ABSTRACT

XU, TANCHUMIN. Advances in Causal Inference and the Study of Interlocus Gene Conversion. (Under the direction of Jeffrey Thorne).

Here we consider two distinct topics, the estimation of average causal effects, and the evolutionary impact of interlocus gene conversion, neither of which is fully treated in existing literature. The identification and characterization of causal effects is a central goal of the discipline of statistics. Our contribution to causal inference is detailed in a thesis chapter introducing a new estimator for causal effects that combines desirable features of two previously proposed estimators and a novel approach to finding tuning parameters in cross validation. Interlocus gene conversion is not commonly treated as a type of genetic mutation that affects duplicated DNA sequences and that has typically been ignored when the evolutionary origins of genetic variation have been considered. Our contributions regarding interlocus gene conversion are detailed in two thesis chapters. One chapter examines duplicated teleost genes to contrast the relative rates of nonsynonymous changes that homogenize and that do not homogenize the amino acids at corresponding positions in different paralogs. The other chapter examines the relationship between paralog divergence and the rate of evolutionary changes that are attributable to interlocus gene conversion.

Advances in Causal Inference and the Study of Interlocus Gene Conversion

by
Tanchumin Xu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics and Statistics (Co-Major)

Raleigh, North Carolina
2023

APPROVED BY:

Shu Yang	Brian Wiegmann
Gavin Conant	Jeffrey Thorne Chair of Advisory Committee

BIOGRAPHY

The author is from Chongqing, China and attended Bashu Middle School for a duration of six years. In June 2018, he successfully earned a Bachelor of Science degree in Statistics from the Mathematics School at Beijing Normal University. Subsequently, in August 2018, the author embarked on a co-major Ph.D. program in Bioinformatics and Statistics, guided by Dr. Jeffrey Thorne and Dr. Shu Yang at North Carolina State University. The primary focus of his research revolves around the study of molecular evolution and causal inference.

ACKNOWLEDGEMENTS

Thanks to my main advisors: Dr. Jeffrey Thorne and Dr. Shu Yang, who gave me support, help, and kindness. Thanks to my family members for your unlimited love. Thanks to my "informal" tutors: Dr. Hirohisa Kishino and Dr. Xiang Ji. Thanks to my committee members: Dr. Brian Wiegmann and Dr. Gavin Conant. Thanks to my collaborators: Dr. Yunshu Zhang and Yixuan Yang. Thanks to all BRC and statistics faculty members and staff who helped me. Thanks to all my teachers, friends, and classmates from Bashu middle school, Beijing Normal University, North Carolina State University, and other groups.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Background

Statistics is the science of data collection, modeling, and analysis. It is a discipline that has drawn more attention in recent years due to the advent of computers and the information age (Hastie et al. 2009). Cutting-edge technology, such as artificial intelligence and machine learning, can perform previously challenging tasks and can make use of exceptionally large data sets.

Statistical and computational problems in biology and genetics have created bioinformatics, which tries to implement sophisticated statistical models to answer basic biological

questions. One topic in bioinformatics concerns how species evolved from their common ancestors. In *On the Origin of Species*, Charles Darwin first proposed the evolutionary theory that life's diversity arose through common descent via a branching pattern of evolution (Darwin 1859). However, Darwin was hampered by a paucity of data, the unavailability of quantitative technologies, and ignorance of the mechanisms of genetic inheritance. All of the limitations faced by Darwin are being overcome. For example, advanced sequencing techniques such as short-read sequencing methods (Voelkerding et al. 2009) and whole-genome sequencing facilitate the collection of rich data sets for studying evolution at the molecular level. Hence, combining new data with novel inference procedures advances the study of evolutionary processes and evolutionary history.

Understanding the evolutionary history of species can illuminate diverse scientific questions. For example, why were there species explosions and mass extinctions? Meanwhile, evolutionary medicine can help us understand the mechanisms of cancer and autoimmune disease. As another example, the COVID-19 pandemic has arguably been the most important event in recent human history, and it falls squarely into the topic of virus evolution.

Another critical question in the big data era is understanding causality in the real world. One challenge in forecasting directional machine learning algorithms is that while combining sophisticated models (such as deep neural networks) with vast datasets can result in precise predictions, it may not provide much insight into the underlying mechanisms. Two important recent advances are having a big impact on methods for investigating causality. One advance came from Rubin (1976), who introduced the idea of imputing potential outcomes to generate a quasi-random experiment to evaluate causal effects. Meanwhile, Pearl (1988) has shown how to do causal reasoning and thereby study the question of "why" via probability graphs.

Causal inference approaches are heavily used in pharmacy to understand new medicines or therapies. To avoid bias, economists evaluate the outcome of a policy or measurement

with causality models. Also, causal inference provides a set of tools to help optimize policies in Reinforcement Learning (Zhang 2020).

This chapter briefly overviews the basic probabilistic models used in molecular evolution. I also review a model-based quantitative method for studying the evolutionary impact of interlocus gene conversion (IGC). In addition, I introduce the fundamental approaches of causal inference technologies for investigating causal effects.

1.2 Phylogenetic trees

A phylogenetic tree describes the evolutionary relationships between species or genes. From a mathematical perspective, phylogenies can be interpreted as acyclic graphs. From a statistical perspective, phylogenies can be viewed as representing the correlation structure in data that is attributable to common ancestry. Yang (2006) and Felsenstein and Felenstein (2004) have published excellent books that introduce statistical approaches for studying phylogenies and molecular evolution.

Phylogenetic data sets mostly consist of DNA, RNA, or protein sequence data. The focus of this thesis is mainly on DNA sequence data. DNA sequences carry inherited genetic information with individual sequence positions that are each occupied by one of four nucleotide types: A (adenine), C (cytosine), G (guanine), and T (thymine). In protein-coding DNA, a codon is a sequence of three consecutive nucleotides that either encodes a particular amino acid or a "stop" codon that signals the termination of the amino acid chain that forms a protein sequence. Although there is some variation across the tree of life of the genetic code that converts codon triplets to amino acids, this thesis concentrates on analyses of protein-coding genes that rely on the "universal" or standard genetic code (Brown 2012).

To represent potential functional, structural, or evolutionary correspondence among positions in molecular sequences, homologous sequences are conventionally organized

into a matrix format with different sequences in different rows (Gagniuc 2021). Positions in different sequences that are hypothesized to correspond are placed in the same column of the matrix. In this thesis, positions from different sequences that are in the same column are assumed to have evolutionary correspondence. When a sequence is hypothesized to lack a corresponding residue, a gap is placed in the column. When residue types in the same column of aligned DNA sequences differ, the difference is assumed to result from a nucleotide substitution. When residue types in the same column of aligned protein sequences differ, the difference is assumed to result from an amino acid replacement.

Because gaps and nucleotide substitutions (or amino acid replacements) are attributable to biological phenomena, it is important to consider their potential underlying causes. Gaps in sequence alignments can be due to insertion mutations or deletion mutations. In addition, alignment gaps can represent situations where not all sequence information has been included in a data set. For example, sometimes an alignment gap is attributable simply to the fact that some portion of a DNA sequence was not experimentally determined. Alignment uncertainty is conventionally disregarded during evolutionary inference. However, neglecting alignment uncertainty can sometimes be problematic (Wong et al. 2008). Fortunately, promising (albeit computationally demanding) evolutionary inference procedures are being developed (Redelings and Suchard 2005; Bradley et al. 2009). In this thesis, we adopt the conventional approach of ignoring alignment uncertainty.

In the context of evolutionary biology, a node represents a specific point in time when a common ancestor split into multiple descendant lineages (the branches above the node), resulting in a divergence of ancestral lineages (the branch below the node). There are two types of nodes in a phylogenetic tree. The “leaves” of a phylogenetic tree are also referred to as tip nodes or terminal nodes. Depending on the biological context, the tip nodes may represent extant living species or sampled genes or genomes, but the tips also may represent fossils or the ending of now-extinct lineages (Kapli et al. 2020). The internal nodes of a

phylogeny typically represent unobserved common ancestors of different tips.

Phylogeny can be either rooted or unrooted. In a rooted phylogeny, the root node represents all tip nodes' most recent common ancestor. With a rooted tree, the time directionality of every branch is known.

An unrooted tree does not indicate the root node on the tree. The position of the root node is often difficult to infer accurately solely from the information in a molecular sequence data set (e.g., Swofford 1996). In fact, the position of the root node is not statistically identifiable when some of the most widely-used models of sequence change are employed along with conventional phylogeny inference assumptions (Kinene et al. 2016). Even though the position of the root node can be challenging to infer from molecular sequence data, determining the time directionality on branches of an unrooted tree is often desirable. This is often done by employing prior knowledge to interpret an unrooted tree that is inferred from molecular sequence data.

Specifically, biologists often have information about evolutionary relationships external to the molecular sequence data. For example, it may be clear from other information sources that the species or sequences in a data set can be subdivided into two mutually exclusive groups of tips on a tree. The biologist may be confident *a priori* that the sequences from one group of tips are all more closely related to each other than they are to any of the sequences represented by other tips in the data set. This group of comparatively closely-related tips is referred to as the ingroup. All tips that do not belong to the ingroup are referred to as being in the outgroup.

