$ is the observed Fisher information; hence $\hat{\boldsymbol{\psi}}_{\boldsymbol{c}}$ has a distribution which is approximately multivariate Gaussian, but with covariance matrix $\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^{\mathbf{T}}$. The Wald confidence intervals of Tables 2.4, 2.5 and 2.6 were calculated using this covariance matrix.

Profile likelihood confidence intervals were first computed for the parameterisation $\boldsymbol{\psi} = [M_2 b/\theta \quad \theta_a \quad \ldots]^\intercal$. Then, if $\hat{\mathbf{u}}$ (or $\hat{\mathbf{l}}$) is the vector of estimated upper (or lower) bounds for the parameters in $\boldsymbol{\psi}$, $\mathbf{W}\hat{\mathbf{u}}$ (or $\mathbf{W}\hat{\mathbf{l}}$) will be the vector of estimated upper (or lower) bounds for the converted parameters. This follows from the likelihood ratio invariance – see, for example, Pawitan (2001, p. 47-48).

Table 2.5 Divergence time estimates for the data of Wang and Hey (2010) under the model IIM_3 (values in millions of years ago).

Event	**Time since occurrence**	**95% Confidence intervals**	
		Wald	**Profile likelihood**
Onset of speciation (t_0)	3.624	(3.559, 3.689)	(3.561, 3.691)
Complete isolation (t_1)	1.503	(1.419, 1.588)	(1.419, 1.587)

Note: These are the converted estimates of τ_0 and τ_1 (see Figure 2.1).

Table 2.6 Converted migration rates for the data of Wang and Hey (2010) under the model IIM_3.

Migration parameter	**Point Estimate**	**95% Confidence intervals**	
		Wald	**Profile likelihood**
m_{12}	8.8E-09	(1.1E-10, 1.8E-08)	(3.2E-09, 2.4E-08)
s_{12}	0.064	(0.001, 0.127)	(0.023, 0.172)

Note: These are forward-in-time parameters; m_{12} is the expected fraction of subpopulation 1 (*D. simulans*) which, in each generation, migrates to subpopulation 2 (*D. melanogaster*), during the period of gene flow; s_{12} is the expected number of sequences migrating from subpopulation 1 to subpopulation 2 in each generation, during the same period.

Chapter 3

The generalised isolation-with-migration (GIM) model

3.1 Motivation

In this chapter, we introduce the generalised isolation-with-migration (GIM) model, and derive some theoretical results which enable its estimation by maximum-likelihood. Broadly speaking, a GIM model is an IM model in which population sizes and migration rates are allowed to change at some point in the past, as illustrated in Figure 3.1.

The need for a maximum-likelihood implementation of the GIM model became apparent during the analysis of a set of mRNA sequences in Janko et al. (2016), as part of an effort to reconstruct the speciation history of four species of European loaches (*Cobitis*): *C. elongatoides*, *C. tanaitica*, *C. taenia* and *C. pontica*. In previous studies (Janko et al., 2007; Choleva et al., 2012; Janko et al., 2012), hybrids of *C. elongatoides* with any of the three other species seemed unable to mediate gene flow, as they were found to be either infertile (males) or fertile but clonally reproducing (females). However, the signs of past mitochondrial gene flow between *C. elongatoides* and *C. tanaitica* reported in Choleva et al. (2014) suggested that, at least between these two species, non-clonal hybrids may have existed in the past, and that reproductive isolation may have been accomplished through the initiation of hybrid asexuality. In Janko et al. (2016), the scenario of ancestral gene flow between *C. elongatoides* and any of *C. tanaitica*, *C. taenia*, and *C. pontica* was represented by the IIM model. We were interested in assessing how well it fitted the available data

compared to other models representing two alternative scenarios: divergence without gene flow and divergence with continuous gene flow until the present.

One way to perform this comparison would be to fit the three models depicted in Figure 3.2: a complete isolation model, a standard IM model, and a version of the IIM model in which the sizes of the diverging populations are kept constant. The aim of this latter restriction is to separate, as much as possible, the effect of allowing for a different gene flow pattern from the effect of allowing for population size changes.

In practice, however, one is often ignorant of whether the sizes of the populations during divergence have changed significantly or not, and allowing for population size changes may improve the fit of the models substantially. Therefore, we would like to be able to compare the three gene flow scenarios in a framework which incorporates the full IIM model shown in Figure 2.1. The aim of this chapter is to build such a framework, by developing a maximum-likelihood implementation of the GIM model. This will enable us to compare the three models shown in Figure 3.3, which include the full GIM model (central diagram) and two models nested in it. As in the case of the IIM model, our goal is to enable these models to be fitted to data sets consisting of observations on the number of nucleotide differences between pairs of DNA sequences from a large number of independent, non-recombining loci.

3.2 Theory and methods

From a backward-in-time perspective, the fullest GIM model we consider consists of two successive Wright-Fisher two-island models with migration and one ancestral Wright-Fisher population, as illustrated in Figure 3.1. As before, the subpopulation on the left of the diagram will be referred to as 'subpopulation 1' and the subpopulation on the right as 'subpopulation 2'. The full GIM model makes the same assumptions of the full IIM model (see section 2.1), with one exception: between τ_1 and 0, the two subpopulations evolve according to a two-island Wright-Fisher model with gene flow. More specifically, for $i, j \in \{1, 2\}$, and $i \neq j$, subpopulation i has $[2c_iN]$ sequences and, in each generation, each sequence in subpopulation i migrates to subpopulation j independently with probability m'_{ij}.

As in the case of the IIM model, we are interested in the genealogical process of a random sample of two DNA sequences from the same locus, taken from either of the present populations (or one from each population). Before τ_0 into the past, this process has four possible states: state 1, if there are two

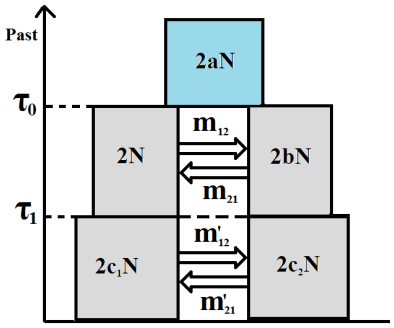

Fig. 3.1 The generalised isolation-with-migration (GIM) model. The size of each population is given by (the integer part of) the value inside its corresponding box. The probabilities of migration of each sequence are given by m_{12} and m_{21}, between τ_0 and τ_1, and by m'_{12} and m'_{21}, between τ_1 and 0.

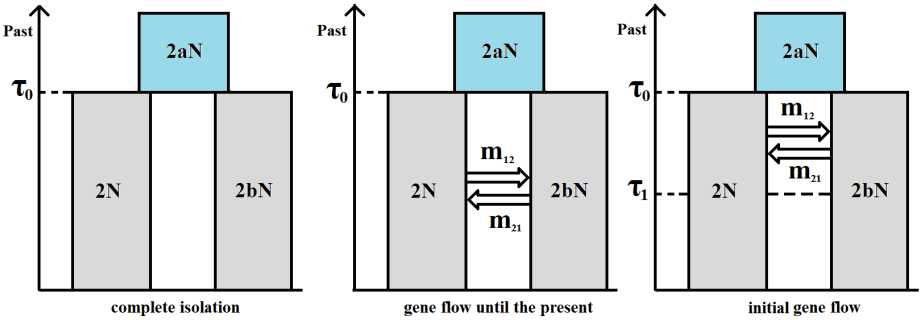

Fig. 3.2 Three models of divergence nested in the isolation-with-initial-migration (IIM) model. The parameters have the same meaning as in Figure 3.1.

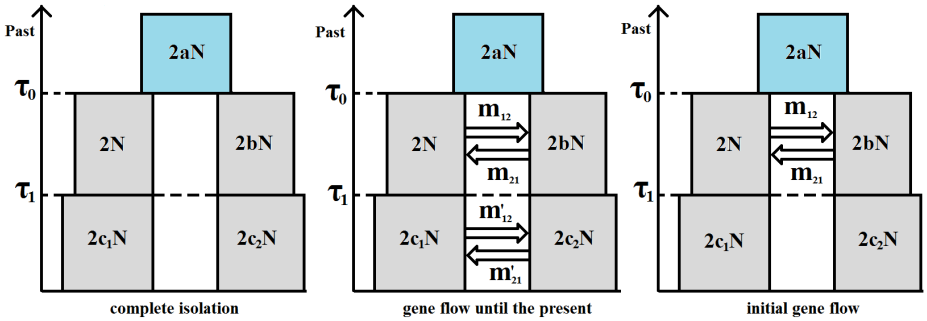

Fig. 3.3 The full GIM model (centre) and two models of divergence nested in it. The parameters have the same meaning as in Figure 3.1.

lineages in population 1; state 2, if there are two lineages in population 2; state 3, if there is one lineage in each population; and state 4, if coalescence has occurred. To facilitate the derivation of the time until the MRCA, or the coalescence time, we let the process have four states even after τ_0 into the past: if the process reaches τ_0 in state $i \in \{1, 2, 3\}$, it remains in this state until the two lineages have coalesced (state 4). As before, the density of the time until coalescence is derived for the *coalescent under the GIM model*, that is, for the continuous-time limit of the genealogical process when time is measured in units of $2N$ generations and N goes to infinity.

3.2.1 The coalescent under the GIM model

The coalescent under the GIM model is defined by the following generator matrices. When $0 \leq t \leq \tau_1$,

$$\mathbf{Q_1} = \begin{array}{c} \\ (1) \\ (3) \\ (2) \\ (4) \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (3) & (2) & (4) \end{array} \\ \begin{bmatrix} -\left(\frac{1}{c_1} + M_1'\right) & M_1' & 0 & \frac{1}{c_1} \\ \frac{M_2'}{2} & -\left(\frac{M_1'+M_2'}{2}\right) & \frac{M_1'}{2} & 0 \\ 0 & M_2' & -\left(\frac{1}{c_2} + M_2'\right) & \frac{1}{c_2} \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{3.1}$$

(Notohara, 1990), where, for $i \in \{1, 2\}$ and $i \neq j$, $M_i'/2 := \lim_{N\to\infty} 2Nm_{ji}' \frac{c_j}{c_i}$ is the rate of backward migration of a single lineage when in subpopulation i. The rate $\frac{1}{c_i}$ is the rate of coalescence of two lineages if both are in subpopulation i. Note again that, for mathematical and notational convenience, state 2 corresponds to row and column 3, whereas state 3 corresponds to row and column 2: this makes $\mathbf{Q_1}$ as symmetric as possible, while reserving states 1 and 2 for the states in which two lineages are present in subpopulation 1 and subpopulation 2 respectively. If $\tau_1 < t \leq \tau_0$,

$$\mathbf{Q_2} = \begin{array}{c} \\ (1) \\ (3) \\ (2) \\ (4) \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (3) & (2) & (4) \end{array} \\ \begin{bmatrix} -(1 + M_1) & M_1 & 0 & 1 \\ \frac{M_2}{2} & -\left(\frac{M_1+M_2}{2}\right) & \frac{M_1}{2} & 0 \\ 0 & M_2 & -\left(\frac{1}{b} + M_2\right) & \frac{1}{b} \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \tag{3.2}$$

where 1 and $\frac{1}{b}$ are the coalescence rates of two lineages in subpopulation 1 and subpopulation 2 respectively, $M_1/2 := \lim_{N\to\infty} 2Nm_{21}b$ and $M_2/2 := \lim_{N\to\infty} 2Nm_{12}/b$.

Finally, for $t > \tau_0$,

$$\mathbf{Q_3} = \begin{array}{c} \\ (1) \\ (3) \\ (2) \\ (4) \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (3) & (2) & (4) \end{array} \\ \begin{bmatrix} -\frac{1}{a} & 0 & 0 & \frac{1}{a} \\ 0 & -\frac{1}{a} & 0 & \frac{1}{a} \\ 0 & 0 & -\frac{1}{a} & \frac{1}{a} \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{3.3}$$

(Kingman, 1982a), where $\frac{1}{a}$ is the rate of coalescence of two lineages in the ancestral population.

The matrix of transition probabilities $\mathbf{P}(t)$ of the coalescent under the GIM model has the following form:

$$\mathbf{P}(t) = \begin{cases} e^{\mathbf{Q_1}t} & \text{for } 0 \leq t \leq \tau_1, \\ e^{\mathbf{Q_1}\tau_1}\, e^{\mathbf{Q_2}(t-\tau_1)} & \text{for } \tau_1 < t \leq \tau_0, \\ e^{\mathbf{Q_1}\tau_1}\, e^{\mathbf{Q_2}(\tau_0-\tau_1)}\, e^{\mathbf{Q_3}(t-\tau_0)} & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Recall that time and population size parameters are assumed strictly positive. In section 2.3.1, we prove that, if both migration rates are also strictly positive, the matrices $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are diagonalisable and have non-positive, real eigenvalues. Moreover, the matrix

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.5}$$

contains a set of four independent right eigenvectors of $\mathbf{Q_3}$, and the corresponding vector of eigenvalues is $(0, -1/a, -1/a, -1/a)$, i.e., $\mathbf{Q_3}$ is also diagonalisable and has non-positive, real eigenvalues. Hence, for $M_1, M_2, M_1', M_2' > 0$, $\mathbf{P}(t)$ can be written as:

$$\mathbf{P}(t) = \begin{cases} \mathbf{G^{-1}} e^{-\mathbf{A}t} \mathbf{G} & \text{for } 0 \leq t \leq \tau_1, \\ \mathbf{G^{-1}} e^{-\mathbf{A}\tau_1} \mathbf{G}\, \mathbf{C^{-1}} e^{-\mathbf{B}(t-\tau_1)} \mathbf{C} & \text{for } \tau_1 < t \leq \tau_0, \\ \mathbf{G^{-1}} e^{-\mathbf{A}\tau_1} \mathbf{G}\, \mathbf{C^{-1}} e^{-\mathbf{B}(\tau_0-\tau_1)} \mathbf{C}\, \mathbf{D^{-1}} e^{-\mathbf{\Gamma}(t-\tau_0)} \mathbf{D} & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise,} \end{cases} \tag{3.6}$$

where $\mathbf{G}$, $\mathbf{C}$ and $\mathbf{D}$ are the matrices of left eigenvectors of $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ respectively, and $-\mathbf{A}$, $-\mathbf{B}$ and $-\mathbf{\Gamma}$ are the corresponding diagonal matrices of non-positive, real eigenvalues. The entries in the main diagonals of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{\Gamma}$ contain the absolute values of the eigenvalues, and are represented by the letters $\alpha_i = (\mathbf{A})_{ii}$, $\beta_i = (\mathbf{B})_{ii}$ and $\gamma_i = (\mathbf{\Gamma})_{ii}$.

If a matrix $\mathbf{Q}$ is a generator matrix of a migration stage in the GIM model, with migration parameters $M_1 = M_2 = 0$ and subpopulation size parameters c_1 and c_2, then its right eigenvectors are those shown in matrix (3.5) and its vector of eigenvalues is $(0, 0, -1/c_1, -1/c_2)$. So when there is no gene flow between τ_0 and τ_1, or no gene flow between τ_1 and the present, $\mathbf{P}(t)$ can still be decomposed as in equation (3.6).

In addition, for all values of M_1 and M_2, the characteristic polynomial of $\mathbf{Q}$, denoted $\mathcal{P}_{\mathbf{Q}}(\beta)$, is of the form $\beta \times \mathcal{P}_{\mathbf{Q}^{(r)}}(\beta)$, where $\mathbf{Q}^{(r)}$ is the three by three upper-left submatrix of $\mathbf{Q}$. So $\mathbf{Q}$ has a zero eigenvalue and its remaining eigenvalues are the eigenvalues of $\mathbf{Q}^{(r)}$. If $\mathbf{Q}$ has migration parameters $M_i = 0$ and $M_j > 0$ ($i, j \in \{1, 2\}$ and $i \neq j$) , $\mathbf{Q}^{(r)}$ becomes triangular. The eigenvalues of $\mathbf{Q}^{(r)}$ will be the entries in its main diagonal. Hence the vector of eigenvalues of $\mathbf{Q}$ will be $\boldsymbol{\lambda} = [-1/c_i \;\; -M_j/2 \;\; -(M_j + 1/c_j) \;\; 0]^\intercal$. If there are no repeated eigenvalues in $\boldsymbol{\lambda}$, we can be sure that $\mathbf{Q}$ is diagonalisable (and its eigenvalues are non-positive and real). In other words, even if there is unidirectional migration between τ_1 and the present, or between τ_0 and τ_1, the probability transition matrix $\mathbf{P}(t)$ can still be decomposed as in (3.6), as long as there are no repeated entries in $\boldsymbol{\lambda}$. Two comments are in order here: first, repeated eigenvalues will occur if and only if $1/c_i = M_j/2$ or $1/c_i = M_j + 1/c_j$; second, the set of parameter values that make these equalities true is negligible when compared to the whole parameter space, so it is very unlikely that the likelihood maximisation procedure chooses values from this set (although one should be careful to avoid using them as initial values).

The probability that, starting in state i ($i \in \{1, 2, 3\}$), the process has reached state 4 by time t is given by the entry corresponding to the i^{th} row and 4^{th} column of $\mathbf{P}(t)$. This is also the cumulative distribution function (*cdf*) of T_i, the time until coalescence, which we denote $F_{T_i}(t)$. If the initial state is i, and $p_{ij}^{(1)}(t)$, $p_{jl}^{(2)}(t)$ and $p_{l4}^{(3)}(t)$ denote transition probability functions of the Markov chains with generator matrices $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ respectively, then:

$$F_{T_i}(t) = \begin{cases} p_{i4}^{(1)}(t) & \text{for } 0 \leq t \leq \tau_1, \\ \displaystyle\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\, p_{j4}^{(2)}(t - \tau_1) & \text{for } \tau_1 < t \leq \tau_0, \\ \displaystyle\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0 - \tau_1)\, p_{l4}^{(3)}(t - \tau_0) & \text{for } \tau_0 < t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

Representing by A_{mn} the (m,n) entry of a matrix $\mathbf{A}$, and by A^{-1}_{mn} the same entry of the matrix $\mathbf{A^{-1}}$, we have that $p^{(1)}_{ij}(t) = \sum_{k=1}^{4} G^{-1}_{ik} G_{kj}\, e^{-\alpha_k t}$, $p^{(2)}_{ij}(t) = \sum_{k=1}^{4} C^{-1}_{ik} C_{kj}\, e^{-\beta_k t}$ and $p^{(3)}_{i4}(t) = \sum_{k=1}^{4} D^{-1}_{ik} D_{k4}\, e^{-\gamma_k t}$.

Differentiating the expression above gives the following density for T_i :

$$f_{T_i}(t) = \begin{cases} f^{(1)}_i(t) & \text{for } 0 \le t \le \tau_1, \\ \sum_{j=1}^{4} p^{(1)}_{ij}(\tau_1)\, f^{(2)}_j(t-\tau_1) & \text{for } \tau_1 < t \le \tau_0, \\ \sum_{j=1}^{4} p^{(1)}_{ij}(\tau_1) \sum_{l=1}^{4} p^{(2)}_{jl}(\tau_0-\tau_1)\, f^{(3)}_l(t-\tau_0) & \text{for } \tau_0 < t < \infty, \\ 0 \quad , & \text{otherwise,} \end{cases} \tag{3.8}$$

where $f^{(1)}_i(t) = \sum_{k=1}^{4} -\alpha_k\, G^{-1}_{ik} G_{k4}\, e^{-\alpha_k t}$, $f^{(2)}_i(t) = \sum_{k=1}^{4} -\beta_k\, C^{-1}_{ik} C_{k4}\, e^{-\beta_k t}$ and $f^{(3)}_i(t) = \sum_{k=1}^{4} -\gamma_k\, D^{-1}_{ik} D_{k4}\, e^{-\gamma_k t}$.