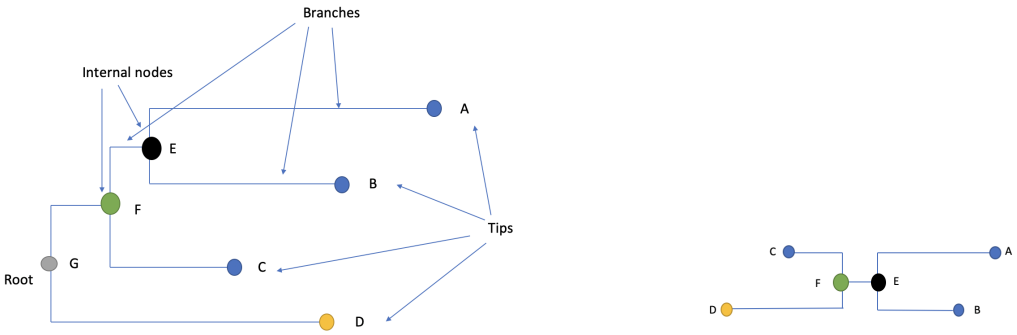
An unrooted tree should have a branch that divides the ingroup sequences from the outgroup sequences. The node at one end of the branch would represent the most recent common ancestor of the ingroup sequences. While the ingroup's most recent common ancestral node does not represent the root of the entire tree, it is sometimes referred to as the ingroup root. The prior knowledge that separates ingroup from outgroup sequences is

often employed in phylogenetic analyses to root the ingroup and assign time directionality to all branches subsequent to the ingroup root. An example of how an outgroup can be used to root an ingroup is employed in Figure 1.1. The analyses in this thesis will use an outgroup to root the ingroup.

Sometimes, the amount of evolution or the time duration of each branch on a phylogeny is of interest. These are referred to as branch lengths. Branch lengths can be measured in chronological time units such as days when a phylogeny relates quickly-evolving and closely-related genomes (e.g., retroviral isolates obtained from an infected patient), or chronological time units such as tens of millions of years when a phylogeny of slowly-evolving species is being considered. Especially when a phylogeny represents intraspecific relationships, the branch lengths may be measured in terms of the number of generations. Often, the branch lengths inferred from molecular sequence data will have units that represent expected amounts of evolution. For models of nucleotide substitution that consider independent change among positions in DNA sequences, branch lengths are typically measured in terms of the expected number of nucleotide substitutions per site. For protein-coding DNA sequence evolution models, the branch lengths are usually measured in terms of the expected number of nucleotide substitutions per codon (i.e., per nucleotide triplet).

1.3 Substitution matrix

Probabilistic models of sequence change can be classified as discrete-state continuous-time Markov processes. The usual situation is to have each sequence position or codon triplet change independently of all other sequence positions or codon triplets but to have all of the independently evolving units change on a shared tree. The instantaneous rates of change between discrete states of an independently evolving sequence position or codon are specified by a substitution matrix Q. This matrix represents the rates at which nucleotides



(a) Rooted tree

(b) Unrooted tree

Figure 1.1: Phylogenetic trees: A rooted tree and its corresponding unrooted tree. (a): In this unrooted tree, the root node (i.e., the most recent common ancestor of all tips is the gray-colored node that is labeled "G". (b): This tree displays the unrooted version of the tree in Part (a). If the yellow-colored node is considered the outgroup and the blue-colored nodes are considered the ingroup, then the green-colored node labeled "F" must be the most recent common ancestor of the ingroup nodes.

(or codon states) at a specific site are replaced by other nucleotide types (or codon states) (Yang 2006).

Most substitution models consider the evolutionary consequences of point mutations. Point mutations result in the nucleotide at a particular sequence being replaced by a nucleotide of a different type. A nucleotide substitution corresponds to a point mutation that eventually becomes fixed (i.e., becomes the common ancestor of all members of a population). While the actual process of mutation and subsequent fixation can take substantial chronological time, nucleotide substitutions are conventionally modeled as instantaneous events when interspecific evolution is being considered.

Assume that each site or codon position evolves independently and identically on an evolutionary tree that is shared among sites or codon positions. We represent the state of a site or codon position at the time (or amount of evolution) t as $X(t)$. Let the entry q_{ij} denote the rate of change from i to j. More specifically, the instantaneous rate q_{ij} for $i \neq j$

is $\lim_{\delta \to 0} P\left\{X(t+\delta) = j | X(t) = i\right\}/\delta$, where $i, j \in \{A, C, G, T\}$ for a nucleotide substitution process. The diagonal elements of Q are $q_{ii} = -\Sigma_{j,i \neq j} q_{ij}$. Therefore, the sum of the row in the matrix Q equals zero. The chain's stationary distribution can be written as a vector denoted by π where each entry π_i represents the stationary (i.e., equilibrium) probability of state i with $\pi Q = \pi$.

Often, models of sequence change are constructed so as to satisfy the time reversibility property (i.e., $\pi_i Q_{ij} = \pi_j Q_{ji}$ for all i and j). When models have the time reversibility property, the position on a phylogeny of the most recent common ancestor is not statistically identifiable. Time reversibility is desirable in that it results in models with fewer free parameters. Also, time reversibility can slightly enhance computational tractability. However, models that violate the time-reversibility assumption often provide improved fits to real data (Sumner et al. 2012).

To obtain the transition probability $p_{ij}(t)$ (i.e., the probability that $X(t) = j$ given Q, the amount of evolution t, and $X(0) = i$), the rate matrix Q is exponentiated by definition as $\exp(tQ) = I + tQ + \frac{(tQ)^2}{2} + ...$, so that

$$p_{ij}(t) = \exp(tQ)_{ij}.$$

This transition probability solves the differential equations $Q = dP(t)/dt$, where $P(t)$ is the transition probability matrix conditioned on time t. Algorithms for numerically computing transition probabilities are detailed in Yang (2007).

1.3.1 DNA substitution

The simplest nucleotide substitution model is the Jukes-Cantor model (JC) (Jukes and Cantor 1969). The JC model assumes all sequence positions evolve independently and identically. The JC model also has equal rates for all possible nucleotide substitutions. This

assumption yields equal frequencies of the four nucleotides at stationarity so that $\pi_i = 0.25$ for $i \in \{A, C, G, T\}$. The associated rate matrix is

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}.$$

While an attractive feature of the JC model is its simplicity, sequence change is substantially more complicated than the JC model. One improvement upon the JC model is to differentiate between the nucleotide types that are purines (Adenine and Guanine) and pyrimidines (Cytosine and Thymine). Changes between the two purines or changes between the two pyrimidines are known as transitions. Changes involving a purine and a pyrimidine are referred to as transversions. Because point mutations that transition tend to have higher rates than point mutations that are transversion (Stoltzfus and Norris 2016), the Kimura (1980) 2-parameter model generalizes the JC model by differentiating between transitions and transversion. A separate sort of improvement upon the JC model was introduced by Felsenstein (1981). The Felsenstein 1981 model generalized the JC model by allowing different nucleotide types to have different stationary probabilities. Hasegawa et al. (1985) combine the innovations of the Kimura 2-parameter and Felsenstein 1981 models by introducing this rate matrix parameterization:

$$Q = \begin{pmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{pmatrix}, \tag{1.1}$$

where . at each row presents the negative sum of the other three elements.

The general time reversible model (GTR) (Yang 1994) is a widely-used nucleotide substitution model of which all four aforementioned substitution models (Jukes and Cantor 1969; Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985) are special cases:

$$Q = \begin{pmatrix} \cdot & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \cdot & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \cdot & \pi_T \\ c\pi_A & ea\pi_C & \pi_G & \cdot \end{pmatrix}. \tag{1.2}$$

1.3.2 Codon substitution

Due to the triplet structure of the genetic codon, the evolution of sequence positions within a codon are correlated. Muse and Gaut (1994) (MG94) and Goldman and Yang (1994) (YN98) approached the evolution of these sequences by treating each codon triplet as a fundamental unit. The dimensions of the codon substitution matrix, excluding the three stop codons, are 61×61. Although codon codes can be mapped to corresponding amino acids, the genetic code is degenerate in that multiple codon triplets can specify the same amino acid. When a codon substitution does not change the corresponding amino acid, it is referred to as a synonymous change. Conversely, if the amino acid changes, it is classified as a nonsynonymous change. The MG94 model introduces an additional parameter, ω, to differentiate between synonymous and nonsynonymous rates.

The MG94 model assumes the instantaneous rate is zero if two or three codon positions are instantaneously changed. This is because the MG94 model describes the evolution of codon substitutions that arise as point mutations and then get fixed to result in a codon substitution. With the MG94 model, only changes involving a single codon position can have positive rates. The rates below correspond to codon substitutions that result in a nucleotide

of type h ($h \in \{A, C, G, T\}$) being introduced into one of the three codon positions. We have π_h denote what would be the stationary frequency of nucleotide type h if there were no natural selection ($0 \leq \pi_h \leq 1$, $\pi_A + \pi_C + \pi_G + \pi_T$). The MG94 model has rates of codon substitution from triplet i^* to j^* being

$$q_{i^* j^*} = \begin{cases} 0, \text{ at least two positions are different} \\ \pi_h, \text{synonymous change} \\ \omega\pi_h, \text{nonsynonymous change.} \end{cases} \tag{1.3}$$

In the study by Yang and Nielsen (2000), the parameter space was extended to improve the accuracy of the codon substitution process description. Unlike previous approaches that only considered nucleotide-level changes, YN98 introduced frequency parameters to account for variations between different codons. Specifically, they denoted the stationary distribution for codon j as π_{j^*} and used it to determine the rate at which codon states transitioned instantaneously from i^* to j^*:

$$q_{i^* j^*} = \begin{cases} 0, \text{ at least two positions are different} \\ \pi_{j^*}, \text{ synonymous and transversion change} \\ \kappa\pi_{j^*}, \text{synonymous and transition change} \\ \omega\pi_{j^*}, \text{nonsynonymous and transversion change} \\ \omega\kappa\pi_{j^*}, \text{nonsynonymous and transition change.} \end{cases} \tag{1.4}$$

1.4 Inference for phylogenetic tree

Utilizing probabilistic models of sequence change to account for unobserved historical molecular data when only the sequences (DNA, RNA, or protein sequences) at the tips of a tree are observed, the likelihood-based approach represents a breakthrough in the

inference of phylogeny trees. The pruning algorithm, developed as a part of this approach, can be employed to calculate the likelihood of a particular phylogenetic tree when only sequence data at the tips of the tree are observed (Felsenstein 1981). The pruning algorithm can be classified as a dynamic programming approach, and it allows likelihood-based statistical techniques to be applied to evolutionary inference.

The likelihood for an aligned data set is a product across the nucleotide or codon positions that are assumed to be independently evolving. This involves multiplying the columns of alignment sequences to obtain the joint likelihood. However, some processes such as gene conversion may cause dependencies between sites, which violates the assumption of independent evolution. If the assumption of independent evolution holds, the overall likelihood for an entire data set can be determined by calculating the independent-evolving units' probability.

The alignment matrix is denoted as $\mathbb{X} = X_i,\ i \in N$, where X_i is the i^{th} sequence. Conditioned on nodes $N = (A, B, C, D, E, F, R)$ in Figure 1.1.(a) with parameter θ, assuming stationarity to handle the root, the marginal likelihood $L(\theta)$ includes a summation over potentially unobserved nodes such as X_R, X_E, and X_F by

$$L(\theta) = \sum_{X_R} \sum_{X_E} \sum_{X_F} f(X_R, X_E, X_F, X_A, X_B, X_C, X_D \mid \theta). \tag{1.5}$$

Specifically, the branch lengths between root and F, root and D cannot be identified. The pruning algorithm assumes each branch is uncorrelated from the others, implying that the nucleotide substitution process follows the Markov property. The equation (1.5) can be transformed into:

$$\begin{aligned} L(\theta) = \sum_{X_R} f(X_R \mid \theta) f(X_D \mid X_R, \theta) \sum_{X_F} f(X_F \mid X_R, \theta) f(X_C \mid X_F, X_R, \theta) \\ \sum_{X_E} f(X_E \mid X_F, X_R, \theta) f(X_A \mid X_E, X_F, X_R, \theta) f(X_B \mid X_E, X_F, X_R, \theta). \end{aligned} \tag{1.6}$$

The algorithm begins by defining the transition probability and equilibrium probability using the model of the sequence change. Next, the conditional likelihood of the observed data (tips) is calculated based on the internal nodes associated with those tips. The likelihood is then computed sequentially by iteratively calculating the conditional likelihood.

1.5 Interlocus gene conversion

In an organism's genome, a point mutation refers to a genetic mutation that involves the replacement of a single nucleotide base in a DNA or RNA sequence with another single nucleotide base. Gene duplication is a significant process for creating new genetic material during molecular evolution. Gene duplication can occur due to various types of errors in DNA replication and repair machinery, as well as accidental capture by selfish genetic elements (Zhang 2003).

In evolutionary biology, the word homology is employed to mean shared common ancestry. Homologous sequences can be categorized according to whether they are orthologous or paralogous by considering the evolutionary tree that relates the sequences. The sequences are termed orthologous when the node representing two sequences' most recent common ancestor represents a speciation event. Alternatively, sequences are termed paralogous when the most recent common ancestral node represents a duplication event.

A duplication tree (Figure 1.2) is a specific phylogenetic tree that illustrates the homologous sequence of the multigene family. The Human A and Chimp A in Figure 1.2 are orthologs of genes in different species with shared ancestor. The Human A and Human B in Figure 1.2 are paralogs, duplicated genes in one species.

Following gene duplication, there are several possible fates for the two resulting paralogs. These possible fates are particularly of interest when the duplicated gene has an important

biological function. Sometimes, duplication of a functional gene may simply generate functional redundancy. Other times having two copies of a functionally important gene may be selectively deleterious (e.g. because too much gene product is made). One possibility is that one of the paralogs could subsequently be lost from the genome.

The loss of one paralog could be due to a deletion event. Alternatively, it's conceivable that a paralog may lose its gene function and become a pseudogene as a result of one or more point mutations. These mutations result in nonfunctional segments of DNA that mimic functional genes. Another possibility is that one paralog retains the original biological function, and the other paralog experiences point mutations or other mutations that yield a new function. This possibility is known as neofunctionalization (Ohno 2013).

A third possibility is referred to as subfunctionalization (Lynch and Force 2000). With subfunctionalization, the ancestral gene function eventually becomes subdivided between the two paralogs so that each paralog represents a gene that performs some but not all of the ancestral gene functions. For example, an ancestral gene may function in all tissues. If paralogs that result from a duplication of this gene become subfunctionalized, then one paralog might be expressed in one set of tissues and another paralog might be expressed in another set of tissues. Subfunctionalization can allow paralog function to become more specialized and functioned than ancestral gene function. Conant and Wolfe (2008) pointed out that "After preservation, duplicate genes continue to evolve, meaning that subfunctionalization can contribute to novelty simply by enabling duplicate genes to survive for long periods, increasing the chances of a neofunctionalizing mutation."

Gene conversion is a phenomenon where a contiguous stretch of DNA sequence from one gene copy overwrites and replaces the corresponding stretch from a homologous gene copy, resulting in the affected sequence region becoming identical between the donor and recipient copies. Because the result of a gene conversion event is to erase genetic information from the recipient copy, gene conversion is sometimes referred to as a unidirectional

information transfer. The most frequent gene conversion event is known as allelic gene conversion because it involves donor and recipient gene copies representing alleles from the same genetic locus in two different chromosomes. The molecular mechanism of allelic gene conversion has been carefully studied (Chen et al. 2007) as have its evolutionary consequences (Lorenz and Mpaulo 2022). Because allelic gene conversion is such a common and ubiquitous phenomenon among living organisms, it is often referred to as gene conversion rather than allelic gene conversion.

Interlocus gene conversion (IGC), also known as non-allelic gene conversion, is primarily initiated by DNA double-strand breaks (DSBs) and involves recombination between paralogous gene copies (Mehta and Haber 2014), either within the same chromosome (intrachromosomal) or between different chromosomes (interchromosomal). Although IGC mutations are believed to be substantially less common than allelic gene conversion events, IGC mutations are potentially important for genomic evolution because they transfer genetic information from one paralog to another.

My research focuses on studying IGC mutations that cause substitutions– i.e., mutations that survive. Figure 1.2 illustrates the evolutionary consequences of IGC events on a phylogeny. It depicts a single ancestral gene that is duplicated to yield a genome with two gene copies (Paralog A and Paralog B). After the duplication event, the two resulting paralogs are expected to diverge due to nucleotide substitutions that originate with point mutations. Therefore, the Paralog A and B sequences will likely differ at the time of the human-chimpanzee speciation. However, the Paralog A sequence from chimpanzees is expected to be identical to the Paralog A sequence from humans immediately after the speciation events. Similarly, the Paralog B sequences from humans and chimpanzees are expected to be identical immediately after the speciation event.

After the speciation event, the human and chimpanzee version of the paralogs can continue to diverge from each other due to nucleotide substitutions that originate with point

mutations. However, homogenization from IGC mutations can counteract the divergence between paralogs that is caused by point mutations. An IGC event in the human lineage is depicted in Figure 1.2. Paralog A is the IGC recipient and Paralog B is the IGC donor, so DNA sequence from a portion of Paralog B overwrites the corresponding stretch of DNA from Paralog A. This IGC event results in the Paralog A and Paralog B sequences being homogenized in the sequence region that experiences the IGC event.

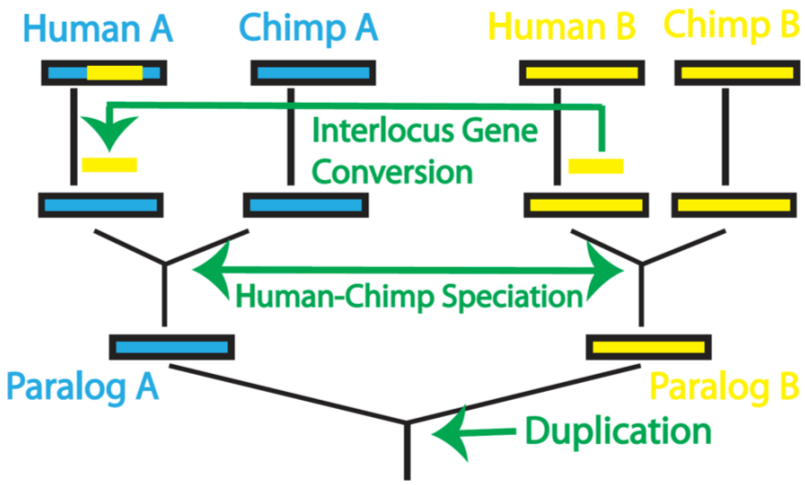


Figure 1.2: Duplicate tree.

Likelihood approach for IGC model

Ji and colleagues developed a likelihood-based approach to investigate the impact of IGC by incorporating the IGC parameter τ in the substitution matrix (Ji et al. 2016). Additionally, the model presupposes that every site, whether a nucleotide or codon, evolves independently of all others. However, observations from empirical research indicate that IGC tracts (i.e., the section of a sequence that experiences the same IGC event) can have lengths that are as much as 4 kilobases (Judd and Petes 1988). Genetic recombination results in the production of offspring that possess unique combinations of traits not present in either parent. High

recombination rates may reduce the length of the IGC tract. In other words, the substitutions due to IGC and because of recombination, not all sequence differences that are erased due to an IGC event, will be fixed. Ji et al. (2016) approach can be considered either a composite likelihood procedure or as inference with a model that has high recombination levels. Moreover, the model assumes that the rate of IGC is fixed and not dependent on the divergence of paralogs. However, subsequent chapters will introduce modifications to relax these assumptions.