3.2.2 The distribution of the number of pairwise nucleotide differences

Assuming the infinite-sites model, the number of nucleotide differences S_i between two sequences from the same locus, given the time T_i until their most-recent common ancestor, follows a Poisson distribution with mean θT_i (see section 2.4.1): as before, the parameter θ denotes the scaled mutation rate of the locus and is defined as twice the expected number of mutations hitting a lineage in $2N$ generations. Denoting $g_s(t) := \frac{(\theta t)^s}{s!} e^{-\theta t}$, we have, for $s \in \{0, 1, 2, ...\}$,

$$\begin{aligned} \mathrm{P}(S_i = s) &= \mathrm{E}[g_s(T_i)] \\ &= \int_0^{\tau_1} g_s(t)\, f^{(1)}_i(t)\mathrm{d}t + \sum_{j=1}^{4} p^{(1)}_{ij}(\tau_1) \int_{\tau_1}^{\tau_0} g_s(t)\, f^{(2)}_j(t-\tau_1)\, \mathrm{d}t \\ &\quad + \sum_{j=1}^{4} p^{(1)}_{ij}(\tau_1) \sum_{l=1}^{4} p^{(2)}_{jl}(\tau_0-\tau_1) \int_{\tau_0}^{\infty} g_s(t)\, f^{(3)}_l(t-\tau_0)\mathrm{d}t \quad , \end{aligned} \tag{3.9}$$

where i is again the initial state of the coalescent, corresponding to the sampling locations of the two sequences. Changing the limits of integration, equation (3.9) becomes:

$$\mathrm{P}\,(S_i = s) = \int_0^{\tau_1} g_s(t)\, f^{(1)}_i(t)\mathrm{d}t + \sum_{j=1}^{4} p^{(1)}_{ij}(\tau_1) \int_0^{\tau_0-\tau_1} g_s(\tau_1+t)\, f^{(2)}_j(t)\, \mathrm{d}t$$

$$+\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0-\tau_1)\int_0^{\infty} g_s(\tau_0+t)\, f_l^{(3)}(t)\mathrm{d}t \quad .$$

Denoting by W_i, Y_j and Z_l the random variables with *pdf*'s $f_i^{(1)}$, $f_j^{(2)}$ and $f_l^{(3)}$ respectively, the above equation can be written as:

$$\begin{aligned}\mathrm{P}\left(S_i=s\right) &= \mathrm{E}[g_s(W_i)|W_i\le\tau_1]\mathrm{P}[W_i\le\tau_1] \\ &+\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\,\mathrm{E}[g_s(\tau_1+Y_j)|\tau_1+Y_j\le\tau_0]\,\mathrm{P}[\tau_1+Y_j\le\tau_0] \\ &+\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0-\tau_1)\,\mathrm{E}[g_s(\tau_0+Z_l)] \quad .\end{aligned}$$

Making use of the law of total expectation and rearranging the previous equation gives:

$$\begin{aligned}\mathrm{P}\left(S_i=s\right) &= \mathrm{E}[g_s(W_i)] - \mathrm{E}[g_s(W_i)|W_i>\tau_1]\mathrm{P}[W_i>\tau_1] \\ &+\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\ \{\mathrm{E}[g_s(\tau_1+Y_j)] \\ &-\ \mathrm{E}[g_s(\tau_1+Y_j)|\tau_1+Y_j>\tau_0]\,\mathrm{P}[\tau_1+Y_j>\tau_0]\} \\ &+\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0-\tau_1)\,\mathrm{E}[g_s(\tau_0+Z_l)]\end{aligned}$$

Recall that $f_i^{(1)}(t) = \sum_{k=1}^{4} -\alpha_k\, G_{ik}^{-1}G_{k4}\, e^{-\alpha_k t}$, $f_i^{(2)}(t) = \sum_{k=1}^{4} -\beta_k\, C_{ik}^{-1}C_{k4}\, e^{-\beta_k t}$ and $f_i^{(3)}(t) = \sum_{k=1}^{4} -\gamma_k\, D_{ik}^{-1}D_{k4}\, e^{-\gamma_k t}$, and that some eigenvalues of $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ are equal to zero, i.e. some of the $-\alpha_k$, $-\beta_k$ and $-\gamma_k$ are zero. For those α_k, β_k and γ_k that are strictly positive, we let W_k^*, Y_k^* and Z_k^* denote exponentially distributed random variables with rates α_k, β_k and γ_k respectively. The equation above can then be written as:

$$\begin{aligned}\mathrm{P}\left(S_i=s\right) = &-\sum_{k:\alpha_k>0} G_{ik}^{-1}G_{k4}\,\{\mathrm{E}[g_s(W_k^*)] - \mathrm{E}[g_s(W_k^*)|W_k^*>\tau_1]\mathrm{P}[W_k^*>\tau_1]\} \\ &-\sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1)\sum_{k:\beta_k>0} C_{jk}^{-1}C_{k4}\,\{\mathrm{E}[g_s(\tau_1+Y_k^*)] \\ &-\mathrm{E}[g_s(\tau_1+Y_k^*)|\tau_1+Y_k^*>\tau_0]\,\mathrm{P}[\tau_1+Y_k^*>\tau_0]\}\end{aligned}$$

$$- \sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0 - \tau_1) \sum_{k:\gamma_k>0} D_{lk}^{-1} D_{k4}\, \mathrm{E}[g_s(\tau_0 + Z_k^*)] \quad .$$

Finally, making use of the lack of memory property of the exponential distribution gives:

$$\mathrm{P}(S_i = s) = - \sum_{k:\alpha_k>0} G_{ik}^{-1} G_{k4} \left\{ \mathrm{E}[g_s(W_k^*)] - \mathrm{E}[g_s(\tau_1 + W_k^*)]\, e^{-\alpha_k \tau_1} \right\} \tag{3.10}$$

$$- \sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1) \sum_{k:\beta_k>0} C_{jk}^{-1} C_{k4} \left\{ \mathrm{E}[g_s(\tau_1 + Y_k^*)] \right. \tag{3.11}$$

$$\left. - \mathrm{E}[g_s(\tau_0 + Y_k^*)]\, e^{-\beta_k(\tau_0 - \tau_1)} \right\} \tag{3.12}$$

$$- \sum_{j=1}^{4} p_{ij}^{(1)}(\tau_1) \sum_{l=1}^{4} p_{jl}^{(2)}(\tau_0 - \tau_1) \sum_{k:\gamma_k>0} D_{lk}^{-1} D_{k4}\, \mathrm{E}[g_s(\tau_0 + Z_k^*)] \quad . \tag{3.13}$$

To give an explicit statement of the expectations in this probability mass function, we use the results of equations (16) and (17) in Wilkinson-Herbots (2012): for a random variable U following an exponential distribution with rate λ,

$$\mathrm{E}[g_s(U)] \quad = \left(\tfrac{\theta}{\lambda+\theta}\right)^s \left(\tfrac{\lambda}{\lambda+\theta}\right) \tag{3.14}$$

and

$$\mathrm{E}[g_s(\tau + U)] \quad = \left(\tfrac{\theta}{\lambda+\theta}\right)^s \left(\tfrac{\lambda}{\lambda+\theta}\right) e^{-\theta\tau} \textstyle\sum_{l=0}^{s} \frac{(\lambda+\theta)^l \tau^l}{l!} \quad . \tag{3.15}$$

As in section 2.4.2, the assumption of free recombination between loci can be used to obtain the *pmf* of a vector of pairwise differences. Redefining θ as the average mutation rate over all loci in a multilocus data set, we can express this *pmf* as

$$\prod_i \mathrm{p}(s_i\,; \boldsymbol{\gamma}, r_i\theta, u_i) \quad ,$$

where

$$\boldsymbol{\gamma} = [\, a \quad b \quad c_1 \quad c_2 \quad \tau_1 \quad \tau_0 \quad M_1 \quad M_2 \quad M_1' \quad M_2' \,] \quad ,$$

$r_i = \theta_i/\theta$ is the relative mutation rate of locus i, u_i represents the initial state of the process (two sequences from subpopulation 1 or 2, or one from each), and $\mathrm{p}\,(\cdot\,; \boldsymbol{\gamma}, r_i\theta, u_i)$ is the *pmf* of S_i under the GIM model. As in the IIM model, the estimation of parameters is based on the estimated likelihood expression given by

$$\prod_i \mathrm{L}\,(\boldsymbol{\gamma}, \theta; s_i, \hat{r}_i, u_i\,) = \prod_i \mathrm{p}(s_i\,; \boldsymbol{\gamma}, \hat{r}_i\theta, u_i\,) \quad , \tag{3.16}$$

i.e. on the expression obtained by introducing estimates of the relative mutation rates r_i into the likelihood function. These estimates can be computed using outgroup sequences and the estimator of Yang (2002) described in section 2.6.1.

3.3 A Cobitis fish data set

As mentioned in section 3.1, in Janko et al. (2016) we were interested in making inferences about the speciation history of four *Cobitis* species: *C. elongatoides*, *C. tanaitica*, *C. taenia* and *C. pontica*. For each pair of species, we had approximately 2500 observations available and fitted the models shown in Figure 3.4. As the sample size was relatively small, and it was not of particular interest to determine the magnitude of gene flow in each direction, all models of speciation with gene flow assume symmetric migrations rates. The point estimates and confidence intervals for the best-fitting models, according to AIC score, are shown in Tables 3.1, 3.2 and 3.3. In these tables, t_1 and t_0 denote time parameters in millions of years ago (Mya). Their estimates are obtained from the estimates of τ_1 and τ_0 respectively, using the formulae in section 2.6.4.

Due to the relatively small sample size, the difference between Wald-type and profile likelihood confidence intervals was not negligible. Only confidence intervals of this last type are shown in Tables 3.2 and 3.3, since they are, in any case, more accurate than the ones of Wald type. It was not always possible to obtain the precise upper or lower bound of the interval (the computational procedure broke off before the required confidence level was reached). In such cases the upper or lower bounds of the intervals are indicated as less than ($<$) or greater than ($>$) the nearest value that could be obtained.

Initially, the best-fitting model for all pairs of species was the IIM model, either in the 'IIM_7' or the 'IIM_8' version. However, for those pairs that do not include *C. elongatoides*, the estimate of M was above 2, giving an indication that, between t_0 and t_1, the mating structure of the population could be better described by a single panmictic population, rather than by two populations exchanging genes. This motivated the inclusion of the I_6 model in the list of models to be compared. Further supporting the scenario of divergence without gene flow in the group *C. taenia*/*C. tanaitica*/*C. pontica*, the I_6 model turned out to have the best AIC score for the pairs *C. taenia*/*C. pontica* and *C. tanaitica*/*C. pontica* and the second-best AIC score for the pair *C. taenia*/*C. tanaitica* (see Table 3.1).

In Figure 3.5, we summarise the inference about the speciation history of the four *Cobitis* species. It depicts the onset of the speciation process between

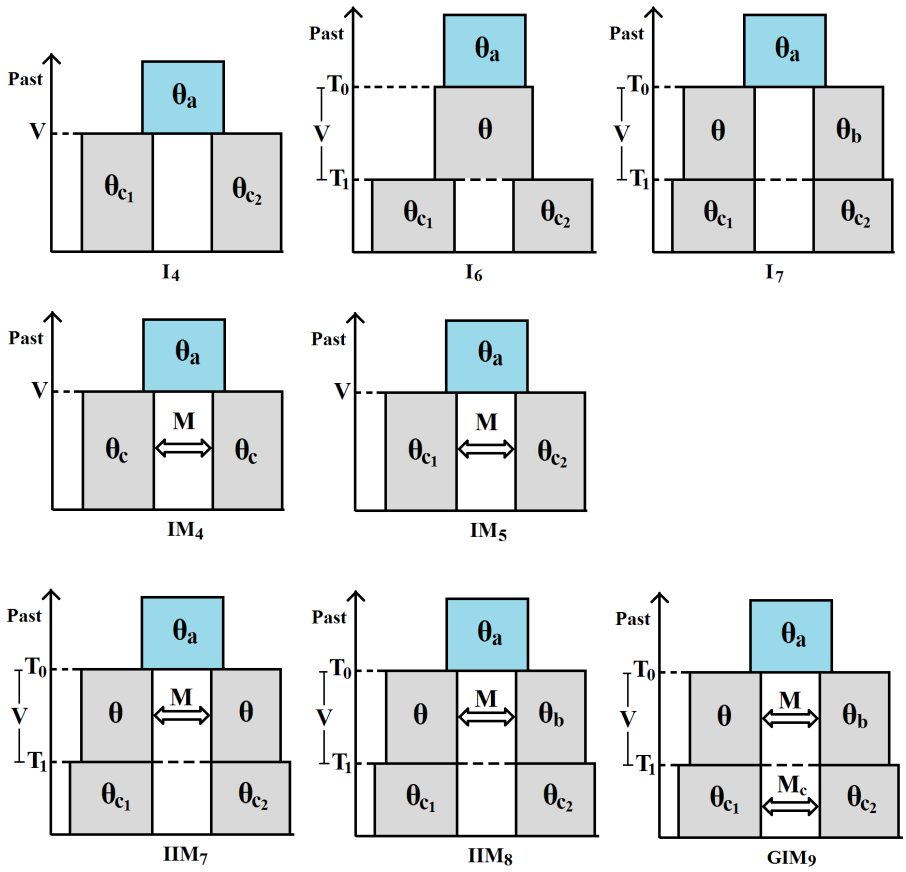

Fig. 3.4 Models fitted to the data of Janko et al. (2016): $\theta_a = \theta a$, $\theta_b = \theta b$, $\theta_{c_1} = \theta c_1$, $\theta_{c_2} = \theta c_2$, $V = T_0 - T_1 = \theta(\tau_0 - \tau_1)$ and $T_1 = \theta\tau_1$.

Table 3.1 Results for the data of Janko et al. (2016): best model fitted to each pair of species and maximum-likelihood estimates.

Species	**Model**	θ_a	θ	θ_b	θ_{c_1}	θ_{c_2}	t_1	t_0	**M**
E-T	IIM_8	0.989	1.336	3.471	0.287	0.122	0.678	8.782	0.137
E-N	IIM_8	0.827	1.294	7.972	0.474	0.289	1.490	9.151	0.166
E-P	IIM_7	1.115	1.223	-	0.257	0.315	0.495	8.811	0.055
N-P	I_6	2.898	1.227	-	0.251	0.370	0.669	5.878	-
T-P	I_6	2.627	1.149	-	0.080	0.254	0.294	5.518	-
T-N	IIM_7	2.907	0.713	-	0.113	0.213	0.458	7.898	2.139

Note: E=*C. elongatoides*, T=*C. taenia*, N=*C. tanaitica*, P=*C. pontica*.

Table 3.2 Results for the data of Janko et al. (2016): profile likelihood confidence intervals for population sizes.

Data	**Model**	θ_a	θ	θ_b	θ_{c_1}	θ_{c_2}
E-T	IIM_8	0.530,1.378	0.920,1.910	1.512,35.370	0.190,0.374	0.080,0.161
E-N	IIM_8	0.189,1.276	0.898,2.015	1.847,>63.289	0.326,0.584	0.193,0.367
E-P	IIM_7	0.754,1.533	0.968,1.671	-	0.061,0.410	0.081,0.481
N-P	I_6	2.195,3.976	1.052,1.395	-	0.193,0.316	0.285,0.470
T-P	I_6	1.874,3.965	0.989,1.307	-	0.049,0.117	0.153,0.380
T-N	IIM_7	1.990,>3.450	0.554,0.853	-	0.071,0.151	0.142,0.278

Note: E=*C. elongatoides*, T=*C. taenia*, N=*C. tanaitica*, P=*C. pontica*. It was not always possible to obtain the precise bounds of the interval. In these cases, the upper or lower bounds of the intervals are indicated as less than (<) or greater than (>) the nearest value that could be obtained.

C. elongatoides and the common ancestral species of *C. tanaitica*, *C. taenia*, and *C. pontica*, around roughly 9 Mya, as well as the relatively recent split of *C. tanaitica*, *C. taenia* and *C. pontica*, around roughly 0.5 Mya. For the process of divergence between *C. elongatoides* and the other species, estimates of the scaled migration M vary between 0.055 and 0.166. Using the formulae in section 2.6.4, this translates to estimates between 0.03 and 0.51 for the expected number of migrant sequences per generation, where the large value at the top of this range is due to the large estimate of θ_b obtained for the species pair *C. elongatoides*/*C. tanaitica*. The establishment of reproductive isolation between *C. elongatoides* and the group of *C. tanaitica*, *C. taenia*, and *C. pontica* has estimates ranging from 0.5 to 1.5 Mya, with large confidence intervals, and is not represented in Figure 3.5.

Our best-fitting model for the species pair *C. elongatoides*/*C. tanaitica* was an IIM model, which reinforces the evidence of ancestral gene flow reported in Choleva et al. (2014) for this species pair. In addition, the current isolation of *C. elongatoides* from the remaining *Cobitis* species is also supported by other types of data. In Janko et al. (2016), an analysis of microsatellite and allozyme data from individuals captured in the *C.elongatoides*-*C. tanaitica* hybrid zone, using the computer programs *Structure 2.3.3* (Pritchard et al.,

Table 3.3 Results for the data of Janko et al. (2016): profile likelihood confidence intervals for speciation times and migration rates.

Data	Model	t_1	t_0	M
E-T	IIM_8	0.376,0.997	7.212,10.151	0.050,0.308
E-N	IIM_8	0.672,2.149	7.811,11.011	0.042,0.576
E-P	IIM_7	0.066,1.184	7.871,10.165	0.022,0.117
N-P	I_6	0.485,0.875	4.151,7.964	-
T-P	I_6	0.166,0.451	3.449,8.269	-
T-N	IIM_7	<0.362,0.675	<7.447,11.891	<2.139,18.504

Note: E=*C. elongatoides*, T=*C. taenia*, N=*C. tanaitica*, P=*C. pontica*. It was not always possible to obtain the precise bounds of the interval. In these cases, the upper or lower bounds of the intervals are indicated as less than (<) or greater than (>) the nearest value that could be obtained.

2000) and *NewHybrids 1.1* (Anderson and Thompson, 2002), pointed towards the inexistence of admixed individuals. The same type of analysis of the *C.elongatoides-C. taenia* hybrid zone, carried out in Janko et al. (2012), arrived at the same conclusion. In addition, experimental crossings of *C.elongatoides-C. tanaitica* (Janko et al., 2016) and *C.elongatoides-C. taenia* (Janko et al., 2012) resulted either in infertile hybrid males or hybrid females unable to produce recombinant progeny (i.e., that reproduce clonally and hence are unable to mediate gene flow).