When considering IGC, the pair of paralogs will affect one another. The HKY or the MG94 model can be applied to study substitutions arising from point mutation. However, to parameterize the dependence between multigene families, the substitution matrix considers the joint states i and i' at the corresponding codons of two paralogs. The entry $Q_{(i,i'),(j,j')}$ denotes the rate of change from paralog states (i,i') to (j,j'). When $\tau = 0$ there are no changes due to IGC and therefore the two paralogs evolve independently due to the codon substitutions that originate with point mutation and that have instantaneous rates that are described by the MG94 model rates of equation 1.3. The $61^2 * 61^2$ codon substitution matrix is sparse:

$$Q_{(i,i'),(j,j')} = \begin{cases} 0 & i \neq j, i' \neq j' \\ Q_{i,j} & i \neq j, i' = j', j \neq j' \\ Q_{i',j'} & i = j, i' \neq j', j \neq j' \\ Q_{i,j} + \nu & i \neq j, i' = j', j = j' \\ Q_{i',j'} + \nu & i = j, i' \neq j', j = j', \end{cases} \tag{1.7}$$

where $\nu = \omega * \tau$ for nonsynonymous change, $\nu = \tau$ for synonymous change. To differentiate between synonymous and nonsynonymous changes, the Ji's IGC requires ω to study how natural selection may affect nonsynonymous change. If $\omega < 1$, it suggests diversifying positive selection, which increases the likelihood of fixing synonymous changes. On the

other hand, if $\omega > 1$, it implies purifying selection, increasing the likelihood of fixing non-synonymous changes. A value of $\omega = 1$ indicates that the variation is fixed at the same rate, known as neutral protein evolution.

IGC is often disregarded in evolutionary studies. However, a research study carried out by (Ji et al. 2016) on 14 yeast data sets of duplicated ribosomal protein-coding genes from yeast species suggested that a large proportion of codon substitutions originated with IGC rather than point mutations. The results demonstrated that incorporating IGC greatly improved the fit of the probabilistic evolutionary models.

IGC model extensions–tract model

Ji and Thorne (2019), and Harpak et al. (2017) introduced novel approaches to address the impact of IGC on paralog evolution. Instead of assuming that each site changes independently of others but depends on the corresponding site in another paralog, their models relax this assumption and regard IGC as the average rate for sites experiencing it. Ji's model, the pair-site (PS) approach, explains the additional rate τ as the average rate for sites undergoing IGC. For k sites under IGC with rate τ, the tract length parameter k is distributed as the geometric distribution $p(1-p)^{k-1}$ with probability p, which represents the probability of a fixed IGC event. The PS approach modifies the HKY model by adding a position parameter r to better represent codon-based sequences undergoing IGC. The hidden Markov model is also appreciated for the tract model.

1.6 Causal inference basic setting

Throughout this thesis, we adopt the potential outcome framework. Assume $\{X_i, A_i, Y_i(0), Y_i(1)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \{X, A, Y(0), Y(1)\}$, and let A_i represent the binary treatment, X_i be the pre-treatment covariates and $Y_i(a)$ denote the potential outcome conditioned on the treatment

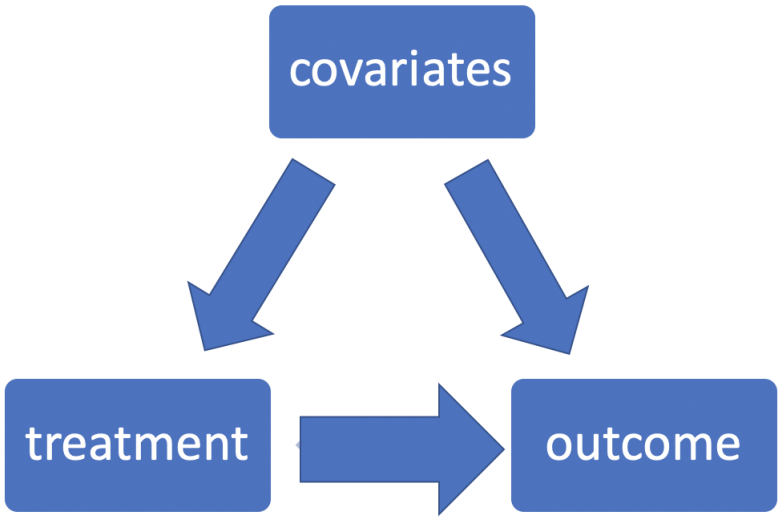


Figure 1.3: The relationship between treatment, covariates, and outcomes in the observational study.

a ($a = 0, 1$). If the treatment is independent of covariates (i.e., a randomized experimental study), statistical hypothesis tests such as the t-test can be applied to examine the effects.

However, observational studies (Figure 1.3) are more prevalent in sociology, economics, and epidemiology, where treatment may correlate with covariates associated with potential outcomes. The confounders are special variables that control both the treatment and outcomes. For instance, investigators might be interested in exploring the correlation between smoking and lung cancer and affirming that smokers are at a greater risk of developing lung cancer than those who don't smoke. Nonetheless, other variables that could affect an individual's decision to smoke may also play a role in causing lung cancer. The goal of reliable causal inference research is to remove the impact of these confounding factors and obtain an unbiased estimate of causal effects. Two commonly-used estimands: the average treatment effect (ATE) $\tau = E\{Y(1) - Y(0)\}$ and the average treatment effect on treated (ATT) $\tau^t = E\{Y(1) - Y(0) \mid A = 1\}$ are developed to study causal effects.

According to the Stable Unit Treatment Value Assumption (SUTVA) (Rosenbaum and Rubin 1983), the observed outcome Y_i is equal to $A_i Y_i(1)+(1-A_i)Y_i(0)$, in which the treatments assigned to one unit do not affect the potential outcomes of other units. The SUTVA is sometimes inappropriate. For example, in assessing the effectiveness of a new flu vaccine, this assumption is violated since individuals who receive the vaccine (i.e., the treatment group) may indirectly protect those in the control group. However, in the case of studying the effects of a new medicine for high blood pressure, the assignment of units to the treatment or control group is unlikely to affect the outcomes of the opposite group.

Let V be a generic variable, and let $e(V) = P(A = 1 \mid V)$ and $u_a(V) = E\{Y(a) \mid V\}$ be the mean functions for treatment and outcome, respectively. When V represents the pretreatment covariate X, $e(X)$ is referred to as the propensity score (Rosenbaum and Rubin 1983), and $u(X)$ is called the prognostic score or the outcomes (Hansen 2008).

Both the $e(X)$ and $u(X)$ can be modeled parametrically (e.g., linear regression, logistic regression), semiparametrically (e.g., general additive model), or nonparametrically (e.g., random forest, boosting).

To estimate causal effects, the unconfoundedness assumption is employed, which posits that $\{Y(0), Y(1)\} \perp A \mid X$, indicating that the treatment is independent of outcomes when conditioned on X. Therefore, by controlling for X, confounding variables associated with the treatments can be eliminated. Although this assumption cannot be tested, it remains fundamental in most observational studies even with a large sample size.

The propensity score, denoted as $e(X)$, refers to the likelihood of a unit belonging to the treatment group, while $1-e(X)$ indicates the likelihood of the unit being in the control group. Rosenbaum and Rubin (1983) shows that the propensity scores have the ability to equalize the two groups, ensuring a balance between them:

$$\{Y(0), Y(1)\} \perp A \mid e(X). \tag{1.8}$$

The meaning of equation (1.8) is that the treatment has no association with the outcomes once it is conditioned on $e(x)$. Therefore, achieving a balance between the groups by using propensity scores is adequate, instead of relying on individual covariates for balancing.

The Overlap assumption posits that the propensity score has two constants, c_1 and c_2, such that $0 < c_1 \leq e(X) \leq c_2 < 1$ is true almost all the time. Without this assumption, units with $e(x)$ equal to zero or one cannot be assigned to either the treatment or control group. In cases where a unit cannot be assigned to a specific group, extrapolation becomes necessary for that particular subpopulation. Researchers often visualize the distribution of covariates between treatment groups to identify any lack of overlap.

1.7 Basic approaches in causal inference

Let the treatment group ($A_i = 1$) have n_t units, and control group ($A_i = 0$) have n_c units. The naive estimator, the regression estimator, for causal effect is

$$\hat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{u}_1(X_i) - \hat{u}_0(X_i)\}, \tag{1.9}$$

where $\hat{u}_1(X_i)$ refers to the modeled estimate for the treatment group, and $\hat{u}_0(X_i)$ for pertains to the modeled estimate for the control group.

The consistency of causal estimands depends on how the propensity score or the outcome is specified. The single robust estimator is consistent with only one model, while the doubly robust estimator is consistent with both the propensity score and outcome models.

Matching

Matching is a technique used to improve the balance between treatment and control groups. This technique creates a "pseudo-outcome" (not listed in Table 1.1) by identifying

targets in the opposite group with similar covariates or other characteristics to those with corresponding missing values. The basic approach involves matching with replacement and setting the number of matches K to be constant at 1, conditional on X.

Table 1.1: Matching scenario.

Covariates	Treatment	Control
X_1	$Y_1(1)$	$\text{Missing}_1(1)$
$\vdots$	$\vdots$	$\vdots$
X_{n_t}	$Y_{n_t}(1)$	$\text{Missing}_{n_t}(1)$
X_{n_1+1}	$\text{Missing}_1(0)$	$Y_1(0)$
$\vdots$	$\vdots$	$\vdots$
$X_{n_t+n_c}$	$\text{Missing}_{n_c}(0)$	$Y_{n_c}(0)$

The procedure for substituting the unit in the opposite group involves selecting the observation with covariates that are most similar to those of the target unit. Typically, either the Mahalanobis or Euclidean distance metric is employed for this purpose. This technique is comparable to the nearest neighbor (NN) method:

$$\hat{Y}_i(1) = \begin{cases} I_{j\in J_V(i)} Y_j & \text{if } A_i = 0, \\ Y_i & \text{if } A_i = 1, \end{cases} ; \; \hat{Y}_i(0) = \begin{cases} Y_i & \text{if } A_i = 0, \\ I_{j\in J_V(i)} Y_j & \text{if } A_i = 1, \end{cases} \tag{1.10}$$

where $J_V(i)$ is the index set of the nearest neighbors for unit i in its opposite group. The naive matching estimator is

$$\hat{\tau}_{mat} = \frac{1}{n_t + n_c} \sum_{i=1}^{n_t+n_c} \hat{Y}_i(1) - \frac{1}{n_t + n_c} \sum_{i=1}^{n_t+n_c} \hat{Y}_i(0) \tag{1.11}$$

As the dimension of X grows, the effectiveness of the matching estimator diminishes. Hence, matching the entire set of covariates may not entirely eliminate confounding bias due to the curse of dimensionality. To overcome the limitations associated with X, the propensity score can be used as a substitute matching criterion, provided assumption 1.8 holds.

Weighting

The idea of weighting stems from survey sampling, where the goal is to ensure that survey samples accurately reflect the larger population. To achieve this, each unit in the sample is given a weight, which is calculated in a way that assigns larger weights to underrepresented units and smaller weights to overrepresented units. The weighting approach aims to approximate the characteristics of the broader population within the survey sample.

Inverse probability weighting (IPW) (Horvitz and Thompson 1952) is a method that utilizes the estimated propensity scores to calculate the inverse of the weights, namely $1/\hat{e}(X)$ and $1/(1-\hat{e}(X))$:

$$\hat{\tau}_{\mathrm{IPW}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}} \frac{\mathrm{Y}_i A_i}{\hat{e}(X_i)} - \frac{1}{\mathrm{n}} \sum_{\mathrm{i}} \frac{\mathrm{Y}_i (1-A_i)}{1-\hat{e}(X_i)}. \tag{1.12}$$

The IPW estimator has a property known as single robustness, which implies that the accuracy of the estimated treatment effect $\hat{\tau}_{IPW}$ heavily depends on the choice of the propensity model. To incorporate outcome information and enhance the robustness of the estimator, the augmented IPW (AIPW) estimator was developed, which has the property of double robustness. As a result, the AIPW estimator can still provide consistent estimates even if either the propensity score or outcome model is not properly specified:

$$\hat{\tau}_{\mathrm{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{(1-A_i) Y_i}{1-\hat{e}(X_i)} - \frac{\{A_i - \hat{e}(X_i)\}}{\hat{e}(X_i)} \hat{u}_1(X_i) + \frac{\{A_i - \hat{e}(X_i)\}}{1-\hat{e}(X_i)} \hat{u}_0(X_i) \right]. \tag{1.13}$$

Extensions

Subclassifying (Imbens 2000; Yang et al. 2016) involves stratifying the propensity scores into different sets and then comparing the treatment and control units within each set to estimate the treatment effects. By subclassifying the propensity score, several sets of units with similar scores are constructed, where the covariates are assumed to be independent of treatment. This technique involves partitioning the range of the propensity scores into K blocks and selecting boundary points using an iterative procedure. The block-specific average effect of treatment is estimated for each set, and the overall ATE is obtained by summarizing the effects of the different sets.

Causal inference encompasses a broad range of topics and research areas. The instrumental variables are variables only correlated with treatments. Instrumental variables can be used to examine intention-to-tdeatle@ncsu.edutreat analysis in noncompliance's presence of noncompliance (Angrist et al. 1996). Difference-in-differences approaches measure two differences between group means in a specific manner to estimate effects. Trimming approaches (Yang and Ding 2018) discard the extreme propensity scores to minimize large variations for causal effects. Random forest, originally a machine learning technique, has been adapted to analyze heterogeneous effects in causal inference (Athey et al. 2019).

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests.

Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., and Pachter, L. (2009). Fast statistical alignment. *PLoS computational biology*, 5(5):e1000392.

Brown, T. A. (2012). *Introduction to genetics: a molecular approach*. Garland Science.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950.

Darwin, C. (1859). *The origin of species by means of natural selection*, volume 247. EA Weeks.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. and Felenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Gagniuc, P. A. (2021). *Algorithms in Bioinformatics: Theory and Implementation*. John Wiley & Sons, Incorporated.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.

Harpak, A., Lan, X., Gao, Z., and Pritchard, J. K. (2017). Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proceedings of the National Academy of Sciences*, 114(48):12779–12784.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Ji, X., Griffing, A., and Thorne, J. L. (2016). A phylogenetic approach finds abundant interlocus gene conversion in yeast. *Molecular biology and evolution*, 33(9):2469–2476.

Ji, X. and Thorne, J. L. (2019). A phylogenetic approach disentangles interlocus gene conversion tract length and initiation rate. *arXiv preprint arXiv:1908.08608*.

Judd, S. R. and Petes, T. (1988). Physical lengths of meiotic and mitotic gene conversion tracts in saccharomyces cerevisiae. *Genetics*, 118(3):401–410.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism. Academic Press, New York*, pages 21–123.

Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

Kinene, T., Wainaina, J., Maina, S., and Boykin, L. (2016). Rooting trees, methods for. *Encyclopedia of Evolutionary Biology*, page 489.

Lorenz, A. and Mpaulo, S. J. (2022). Gene conversion: a non-mendelian process integral to meiotic recombination. *Heredity*, 129(1):56–63.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473.

Mehta, A. and Haber, J. E. (2014). Sources of dna double-strand breaks and models of recombinational dna repair. *Cold Spring Harbor perspectives in biology*, 6(9):a016428.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724.

Ohno, S. (2013). *Evolution by gene duplication*. Springer Science & Business Media.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Redelings, B. D. and Suchard, M. A. (2005). Joint bayesian estimation of alignment and phylogeny. *Systematic biology*, 54(3):401–418.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Stoltzfus, A. and Norris, R. W. (2016). On the causes of evolutionary transition: transversion bias. *Molecular biology and evolution*, 33(3):595–602.

Sumner, J. G., Jarvis, P. D., Fernández-Sánchez, J., Kaine, B. T., Woodhams, M. D., and Holland, B. R. (2012). Is the general time-reversible model bad for molecular phylogenetics? *Systematic biology*, 61(6):1069–1074.

Swofford, D. L. (1996). Phylogenic inference. *Molecular systematic*.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4):641–658.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39:105–111.

Yang, Z. (2006). *Computational molecular evolution*. OUP Oxford.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591.

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution*, 17(1):32–43.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18(6):292–298.

Zhang, J. (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR.

CHAPTER

2

AUGMENTED MATCH WEIGHTED ESTIMATORS FOR AVERAGE TREATMENT EFFECTS

2.1 Abstract

Propensity score matching (PSM) and augmented inverse propensity weighting (AIPW) are used in observational studies to estimate causal effects. The AIPW estimator is doubly robust and locally efficient but can be unstable when the propensity scores are close to zero

or one. PSM circumvents the instability of propensity score weighting but it hinges on the correctness of the propensity score model and cannot attain the semiparametric efficiency bound. The fixed number of matches, K, renders PSM nonsmooth and thus invalidates standard bootstrap inference. This chapter presents novel augmented match weighted (AMW) estimators that combine the advantages of matching and weighting estimators. AMW adheres to the form of AIPW for its double robustness and local efficiency but it mitigates the instability. We replace inverse propensity weights with matching weights resulting from PSM with unfixed K. Meanwhile, we propose a new cross-validation procedure to select K that minimizes the mean squared error anchored around an unbiased estimator of the causal estimand. We derive the limiting distribution showing AMW estimators enjoy the double robustness property and can achieve the semiparametric efficiency bound if both nuisance models are correct. As a byproduct of unfixed K which smooths AMW estimators, nonparametric bootstrap can be adopted for variance estimation. Furthermore, simulation and real-data applications support that the AMW estimators are stable and their variances can be obtained by nonparametric bootstrap.

2.2 Introduction

Causal inference has garnered significant attention in recent years due to its potential applications in various fields, including sociology, economics, and medicine. Causal inference aims to identify the cause-and-effect relationships between variables of interest. In observational studies, it is often crucial to estimate the average treatment effect (ATE) and the average treatment effect on the treated (ATT). However, this can be challenging since treatment assignment may be correlated with various covariates associated with potential outcomes. Moreover, due to the non-random assignment of treatments, there may be an imbalance in both treatment and control groups (Imbens and Rubin 2015).

These challenges arise because researchers in observational studies have no control over the treatment assignment to units. As a result, causal inference methods are employed to adjust for confounding variables and estimate the causal effects of treatments.

Matching (Cochran and Rubin 1973; Rubin 1973) has a rich history for conducting causal inference in observational studies. These methods involve pairing each treated unit to untreated units that have similar covariates, aiming to remove the confounding bias and balance the two groups (Stuart 2010). Intuitively, the nearest neighbor (NN) matching approach matches based on covariates X, with a fixed number of matches, $K = 1$. Some approaches also allow K to change with varying sample sizes. Heckman et al. (1997) applied kernel matching and local linear kernel matching estimators conditioned on covariates X, which reduced the confounding bias substantially. Note that by allowing K to diverge with different samples, the K-nearest neighbor (KNN) matching estimator with covariates X can be doubly robust and semiparametrically efficient. However, due to the curse of dimensionality, matching on the full set of covariates may not entirely remove the confounding bias in high-dimensional settings. To overcome this challenge, several dimension reduction techniques have been developed. Rosenbaum and Rubin (1983) illustrated the central role of the propensity score to balance the two groups and hence proposed propensity score matching (PSM) to estimate causal effects. PSM with a fixed number of matches K has been extensively studied (Abadie and Imbens 2006, 2012, 2016; Stuart 2010). However, PSM is only singly robust, as it yields consistent estimators only with the correct propensity score models. Leacy and Stuart (2014) and Yang and Zhang (2020) proposed doubly robust matching estimators and explored their asymptotic properties. These estimators are consistent for causal effects if either a model for the propensity score or a model for the outcomes is correctly specified. However, the matching estimators with a fixed K are inefficient, and the nonparametric bootstrap is not amenable to estimating their corresponding variances. The failure of bootstrap is because the matching estimators with a fixed K lack smoothness,

and the distribution of the number of times when each unit is utilized as a match cannot be replicated by the bootstrap procedure(Abadie and Imbens 2008). To address this issue, Otsu and Rai (2017) developed a weighted bootstrap approach by resampling based on certain linear forms of the estimators when matching on X with a fixed number of matches. They demonstrated the validity of their approach, but it did not extend to the PSM estimator.