As already mentioned, the best-fitting speciation models for the group *C. tanaitica*/*C. taenia*/*C. pontica* suggest a relatively recent speciation (≈ 0.5 Mya), followed by an absence of gene flow until the present. This scenario of speciation agrees well with the fact that there is no record either of range overlap or of hybrids in nature for these three species (Janko et al., 2007), even though, at least for the pair *C. taenia*/*C. pontica*, no reproductive incompatibility has emerged (in Janko et al., 2016, artificial crossings between these two species produced mostly viable and fertile hybrids with recombinant gametes).

The balance hypothesis

The 'speciation clock' theory postulates that, as two populations diverge genetically, their reproductive incompatibilities increase gradually (see, for example, Coyne and Orr, 1998). It is based on the pervasive observation that distantly related pairs of taxa are less able to hybridise than closely related ones (Rykena, 2001; Russell, 2003; Sánchez-Guillén et al., 2014). According to this theory, postzygotic isolation is first accomplished by hybrid infertility, and then aggravated by hybrid inviability.

The analysis of the *Cobitis* data, as well as the artificial crossing experiments carried out, suggest that indeed the hybridization capability of the studied species depends on the genetic distance: hybrids of *C. elongatoides*-*C. tanaitica* and *C. elongatoides*-*C. taenia* must have been able to mediate gene flow in the past, but they can no longer do so in the present (as they are infertile, or fertile but reproducing clonally); hybrids of the more closely related *C. taenia*-*C. pontica* are still mostly able to mediate gene flow (even though there is no evidence that gene flow is actually taking place).

In studies of speciation, the loss of sexual reproduction is not commonly considered as a pathway to reproductive isolation (for example, Coyne and Orr, 1998; Russell, 2003). However, the analysis of the *Cobitis* data suggests that the postzygotic isolation of *C. elongatoides* from the group *C. tanaitica/C.taenia/C.pontica* began when the genetic distance between populations was enough to disrupt meiosis, but insufficient to affect the fertility of female hybrids, resulting in clonally reproducing hybrids unable to mediate gene flow. In addition, the fact that some (artificially produced) female hybrids of *C. taenia* and *C. pontica* are beginning to reproduce clonally (Janko et al., 2016) suggests that clonality is likely to become an effective barrier to postzygotic gene flow between these taxa.

The inferred speciation history of the four species of *Cobitis* conforms to the 'balance hypothesis' of Moritz et al. (1989), according to which the loss of sexual reproduction is an intermediate step in the divergence process, occurring at intermediate levels of genetic distance. The relevance of hybrid asexuality as a reproductive isolation mechanism, although seldom recognised in the literature, should extend well beyond the *Cobitis* genus, as suggested by a comparative study of hybridising fish species included in Janko et al. (2016): for each of several fish genera, this study found that the genetic distance between species whose hybrids reproduce asexually is generally lower than the distance between species producing infertile hybrids of both sexes, and higher than the distance between (sub)species producing fertile hybrids only.

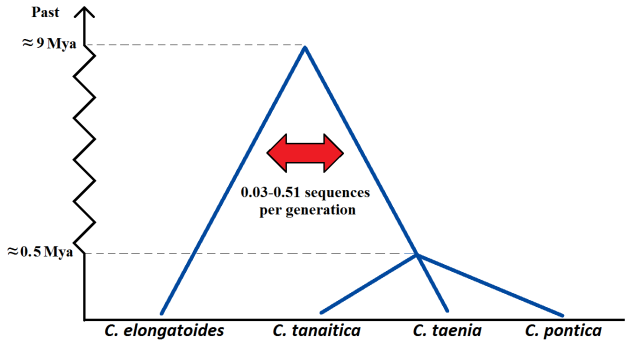

Fig. 3.5 Estimated splitting times and gene flow levels for *C. elongatoides*, *C. tanaitica*, *C. taenia* and *C. pontica* (Janko et al., 2016).

Chapter 4

Improved inference

In the present chapter, we focus on two inference issues that can have a substantial impact on estimation and model selection, and which came into sight during the implementation of IIM and GIM models. The first issue is that of model misspecification. To be able to gain insight into the behaviour of very complex systems, we need to rely on models whose assumptions do not hold exactly and which at best provide good approximations to the true distribution of the data. In the case of the simplest IM or isolation models, the approximation may even be quite imprecise. This is what the extremely low p-values obtained in the likelihood ratio tests for the Wang and Hey (2010) data (section 2.6.2) seem to suggest. For the sake of caution, we should consider that none of the models of speciation that we have implemented matches the reality exactly. Dropping this assumption of correct model specification affects the large-sample distributions of both the maximum-likelihood estimator and of the likelihood ratio statistic. Moreover, it changes the definition of the target value of estimation and the definition of null hypothesis in the likelihood ratio test. In section 4.1 of the present chapter, we state some results available in the literature regarding these issues and illustrate their possible impact by re-computing some confidence intervals and hypothesis tests for the data of Wang and Hey (2010).

The second inference issue considered in this chapter is that of likelihood ratio tests when the true parameter vector lies on the boundary of the parameter space. In chapter 2, in the analysis of the Wang and Hey (2010) data, the use of the 'naive' χ^2 distribution, with degrees of freedom equal to the number of linear constraints imposed by the null hypothesis, was justified by its conservativeness. For that particular data set, using a conservative distribution, rather than the true distribution, does not change the result of any of the model comparisons,

as all the test statistics are significant. In reality, however, this situation is not ideal: for other data sets, with less extreme p-values, it could easily lead to a type II error. The distribution of the likelihood ratio statistic when the true parameter vector lies on the boundary does not have a closed form, but can be estimated efficiently. In section 4.2, we demonstrate the validity of our own method and computer program to perform this estimation, by articulating a number of theoretical results from the literature.

4.1 Model misspecification

4.1.1 Point estimation and Wald confidence intervals

Every statistical inference procedure in the present thesis has relied on the following two assumptions: a) if loci are chosen appropriately, the actual history of molecular evolution at a single locus can be seen as an independent realisation of a stochastic process; and b) this stochastic process is defined according to a version of the IIM or GIM model. A sequence of random variables S_i, $i \in \{1, 2, ..., n\}$, where each of them is a function of the molecular evolution of a different locus, is therefore a sequence of independent random variables. However, as we have seen in the previous two chapters, if S_i represents the number of nucleotide differences between two sequences from locus i, this sequence is in general not identically distributed. In particular, the distribution of S_i depends on the mutation rate of the locus and on the origin of the pair of sequences (subpopulation 1, subpopulation 2 or both subpopulations).

Let $\mathbf{S}_n$ denote a random vector of size n, whose elements are the random variables S_1, S_2,..., S_n defined in the previous paragraph; similarly, let $\mathbf{s}_n$ denote the observed value of $\mathbf{S}_n$ with components s_1, s_2,..., s_n. For $p \in \mathbb{N}$ and a parameter space $\Omega = [0, \infty)^p$, $\boldsymbol{\psi} \in \Omega$ represents the vector of quantities we wish to estimate, and

$$Q_n(\boldsymbol{\psi}; \mathbf{s}_n) = -\sum_{i=1}^{n} \mathrm{p}(s_i; \boldsymbol{\psi}, u_i, r_i)$$

denotes the negated log-likelihood function, with u_i representing the initial state of the process (two sequences from subpopulation 1 or 2, or one from each), r_i the relative mutation rate of locus i, which is assumed to be known, and $\mathrm{p}(\cdot\,; \boldsymbol{\psi}, u_i, r_i)$ the *pmf* of S_i under the model. Suppose further that the true *pmf* of S_i is represented by $\mathrm{q}_i(\cdot)$. If the model is correctly specified, there is a vector $\boldsymbol{\psi} \in \Omega$, termed the *true parameter*, such that, for all $i \in \{1, 2, ..., n\}$ and

$s \in \mathbb{N}$, $\mathrm{q}_i(s) = \mathrm{p}\,(s; \boldsymbol{\psi}, u_i, r_i)$. Denoting this true parameter vector by $\boldsymbol{\psi}^*$, and the Hessian of the negated log-likelihood evaluated at $\boldsymbol{\psi}^*$ by $\mathbf{G}_n\,(\boldsymbol{\psi}^*; \mathbf{s}_n)$, we have that, for large n, and under suitable regularity conditions, the distribution of the maximum-likelihood estimator $\hat{\boldsymbol{\psi}}_n$ is approximately multivariate normal with mean $\boldsymbol{\psi}^*$ and variance $[\mathbf{G}_n^*]^{-1} := \{\mathrm{E}\,[\mathbf{G}_n\,(\boldsymbol{\psi}^*; \mathbf{S}_n)]\}^{-1}$ (theorem 6.4 of White, 1996, with the additional assumption of correct model specification). The vector $\boldsymbol{\psi}^*$ can also be defined as the parameter value which reduces to zero the sum of Kullback-Leibler divergences, i.e.:

$$\sum_{i=1}^{n} D_{\mathrm{KL}}\left[\,\mathrm{q}_i\,(\cdot) \parallel\ \mathrm{p}\,(\cdot\,; \boldsymbol{\psi}^*, u_i, r_i)\,\right] = \sum_{i=1}^{n} \mathrm{E}\left[\log \frac{\mathrm{q}_i\,(S_i)}{\mathrm{p}\,(S_i; \boldsymbol{\psi}^*, u_i, r_i)}\right] = 0 \quad .$$

Let $\mathbf{g}_n\,(\boldsymbol{\psi}; \mathbf{s}_n)$ denote the negated score function. If we allow for model misspecification, i.e., if we admit as possible that there is no $\boldsymbol{\psi}$ such that $\mathrm{q}_i\,(\cdot) = \mathrm{p}\,(\cdot\,; \boldsymbol{\psi}, u_i, r_i)$ for all i, then the distribution of the maximum-likelihood estimator is well approximated, for large n, by a multivariate normal distribution with mean

$$\boldsymbol{\psi}_n^* := \operatorname*{arg\,min}_{\boldsymbol{\psi} \in \Omega} \sum_{i=1}^{n}\ D_{\mathrm{KL}}\left[\,\mathrm{q}_i\,(\cdot) \parallel\ \mathrm{p}\,(\cdot\,; \boldsymbol{\psi}, u_i, r_i)\,\right]$$

and variance

$$\boldsymbol{\Gamma}_n := (\mathbf{G}_n^*)^{-1}\,\mathrm{Var}\,[\mathbf{g}_n\,(\boldsymbol{\psi}_n^*\,; \mathbf{S}_n)]\left[(\mathbf{G}_n^*)^{-1}\right]^{\intercal} \quad ,$$

where $\mathbf{G}_n^*$ is redefined as $\mathrm{E}\,[\mathbf{G}_n\,(\boldsymbol{\psi}_n^*; \mathbf{S}_n)]$ (theorem 6.4 of White, 1996). It can be seen that, once the assumption of correct model specification is dropped, the mean $\boldsymbol{\psi}_n^*$ of the approximative distribution is allowed to vary with n and will not in general reduce the sum of Kullback-Leibler divergences to zero (as also mentioned in section 4 of Freedman, 2006).

The results that have just been stated can be used to build so-called *quasi-Wald* confidence intervals. Let $(\boldsymbol{\psi})_i$ denote the i^{th} element of $\boldsymbol{\psi}$, $\left(\boldsymbol{\Gamma}_n\right)_{ii}$ denote the (i, i) entry of $\boldsymbol{\Gamma}_n$, and $z_{\alpha/2}$ denote the $1 - \alpha/2$ quantile of the standard normal distribution. Then, for large n,

$$\mathrm{P}\left[\left|\frac{\left(\hat{\boldsymbol{\psi}}_n\right)_i - \left(\boldsymbol{\psi}_n^*\right)_i}{\left(\boldsymbol{\Gamma}_n\right)_{ii}^{\frac{1}{2}}}\right| \leq z_{\alpha/2}\right] \approx 1 - \alpha \quad ,$$

and hence

$$\mathrm{P}\left\{\left[\left(\hat{\boldsymbol{\psi}}_n\right)_i - z_{\alpha/2}\left(\boldsymbol{\Gamma}_n\right)_{ii}^{\frac{1}{2}},\ \left(\hat{\boldsymbol{\psi}}_n\right)_i + z_{\alpha/2}\left(\boldsymbol{\Gamma}_n\right)_{ii}^{\frac{1}{2}}\right] \ni \left(\boldsymbol{\psi}_n^*\right)_i\right\} \approx 1 - \alpha \quad . \quad (4.1)$$

Table 4.1 Results for the data of Wang and Hey (2010): point estimates and confidence intervals under the model IIM_3.

Parameter	**Estimate**	**95% Wald CI's**	
		Fisher	**Godambe**
θ_a	3.273	(3.101, 3.445)	(3.076, 3.470)
θ	3.357	(3.139, 3.575)	(3.091, 3.624)
θ_b	1.929	(0.079, 3.779)	(-0.375, 4.232)
θ_{c_1}	6.623	(6.407, 6.839)	(6.382, 6.864)
θ_{c_2}	2.647	(2.304, 2.990)	(2.287, 3.006)
T_1	6.930	(6.540, 7.320)	(6.249, 7.611)
V	9.778	(9.457, 10.099)	(9.347, 10.208)
M_2	0.223	(0.190, 0.256)	(0.181, 0.276)

In this formula, $\boldsymbol{\Gamma}_n$ is unknown and must be replaced by an estimate: the matrix $\mathbf{G}_n^* := \mathrm{E}\left[\mathbf{G}_n\left(\boldsymbol{\psi}_n^*; \mathbf{S}_n\right)\right]$ is replaced by $\mathbf{G}_n\left(\hat{\boldsymbol{\psi}}_n; \mathbf{s}_n\right)$, which can be obtained by numerical differentiation of $\mathbf{g}_n\left(\cdot\,; \mathbf{s}_n\right)$ at $\hat{\boldsymbol{\psi}}_n$; and $\mathrm{Var}\left[\mathbf{g}_n\left(\boldsymbol{\psi}_n^*; \mathbf{S}_n\right)\right]$ is replaced by

$$\sum_{i=1}^{n} \mathbf{g}^{(i)}\left(\hat{\boldsymbol{\psi}}_n; s_i\right)\left[\mathbf{g}^{(i)}\left(\hat{\boldsymbol{\psi}}_n; s_i\right)\right]^{\intercal} \quad ,$$

where $\mathbf{g}^{(i)}\left(\hat{\boldsymbol{\psi}}_n; s_i\right)$ is the observed score vector for the i^{th} observation, evaluated at the maximum-likelihood estimate (section 3.4 of Jesus and Chandler, 2011). The quasi-Wald confidence intervals shown in the rightmost column of Table 4.1 refer to the best-fitting IIM model for the data of Wang and Hey (2010), and were computed using formula 4.1 and the estimate of $\boldsymbol{\Gamma}_n$ just mentioned: they appear under the heading 'Godambe' since the inverse of $\boldsymbol{\Gamma}_n$ is sometimes called the Godambe information. As can be seen from this table, correcting for misspecification does not widen substantially the confidence intervals that were calculated in section 2.6.3 (also shown here for ease of reference, under the heading 'Fisher'). The only exception is the parameter θ_b, most likely because its estimation relies on the much smaller 'Hutter subset' (see section 2.6.1).

4.1.2 Likelihood ratio tests and profile-likelihood confidence intervals

Recall that the vector $\boldsymbol{\psi}_n^*$ is defined as the value of the parameter that minimises the sum of Kullback-Leibler divergences between the true distributions of the random variables $S_1, S_2, ..., S_n$ and their respective distributions according to a given speciation model. Suppose that we wish to test whether $\boldsymbol{\psi}_n^* \in \Omega_0$, where $\Omega_0 = \{\boldsymbol{\psi} : \boldsymbol{\Xi}\boldsymbol{\psi} = \boldsymbol{\xi}_0, \boldsymbol{\psi} \in \Omega\}$, $\boldsymbol{\Xi}$ is a $q \times p$ matrix of rank $q \leq p$ and

$\boldsymbol{\xi}_0$ is a column vector of size q. Suppose further that $\tilde{\boldsymbol{\psi}}_n$ is the vector in Ω_0 that minimises the negated log-likelihood, and that, as before, $\hat{\boldsymbol{\psi}}_n$ denotes the (unrestricted) maximum-likelihood estimator. Then, a fundamental large-sample theoretical result states that, under suitable regularity conditions, the quantity

$$2\left[Q_n\left(\tilde{\boldsymbol{\psi}}_n; \mathbf{S}_n\right) - Q_n\left(\hat{\boldsymbol{\psi}}_n; \mathbf{S}_n\right)\right]$$

has approximately the same distribution as $\mathbf{Z}^\intercal \mathbf{A}^{-1} \mathbf{Z}$, where $\mathbf{A} = \boldsymbol{\Xi}\left[\mathbf{G}_n^*\right]^{-1}\boldsymbol{\Xi}^\intercal$, and where $\mathbf{Z}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{\Xi}\boldsymbol{\Gamma}_n\boldsymbol{\Xi}^\intercal$ (theorem 8.10 (iii) of White, 1996).

If we now define $\boldsymbol{\Xi}$ to be a row vector with a 1 in position i and zeros everywhere else, then assuming a null hypothesis of the form $\boldsymbol{\psi}_n^* \in \{\boldsymbol{\psi} : \boldsymbol{\Xi}\boldsymbol{\psi} = \xi_0, \boldsymbol{\psi} \in \Omega\}$ is the same as claiming that the true value of $\left(\boldsymbol{\psi}_n^*\right)_i$ is ξ_0. Under such a hypothesis, it follows from the previous result that, approximately,

$$2\left[Q_n\left(\tilde{\boldsymbol{\psi}}_n; \mathbf{S}_n\right) - Q_n\left(\hat{\boldsymbol{\psi}}_n; \mathbf{S}_n\right)\right] \sim \frac{\left(\boldsymbol{\Gamma}_n\right)_{ii}}{\left[\left(\mathbf{G}_n^*\right)^{-1}\right]_{ii}} \chi_1^2 \quad ,$$

(Jesus and Chandler, 2011, p. 877). A $100\,(1-\alpha)\,\%$ profile likelihood confidence interval for $\left(\boldsymbol{\psi}_n^*\right)_i$ is therefore the set of points $\{\xi_0\}$ such that

$$2\left[Q_n\left(\tilde{\boldsymbol{\psi}}_n; \mathbf{S}_n\right) - Q_n\left(\hat{\boldsymbol{\psi}}_n; \mathbf{S}_n\right)\right] \leq \frac{\left(\boldsymbol{\Gamma}_n\right)_{ii}}{\left[\left(\mathbf{G}_n^*\right)^{-1}\right]_{ii}} \chi_1^2\,(1-\alpha) \quad , \qquad (4.2)$$

where $\chi_1^2\,(1-\alpha)$ is the $1-\alpha$ quantile of the χ_1^2 distribution (expression 3.7 of Jesus and Chandler, 2011).

The 95% profile-likelihood confidence intervals for the model IIM_3 and the data of Wang and Hey (2010), based on the estimated inverse of the Godambe information, are listed in Table 4.2. For ease of reference, we also present point estimates and profile-likelihood confidence intervals based on the inverse of the observed Fisher information, as calculated in section 2.6.3. Once more, with the exception of the confidence interval for θ_b, adjusting for model misspecification does not result in substantial changes.