In addition to matching, several other ways exist to obtain valid estimations for treatment effects. The inverse probability of weighting (IPW) (Imbens and Rubin 2015) estimator addresses the confounding issue by assigning the inverse of the propensity scores as weights to all units. But it is only singly robust and may be unstable even when the propensity score model is correctly specified. Besides, Robins et al. (1994); Bang and Robins (2005); Cao et al. (2009) and Cao et al. (2009) provided doubly robust weighting estimators, i.e., the augmented IPW (AIPW) estimators, which achieve the semiparametric efficiency bound (Hahn 1998; Tsiatis 2006). However, such weighting estimators can still be unstable when the propensity scores are close to one or zero.

In this paper, we provide a comprehensive review of the PSM and AIPW estimators and revisit their strengths and weaknesses. Matching has advantages over weighting: first, and most significantly, matching does not involve the potentially unstable inverse of propensity scores (Frölich 2004) and second, matching is intuitively appealing to replicate a randomized experiment (Heckman et al. 1997; Dehejia and Wahba 2002; Rubin 2006; Stuart 2010). However, AIPW is easier to implement and can achieve the semiparametric efficiency bound.

Motivated by the distinguishing features of PSM and AIPW, we propose a new type of estimator called the augmented match weighted (AMW) estimator, which is stable, efficient, and doubly robust for estimating the ATE and ATT. The AMW estimator introduces a fresh perspective to reveal the connections between matching and weighting. On the one hand, they can be viewed as an augmentation of PSM estimators with outcome models. On the

other hand, it is connected to the AIPW estimators by replacing inverse propensity score weights with matching weights from PSM. Hence, the AMW estimators can be seen as a combination of the PSM and AIPW estimators. In addition, the AMW estimator has a smooth property, as the number of matches is unfixed, and the standard bootstrap is valid for estimating the variance. The AMW estimator has three key advantages. Firstly, it is stable, unlike the weighting type estimators. The kernel estimations for weights are less extreme than the propensity scores from parametric models, making the AMW estimator more stable when propensity scores are extreme. Secondly, if both nuisance models are correct, the AMW estimator achieves the semiparametric efficiency bound. Thirdly, the AMW estimator enjoys the double robustness property, which means it is still consistent when either the propensity score model or outcome model is misspecified.

Besides, selecting the tuning parameter K is an important aspect of the AMW estimator. Cross-validation (CV) is commonly used for model selection in causal inference. Cui and Tchetgen (2019) minimized the CV-based pseudo-risk for doubly robust, semiparametric estimating functions from the nuisance models. CV can also be used to choose different estimators for conditional treatment effects (Rolling and Yang 2014). Brookhart and Van Der Laan (2006) and Rothenhäusler (2020) developed CV approaches to select tuning parameters based on benchmark estimators using baseline parameters. For example, Ju et al. (2019) optimized a closed-form mean squared error (MSE) based on the sum of the bias and variance terms to select cutoff parameters for the propensity score truncation problem where the IPW estimator truncated at $0th$ percentile is treated as the reference estimator. To select the unfixed K for the AMW estimator, we propose a new CV approach in practice. The idea is to treat the AMW estimator with $K = 1$ as an unbiased benchmark estimator and study the bias of the AMW estimator with candidate parameter K. The variance can be estimated by naive bootstrap, and we can minimize the MSE for an appropriate parameter K.

We present the theoretical properties in two steps. In the first step, we study their large sample distributions with known parameters under the KNN framework by Mack and Rosenblatt (1979) and Yang and Kim (2017). We derive their asymptotic linear expansions for AMW and show that they have approximately normal distributions. The asymptotic variances indicate that the AMW estimators attain the semiparametric efficiency bounds when the propensity score and outcome models are correct. In the second step, we extend the results of Andreou and Werker (2012) and Abadie and Imbens (2016) to derive the large sample properties of the AMW estimators when nuisance parameters are estimated. Then we approximate their limiting normal distributions based on the estimated scores, building on the works of Abadie and Imbens (2016) and Yang and Zhang (2020).

The article is organized as follows. Section 2.3 provides a basic setup and briefly reviews the matching and AIPW estimators to motivate the proposed estimators. Section 2.4 gives a basic framework and algorithms to construct the AMW estimators and corresponding variance estimators. In Section 2.5, main theories are established for the AMW estimators with both known and estimated scores. Section 2.6 reports simulation studies to investigate the performances of different estimators. Section 2.7 showcases the proposed approach through three real data applications. Section 2.8 concludes with a discussion.

2.3 A review of methodologies

2.3.1 Basic setup

We follow the potential outcomes framework throughout this paper. Assume $\{X_i, A_i, Y_i(0), Y_i(1)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \{X, A, Y(0), Y(1)\}$ in the observational study, and let A_i be the binary treatment, X_i be the pre-treatment covariates and $Y_i(a)$ be the potential outcome conditioned on the treatment a $(a = 0, 1)$. Given the standard Stable Unit Treatment Value Assumption

(Rosenbaum and Rubin 1983), the observed outcome is $Y_i = A_i\, Y_i(1) + (1 - A_i) Y_i(0)$. Since $\{X_i, A_i, Y_i(0), Y_i(1)\}_{i=1}^n$ are i.i.d, the observations $\{X_i, A_i, Y_i\}_{i=1}^n$ are also i.i.d. The sample size is denoted as n, while n_1 and n_0 represent the sizes of the treated and control subpopulation, respectively. This paper focus on two commonly-used estimands: the ATE $\tau = \mathbb{E}\{Y(1) - Y(0)\}$ and the ATT $\tau^t = \mathbb{E}\{Y(1) - Y(0) \mid A = 1\}$. To simplify the presentation, for a generic variable V, denote $u_a(V) = \mathbb{E}\{Y(a) \mid V\}$, $\sigma_a^2(V) = \mathbb{V}\{Y(a)|V\}$, $e(V) = \mathbb{P}(A = 1 \mid V)$, where $u_a(V)$ represents the mean function for the outcome, $\sigma_a^2(V)$ represents the variance function, and $e(V)$ denotes the mean function for the treatment. When V corresponds to the pre-treatment covariates X, $e(X)$ is commonly referred to as the propensity score.

The basic framework relies on several fundamental assumptions:

Assumption 1 *There exist two constants c_0 and c_1 such that $0 < c_0 \le e(X) \le c_1 < 1$ almost surely. $\{Y(0), Y(1)\} \perp A \mid X$.*

The first part for Assumption 1 implies that all units can be allocated to either the treatment or control group, indicating a considerable overlap in the distributions of pre-treatment variables. The second part is also referred to as strong ignorability of treatment assignment, suggesting that treatments and potential outcomes are independent given adequate covariates. While researchers cannot test this assumption, they can enhance the reliability of their inferences by collecting more covariates.

2.3.2 K-nearest neighbor matching

In order to explain the concept of AMW estimators, we will review the class of matching estimators. To fix ideas, we first consider the case of matching with replacement, where the number of matches to be fixed as K ($K \ge 1$). This approach is equivalent to the K-nearest neighbor (KNN) approach (Here, KNN is the conventional name, so we will use K in our context). Generally, it is recommended to use $K = 1$, namely the nearest neighbor (NN)

approach (Rubin 1973; Mack and Rosenblatt 1979) to gain fewer biases. The matching estimators with fixed K are one of the most commonly used estimators in observational studies, and their statistical properties have been extensively studied in a series of works (Abadie and Imbens 2006, 2011, 2016). Still, we will later propose an algorithm to specify the matching estimators with unfixed K in Section 2.4.2. Based on the matching variable V, let $\mathscr{J}_V(i)$ denote the set of indices for the nearest K neighbors of unit i in the opposite group. $Y_i(A_i)$ is observed, the counterfactual outcome $Y_i(1-A_i)$ is missing but can be estimated by averaging the observed outcomes of the K matched units in the corresponding group. Specifically, the potential outcomes for the unit i can be imputed as

$$\hat{Y}_i(1)=\begin{cases} K^{-1}\sum_{j\in\mathscr{J}_V(i)} Y_j & \text{if } A_i=0, \\ Y_i & \text{if } A_i=1, \end{cases}; \ \hat{Y}_i(0)=\begin{cases} Y_i & \text{if } A_i=0, \\ K^{-1}\sum_{j\in\mathscr{J}_V(i)} Y_j & \text{if } A_i=1. \end{cases}$$

Let $M_{V,i}=\sum_l I_{(i\in\mathscr{J}_V(l))}$ be the number of times that unit i is being matched to other units. We use the Euclidean distance for matching, $||\cdot||$, although our discussion can be applied to other distance measures. A simple matching estimator $\hat{\tau}^{(0)}_{\text{mat}}$ for ATE can be given:

$$\hat{\tau}^{(0)}_{\text{mat}}=\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i(1)-\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i(0)=\frac{1}{n}\sum_{i=1}^{n}(2A_i-1)(1+K^{-1}M_{V,i})\hat{Y}_i. \tag{2.1}$$

Then $\hat{\tau}^{t}_{\text{mat}}$ for ATT is

$$\hat{\tau}^{t,(0)}_{\text{mat}}=\frac{1}{n_1}\sum_{i=1}^{n}Y_iA_i-\frac{1}{n_1}\sum_{i=1}^{n}\hat{Y}_i(0)A_i=\frac{1}{n_1}\sum_{i=1}^{n}\{A_i-(1-A_i)K^{-1}M_{V,i}\}\hat{Y}_i. \tag{2.2}$$