The q-q plot in Figure 4.1 illustrates the impact of model misspecification on the distribution of the likelihood ratio test of the model IIM_1 (null model) versus the model IIM_2, for the data of Wang and Hey (2010). The y-coordinates of the plot represent the percentiles of the χ^2 distribution with two degrees of freedom: the number of linear constraints imposed on $\boldsymbol{\psi}$ by the null hypothesis. This is the approximate distribution of the likelihood ratio statistic under

Table 4.2 Results for the data of Wang and Hey (2010): point estimates and 95% profile likelihood confidence intervals under the model IIM_3.

Parameter	**Estimate**	**95% profile likelihood CI's**	
		Fisher	**Godambe**
θ_a	3.273	(3.100, 3.444)	(3.074, 3.468)
θ	3.357	(3.097, 3.578)	(3.096, 3.627)
θ_b	1.929	(0.672, 5.010)	(0.493, 6.508)
θ_{c_1}	6.623	(6.415, 6.843)	(6.389, 6.871)
θ_{c_2}	2.647	(2.331, 3.021)	(2.317, 3.040)
T_1	6.930	(6.542, 7.319)	(6.252, 7.611)
V	9.778	(9.456, 10.098)	(9.347, 10.207)
M_2	0.223	(0.186, 0.259)	(0.181, 0.276)

the assumption that the data were generated from the IIM_1 model. The x-coordinates correspond to the percentiles of the distribution of $1.39\,X + 0.08$, where $X \sim \chi^2_{1.81}$, which approximates the large-sample distribution of the likelihood ratio statistic for the same model comparison, under the weaker assumption that

$$\arg\min_{\boldsymbol{\psi}\in\Omega} \sum_{i=1}^{n} D_{\mathrm{KL}}\left[\,\mathrm{q}_i\left(\cdot\right) \parallel \mathrm{p}\left(\cdot\,;\boldsymbol{\psi}, u_i, r_i\right)\right] \in \Omega_0 \quad ,$$

i.e. that, within the set of IIM_1 and IIM_2 models, the one closest to the true unknown model of the data is an IIM_1 model. This approximation is based on formula 3.6 of Jesus and Chandler (2011). The plot clearly suggests that using the naive χ^2_2 distribution is likely to lead to the underestimation of p-values. For the Wang and Hey (2010) data, however, there is overwhelming evidence to reject the IIM_1 model in favour of the IIM_2 even if we correct for misspecification: the p-value of the likelihood ratio statistic is extremely small using the naive χ^2_2 distribution (1.187 E-74), and increases to a larger, but still extremely small, p-value under the corrected χ^2 distribution (2.589 E-54).

4.2 Parameters on the boundary

4.2.1 Setting

Theorems 6.4 and 8.10 of White (1996), which were invoked in the previous section, rely on the standard assumption that the true parameter $\boldsymbol{\psi}^*$ (or, more generally, the 'target' vector $\boldsymbol{\psi}^*_n$) is an interior point of the parameter space Ω. However, when performing pairwise comparison of models, by means of

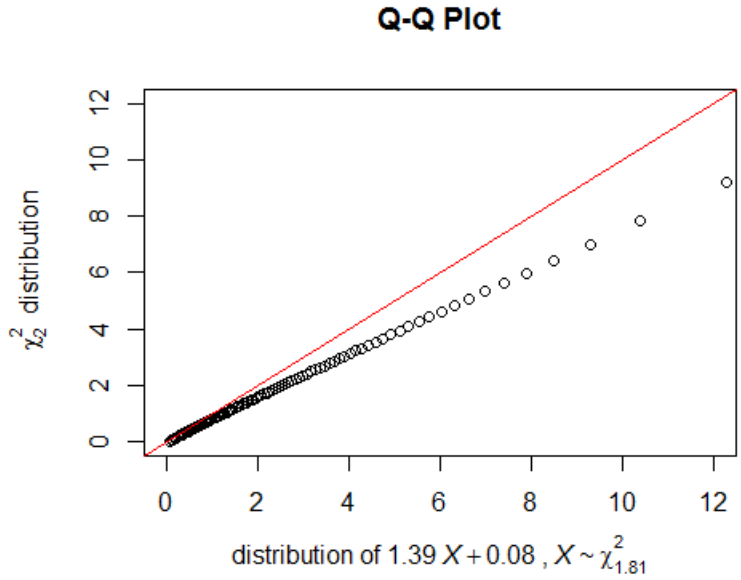

Fig. 4.1 A q-q plot of the percentiles of the χ^2_2 distribution against the percentiles of the distribution of $1.39\,X + 0.08$, where $X \sim \chi^2_{1.81}$. The χ^2_2 distribution is the large-sample distribution of the likelihood ratio statistic of IIM_1 (H_0) versus IIM_2 (H_1) when the true model is IIM_1. The distribution of $1.39\,X + 0.08$ approximates the large-sample distribution of the likelihood ratio statistic for the same model comparison, under the weaker assumption that the IIM_1 model is closer to the true unknown model of the Wang and Hey (2010) data than the IIM_2 model, in the sense of the Kullback-Leibler divergence.

the likelihood ratio statistic, this assumption must sometimes be dropped. For some model comparisons, the null hypothesis actually states that $\boldsymbol{\psi}^*$ is on the boundary of Ω. This is the case if, for example, we wish to compare the model 'ISO' as the null model against the model 'IM_1' as the alternative model (see Figure 2.9). Here the vector of parameters is

$$\boldsymbol{\psi} = \begin{bmatrix} M_1 & M_2 & V & \theta_a & \theta & \theta_b \end{bmatrix}^\intercal \quad ,$$

and the parameter space is $\Omega = [0, \infty)^2 \times (0, \infty)^4$. According to the null hypothesis, $\boldsymbol{\psi}^* \in \Omega_0$, with $\Omega_0 = \{0\}^2 \times (0, \infty)^4$, and, according to the alternative hypothesis, $\boldsymbol{\psi}^* \in \Omega_1$, with $\Omega_1 = \left\{[0, \infty)^2 \setminus \{0\}^2\right\} \times (0, \infty)^4$. The parameters V, θ_a, θ and θ_b are thus nuisance parameters.

The present section has two main goals. One is to state the main large-sample results regarding the distribution of the likelihood ratio statistic in settings as the one just described. To the best of our knowledge, analytic expressions are only available for the simplest cases, involving at most two parameters of interest on the boundary (see, for example, Self and Liang, 1987;

Liang and Self, 1996; Chen and Liang, 2010). The limiting distribution can always be represented as a mixture of χ^2 distributions, but, for more than two parameters of interest on the boundary, the mixing coefficients of this mixture need to be estimated by simulation. The other goal, very closely related to the first, is to demonstrate the validity of a simulation method we developed to estimate this limiting distribution. We start from the large-sample results described in Self and Liang (1987), which extend the work of Chernoff (1954), and hence assume the regularity conditions stated in their paper (in the last paragraph of the introduction). Of these conditions, the only one we shall use explicitly concerns the positive-definiteness of the Fisher information matrix. For our models and data, the setting and the assumptions described in Self and Liang (1987) are not the most realistic, as they do not allow for model misspecification and non-identically distributed observations. An extension to this more general setting should be possible but would involve redoing the derivations in Self and Liang (1987) and Chernoff (1954) under more general conditions. This had to be left for future work.

Suppose that, for $q, r \geq 1$, $\boldsymbol{\psi}$ is a vector of size $p = q + r$ taking values in $\Omega = [0, \infty)^q \times (0, \infty)^r$. When comparing two speciation models, the problem of parameters on the boundary arises when we have a null hypothesis of the kind $\boldsymbol{\psi}^* \in \Omega_0 = \{0\}^q \times (0, \infty)^r$. Recall that

$$\hat{\boldsymbol{\psi}}_n := \arg\min_{\boldsymbol{\psi} \in \Omega} Q_n \left(\boldsymbol{\psi}; \mathbf{S}_n \right)$$

and

$$\tilde{\boldsymbol{\psi}}_n := \arg\min_{\boldsymbol{\psi} \in \Omega_0} Q_n \left(\boldsymbol{\psi}; \mathbf{S}_n \right) \quad .$$

From Self and Liang (1987), the likelihood ratio statistic

$$2 \left[Q_n \left(\tilde{\boldsymbol{\psi}}_n; \mathbf{S}_n \right) - Q_n \left(\hat{\boldsymbol{\psi}}_n; \mathbf{S}_n \right) \right]$$

is asymptotically equivalent to

$$\min_{\boldsymbol{\psi} \in \Omega_0} [\, \mathbf{W} - \boldsymbol{\psi}]^\intercal \, \mathbf{M}_0 \, [\, \mathbf{W} - \boldsymbol{\psi}] - \min_{\boldsymbol{\psi} \in \Omega} [\, \mathbf{W} - \boldsymbol{\psi}]^\intercal \, \mathbf{M}_0 \, [\, \mathbf{W} - \boldsymbol{\psi}] \quad , \qquad (4.3)$$

where $\mathbf{M}_0$ is the Fisher information matrix, $\mathbf{W}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{M}_0^{-1}$, and Ω and Ω_0 are redefined respectively as $[0, \infty)^q \times \mathbb{R}^r$ and $\{0\}^q \times \mathbb{R}^r$. If $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\intercal$ is the eigen-decomposition

of $\mathbf{M}_0$, expression (4.3) can be rewritten as

$$D\left(\mathbf{Z}\right) := \min_{\boldsymbol{\theta}\in\tilde{\Omega}_0} \left[\ \mathbf{Z}-\boldsymbol{\theta}\right]^{\intercal}\ \left[\ \mathbf{Z}-\boldsymbol{\theta}\right]\ -\min_{\boldsymbol{\theta}\in\tilde{\Omega}} \left[\ \mathbf{Z}-\boldsymbol{\theta}\right]^{\intercal}\ \left[\ \mathbf{Z}-\boldsymbol{\theta}\right]\ , \tag{4.4}$$

where $\mathbf{Z}$ is a standard normal random vector of size p, $\tilde{\Omega} = \left\{\boldsymbol{\Lambda}^{\frac{1}{2}}\,\mathbf{P}^{\intercal}\boldsymbol{\psi} : \boldsymbol{\psi}\in\Omega\right\}$ and $\tilde{\Omega}_0 = \left\{\boldsymbol{\Lambda}^{\frac{1}{2}}\,\mathbf{P}^{\intercal}\boldsymbol{\psi} : \boldsymbol{\psi}\in\Omega_0\right\}$ (Self and Liang, 1987).

4.2.2 An alternative description of the parameter space

More insightful representations of the limiting distribution of the likelihood ratio statistic have been given in the literature. To understand how they follow from the results in Self and Liang (1987) just given, it is helpful to use an alternative description of the parameter space and to clarify some of its properties.

Let $\mathbf{C}_q$ denote the matrix composed of the q leftmost columns of $\boldsymbol{\Lambda}^{\frac{1}{2}}\,\mathbf{P}^{\intercal}$, and $\mathbf{C}_r$ denote the matrix with its r rightmost columns. The set $\tilde{\Omega}_0$ is then simply the subspace of $\mathbb{R}^p$ spanned by the columns of $\mathbf{C}_r$, which we shall denote as $S(\mathbf{C}_r)$. The set $\tilde{\Omega}$ can be rewritten as

$$\begin{aligned}\tilde{\Omega} &= \left\{[\mathbf{C}_q\,|\,\mathbf{C}_r]\ \boldsymbol{\psi} : \boldsymbol{\psi}\in[0,\infty)^q\times\mathbb{R}^r\right\}\\ &= \left\{[\mathbf{C}_q\,|\,\mathbf{C}_r\,|\,-\mathbf{C}_r]\ \boldsymbol{\rho} : \boldsymbol{\rho}\in[0,\infty)^{q+2r}\right\}\quad ,\end{aligned}$$

where $[\mathbf{C}_q\,|\,\mathbf{C}_r]$ denotes a matrix built by setting matrices $\mathbf{C}_q$ and $\mathbf{C}_r$ side by side. It becomes clear that $\tilde{\Omega}$ consists of the non-negative linear combinations of a finite set of column vectors, and hence it makes up a polyhedral cone (Borovik and Borovik, 2010, p. 25, for the definition of polyhedral cone). We shall denote a polyhedral cone generated by the non-negative linear combinations of the columns of a matrix $\mathbf{A}$ by $C(\mathbf{A})$, so, for example, $\tilde{\Omega} = C([\mathbf{C}_q\,|\,\mathbf{C}_r\,|\,-\mathbf{C}_r])$. As $\mathbf{M}_0 = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\intercal}$ is symmetric, $\mathbf{P}^{\intercal}$ has full rank; since $\mathbf{M}_0$ is also positive definite, the diagonal of $\boldsymbol{\Lambda}$ contains only strictly positive eigenvalues (Strang, 2009, pp. 334, 342). To obtain the matrix $[\mathbf{C}_q\,|\,\mathbf{C}_r] = \boldsymbol{\Lambda}^{\frac{1}{2}}\,\mathbf{P}^{\intercal}$, each row vector of $\mathbf{P}^{\intercal}$ is multiplied by one of the strictly positive elements in the diagonal of $\boldsymbol{\Lambda}^{\frac{1}{2}}$. Hence $[\mathbf{C}_q\,|\,\mathbf{C}_r]$ must have full rank as well.

Suppose also that from each column vector of $\mathbf{C}_q$ we subtract its projection onto $S(\mathbf{C}_r)$. We obtain the matrix $\mathbf{X} = \mathbf{C}_q - \mathbf{P}_q$, where $\mathbf{P}_q$ is a $p\times q$ matrix whose column vectors are the projections onto $S(\mathbf{C}_r)$. Each column vector of $\mathbf{X}$ is the error vector of a projection onto $S(\mathbf{C}_r)$, and is by definition orthogonal to this subspace (Strang, 2009, pp. 209-210). Let $\mathbf{x}$ denote any one of these

column vectors, and let the matrices $\mathbf{X}$, $\mathbf{C}_q$ and $\mathbf{P}_q$ be partitioned to give

$$\mathbf{X} = [\mathbf{X}_{q-1} \vdash \mathbf{x}] = [\mathbf{C}_{q-1} \vdash \mathbf{c}] - [\mathbf{P}_{q-1} \vdash \mathbf{p}] \quad .$$

The matrix $\mathbf{X}$ has rank lower than q if and only if we can pick a column $\mathbf{x}$ such that, for some $\mathbf{b} \in \mathbb{R}^{q-1}$,

$$\begin{aligned} & \mathbf{x} &&= \mathbf{X}_{q-1}\mathbf{b} \\ \Leftrightarrow \quad & \mathbf{c} - \mathbf{p} &&= [\mathbf{C}_{q-1} - \mathbf{P}_{q-1}]\,\mathbf{b} \\ \Leftrightarrow \quad & \mathbf{c} &&= [\mathbf{C}_{q-1} \vdash \mathbf{P}_{q-1} \vdash \mathbf{p}] \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \\ 1 \end{bmatrix} \\ \Rightarrow \quad & \mathbf{c} &&\in S\,[\mathbf{C}_{q-1} \vdash \mathbf{P}_q] \quad . \end{aligned} \tag{4.5}$$

Since the column vectors of $\mathbf{P}_q$ are projections onto $S(\mathbf{C}_r)$, the last line of (4.5) implies that $\mathbf{c} \in S\,[\mathbf{C}_{q-1} \vdash \mathbf{C}_r]$, which could only happen if $[\mathbf{C}_{q-1} \vdash \mathbf{c} \vdash \mathbf{C}_r] = [\mathbf{C}_q \vdash \mathbf{C}_r]$ had less than full rank. Hence $\mathbf{X}$ has rank q, i.e. full column rank.

The set

$$\tilde{\Omega} \;=\; \{[\mathbf{C}_q \vdash \mathbf{C}_r]\ \boldsymbol{\psi} : \boldsymbol{\psi} \in [0,\infty)^q \times \mathbb{R}^r\}$$

can also be written as

$$\{[\mathbf{X} \vdash \mathbf{C}_r]\ \boldsymbol{\psi} : \boldsymbol{\psi} \in [0,\infty)^q \times \mathbb{R}^r\} \quad .$$

To see why this is true, consider a vector $\boldsymbol{\theta} \in \mathbb{R}^p$. If $\boldsymbol{\theta} \in \tilde{\Omega}$, then it can be expressed, for some $\mathbf{a} \in [0,\infty)^q$ and for some $\mathbf{b}$, $\mathbf{c}\ \in \mathbb{R}^r$, as

$$\begin{aligned} \boldsymbol{\theta} &= [\mathbf{C}_q]\,\mathbf{a} + [\mathbf{C}_r]\,\mathbf{b} \\ &= [\mathbf{C}_q - \mathbf{P}_q]\,\mathbf{a} + [\mathbf{C}_r]\,\mathbf{b} + [\mathbf{P}_q]\,\mathbf{a} \\ &= \mathbf{X}\mathbf{a} + [\mathbf{C}_r]\,\mathbf{c} \quad , \end{aligned}$$

where $[\mathbf{C}_r]\,\mathbf{c} = [\mathbf{C}_r]\,\mathbf{b} + [\mathbf{P}_q]\,\mathbf{a}$: there is always a vector $\mathbf{c} \in \mathbb{R}^r$ that satisfies this equality, since $[\mathbf{C}_r]\,\mathbf{b} + [\mathbf{P}_q]\,\mathbf{a}$ belongs to $S(\mathbf{C}_r)$. Conversely, any vector $\boldsymbol{\theta}$ belonging to

$$\{[\mathbf{X} \vdash \mathbf{C}_r]\ \boldsymbol{\psi} : \boldsymbol{\psi} \in [0,\infty)^q \times \mathbb{R}^r\} \quad ,$$

can be expressed, for some $\mathbf{a} \in [0,\infty)^q$ and $\mathbf{b}$, $\mathbf{c}\ \in \mathbb{R}^r$, as

$$\begin{aligned} \boldsymbol{\theta} &= \mathbf{X}\mathbf{a} + [\mathbf{C}_r]\,\mathbf{b} \\ &= [\mathbf{C}_q - \mathbf{P}_q]\,\mathbf{a} + [\mathbf{C}_r]\,\mathbf{b} \\ &= [\mathbf{C}_q]\,\mathbf{a} + [\mathbf{C}_r]\,\mathbf{c} \quad , \end{aligned}$$

since again there is always a $\mathbf{c} \in \mathbb{R}^r$ that satisfies $[\mathbf{C}_r]\,\mathbf{c} = [\mathbf{C}_r]\,\mathbf{b} - [\mathbf{P}_q]\,\mathbf{a}$.

It should also be noted that the subspaces $S(\mathbf{X})$ and $S(\mathbf{C}_r)$ are mutually orthogonal, and that their sum, i.e., the vector space $S\left[\mathbf{X} \vdots \mathbf{C}_r\right]$, is equal to $\mathbb{R}^p$. In other words, the subspace spanned by the columns of $\mathbf{C}_r$ is the orthogonal complement of the subspace spanned by the columns of $\mathbf{X}$, and vice-versa (Gentle, 2007, p. 23). We shall therefore refer to $\mathbf{C}_r$ as $\mathbf{X}^\perp$ and write

$$\begin{aligned}\tilde{\Omega} &= \left\{\left[\mathbf{X} \vdots \mathbf{X}^\perp\right]\,\boldsymbol{\psi} : \boldsymbol{\psi} \in [0,\infty)^q \times \mathbb{R}^r\right\} \\ &= \left\{\left[\mathbf{X} \vdots \mathbf{X}^\perp \vdots -\mathbf{X}^\perp\right]\,\boldsymbol{\psi} : \boldsymbol{\psi} \in [0,\infty)^{q+2r}\right\} \quad .\end{aligned}$$

4.2.3 The asymptotic distribution of the likelihood ratio statistic

Suppose that $\mathbf{Y}$ is a standard normal random vector with p components, partitioned as $\left[\mathbf{Y}_q^\intercal \vdots \mathbf{Y}_r^\intercal\right]^\intercal$, where $\mathbf{Y}_q$ has length q and $\mathbf{Y}_r$ has length r. Suppose further that $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}^\perp$ are orthonormal bases for the subspaces spanned by $\mathbf{X}$ and $\mathbf{X}^\perp$ respectively. Since the $p \times p$ matrix $\left[\bar{\mathbf{X}} \vdots \bar{\mathbf{X}}^\perp\right]$ is orthonormal, i.e., since it belongs to the group of rotations of $\mathbb{R}^p$, and $\mathbf{Y}$ is rotation invariant, the random vector $\mathbf{Z} = \left[\bar{\mathbf{X}} \vdots \bar{\mathbf{X}}^\perp\right]\mathbf{Y}$ is also a standard normal random vector with p components (Bryc, 1995, proof of theorem 4.1.2). Defining $\mathbf{Z} = \mathbf{Z}_q + \mathbf{Z}_r$, where $\mathbf{Z}_q = \bar{\mathbf{X}}\,\mathbf{Y}_q$ and $\mathbf{Z}_r = \bar{\mathbf{X}}^\perp\,\mathbf{Y}_r$, the first term in expression (4.4) can be rewritten as

$$\begin{aligned}\min_{\boldsymbol{\theta} \in S(\mathbf{X}^\perp)} \|\,\mathbf{Z} - \boldsymbol{\theta}\|^2 &= \min_{\boldsymbol{\theta} \in S(\mathbf{X}^\perp)} \|\,\mathbf{Z}_q + \mathbf{Z}_r - \boldsymbol{\theta}\|^2 \\ &= \|\,\mathbf{Z}_q\|^2 + \min_{\boldsymbol{\theta} \in S(\mathbf{X}^\perp)} \|\,\mathbf{Z}_r - \boldsymbol{\theta}\|^2 \\ &= \|\,\mathbf{Z}_q\|^2 \\ &= \|\,\mathbf{Y}_q\|^2 \quad ,\end{aligned}$$

where the second equality follows from the fact that $\mathbf{Z}_q$ and $\mathbf{Z}_r - \boldsymbol{\theta}$ are orthogonal. To simplify the second term in expression (4.4), $\tilde{\Omega}$ is rewritten as

$$\begin{aligned}\tilde{\Omega} &= \left\{\left[\mathbf{X} \vdots \mathbf{X}^\perp\right]\,\boldsymbol{\psi} : \boldsymbol{\psi} \in [0,\infty)^q \times \mathbb{R}^r\right\} \\ &= \left\{\mathbf{X}\,\boldsymbol{\psi}_q + \mathbf{X}^\perp\,\boldsymbol{\psi}_r : \boldsymbol{\psi}_q \in [0,\infty)^q\,,\,\boldsymbol{\psi}_r \in \mathbb{R}^r\right\} \\ &= \left\{\boldsymbol{\theta}_q + \boldsymbol{\theta}_r : \boldsymbol{\theta}_q \in C\left(\mathbf{X}\right)\,,\,\boldsymbol{\theta}_r \in S(\mathbf{X}^\perp)\right\} \quad .\end{aligned}$$

We then get

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\tilde{\Omega}}\| \mathbf{Z}-\boldsymbol{\theta}\|^2 &= \min_{\boldsymbol{\theta}_q\in C(\mathbf{X}),\,\boldsymbol{\theta}_r\in S(\mathbf{X}^{\perp})}\| \mathbf{Z}_q-\boldsymbol{\theta}_q+\mathbf{Z}_r-\boldsymbol{\theta}_r\|^2 \\ &= \min_{\boldsymbol{\theta}_q\in C(\mathbf{X})}\| \mathbf{Z}_q-\boldsymbol{\theta}_q\|^2+\min_{\boldsymbol{\theta}_r\in S(\mathbf{X}^{\perp})}\| \mathbf{Z}_r-\boldsymbol{\theta}_r\|^2 \\ &= \min_{\boldsymbol{\theta}_q\in C(\mathbf{X})}\| \mathbf{Z}_q-\boldsymbol{\theta}_q\|^2 \quad ,\end{aligned}$$

where the second equality follows from the fact that $\mathbf{Z}_q-\boldsymbol{\theta}_q$ and $\mathbf{Z}_r-\boldsymbol{\theta}_r$ are orthogonal. Expression (4.4) then becomes

$$D\left(\mathbf{Z}\right)=D\left(\mathbf{Y}_q\right)=\| \mathbf{Y}_q\|^2-\min_{\boldsymbol{\theta}_q\in C(\mathbf{X})}\| \mathbf{Z}_q-\boldsymbol{\theta}_q\|^2 \quad . \tag{4.6}$$

If $\mathbf{X}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\intercal}$ is the reduced singular value decomposition of $\mathbf{X}$ (Strang, 2009, p. 363-364), and $\bar{\mathbf{X}}$ is set equal to $\mathbf{U}$, expression (4.6) can be rewritten as

$$D\left(\mathbf{Z}\right)=D\left(\mathbf{Y}_q\right)=\| \mathbf{Y}_q\|^2-\min_{\boldsymbol{\eta}\in C(\mathbf{K})}\|\mathbf{Y}_q-\boldsymbol{\eta}\|^2 \quad , \tag{4.7}$$

where $\mathbf{K}:=\boldsymbol{\Sigma}\mathbf{V}^{\intercal}$. The vector $\boldsymbol{\eta}\in C(\mathbf{K})$ that minimises $\|\mathbf{Y}_q-\boldsymbol{\eta}\|^2$ is the unique minimum-distance projection of $\mathbf{Y}_q$ onto $C\left(\mathbf{K}\right)$ (Dattorro, 2005, p. 576), and we shall represent it by $p(\mathbf{Y}_q)$. Since the collection of the relative interiors of the faces of $C\left(\mathbf{K}\right)$ forms a partition of $C\left(\mathbf{K}\right)$ (Luc, 2016, corollary 2.3.7), $p(\mathbf{Y}_q)$ must lie on the relative interior of one (and only one) of these faces. If $\mathcal{F}_k$ denotes the union of the relative interiors of all k-dimensional faces of $C(\mathbf{K})$, the asymptotic distribution of the likelihood ratio statistic can thus be written as

$$\mathrm{P}\left[D\left(\mathbf{Y}_q\right)\leq d\right]=\sum_{k=0}^{q}\mathrm{P}\left[D\left(\mathbf{Y}_q\right)\leq d\,|\,p\left(\mathbf{Y}_q\right)\in\mathcal{F}_k\right]\mathrm{P}\left[p\left(\mathbf{Y}_q\right)\in\mathcal{F}_k\right]$$

which, from theorem 3.4.2 and proposition 3.6.1.1 in Silvapulle and Sen (2011), is equivalent to

$$\mathrm{P}\left[D\left(\mathbf{Y}_q\right)\leq d\right]=\sum_{k=0}^{q}\mathrm{P}\left(\chi^2_k\leq d\right)\mathrm{P}\left[p\left(\mathbf{Y}_q\right)\in\mathcal{F}_k\right] \quad . \tag{4.8}$$

Also from Silvapulle and Sen (2011, p. 78), the probability in (4.8) can be estimated by simulating $D\left(\mathbf{Y}_q\right)$, typically thousands of times, and then finding the proportion of times that it turns out to be smaller than d. Each simulated value of $D\left(\mathbf{Y}_q\right)$ is obtained by first generating a random vector $\mathbf{Y}_q$, then finding its projection $p(\mathbf{Y}_q)$ onto $C(\mathbf{K})$, and finally evaluating expression

(4.7). The mixing probability of each χ^2_k distribution in (4.8) can also be estimated by the proportion of times that $p(\mathbf{Y}_q)$ happens to lie on the relative interior of a k-dimensional face of $C(\mathbf{K})$ (Silvapulle and Sen, 2011, p. 79). Our program for estimating the mixing probabilities in (4.8), available at https://github.com/ruibarrigana/boundary, does not compute the dimension of the face of $C(\mathbf{K})$ containing $\mathbf{Y}_q$, nor finds the projection $p(\mathbf{Y}_q)$. Instead, it partitions $\mathbb{R}^q$ into subsets $\mathcal{R}_0, \mathcal{R}_1, ..., \mathcal{R}_q$, where $\mathcal{R}_k$, $k \in \{0, 1, 2, ..., q\}$, contains the vectors whose projections lie on the relative interior of any of the k-dimensional faces of $C(\mathbf{K})$. The mixing probability associated with χ^2_k is thus estimated by the proportion of times that $\mathbf{Y}_q \in \mathcal{R}_k$ rather than the proportion of times that $p(\mathbf{Y}_q) \in \mathcal{F}_k$. To define the subsets $\mathcal{R}_k$, we use the following argument.

Let the q column vectors of $\mathbf{K}$ be labelled $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$,..., $\boldsymbol{\rho}_q$. Since $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal$ has full column rank, $\mathbf{V}$ is a $q \times q$ orthonormal matrix, and $\boldsymbol{\Sigma}$ is a $q \times q$ diagonal matrix with positive values on the main diagonal (Strang, 2009, p. 363). Hence $\mathbf{K} = \boldsymbol{\Sigma}\mathbf{V}^\intercal$ is non-singular and $C(\mathbf{K})$ is a simplicial cone in $\mathbb{R}^q$ (Borovik and Borovik, 2010, p. 30). Denote by $\mathbf{K}^*$ a $q \times q$ matrix whose column vectors $\boldsymbol{\rho}^*_1$, $\boldsymbol{\rho}^*_2$,..., $\boldsymbol{\rho}^*_q$ belong to $\mathbb{R}^q$ and satisfy the conditions

$$(\boldsymbol{\rho}_i)^\intercal \boldsymbol{\rho}^*_j = \begin{cases} -1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{4.9}$$

Then $C(\mathbf{K}^*)$ is the polar cone of the simplicial cone $C(\mathbf{K})$ (Borovik and Borovik, 2010, pp. 30-31). To describe the faces of $C(\mathbf{K})$ and $C(\mathbf{K}^*)$, we let $\boldsymbol{\rho}_0 = \boldsymbol{\rho}^*_0$ denote the origin of $\mathbb{R}^p$, and represent the power set of $I = \{1, 2, ..., q\}$ by $\mathcal{P}(I)$. For any set $J \in \mathcal{P}(I)$, let $J_0 = J \bigcup \{0\}$ and denote by $\mathbf{K}_{J_0}$ the matrix composed of the column vectors $\boldsymbol{\rho}_i$, $i \in J_0$. Similarly, denoting $I_0 = I \bigcup \{0\}$, the matrix $\mathbf{K}^*_{I_0 \setminus J}$ is composed of the column vectors $\boldsymbol{\rho}^*_i$, $i \in I_0 \setminus J$. Suppose now that F is a subset of $\mathbb{R}^p$. We know that F is a face of $C(\mathbf{K})$ if and only if it can be defined, for some $J \in \mathcal{P}(I)$, as $C(\mathbf{K}_{J_0})$; similarly, F is a face of $C(\mathbf{K}^*)$ if and only if it can be defined, for some $J \in \mathcal{P}(I)$, as $C(\mathbf{K}^*_{I_0 \setminus J})$ (Borovik and Borovik, 2010, section 4.5). Furthermore, the face $C(\mathbf{K}_{J_0})$ has dimension equal to the cardinality of J, denoted $|J|$, just as the face $C(\mathbf{K}^*_{I_0 \setminus J})$ has dimension $q - |J|$. The relative interior of $C(\mathbf{K}_{J_0})$ is denoted 'relint $C(\mathbf{K}_{J_0})$' and consists of all vectors of the form $\sum_{i \in J_0} a_i \boldsymbol{\rho}_i$, where the scalars a_i are strictly positive (Dattorro, 2005, definition 2.13.6.0.1).

A point $\mathbf{b}$ is the unique minimum-distance projection of $\mathbf{y}_q$ onto $C(\mathbf{K})$, denoted $p(\mathbf{y}_q)$, if and only if

$$\mathbf{b} \in C(\mathbf{K}) \quad , \tag{4.10}$$

$$[\mathbf{y}_q - \mathbf{b}] \in C(\mathbf{K}^*) \tag{4.11}$$

and

$$[\mathbf{y}_q - \mathbf{b}]^\mathsf{T} \mathbf{b} = 0 \tag{4.12}$$

(Dattorro, 2005, p. 580). Since any vector in the cone $C(\mathbf{K})$ must belong to the relative interior of one (and only one) of its faces, requiring that $\mathbf{b}$ satisfies condition (4.10) is equivalent to requiring that

$$\text{for some set } J \in \mathcal{P}(I)\,,\ \mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0}) \quad , \tag{4.10a}$$

or that

$$\text{for some set } J \in \mathcal{P}(I) \text{ and } a_i > 0\,,\ \mathbf{b} = \sum_{i \in J_0} a_i \boldsymbol{\rho}_i \quad . \tag{4.10b}$$

Thus if condition (4.10)/(4.10a)/(4.10b) holds, $\mathbf{b}$ satisfies condition (4.12) if and only if, for the same set J and the same scalars a_i that make (4.10a)/(4.10b) true,

$$\sum_{i \in J_0} a_i \, [\mathbf{y}_q - \mathbf{b}]^\mathsf{T} \boldsymbol{\rho}_i = 0 \quad .$$

Furthermore, if both condition (4.10)/(4.10a)/(4.10b) *and* condition (4.11) hold, then $\mathbf{b}$ satisfies condition (4.12) if and only if, for the same set J and the same scalars a_i that make (4.10a)/(4.10b) true,

$$[\mathbf{y}_q - \mathbf{b}]^\mathsf{T} \boldsymbol{\rho}_i = 0$$

for all $i \in J_0$: this must be the case because the dot product between a vector in $C(\mathbf{K}^*)$ (such as $[\mathbf{y}_q - \mathbf{b}]$) and a vector in $C(\mathbf{K})$ (such as $\boldsymbol{\rho}_i$) is always non-positive (Luc, 2016, p. 31). Summarising, we have that $\mathbf{b} = p(\mathbf{y}_q)$ if and only if, for some set $J \in \mathcal{P}(I)$,

$$\mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0}) \quad , \tag{4.10a}$$

$$[\mathbf{y}_q - \mathbf{b}] \in C(\mathbf{K}^*) \tag{4.11}$$

and

$$[\mathbf{y}_q - \mathbf{b}]^\intercal \boldsymbol{\rho}_i = 0 \text{ for all } i \in J_0 \quad . \tag{4.13}$$

Requiring that, for $i \in J_0$, the dot product $[\mathbf{y}_q - \mathbf{b}]^\intercal \boldsymbol{\rho}_i$ is equal to zero is the same as requiring that $[\mathbf{y}_q - \mathbf{b}]$ belongs to the orthogonal complement of $S(\mathbf{K}_{J_0})$ in $\mathbb{R}^q$, i.e., that it belongs to the subspace containing all the vectors in $\mathbb{R}^q$ that are orthogonal to the subspace spanned by $\{\boldsymbol{\rho}_i\}_{i \in J_0}$. Hence we have that $\mathbf{b} = p(\mathbf{y}_q)$ if and only if, for some set $J \in \mathcal{P}(I)$,

$$\mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0}) \tag{4.10a}$$

$$[\mathbf{y}_q - \mathbf{b}] \in C(\mathbf{K}^*) \tag{4.11}$$

and

$$[\mathbf{y}_q - \mathbf{b}] \in S(\mathbf{K}_{J_0})^\perp \quad , \tag{4.14}$$

where $S(\mathbf{K}_{J_0})^\perp$ denotes the orthogonal complement of $S(\mathbf{K}_{J_0})$ in $\mathbb{R}^q$. A final reformulation of these conditions is possible if we take into account that $C(\mathbf{K}^*_{I_0 \setminus J})$ is composed precisely of those vectors that belong both to $C(\mathbf{K}^*)$ and $S(\mathbf{K}_{J_0})^\perp$ (Borovik and Borovik, 2010, p. 32). Therefore $\mathbf{b} = p(\mathbf{y}_q)$ if and only if

$$\mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0}) \tag{4.10a}$$

and

$$[\mathbf{y}_q - \mathbf{b}] \in C(\mathbf{K}^*_{I_0 \setminus J}) \tag{4.15}$$

for some set $J \in \mathcal{P}(I)$.

Using this last definition of projection, it is not difficult to show that, for any $J \in \mathcal{P}(I)$, the set

$$\{\mathbf{a} : \mathbf{a} \in \mathbb{R}^q, p(\mathbf{a}) \in \operatorname{relint} C(\mathbf{K}_{J_0})\}$$

can also be defined as

$$\left\{\mathbf{a} : \mathbf{a} = \mathbf{b} + \mathbf{c}, \mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0}), \mathbf{c} \in C(\mathbf{K}^*_{I_0 \setminus J})\right\} .$$

Suppose indeed that $\mathbf{a} \in \mathbb{R}^q$ and $p(\mathbf{a}) \in \operatorname{relint} C(\mathbf{K}_{J_0})$. Then $[\mathbf{a} - p(\mathbf{a})] \in C(\mathbf{K}^*_{I_0 \setminus J})$ from condition (4.15), and $\mathbf{a} = \mathbf{b} + \mathbf{c}$, with $\mathbf{b} = p(\mathbf{a}) \in \operatorname{relint} C(\mathbf{K}_{J_0})$, and $\mathbf{c} = [\mathbf{a} - p(\mathbf{a})] \in C(\mathbf{K}^*_{I_0 \setminus J})$. The converse is also true: if there is a $\mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0})$ and a $\mathbf{c} \in C(\mathbf{K}^*_{I_0 \setminus J})$ such that $\mathbf{a} = \mathbf{b} + \mathbf{c}$, then, because all the columns of $\mathbf{K}$ and $\mathbf{K}^*$ are in $\mathbb{R}^q$, the vector $\mathbf{a}$ must also belong to this

subspace. In addition, because $\mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0})$ and $\mathbf{a} - \mathbf{b} = \mathbf{c}$ belongs to $C(\mathbf{K}^*_{I_0 \setminus J})$, it follows that $\mathbf{b} = p(\mathbf{a})$, and hence $p(\mathbf{a}) \in \operatorname{relint} C(\mathbf{K}_{J_0})$.