Matching has the advantage of producing intuitive estimators. It is also straightforward to evaluate the matching performance by examining the balance of the covariates. A small difference between the treatment and control groups may indicate a successful matching process. Nevertheless, one of the drawbacks of matching is that the estimators can be

biased due to matching discrepancies. To illustrate, Abadie and Imbens (2006) presented the decomposition for $\hat{\tau}^{(0)}_{\text{mat}}$ to investigate asymptotic properties: $n^{1/2}(\hat{\tau}^{(0)}_{\text{mat}}-\tau)=B_n+D_n$,

$$\begin{aligned} B_n &= n^{-1/2}\sum_{i=1}^{n}(2A_i-1)[K^{-1}\sum_{l\in\mathscr{J}_V(i)}\left\{u_{1-A_i}(V_i)-u_{1-A_i}(V_l)\right\}], \\ D_n &= n^{-1/2}\sum_{i=1}^{n}[u_1(V_i)-u_0(V_i)-\tau+(2A_i-1)(1+K^{-1}M_{V,i})\left\{Y_i-u_{A_i}(V_i)\right\}]. \end{aligned} \tag{2.3}$$

The difference between $u_{1-A_i}(V_i)$ and $u_{1-A_i}(V_l)$ is responsible for the matching discrepancy. The term D_n has a mean of zero, while B_n contributes to the asymptotic bias of the matching estimator. Abadie and Imbens (2006) showed that the asymptotic bias B_n has an order of $O_{\mathbb{P}}(N^{1/2-1/d})$, where d represents the dimension of the matching variable V. For example, when matching is directly based on the pre-treatment covariates (i.e., $V=X$), the bias is considerable unless the dimension of X is one. A bias-corrected matching estimator is given as $\hat{\tau}_{\text{mat}}=\hat{\tau}^{(0)}_{\text{mat}}-\hat{B}_n$. The term $\hat{u}_a(V_i)$ can be obtained either parametrically (e.g., by a linear regression estimator), or nonparametrically. Equivalently, we update the imputed potential outcomes for unit i as

$$\widetilde{Y}_i(a)=\begin{cases} K^{-1}\sum_{j\in\mathscr{J}_V(i)}\left\{Y_j+\hat{u}_a(V_i)-\hat{u}_a\left(V_j\right)\right\} & \text{if } A_i\neq a, \\ Y_i & \text{if } A_i=a. \end{cases};$$

Then the bias-corrected matching estimator is

$$\hat{\tau}_{\text{mat}}=\frac{1}{n}\sum_{i=1}^{n}\{\widetilde{Y}_i(1)-\widetilde{Y}_i(0)\}. \tag{2.4}$$

However, correcting the matching bias by relying solely on the pre-treatment covariates X may not always remove all sources of confounding bias due to the curse of dimensionality. As a result, it is necessary to use sufficient statistics as the matching elements. Rosenbaum

and Rubin (1983) demonstrated the central role of the score $e(X)$ as a balancing score:

Lemma 1 *Under Assumptions 1,* $\{Y(1), Y(0)\} \perp A \mid e(X)$*, which implies that* $\tau = \mathbb{E}[\mathbb{E}\{Y \mid A = 1, e(X)\} - \mathbb{E}\{Y \mid A = 0, e(X)\}]$ *and* $\tau^t = \mathbb{E}[\mathbb{E}\{Y \mid A = 1, e(X)\} - \mathbb{E}\{Y \mid A = 0, e(X)\} | A = 1]$.

Lemma 1 suggests that the scores can reduce the dimension of the matching variables. When the lemma holds true, PSM controls for $e(X)$ instead of X to estimate causal effects τ and τ^t via matching the treatment and control subjects with similar scores. Under a correct model, the PSM estimator is consistent and can sufficiently eliminate biases associated with a non-random treatment assignment. However, PSM suffers from bias due to possible model misspecification, and it may be biased if the model is incorrect. This limitation motivates us to improve PSM's robustness by constructing our AMW estimator in Section 2.4.

Before proceeding, we would like to point out that the bias-corrected matching estimator, which is derived from the pre-treatment covariates X, can be rewritten in the following form:

$$\hat{\tau}_{\text{mat},X} = \hat{\tau}_{\text{reg}} + \frac{1}{n}\sum_{i=1}^{n}\left\{A_i\left(1 + \frac{M_{X,i}}{K}\right)\hat{R}_i - (1-A_i)\left(1 + \frac{M_{X,i}}{K}\right)\hat{R}_i\right\}, \tag{2.5}$$

where $\hat{\tau}_{\text{reg}} = n^{-1}\sum_{i=1}^{n}\{\hat{u}_1(X_i) - \hat{u}_0(X_i)\}$ is the outcome regression estimator of τ, and $\hat{R}_i = Y_i - \hat{u}_{A_i}(X_i)$ is the residual of unit i. Specifically, the bias-corrected matching estimator based on the score $e(X)$ can also be expressed similarly

$$\begin{aligned}\hat{\tau}_{\text{mat},ps} = &\frac{1}{n}\sum_{i=1}^{n}[\hat{u}_1\{e(X_i)\} - \hat{u}_0\{e(X_i)\}] \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left\{A_i\left(1 + \frac{M_{e(X),i}}{K}\right)\hat{R}_i - (1-A_i)\left(1 + \frac{M_{e(X),i}}{K}\right)\hat{R}_i\right\}.\end{aligned} \tag{2.6}$$

Note that the first term on the right-hand side is not equivalent to $\hat{\tau}_{\text{reg}}$. This is why PSM cannot benefit from the double robustness property. In the section 2.3.3, we will contrast this form with the AIPW estimator and utilize these concepts to build our AMW estimator.

2.3.3 Augmented inverse probability weighting (AIPW) estimator

To enhance the effectiveness and robustness of prior matching estimators, we aim to review some double robustness efficient estimators to gain insights. In practice, to address the weakness of unknown score and outcome models, researchers have been developing more robust estimators (Yang and Zhang 2020; Han and Wang 2013). Among them, the augmented inverse probability weighting (AIPW) estimator (Robins et al. 1994) is most commonly used because of its double robustness and semiparametric efficiency. Let $\hat{u}_a(X)$ and $\hat{e}(X)$ be some parametric estimators for the outcome $u_a(X)$ and propensity score $e(X)$. The AIPW estimator for the ATE is considered as follows and can be rewritten as:

$$\begin{aligned}\hat{\tau}_{\text{AIPW}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{(1-A_i)Y_i}{1-\hat{e}(X_i)} - \frac{\{A_i-\hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{u}_1(X_i) - \frac{\{A_i-\hat{e}(X_i)\}}{1-\hat{e}(X_i)}\hat{u}_0(X_i)\right] \\ &= \hat{\tau}_{\text{reg}} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i\hat{R}_i}{\hat{e}(X_i)} - \frac{(1-A_i)\hat{R}_i}{1-\hat{e}(X_i)}\right\}.\end{aligned} \tag{2.7}$$

Note that (2.7) bears resemblance to (2.5), as they both utilize the outcome regression term with some adjustments based on the residuals $\hat{R}_i$. This results in the doubly robust property. However, the AIPW estimator may be unstable when the inverse propensity scores take values near zero or one (Kang and Schafer 2007; Guo and Fraser 2014). Conversely, matching estimators avoid the limitation of extreme propensity scores by employing $1+M_X/K$ in lieu of the inverse probabilities. To conclude, we summarize:

- The matching estimator based on the pre-treatment covariates in (2.5) is robust to extreme propensity scores, but matching on X suffers from the curse of dimensionality;
- The PSM estimator with fixed $K=1$ in (2.6) is asymptotically unbiased in high dimensional settings and robust to extreme propensity scores, but it suffers from model misspecifications and is inefficient;

- The AIPW estimator in (2.7) is doubly robust and efficient, but the inverse probabilities suffer from extreme propensity scores.

Given these considerations, we are ready to leverage the respective advantages of the matching estimator and the AIPW estimator to construct the AMW estimator in the following section. Before proceeding, we also intend to rephrase the AIPW estimator for the ATT:

$$\hat{\tau}^t_{\text{AIPW}} = \hat{\tau}^t_{\text{reg}} + \frac{1}{n_1}\sum_{i=1}^{n}\left\{A_i\hat{R}_i - \frac{(1-A_i)\hat{e}(X_i)\hat{R}_i}{1-\hat{e}(X_i)}\right\}. \tag{2.8}$$

2.4 Proposed methodology

2.4.1 AMW estimators

Motivated by (2.5)-(2.7), we propose the AMW estimators based on the advantages of matching and weighting:

$$\hat{\tau}_{\text{AMW}} = \hat{\tau}_{\text{reg}} + \frac{1}{n}\sum_{i=1}^{n}\left[A_i\left\{1+\frac{M_{e(X),i}}{K}\right\}\hat{R}_i - (1-A_i)\left\{1+\frac{M_{e(X),i}}{K}\right\}\hat{R}_i\right], \tag{2.9}$$

$$\hat{\tau}^t_{\text{AMW}} = \hat{\tau}^t_{\text{reg}} + \frac{1}{n_1}\sum_{i=1}^{n}\left[A_i\hat{R}_i - (1-A_i)\left\{1+\frac{M_{e(X),i}}{K}\right\}\hat{R}_i\right]. \tag{2.10}$$

Kang and Schafer (2007) found that weighting type estimators tend to have high variability due to the denominator of the weights being close to zero. To address this issue, AMW uses the K-nearest neighbor method to estimate the propensity score, treating $\{1+K^{-1}M_{e(X),i}\}^{-1}$ as an estimator for the propensity score. Matching, which avoids inverting the estimated probability of treatment, makes the resulting AMW estimator more stable than the AIPW estimator. In this way, $1+K^{-1}M_{e(X),i}$ in (2.9) serves the same purpose as $\{P(A=A_i \mid X_i)\}^{-1}$ in (2.7). Thus, AMW can be viewed as an alternative version of AIPW. On

the other hand, the regression estimator $\hat{\tau}_{\text{reg}}$ in the AMW estimator acts as a bias correction term for the PSM estimator in (2.6), bringing protection against incorrect modeling of the propensity score model. As a result, the AMW estimator is doubly robust in the large sample setting, making it more reliable than the PSM estimator when the propensity score model is misspecified and more robust than $\hat{\tau}_{\text{reg}}$ when the outcome model is misspecified. Similarly, $K^{-1}M_{e(X),i}$ in (2.10) serves the same role as $\{P(A=1 \mid X_i)\} \times \{P(A=0 \mid X_i)\}^{-1}$. The ATT estimator $\hat{\tau}^t_{\text{AMW}}$ in (2.10) shares the same structures and properties as $\hat{\tau}_{\text{AMW}}$.