Recall that $\mathcal{F}_k$ denotes the union of all the relative interiors of k-dimensional faces of $C(\mathbf{K})$, and that $p : \mathbb{R}^q \to C(\mathbf{K})$ is the minimum-distance projection function. We can then write:

$$
\begin{aligned}
p^{-1}(\mathcal{F}_k) &= p^{-1}\left[\bigcup_{J \in \mathcal{P}(I), |J|=k} \operatorname{relint} C(\mathbf{K}_{J_0})\right] \\
&= \bigcup_{J \in \mathcal{P}(I), |J|=k} p^{-1}\left\{ \operatorname{relint} C(\mathbf{K}_{J_0}) \right\} \\
&= \bigcup_{J \in \mathcal{P}(I), |J|=k} \left\{ \mathbf{a} : \mathbf{a} \in \mathbb{R}^q, p(\mathbf{a}) \in \operatorname{relint} C(\mathbf{K}_{J_0}) \right\} \\
&= \bigcup_{J \in \mathcal{P}(I), |J|=k} \left\{ \mathbf{a} : \mathbf{a} = \mathbf{b} + \mathbf{c}, \mathbf{b} \in \operatorname{relint} C(\mathbf{K}_{J_0}), \mathbf{c} \in C(\mathbf{K}^*_{I_0 \setminus J}) \right\} \\
&=: \mathcal{R}_k \quad ,
\end{aligned}
$$

where the second equality follows from the properties of the inverse image (see, for example, proposition 1.6 of McDonald and Weiss, 1999). This finally allows us to rewrite the asymptotic distribution of the likelihood ratio statistic given above in (4.8) as

$$
\begin{aligned}
\mathrm{P}\left[D\left(\mathbf{Y}_q\right) \leq d\right] &= \sum_{k=0}^{q} \mathrm{P}\left(\chi^2_k \leq d\right) \mathrm{P}\left[\, p(\mathbf{Y}_q) \in \mathcal{F}_k \,\right] \\
&= \sum_{k=0}^{q} \mathrm{P}\left(\chi^2_k \leq d\right) \mathrm{P}\left[\, \mathbf{Y}_q \in \mathcal{R}_k \,\right] \quad .
\end{aligned}
\tag{4.16}
$$

4.2.4 Tests on simulated data

To verify that the asymptotic *cdf* of the likelihood ratio statistic can be expressed as in equation (4.16), we repeated the following procedure. First, we randomly generated a positive definite matrix to serve as $\mathbf{M}_0$, the variance-covariance matrix of the score statistic; second, we generated thousands of observations from $\mathbf{W}$ in expression (4.3); third, for some parameter space of the form $[0, \infty)^q \times \mathbb{R}^r$, we found the minimum of expression (4.3) for each observed value of $\mathbf{W}$; fourth, we generated thousands of observations from $\mathbf{Y}_q$, and estimated $\mathrm{P}\left\{\mathbf{Y}_q \in \mathcal{R}_k\right\}$ using the proportion of observations of $\mathbf{Y}_q$ in $\mathcal{R}_k$; fifth, we generated approximately one million observations from a mixture of χ^2 distributions with the estimated mixing probabilities; finally, we built

q-q plots of the sample percentiles of the simulated minima of expression (4.3) against the sample percentiles of the observations simulated from the mixture of χ^2 distributions. The R code used to perform these tasks can be found in https://github.com/ruibarrigana/boundary .

We found that regardless of the matrix $\mathbf{M}_0$ used to perform the procedure just described, the q-q plot would show that the estimated distributions of expression (4.3) and of $\sum_{k=0}^{q} \mathbf{1}\Big(\mathbf{Y}_q \in \mathcal{R}_k\Big)\chi^2_k$ are extremely close. This is illustrated through the following two examples. In the first example, for a randomly generated matrix

$$\mathbf{M}_0 = \begin{bmatrix} 4.169 & -1.454 & 0.059 & 0.060 \\ -1.454 & 0.865 & -0.276 & 0.095 \\ 0.059 & -0.276 & 2.426 & -0.534 \\ 0.060 & 0.095 & -0.534 & 1.433 \end{bmatrix} , \tag{4.17}$$

for $\Omega = [0, \infty)^3 \times (0, \infty)$ and $\Omega_0 = \{0\}^3 \times (0, \infty)$, we estimated expression (4.3) to be asymptotically equivalent to $0.04\,\chi^2_0 + 0.27\,\chi^2_1 + 0.46\,\chi^2_2 + 0.23\,\chi^2_3$. The associated q-q plot is shown in Figure 4.2. In a second example, we took the

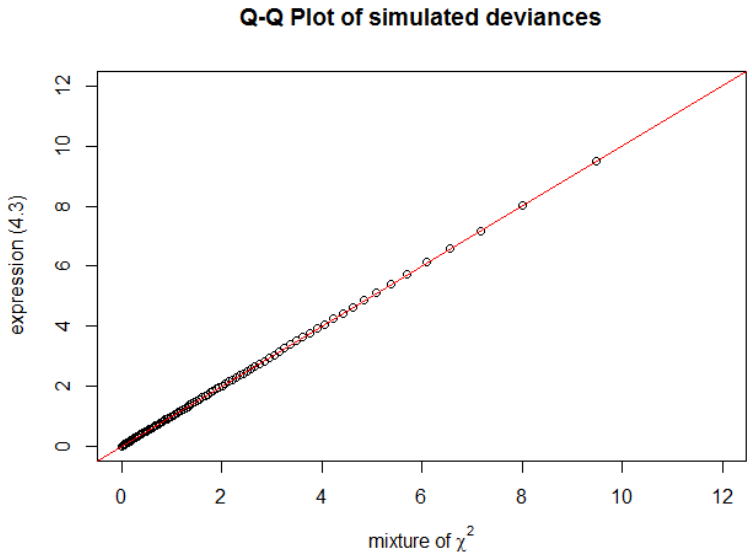

Fig. 4.2 A q-q plot of the sample percentiles of simulated observations from expression (4.3) against the sample percentiles of simulated observations from $\sum_{k=0}^{q} \mathbf{1}(\mathbf{Y}_q \in \mathcal{R}_k)\chi^2_k$. The matrix $\mathbf{M}_0$ is given by equation 4.17, $\Omega = [0, \infty)^3 \times (0, \infty)$ and $\Omega_0 = \{0\}^3 \times (0, \infty)$.

variance of the score statistic to be

$$\mathbf{M}_0 = \begin{bmatrix} 2.678 & 0.085 & 1.824 & -2.445 & -0.571 \\ 0.085 & 4.023 & 1.015 & -0.182 & 0.174 \\ 1.824 & 1.015 & 6.245 & -1.677 & -2.136 \\ -2.445 & -0.182 & -1.677 & 2.589 & 1.783 \\ -0.571 & 0.174 & -2.136 & 1.783 & 5.525 \end{bmatrix}, \tag{4.18}$$

and defined Ω and Ω_0 respectively as $\Omega = [0, \infty)^4 \times (0, \infty)$ and $\Omega_0 = \{0\}^4 \times (0, \infty)$. In this case, expression (4.3) was estimated to be asymptotically equivalent to $0.010\,\chi^2_0 + 0.203\,\chi^2_1 + 0.454\,\chi^2_2 + 0.299\,\chi^2_3 + 0.034\,\chi^2_4$. Once more, the q-q plot confirms the theoretical results (Figure 4.3).

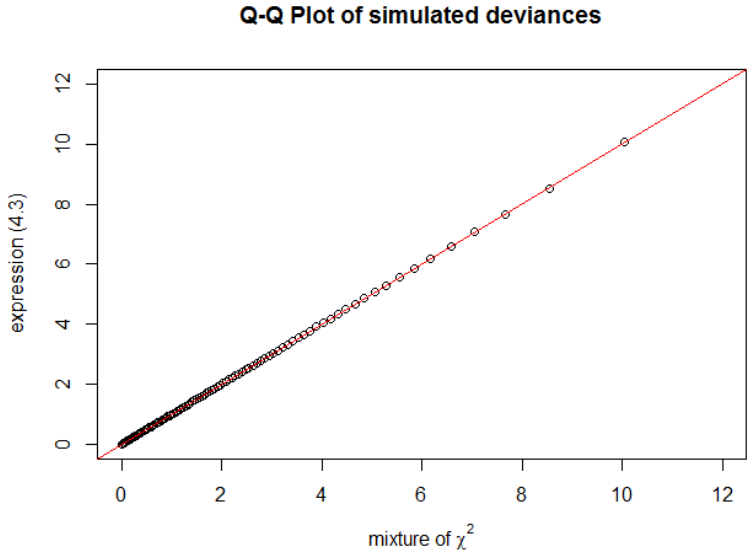

Fig. 4.3 A q-q plot of the sample percentiles of simulated observations from expression (4.3) against the sample percentiles of simulated observations from $\sum_{k=0}^{q} \mathbf{1}(\mathbf{Y}_q \in \mathcal{R}_k)\chi^2_k$. The matrix $\mathbf{M}_0$ is given by equation 4.18, $\Omega = [0, \infty)^4 \times (0, \infty)$ and $\Omega_0 = \{0\}^4 \times (0, \infty)$.

In our code to estimate the mixture of χ^2 distributions, the matrix $\mathbf{M}_0$ is an essential input. It defines the cone $C(\mathbf{K})$, its polar cone $C(\mathbf{K}^*)$, as well as the regions $\mathcal{R}_k$. Given that $\mathbf{M}_0$ is known, the asymptotic distribution of the likelihood ratio statistic can be estimated very accurately – this is what Figures 4.2 and 4.3 show. In a real inference problem, because we must rely on an estimate of $\mathbf{M}_0$, our method will provide an approximation to a mixture of χ^2 distributions which is not exactly the limiting distribution. In addition, this limiting distribution may be more or less distant from the actual distribution of the likelihood ratio statistic, which is based on a finite sample size. To

provide an illustration of how well the estimated mixture of χ^2 distributions may approximate the true distribution of the likelihood ratio statistic, we resorted to simulated data once more. We generated 1000 data sets from a simple isolation model with equal population sizes (model ISO from Figure 2.9 with $\theta_a = \theta = \theta_b$), each set containing 40,000 independent observations on the number of nucleotide differences between a pair of sequences. To each set, both the true model and an IM model with bidirectional migration (but also a single population size) were fitted and the likelihood ratio statistic was computed. Using the sample of 1000 likelihood ratio statistics thus obtained and the function *quantile* in R, we estimated the percentiles of the likelihood ratio statistic distribution. These estimated percentiles correspond to the y-coordinates of the q-q plot of Figure 4.4. In addition, using the observed Fisher information for a single data set, divided by n, as an approximation to $\mathbf{M}_0$, we also estimated the limiting mixture of χ^2 distributions. The estimates of the percentiles of this mixture of χ^2 distributions, computed from a sample of 1000 simulated observations, correspond to the x-coordinates of the q-q plot of Figure 4.4. Clearly the estimated percentiles of both distributions agree quite well. This suggests that, in a real statistical inference scenario, our method can estimate with precision the limiting distribution of the likelihood ratio statistic, as long as the sample size is large enough to enable an accurate estimate of $\mathbf{M}_0$, and as long as this matrix is positive definite. For comparison, in Figure 4.5 we include a q-q plot whose y-coordinates are the same as in Figure 4.4, but whose x-coordinates represent the theoretical percentiles of a χ^2 distribution with degrees of freedom equal to the number of linear constraints imposed by the null hypothesis (i.e. two degrees of freedom).

4.2.5 The data of Wang and Hey (2010)

In chapter 2, in the analysis of the data of Wang and Hey (2010), there were two likelihood ratio tests in which the true parameter vector is assumed to be on the boundary. In the case of IM_1 versus IIM_1, no simulation is needed, as we are just testing the assumption that $T_1 = 0$. The true limiting distribution is simply $0.5\,\chi^2_0 + 0.5\,\chi^2_1$, from 'case 5' of Self and Liang (1987). This means that if we use the naive χ^2_1 distribution instead, the p-value of any given likelihood ratio statistic will be twice as large as the p-value according to the true distribution under the null hypothesis.

Unfortunately, the limiting distribution of the other likelihood ratio test with a true parameter vector on the boundary, the one of the ISO model versus

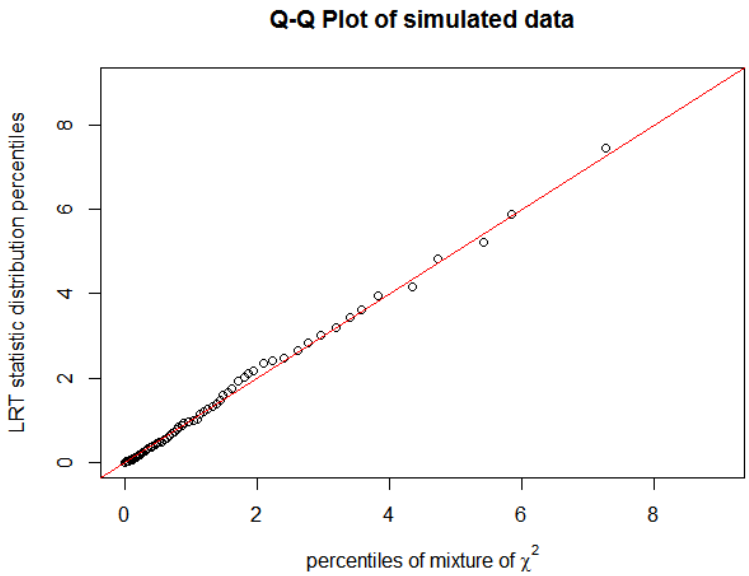

Fig. 4.4 A q-q plot of estimated percentiles of a likelihood ratio statistic distribution, against the estimated percentiles of $0.251\,\chi^2_0 + 0.504\,\chi^2_1 + 0.245\,\chi^2_2$. The likelihood ratio statistics refer to the comparison between the ISO model with $\theta_a = \theta = \theta_b$ (true model) and the IM_1 model with $\theta_a = \theta = \theta_b$ (see Figure 2.9). The χ^2 mixture was estimated using the observed Fisher information (for a single data set), divided by the number of observations, as an approximation to $\mathbf{M}_0$.

the IM_1 model, cannot be estimated by any method or theoretical result relying on the positive-definiteness of the Fisher information (such as our own). This is because the observed Fisher information divided by the number of observations, which replaces the unknown Fisher information, happens to be non-positive definite in this particular case. In fact, when the vector of estimated parameters lies on the boundary of the parameter space, the positive-definiteness of the observed Fisher information is not guaranteed (Gill and King, 2004).

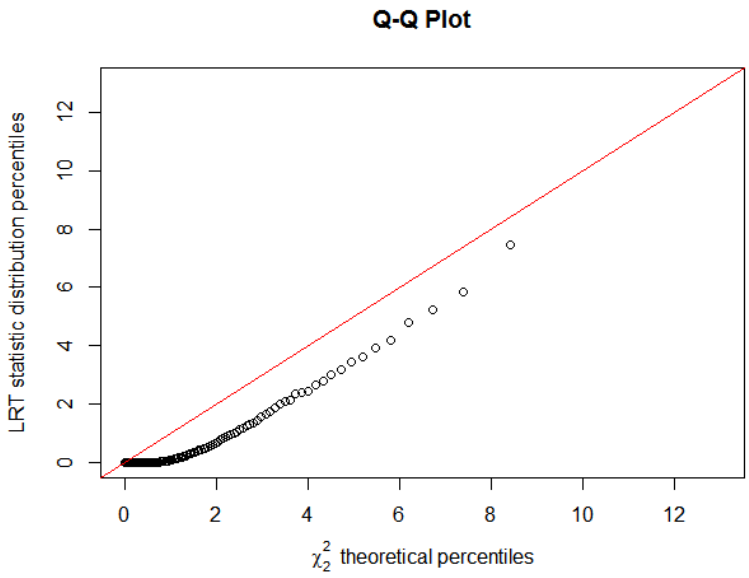

Fig. 4.5 A q-q plot of estimated percentiles of a likelihood ratio statistic distribution, against the theoretical percentiles of the χ^2_2 distribution. The likelihood ratio statistics refer to the comparison between the ISO model with $\theta_a = \theta = \theta_b$ (true model) and the IM_1 model with $\theta_a = \theta = \theta_b$ (see Figure 2.9).

Chapter 5

Discussion

5.1 Notes on our method and results

We have described a fast method to fit a range of demographic models to large data sets of pairwise nucleotide differences at a large number of independent loci. This method relies essentially on the eigendecomposition of the generator matrix of the process during the migration stages of the model: for each set of parameter values, the computation of the likelihood involves this decomposition. Nevertheless, the whole process of estimation takes no more than a couple of minutes for a data set of tens of thousands of loci such as that of Wang and Hey (2010), and it does not require high-performance computing resources.

Using the estimation methods developed in the present thesis, along with formal procedures of model comparison such as hypothesis tests and AIC scores, we are able to tell which of four gene flow scenarios is most consistent with a set of pairwise nucleotide differences. In addition to the scenario of divergence with continuous gene flow until the present, and divergence under complete isolation, we can hope to identify two other important evolutionary scenarios: divergence with ancestral gene flow and divergence with secondary contact. A model of divergence with ancestral gene flow, as represented by the IIM model of chapter 2, is clearly useful when it is known, or at least suspected, that present-day subpopulations have achieved reproductive isolation. A model of divergence with secondary contact, as represented in figure 5.1, is of special interest when trying to determine whether the formation of two sympatric subspecies, which are presently still exchanging genes, could have been caused by a period of isolation in the past, rather than by disruptive selection alone. The importance of being able to fit models representing these divergence scenarios while allowing for unequal subpopulation sizes is illustrated in the robustness

analysis of chapter 2: substantial deviations from reality in this respect can translate into inadequate estimates of gene flow levels and splitting times. As to the ability of estimating the level of gene flow in each direction, it should help us understand how prevalent asymmetric gene flow is, how it relates to the sizes of subpopulations, and its effect on the observed patterns of genetic variation.

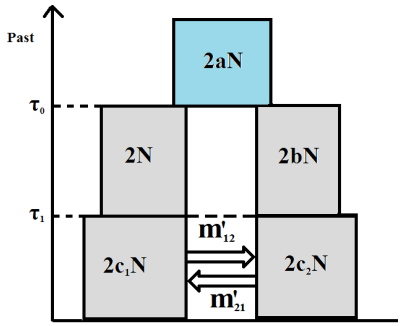

Fig. 5.1 A model of divergence in which current gene flow is preceded by a period of isolation (a GIM model with $m_{12} = m_{21} = 0$). Such a scenario may have been caused, for example, by climatic changes leading to habitat fragmentation and subsequent reconnection of populations.

Due to the number of parameters, it is not feasible to assess the performance of our method systematically over every region of the parameter space. However, our experience with simulated data sets suggests that the variances of the estimators associated with a given period of gene flow may become inflated in two particular cases. One of such cases arises when it is very unlikely that the genealogy of pair of sequences is affected by events that occurred during a given period of gene flow. For example, under the full IIM model, whenever V is very small or T_1 is very large, the precision of $\hat{M}_1$, $\hat{M}_2$, $\hat{\theta}$, $\hat{\theta}_b$ and $\hat{V}$ is likely to be affected. The second case arises when the values of the scaled migration rates are greater than one, so that the two subpopulations during the period of gene flow resemble a single panmictic population. In either of these cases, the very process of model fitting can become unstable, that is, the algorithm of maximisation of the likelihood may have difficulty converging.

It should also be noted that, for sample sizes of just a few thousand loci, the distribution of migration rate estimates may still be far from Gaussian (Figure 2.8). In such cases, computation of confidence intervals should be based on bootstrap methods or on the likelihood (profile likelihood confidence intervals) rather than on the observed Fisher information (Wald confidence intervals).

Table 5.1 Comparison of converted estimates obtained with IM and IIM models

	$\mathbf{IM}_{\mathrm{wh}}$	$\mathbf{IM}_1$	$\mathbf{IIM}_3$
Time since onset of speciation	3.040	3.240	3.624
Time since isolation	-	-	1.503
Size of ancestral population	3.060	4.310	3.549
Current size of *D. sim.* population	5.990	6.120	7.182
Current size of *D. mel.* population	2.440	2.700	2.871
Size of *D. sim.* population during IIM gene flow period	-	-	3.640
Size of *D. mel.* population during IIM gene flow period	-	-	2.092
Migration rate (*D. sim.* $\rightarrow$ *D. mel.*)	0.013	0.012	0.064
Migration rate (*D. mel.* $\rightarrow$ *D. sim.*)	0.000	0.000	-

Note: Times are given in millions of years; population sizes are given in millions of individuals; the migration rates stated represent the number of sequences that migrate per generation, forward in time. The model $\mathrm{IM}_{\mathrm{wh}}$ is the IM model fitted by Wang and Hey (2010).

How many loci are needed to obtain good estimates and confidence intervals will also depend on the region of the parameter space concerned.

It is not the goal of this thesis to draw conclusions regarding the evolutionary history of *Drosophila* species. We used the data of Wang and Hey (2010) with the sole objective of demonstrating that our method can be applied efficiently and accurately to real data. In Table 5.1, we list both our estimates and those of Wang and Hey (2010) for a six-parameter isolation-with-migration model (the IM_1 model – see Figure 2.9). The same table contains the estimates for our best-fitting IIM model. Our parameter estimates for the IM model agree well with those of Wang and Hey (2010). The reason that they do not match exactly lies in the fact that we have omitted the 'screening procedure' described in Wang and Hey (2010) and have therefore not excluded some of the most divergent sequences in the data set. It should also be borne in mind that our model of mutation is the infinite-sites model, whereas Wang and Hey (2010) have worked with the Jukes-Cantor model. Furthermore, our choice of sequence pairs was somewhat different: Wang and Hey (2010) randomly selected a pair of sequences at each locus, whereas we followed the procedure described in section 2.6.1 above. There are some notable differences between the estimates for both IM models and those for the IIM model: under the IIM model, the process of speciation is estimated to have started earlier (3.6 million years ago instead of 3.0 or 3.2 million years ago), to have reached complete isolation before the present time (1.5 million years ago), and to have a higher rate of gene flow (0.064 sequences per generation instead of 0.013 or 0.012 sequences) during a shorter period of time (2.1 million years of gene flow instead of 3.0

or 3.2 million years). As might be expected, the estimates of each descendant population size (*D. simulans* and *D. melanogaster*) in the IM models lie in between the estimates of the corresponding current population size and its size during the gene flow period in the IIM model.

The purpose of presenting the *Cobitis* fish case study in Chapter 3 was to provide an example of how our methods can be applied to specific research questions. Whether or not clonality can be a primary reproduction barrier falls beyond the scope of the present thesis. We also wished to stress the fact that polymorphism-based statistical inference is not a stand-alone research tool. On the contrary, a given evolutionary hypothesis becomes plausible when inferences based on different sources (field observations, laboratory experiments, fossil records, polymorphism data) are in agreement.

The method we used assumes that relative mutation rates are known (see section 2.4.2). In reality, we must deal with estimates of these rates, and this introduces additional uncertainty which is not reflected in the standard errors and confidence intervals obtained. In principle, this uncertainty can be reduced by increasing the number of ingroup and outgroup sequences used to compute the average number of pairwise differences at each locus in equation (2.15). Ideally, estimates of the relative mutation rates should be based on outgroup species only, to avoid any dependence between these estimates and the observations on ingroup pairwise differences (Wang and Hey, 2010).

5.2 Violation of assumptions

Some assumptions of our models, such as the infinite-sites assumption and the assumption of free recombination between loci and no recombination within loci, may not be sensible for some real data sets. The appropriateness of other assumptions, for example those regarding the constant size of populations or the constant rate of gene flow, will depend on the actual evolutionary history of the species or populations involved. Whilst a systematic, in-depth robustness analysis of our method (similar to, for example, the robustness studies by Becquet and Przeworski, 2009, and Strasburg and Rieseberg, 2010, for commonly used IM methods) is beyond the scope of this thesis, we will in this section informally examine the impact of possible violations of some of the main assumptions made.

Misspecification of the demographic model

In order to explore the potential effect of misspecification of the demographic model on inference accuracy, we first simulated 20 data sets of 40,000 loci each from a somewhat more complex evolutionary scenario, depicted in the left-hand side diagram of Figure 5.2, where subpopulation sizes gradually increase and gene flow gradually declines. The precise parameter values assumed for the true model were chosen arbitrarily and are also shown in the left-hand side diagram; in accordance with the reparameterisation used in section 2.5, divergence times are measured on a mutational scale by twice the expected number of mutations per sequence (as an average over all loci), population sizes are represented by scaled mutation rates, and rates of gene flow by scaled migration rates. We then applied our method to fit isolation, IM, IIM and GIM models to each of the simulated data sets and selected the best-fitting model by means of likelihood ratio tests – for 18 out of the 20 data sets generated this was found to be the full IIM model. The average point estimates obtained for each parameter under the full IIM model are shown on the right-hand side diagram of Figure 5.2. In each diagram, the widths of the boxes are proportional to the population sizes and the heights are proportional to the durations of the time periods concerned. It is readily seen that the IIM model reflects the dynamics of the true model quite well. Population sizes, migration rates and splitting times are all estimated at intermediate values.

We also repeated the simulation and estimation procedure for an evolutionary scenario involving a period of secondary gene flow, depicted in the left-hand side diagram of Figure 5.3. Again, for 18 out of 20 simulated data sets, the full IIM model provides the best fit amongst the models considered (isolation, IM, IIM and GIM). Comparing the two diagrams in Figure 5.3 (where the IIM parameter values in the right-hand side diagram are once more the averages of the estimates obtained), we see that the IIM model obtained provides a reasonable approximation to the true model, though of course our method did not detect the initial period of isolation. The estimates of the time since the onset of speciation and the time since complete isolation are, on average, close to the true values, whilst the average estimates of the migration rate and population size parameters are again at intermediate values, compared to the range of true values over time.

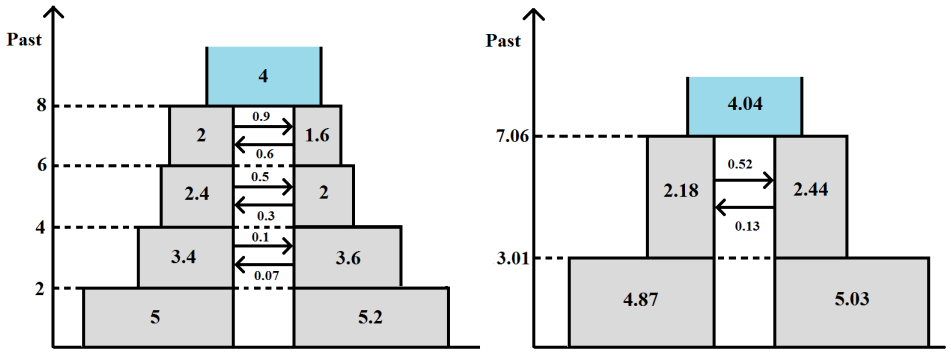

Fig. 5.2 Violation of demographic assumptions. Left-hand side diagram: true model. Right-hand side diagram: best-fitting model. Divergence times are measured by twice the expected number of mutations per sequence, population sizes are represented by scaled mutation rates, and rates of gene flow by scaled migration rates.

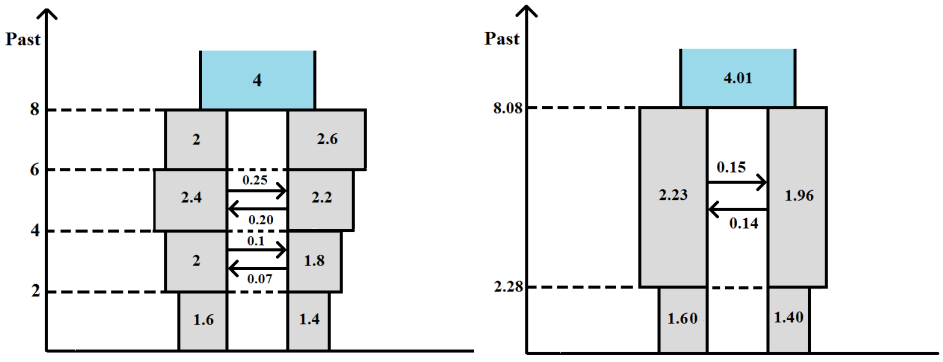

Fig. 5.3 Violation of demographic assumptions. Left-hand side diagram: true model. Right-hand side diagram: best-fitting model. Divergence times are measured by twice the expected number of mutations per sequence, population sizes are represented by scaled mutation rates, and rates of gene flow by scaled migration rates.

Intra-locus recombination

In common with other methods mentioned in this thesis (for example, Wang and Hey, 2010; Lohse et al., 2011), our method assumes free recombination between loci and no recombination within loci. The second of these two assumptions is the most important one, without which our method would not be valid. Recombination within loci mixes up the genealogies of DNA sequences on which our method relies, making pairs of sequences more equidistant: intra-locus recombination does not affect the mean number of nucleotide differences in a pair of sequences but the *variance* decreases with increasing recombination (Griffiths, 1981; Hudson, 1983; Schierup and Hein, 2000), resulting in data sets which contain more intermediate values and fewer extreme values. This can be expected to lead to overestimation of the current population sizes and underestimation of the ancestral population size, whilst the effect on estimates of the other parameters is intuitively somewhat less obvious. The impact of intra-locus recombination on the variance of the number of pairwise differences, and hence on the accuracy of our method, may be expected to be less severe in cases of recombination rate heterogeneity within loci (see Figure 1 in Hudson, 1983, for the extreme case of recombination hotspots separating completely linked regions) .

A simulation study by Strasburg and Rieseberg (2010) found that even relatively low levels of intra-locus recombination can cause substantial bias in estimates of the IM model parameters obtained using the program *IMa* (Hey and Nielsen, 2007), with highest posterior density intervals failing to contain the true parameter values far more often than would be expected by chance. In IM simulations allowing a minimal but realistic amount of intra-locus recombination, Lohse et al. (2016) found that the bias in their parameter estimates was small. Although our method and models are different from those of Hey and Nielsen (2007) and Lohse et al. (2016), the effect of recombination on the underlying genealogies remains the same, and therefore similar biases will occur if the assumption of no intra-locus recombination is violated.

For the *Drosophila* data considered in section 2.6, Wang and Hey (2010) assessed the impact of potential intra-locus recombination on their estimates of the parameters of an IM model by comparison with the estimates obtained from the same sequences but halved in length (i.e. approximately halving the expected number of intra-locus recombination events). Their estimates of the ancestral population size and the migration rate from the half-length data were about 30% larger than those from the full-length data, whilst the differences for

the other parameter estimates were small. In the same spirit, we repeated our previous analysis of the *Drosophila* data but now using the trimmed version of the Wang subset prepared by Lohse et al. (2011), in which the average locus length was reduced by approximately a factor of 3; the Hutter subset (approximately 1% of the total number of loci) was retained in its entirety as we could not afford to further reduce this already very small data set of *D. melanogaster* pairs. Applying the estimation and model selection procedures described in section 2.6 to this trimmed version of the data, the likelihood ratio test of the models IIM_1 versus IIM_2 was no longer significant, i.e., there was no longer significant evidence of an increase in population size at time T_1, and the best-fitting model was a unidirectional version of IIM_1 (i.e. with $M_1 = 0$).

Table 5.2 shows the estimates obtained from the trimmed data; the estimates obtained in section 2.6 from the full data are also listed again for comparison. In line with our expectations regarding the potential effect of intra-locus recombination, it is seen that the full data gave a larger estimate of the current population size of *D. simulans* and a smaller estimate of the ancestral population size; the estimated size of *D. simulans* during the gene flow stage was also smaller than that obtained from the trimmed data. The estimated time since the onset of speciation is nearly identical for the two data sets, but the full data placed the end of gene flow substantially further back into the past (1.5 million years ago compared to 0.93 million years) and estimated a somewhat higher number of migrant sequences per generation (0.064 compared to 0.051) during a shorter period of gene flow (2.12 compared to 2.68 million years). This suggests that, in addition to the impact on population size estimates already discussed, intra-locus recombination may lead to an overestimate of the time since the end of gene flow in an IIM model and (possibly as a consequence) an overestimate of the migration rate. Nevertheless, for both versions of the *Drosophila* data, the likelihood ratio tests of non-zero migration rate and non-zero time since the end of gene flow were significant.

The above considerations imply that, when preparing data for use with our method (or any other method relying on the assumption of no intra-locus recombination), loci should be chosen carefully to try to keep the amount of intra-locus recombination negligible, and some caution may be needed in the interpretation of results. For data sets showing signs of recombination within loci, it may be possible to reduce its effect by trimming or breaking up such loci to form shorter, apparently non-recombining segments of DNA sequence (Hey and Nielsen, 2004; Strasburg and Rieseberg, 2010). An extension of our method to account for recombination within loci would be of interest but is

Table 5.2 Converted estimates for the data of Wang and Hey (2010): full sequences and trimmed sequences.

	trimmed		**full**
	IIM$_1^*$	**IIM$_3$**	**IIM$_3^*$**
Time since onset of speciation	3.614	3.634	3.624
Time since isolation	0.934	0.997	1.503
Size of ancestral population	4.264	4.237	3.549
Current size of *D. sim.* population	-	6.024	7.182
Current size of *D. mel.* population	-	2.984	2.871
Size of *D. sim.* population during gene flow	-	5.956	3.640
Size of *D. mel.* population during gene flow	-	1.891	2.092
Size of *D. sim.* population	5.998	-	-
Size of *D. mel.* population	2.795	-	-
Migration rate (*D. sim.* $\rightarrow$ *D. mel.*)	0.051	0.038	0.064
Migration rate (*D. mel.* $\rightarrow$ *D. sim.*)	0.000	0.000	0.000

Note: Times are given in millions of years; population sizes are given in millions of individuals; the migration rates stated represent the expected number of sequences that migrate per generation, forward in time. The best-fitting model for each data set is marked with an (*).

challenging. An extension to a finite-sites model for use with shorter fragments of DNA sequence would also be of interest – such an extension is relatively straightforward but is yet to be implemented in our method (but see Wang and Hey 2010 and Andersen et al. 2014 for the IM model).

Linkage disequilibrium

If the assumption of free recombination between loci does not hold, then loci are not independent, in which case the likelihood in equations (2.14) and (3.16) is in fact a composite marginal likelihood (also called the 'independence likelihood' in Chandler and Bate, 2007) rather than an ordinary full likelihood (see Varin, 2008, for an overview of composite marginal likelihood methods; see also the Discussion of Lohse et al., 2016). Statistical theory indicates that in that case the maximum composite likelihood estimator (MCLE) is still consistent (Cox and Reid, 2004; Wiuf, 2006, with some minor modifications to account for our slightly different assumptions; Varin, 2008), provided the relative mutation rates at the different loci are bounded. Thus, if linkage between loci cannot be ignored, the MCLE of the parameters of a model obtained with our method will still be approximately unbiased if the number of loci is sufficiently large, and if all our other assumptions hold (including the assumption of no recombination within loci). However, if linkage between loci is not negligible, then standard

errors and confidence intervals computed using the observed Fisher information (as was done in section 2.6.3) will underestimate the true uncertainty about the parameter estimates obtained (Baird, 2015); instead, standard errors and confidence intervals should be based on an estimate of the Godambe information (Godambe, 1960). For a data set made up of a single string of correlated loci, or a small number of such strings, obtaining an accurate estimate of the Godambe information presents some difficulties (see Varin, 2008, and Varin et al., 2011, for a discussion and some possible strategies). A much simpler situation arises if the data consist of a sufficiently large number of 'clusters' of loci, where loci within clusters are correlated but where different clusters can be considered independent. This may be the case, for example, if different clusters of loci are chosen from different chromosomes, or are separated by recombination hotspots or by a large enough distance along the genome. For such data, an empirical estimate of the Godambe information can easily be computed as described in Chandler and Bate (2007) or Varin (2008).

To try to quantify the effect of linkage on the standard errors of our parameter estimates, we conducted the following analysis of a suitable subset of the Wang and Hey (2010) data. We partitioned the 30247 loci of the Wang subset into blocks of 100 consecutive loci and discarded every other block, so that 151 blocks were retained of 100 loci each. Since the individual loci are approximately 500 bp in length and separated by at least 2 kb, this leaves a distance of at least 0.25 Mb between different blocks, and we can reasonably assume that any effect of linkage between blocks of loci this far apart is negligible compared to that within blocks. In the Hutter subset the distance between consecutive loci is on average about 50 kb, and we retained these 378 loci in order to enable estimation of the *D. melanogaster* population size parameters.

To examine the effect of linkage we analysed this reduced data set in two ways in order to compare the results: (i) assuming that loci are independent, and (ii) accounting for any linkage between loci within blocks, i.e. accounting for the bulk of the linkage in the data. In case (i), the model selection procedure described in section 2.6.2 was carried out on the reduced data set. As was the case for the full data, the model IIM_3 provided by far the best fit also for the reduced data set. The p-values computed as part of the model selection procedure were all smaller than 10^{-42} and are shown in the left-hand side column of Table 5.3. The parameter estimates for the best-fitting model, IIM_3, are shown in Table 5.4 and are very close to the estimates obtained from the full Wang and Hey (2010) data (see Table 2.3). Standard errors of the parameter estimates, based on the observed Fisher information, are also shown in Table 5.4

for the reduced data set. As expected, these standard errors are larger than those for the full data set (by a factor of approximately $\sqrt{2}$), except those of the *D. melanogaster* population size parameters, which are largely unchanged.

In case (ii), i.e., in order to account for any linkage within blocks of loci, both the model selection procedure and the computation of standard errors were performed using theoretical results for composite marginal likelihoods. The hypothesis tests in the model selection procedure were carried out using Result 3.5 and approximation (3.6) of Jesus and Chandler (2011), by which the null distribution of the composite likelihood ratio test statistic is approximated by a scaled and shifted χ^2 distribution (see also the comments regarding the distribution of the independence likelihood ratio test statistic in Chandler and Bate, 2007, p.170-171). The p-values obtained in this way for the tests in the model selection procedure are shown in the right-hand side column of Table 5.3. As expected, these p-values are not as small as those obtained when ignoring linkage, and in fact they differ by many orders of magnitude. Nevertheless, they are all still smaller than 10^{-20}, and the model IIM_3 still gives by far the best fit for the reduced Wang and Hey (2010) data (note however that, to the best of our knowledge, it has not been established in the literature whether the approximate null distribution used for the composite likelihood ratio test statistic is still conservative in the case of tests involving parameters on the boundary, although this would seem plausible). Standard errors of the parameter estimates of the IIM_3 model were computed by obtaining an empirical estimate of the inverse of the Godambe information matrix using the method for clustered data described in Chandler and Bate (2007): the covariance matrix of the score vector evaluated at the true parameter vector was estimated by

$$\hat{\mathbf{V}} = \sum_j \mathbf{U}_j \mathbf{U}'_j$$

where $\mathbf{U}_\text{j}$ is the score of the jth block of loci, evaluated at the MCLE, and the sum is over all blocks; an estimate of the inverse of the Godambe information matrix (also referred to as the 'robust' variance estimator) was then computed as

$$\hat{\mathbf{G}}^{-1} = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{H}}^{-1}$$

where $\mathbf{H}$ is the negated Hessian matrix of the log-likelihood function, evaluated at the MCLE. The resulting standard errors are shown in the right-hand side column of Table 5.4. It is seen that, on average, the standard errors based on the Fisher information account for about 80% of the uncertainty

given by the 'robust' standard errors, though this percentage is different for different parameters. The strongest impact is on the standard error of θ_{c_1} (the 'current size' parameter of *D. simulans*), for which the standard error ignoring linkage is only 59% of that which does account for linkage between loci within blocks – one would indeed expect the impact of linkage to be strongest on the standard errors of parameters relating to more recent events, as a shorter time allows less opportunity for recombination between loci (no such effect is seen on the standard error of θ_{c_2} as we continued to treat the Hutter subset as independent loci). An alternative method to account for linkage disequilibrium is by means of a parametric bootstrap (for example, Lohse et al., 2016), but this is computationally intensive and the results will inevitably depend on the recombination rate assumed, and on any other assumptions made such as homogeneity of the recombination rate along the genome.

The 'robust' standard errors in the right-hand side column of Table 5.4 were computed accounting for linkage whilst assuming that all our other assumptions hold. If the latter is not the case, then the individual factors in equations (2.14) and (3.16) may be misspecified so that their product no longer defines a composite marginal likelihood. Instead, the derivative of its logarithm can be regarded as an 'estimating function' and the corresponding statistical theory applied. In that case, our 'robust' calculations of standard errors and p-values in (ii) above still apply (Jesus and Chandler, 2011, Section 3), so that the results in the right-hand columns of Tables 5.3 and 5.4 are still valid. Thus the differences between the left- and right-hand side columns of standard errors and p-values in Tables 5.3 and 5.4 should be interpreted as *upper bounds* on the impact of linkage, since part of these differences may be due to other forms of model misspecification, including from any of the potential sources discussed above: inaccurate estimates of the relative mutation rates, misspecification of the mutation model, misspecification of the demographic model, and intra-locus recombination. In other words, when we allow for the fact that other forms of misspecification, apart from linkage between loci, may exist, the present adjustment is essentially the same as the one described in section 4.1 (model misspecification): the difference is that, in that section, we had not dropped the assumption that the full data consist of completely unlinked loci.

5.3 Further work

In population genetics, as in other areas of probability theory, there is a trade-off between mathematical convenience and realism. Methods that implement

Table 5.3 Results for the data of Wang and Hey (2010), reduced version: p-values for (composite) likelihood ratio tests in model selection.

$\mathbf{H_0}$	$\mathbf{H_1}$	**p-values**	
		(i) χ^2 null distribution	**(ii) 'robust' null distribution**
ISO	IM_1	2.60 E-129	1.39 E-110
IM_1	IIM_1	8.40 E-57	2.11 E-21
IIM_1	IIM_2	1.62 E-43	7.86 E-28

Note: in (i) the usual χ^2 distribution with the appropriate number of degrees of freedom was used as the null distribution; in (ii) the null distribution used is a scaled and shifted χ^2 distribution (Jesus and Chandler, 2011, equation 3.6).

Table 5.4 Results for the data of Wang and Hey (2010), reduced version: point estimates and estimated standard errors under the model IIM_3.

Parameter	**Estimate**	**Standard Errors**	
		(i) Fisher	**(ii) Godambe**
θ_a	3.217	0.130	0.146
θ	3.259	0.155	0.168
θ_b	1.934	0.998	1.251
θ_{c_1}	6.833	0.161	0.271
θ_{c_2}	2.643	0.174	0.182
T_1	7.118	0.273	0.435
V	9.826	0.228	0.286
M_2	0.250	0.026	0.035

Note: 'Fisher' and 'Godambe' standard errors are based on the observed Fisher and on the estimated Godambe information matrices respectively.

more complex models are normally less efficient, but they can also provide more reliable and accurate inferences – as long as there is enough information in the data to estimate the additional parameters. There are many ways of extending and generalising both the GIM model and the infinite-sites model of mutation, making them more realistic. Below we consider some extensions whose implementation is likely to be possible using methods similar to the ones described in the present thesis.

The assumption that relative mutation rates are known is one that could be dropped in the future. If we do so, the likelihood of the parameters under any of the models in this thesis becomes what it actually is: an *estimated* or *pseudo-likelihood*. It should then be possible to incorporate the uncertainty about the relative mutation rates, by applying, for example, the asymptotic results derived by Gong and Samaniego (1981). Furthermore, an extension of our method

to the Jukes-Cantor model of mutation would also be of interest. Under this model of mutation, the *pmf* of the number of pairwise nucleotide differences can be written as a sum of moment generating functions of the coalescence time between two lineages (see Lohse et al., 2011, equation 3). Hence, an explicit expression for this *pmf*, under any of our models of speciation, is likely to be attainable.

If the period between the split of the ancestral population and the present is divided into three or more two-island models with potential migration, instead of just two, the derivations in chapter 3 (GIM model) will still be applicable with minor modifications. Such an extension would allow us to fit, for example, a model of secondary contact followed by a period of isolation, similar to the one depicted in the left-hand side diagram of Figure 5.3. In terms of demographic assumptions, another generalisation of interest would be the inclusion of more subpopulations/species. Our analysis of the *Cobitis* data, together with the other independent evidence mentioned in chapter 3, suggest that there was no gene flow between more than two *Cobitis* species at the same time. In general, however, substantial gene flow with a third species is enough to compromise the accuracy of estimates (Strasburg and Rieseberg, 2010). An extension of our method to models with three subpopulations, and simultaneous gene flow between them, may be challenging, even for data sets consisting of pairwise nucleotide differences; and it can happen that, even in a large data set as that of Wang and Hey (2010), there is not enough information to fit such complex models.

The results presented in section 4.2 also suggest the need of further investigations. The simulation method described in that section can estimate efficiently the asymptotic distribution of the likelihood ratio statistic when the true parameter lies on the boundary of the parameter space. However, it relies on an assumption which would be of interest to drop, namely the assumption of correct model misspecification. Instead of the results in Chernoff (1954) and Self and Liang (1987), an extension of our method to this more general setting could be based on the more complex theory in Liang and Self (1996) and Chen and Liang (2010). Recall also that the simulation method relies on the positive-definiteness of the observed Fisher information. Hence it would be useful to know what can be done, if anything, to solve the problem of not always obtaining a positive-definite observed Fisher information, when the maximum-likelihood estimate happens to lie on the boundary of the parameter space (see section 4.2.5). For example, it should be investigated whether adequate estimates of the asymptotic distribution can still be obtained when

a non-positive definite observed Fisher information is replaced by the nearest positive definite matrix (which can be computed, for example, using the R package ‘Matrix’).

References

Andersen, L., T. Mailund, and A. Hobolth (2014). Efficient computation in the IM model. *Journal of Mathematical Biology 68*(6), 1423–1451.

Anderson, E. C. and E. A. Thompson (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics 160*(3), 1217–1229.

Bahlo, M. and R. Griffiths (2000). Inference from gene trees in a subdivided population. *Theoretical Population Biology 57*(2), 79–95.

Baird, S. J. (2015). Exploring linkage disequilibrium. *Molecular Ecology Resources 15*(5), 1017–1019.

Becquet, C. and M. Przeworski (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research 17*(10), 1505–1519.

Becquet, C. and M. Przeworski (2009). Learning about modes of speciation by computational approaches. *Evolution 63*(10), 2547–2562.

Beerli, P. and J. Felsenstein (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics 152*(2), 763–773.

Borovik, A. V. and A. Borovik (2010). *Mirrors and Reflections*. Springer.

Bryc, W. (1995). *The Normal Distribution*. Springer.

Casella, G. and R. Berger (2001). *Statistical Inference* (2nd ed.). Duxbury.

Chandler, R. E. and S. Bate (2007). Inference for clustered data using the independence loglikelihood. *Biometrika 94*(1), 167–183.

Chen, Y. and K.-Y. Liang (2010). On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika 97*(3), 603.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics 25*(3), 573–578.

Choleva, L., K. Janko, K. De Gelas, J. Bohlen, V. Šlechtová, M. Rábová, and P. Ráb (2012). Synthesis of clonality and polyploidy in vertebrate animals by hybridization between two sexual species. *Evolution 66*(7), 2191–2203.

Choleva, L., Z. Musilova, A. Kohoutova-Sediva, J. Paces, P. Rab, and K. Janko (2014). Distinguishing between incomplete lineage sorting and genomic introgressions: Complete fixation of allospecific mitochondrial DNA in a sexually reproducing fish (*Cobitis*; Teleostei), despite clonal reproduction of hybrids. *PLoS ONE 9*(6), 1–16.

Costa, R. J. and H. Wilkinson-Herbots (2017). Inference of gene flow in the process of speciation: An efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics 205*(4), 1597–1618.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*(3), 729–737.

Coyne, J. A. and H. A. Orr (1998). The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 353*(1366), 287–305.

Dattorro, J. (2005). *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing USA.

Ewens, W. J. (2004). *Mathematical Population Genetics I*. Springer-Verlag, New York.

Excoffier, L. (2004). Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology 13*(4), 853–864.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics 22*(1), 521–565.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician 60*(4), 299–302.

Futuyma, D. (2005). *Evolution*. Sinanuer Associates.

Gentle, J. E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer Science & Business Media.

Gill, J. and G. King (2004). What to do when your hessian is not invertible: Alternatives to model respecification in nonlinear estimation. *Sociological Methods & Research 33*(1), 54–87.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics 31*(4), 1208–1211.

Gong, G. and F. J. Samaniego (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics 9*(4), 861–869.

Griffiths, R. C. (1981). The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. *Journal of Mathematical Biology 12*, 251–261.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009, 10). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics 5*(10), 1–11.

Hein, J., M. H. Schierup, and C. Wiuf (2005). *Gene Genealogies, Variation and Evolution*. Oxford.

Hey, J. (2006). Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development 16*(6), 592–596.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution 27*(4), 905–920.

Hey, J. and R. Nielsen (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics 167*(2), 747–760.

Hey, J. and R. Nielsen (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences 104*(8), 2785–2790.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*(2), 183–201.

Hutter, S., H. Li, S. Beisswanger, D. De Lorenzo, and W. Stephan (2007). Distinctly different sex ratios in african and european populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. *Genetics 177*(1), 469–480.

Innan, H. and H. Watanabe (2006). The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Molecular Biology and Evolution 23*(5), 1040–1047.

Janko, K., J. Bohlen, D. Lamatsch, M. Flajšhans, J. T. Epplen, P. Ráb, P. Kotlík, and V. Šlechtová (2007). The gynogenetic reproduction of diploid and triploid hybrid spined loaches (*Cobitis*: Teleostei), and their ability to establish successful clonal lineages—on the evolution of polyploidy in asexual vertebrates. *Genetica 131*(2), 185–194.

Janko, K., M. Flajšhans, L. Choleva, J. Bohlen, V. ŠLechtová, M. Rábová, Z. Lajbner, V. ŠLechta, P. Ivanova, I. Dobrovolov, M. Culling, H. Persat, J. Kotusz, and P. Ráb (2007). Diversity of European spined loaches (genus *Cobitis* L.): an update of the geographic distribution of the *Cobitis taenia* hybrid complex with a description of new molecular tools for species and hybrid determination. *Journal of Fish Biology 71*, 387–408.

Janko, K., J. Kotusz, K. De Gelas, V. Slechtova, Z. Opoldusova, P. Drozd, L. Choleva, M. Popiolek, and M. Balaz (2012). Dynamic formation of asexual diploid and polyploid lineages: multilocus analysis of *Cobitis* reveals the mechanisms maintaining the diversity of clones. *PLoS One 7*(9), 1–14.

Janko, K., J. Pačes, H. Wilkinson-Herbots, R. J. Costa, J. Röslein, P. Drozd, N. Iakovenko, J. Rídl, J. Kočí, R. Reichová, V. Šlechtová, and L. Choleva (2016). Hybrid asexuality as a primary reproductive barrier: on the interconection between asexuality and speciation. *bioRxiv e-prints*. URL: *http://biorxiv.org/content/early/2016/03/21/038299*.

Jesus, J. and R. E. Chandler (2011). Estimating functions and the generalized method of moments. *Interface Focus 1*(6), 871–885.

Kamm, J. A., J. Terhorst, and Y. S. Song (2016). Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics* (In Press), 1–37.

Kingman, J. F. (1982a). The coalescent. *Stochastic Processes and Their Applications 13*(3), 235–248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability 19*, 27–43.

Kopylev, L. and B. Sinha (2011). On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary. *Sankhya B 73*(1), 20–41.

Liang, K.-Y. and S. G. Self (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society. Series B (Methodological)*, 785–796.

Lohse, K., M. Chmelik, S. H. Martin, and N. H. Barton (2016). Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics 202*(2), 775–786.

Lohse, K., M. Clarke, M. G. Ritchie, and W. J. Etges (2015). Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution 69*(5), 1178–1190.

Lohse, K., R. J. Harrison, and N. H. Barton (2011). A general method for calculating likelihoods under the coalescent process. *Genetics 189*(3), 977–987.

Luc, D. T. (2016). *Multiobjective Linear Programming*. Springer.

Mailund, T., A. E. Halager, M. Westergaard, J. Y. Dutheil, K. Munch, L. N. Andersen, G. Lunter, K. Prüfer, A. Scally, A. Hobolth, and M. H. Schierup (2012, 12). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLOS Genetics 8*(12), 1–19.

Mayr, E. (1997). *Evolution and the Diversity of Life: Selected Essays*. Harvard University Press.

McDonald, J. N. and N. A. Weiss (1999). *A Course in Real Analysis*. Taylor & Francis US.

Moritz, C., W. Brown, L. Densmore, J. Wright, D. Vyas, S. Donnellan, M. Adams, and P. Baverstock (1989). Genetic diversity and the dynamics of hybrid parthenogenesis in Cnemidophorus (Teiidae) and Heteronotia (Gekkonidae). *Evolution and Ecology of Unisexual Vertebrates 466*, 87–112.

Nath, H. and R. Griffiths (1993). The coalescent in two colonies with symmetric migration. *Journal of Mathematical Biology 31*(8), 841–851.

Nath, H. and R. Griffiths (1996). Estimation in an island model using simulation. *Theoretical Population Biology 50*(3), 227–253.

Nielsen, R. (1998). Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology 53*(2), 143–151.

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck, and M. Slatkin (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution 52*(3), 669–677.

Nielsen, R. and M. Slatkin (2000). Likelihood analysis of ongoing gene flow and historical association. *Evolution 54*(1), 44–50.

Nielsen, R. and J. Wakeley (2001). Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics 158*(2), 885–896.

Nordborg, M. (2007). Coalescent theory. In *Handbook of Statistical Genetics*, Volume II. Wiley.

Nosil, P. (2012). *Ecological Speciation*. Oxford University Press.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology 29*(1), 59–75.

Pawitan, Y. (2001). *In All Likelihood*. Oxford University Press.

Pinho, C. and J. Hey (2010). Divergence with gene flow: Models and data. *Annual Review of Ecology, Evolution, and Systematics 41*(1), 215–230.

Powell, J. R. (1997). *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics 155*(2), 945–959.

Rosenberg, N. A. and M. W. Feldman (2002). The relationship between coalescence times and population divergence times. *Modern Developments in Theoretical Population Genetics*, 130–164.

Russell, S. T. (2003). Evolution of intrinsic post-zygotic reproductive isolation in fish. In *Annales Zoologici Fennici*, pp. 321–329. JSTOR.

Rykena, S. (2001). Experimental hybridization in green lizards (lacerta s. str.), a tool to study species boundaries. *Mertensiella 13*, 78–88.

Sánchez-Guillén, R., A. Córdoba-Aguilar, A. Cordero-Rivera, and M. Wellenreuther (2014). Genetic divergence predicts reproductive isolation in damselflies. *Journal of Evolutionary Biology 27*(1), 76–87.

Schierup, M. H. and J. Hein (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics 156*(2), 879–891.

Schiffels, S. and R. Durbin (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics 46*(8), 919–925.

Self, S. G. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(398), 605–610.

Servedio, M. R. (2000). Reinforcement and the genetics of nonrandom mating. *Evolution 54*(1), 21–29.

Silvapulle, M. J. and P. K. Sen (2011). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons.

Steinrücken, M., J. A. Kamm, and Y. S. Song (2015). Inference of complex population histories using whole-genome sequences from multiple populations. *bioRxiv*. URL:*biorxiv.org/content/early/2015/09/16/026591*.

Strang, G. (2009). *Introduction to Linear Algebra*. Wellesley-Cambridge Press.

Strasburg, J. L. and L. H. Rieseberg (2010). How robust are "isolation with migration" analyses to violations of the IM model? A simulation study. *Molecular Biology and Evolution 27*(2), 297–310.

Takahata, N. (1988, 12). The coalescent in two partially isolated diffusion populations. *Genetics Research 52*, 213–222.

Telschow, A., J. Engelstädter, N. Yamamura, P. Hammerstein, and G. Hurst (2006). Asymmetric gene flow and constraints on adaptation caused by sex ratio distorters. *Journal of Evolutionary Biology 19*(3), 869–878.

Teshima, K. M. and F. Tajima (2002). The effect of migration during the divergence. *Theoretical Population Biology 62*(1), 81–95.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis 92*(1), 1–28.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5–42.

Wakeley, J. (1996a). Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology 49*(3), 369 – 386.

Wakeley, J. (1996b). Pairwise differences under a general model of population subdivision. *Journal of Genetics 75*(1), 81–89.

Wakeley, J. (2009). *Coalescent Theory - An Introduction*. Roberts and Company Publishers.

Wakeley, J. (2010). Natural selection and coalescent theory. In M. Bell, D. Futuyma, W. Eanes, and J. Levinton (Eds.), *Evolution Since Darwin: the First 150 Years*. Sinauer Associates.

Wakeley, J. and J. Hey (1997). Estimating ancestral population parameters. *Genetics 145*(3), 847–855.

Wang, Y. and J. Hey (2010). Estimating divergence parameters with small samples from a large number of loci. *Genetics 184*(2), 363–379.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology 7*(2), 256–276.

White, H. (1996). *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Wilkinson-Herbots, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology 37*, 535–585.

Wilkinson-Herbots, H. M. (2008). The distribution of the coalescence time and the number of pairwise nucleotide differences in the isolation with migration model. *Theoretical Population Biology 73*(2), 277–288.

Wilkinson-Herbots, H. M. (2012). The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theoretical Population Biology 82*(2), 92–108.

Wilkinson-Herbots, H. M. (2015). A fast method to estimate speciation parameters in a model of isolation with an initial period of gene flow and to test alternative evolutionary scenarios. *ArXiv e-prints*. URL: *http://arxiv.org/abs/1511.05478* .

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology 53*(5), 821–841.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics 162*(4), 1811–1823.

Zhu, T. and Z. Yang (2012). Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution 29*(10), 3131–3142.