The matching type estimator with a fixed number of matches is not smooth, and consequently, the nonparametric bootstrap (Efron 1979) is unsuitable for estimating variances. To solve this problem, we unfix the number of matches K by letting K increase as the sample size grows. We provide the main algorithm to obtain $\hat{\tau}_{AMW}$ and its corresponding variance estimator.

AMW-Step1 Posit models to obtain the propensity score $\hat{e}(X)$ and the regression estimator $\hat{\tau}_{reg}$.

AMW-Step2 Implement Cross Validation (Section 2.4.2) to determine the optimal value of K for the AMW estimators.

AMW-Step3 For each unit i with treatment A_i, compute $\mid \hat{e}(X_{i'}) - \hat{e}(X_i) \mid_2^2$ for $\forall i'$, choose first K smallest distances to get indexes, and calculate $M_{\hat{e}(X),i}$ as the number of times that unit i is involved in one-time matching.

AMW-Step4 Construct $\hat{\tau}_{AMW}$ based on K, $M_{\hat{e}(X),i}$ and $\hat{\tau}_{reg}$. Estimate variance for $\hat{\tau}_{AMW}$ by naive bootstrap.

2.4.2 Unfix the number of matches in AMW estimators

Cross Validation (CV) is a resampling technique to select a model for a given predictive modeling problem. Initially, the data is randomly partitioned into training and testing subsets, and candidate models are fitted to the training data. Then, evaluate the prediction performance of the model via mean squared error (MSE), which compares prediction results with test data. Lastly, choose a candidate model with minimal MSE to strike a balance between significant variance from overfitting and considerable bias from underfitting. Although CV is highly accurate for prediction tasks, the selection of tuning parameters may not require predictive ability in causal inference (Rothenhäusler 2020). In this section, we propose a new form for the bias in the MSE for matching and discuss the consistency property of the estimated K:

$$MSE\left\{\hat{\tau}_{\mathrm{AMW}}(K)\right\} = Var\left\{\hat{\tau}_{\mathrm{AMW}}(K)\right\} + Bias^2\left\{\hat{\tau}_{\mathrm{AMW}}(K)\right\}, \tag{2.11}$$

The AMW estimator comprises two parts: a function $B(K)$ that depends on the number of matches K and a constant term C, where

$$B(K) = \frac{\sum_{i=1}^{n}\left[A_i\left\{M_{e(X),i}/K\right\}\hat{R}_i - (1-A_i)\left\{M_{e(X),i}/K\right\}\hat{R}_i\right]}{n}.$$

The bias of AMW estimators can be represented by $B(K)$, given that $\mathbb{E}(C) = \tau$. To illustrate the relationship between K and $B(K)$, we assume the propensity scores for unit i is fixed as e_i^*, and the residual term is fixed as $R_i^*(a_i)$. For simplicity, $\hat{m}_{a_i}(e_i^*)$ is denoted as $K^{-1}\sum_{j\in\mathcal{J}_K(i)} D_{R_{e_i}}\left(e_i^* - e_j^*\right)R_j^*(a_i)I_{(A_j=1-a_i)}$, where $D_d(x) = \frac{1}{d}D\left(\frac{x}{d}\right)$, and $D(x) = I_{(\|x\|\le 1)}$. Take bandwidth $d = R_{e_i}$, which is the random distance between e_i^* and its furthest element among the K nearest neighbors.

Let $f(e_i^*)$ be density function of e_i^*. We apply theorems from Mack and Rosenblatt (1979)

in Appendix A to study the bias and variance of $B(K)$ by Lemma 2.

Lemma 2 *Under regularity conditions, as $n \to \infty$, $K \to \infty$, $K/n \to 0$, we have*

$$n^2 \times Var\{B(K)\} = \frac{n_0 \mathbb{V}\{AR^*(1)|e^*\} + n_1 \mathbb{V}\{(1-A)R^*(0)|e^*\}}{K}, \tag{2.12}$$

$$Bias\{B(K)\} = \frac{K^2}{24nf^3(e^*)}\left[\left\{(m_1 f)''(e^*) - m_1(e^*)f''(e^*)\right\}\frac{n_0}{n_1^2} - \left\{(m_0 f)''(e^*) - m_0(e^*)f''(e^*)\right\}\frac{n_1}{n_0^2}\right], \tag{2.13}$$

where $(m_a f)(e^*) = m_a\left\{f(e^*)\right\}$. Choosing multiple matches from the opposite groups can introduce bias for the unit, as the second, third, and fourth closest matches are further away from the nearest neighbors (Stuart 2010). However, obtaining the theoretical value of the optimal K is difficult in practice because Equation (2.12) and (2.13) have complex forms. Hence, we propose a new data-driven CV algorithm to choose K to minimize the MSE when constructing the AMW estimator. The idea is motivated from the relationship between $Bias\{B(K)\}$ and K, where the bias will increase at a rate of K^2, as shown from Equation (12). Therefore, the AMW estimator has the smallest bias when $K = 1$. For a set of candidate parameters $K^{(1)}, ..., K^{(p)}$, we set $K^{(1)} = 1$ and treat $\hat{\tau}_{\text{AMW}}(K^{(1)})$ as unbiased. The $Bias\{\hat{\tau}_{\text{AMW}}(K^{(j)})\}$ is calculated as the difference between $\hat{\tau}_{\text{AMW}}(K^{(j)})$ and $\hat{\tau}_{\text{AMW}}(K^{(1)})$. Specially, $Bias\{\hat{\tau}_{\text{AMW}}(K^{(1)})\} = 0$ and reducing K can help to decrease the bias of the AMW estimator.

Conversely, the variance of $B(K)$ in (2.12) decreases as K increases, indicating that a larger K can reduce variance. Thus, a trade-off problem exists between the variance and bias when constructing the AMW estimator and estimating appreciation K to gain a balance between these factors. Besides, we propose a new data-driven CV algorithm to choose K to minimize the MSE when constructing the AMW estimator.

CV-Step1 For the j^{th} candidate parameter $K^{(j)}$, compute $Var\{\hat{\tau}_{AMW}(K^{(j)})\}$ by naive bootstrap.

CV-Step2 Split the dataset randomly into two equal halves, compute $\hat{\tau}_{AMW}(K^{(j)})_1$ for one half, and $\hat{\tau}_{AMW}(K^{(1)})_2$ for the rest. Obtain bias as $\hat{\tau}_{AMW}(K^{(j)})_1 - \hat{\tau}_{AMW}(K^{(1)})_2$.

CV-Step3 Repeat CV-Step2 multiple times to obtain several biases and then take the average of these estimates to obtain a robust estimate of the bias $Bias\{\hat{\tau}_{\text{AMW}}(K^{(j)})\}$. Compute the MSE by adding the variance of the estimate to the square of the bias.

CV-Step4 Select the value of K that has the smallest MSE among all options.

Similarly, the algorithms to get $\hat{\tau}^t_{AMW}$ and its variance estimator can mimic the above strategies.

2.5 Main theory

This section focuses on investigating the asymptotic properties of $\hat{\tau}_{\text{AMW}}$ and $\hat{\tau}^t_{\text{AMW}}$. Suppose that (X_i, A_i, Y_i) is independently observed from $\mathbb{P}_\theta$, indexed by $\theta^\top = (\alpha^\top, \beta_0^\top, \beta_1^\top)$, where α refers to the propensity score model parameter, and β_a controls the distribution of $Y(a)$. We assume that θ is distributed over on an open ball of $\mathbb{R}^k$.

2.5.1 Asymptotic distribution based on fixed θ

To start with, we treat θ as fixed parameters θ^*, which can be represented as known parameters, but may not be true parameters. Extract the similar assumptions (Mack and Rosenblatt 1979; Abadie and Imbens 2016), and impose some regularity conditions to derive the asymptotic distributions of $\hat{\tau}^{\theta^*}_{\text{AMW}}$ and $\hat{\tau}^{t\,\theta^*}_{\text{AMW}}$ as follows.

Assumption 2 *Set $f(e_i^*)$ and $\mathbb{E}\{R_i^*(A_i)|\theta^*\}$ are continually differential and bounded. Besides, $R_i^*(A_i)$ and $\mathbb{E}\{R_i^*(A_i)^3|\theta^*\}$ are uniformly bounded. Meanwhile, let $u_{A_i}(X_i, \beta^*_{A_i})$ and $\sigma^2_{A_i}(\beta^*_{A_i})$ satisfy Lipschitz continuity conditions.*

Abadie and Imbens (2016) and Yang and Zhang (2020) demonstrated these assumptions for PSM and double robust matching estimators, respectively. They impose regularity conditions on moments and smoothness for inference. We establish the theorems as follows.

Theorem 1 *Under Assumptions 1-2, as $n \to \infty$, if either the propensity score model or the outcome model is correctly specified, let $\pi_a(e^*) = P(a \mid e^*)$, we have*

$$n^{1/2}\left(\hat{\tau}^{\theta^*}_{\text{AMW}} - \tau\right) \to N\left(0, \Sigma^{\theta^*}_{\tau}\right), \tag{2.14}$$

$$\begin{aligned}\Sigma^{\theta^*}_{\tau} = \mathbb{E}\left[\left\{\frac{1}{\pi_0(e^*)}\right\}^2 \times \mathbb{V}\{(1-A)R^*(0)|X\} + \left\{\frac{1}{\pi_1(e^*)}\right\}^2 \times \mathbb{V}\{AR^*(1)|X\}\right] \\ + \mathbb{E}[\mathbb{V}\{u_1(X,\beta^*_1) - u_0(X,\beta^*_0) - \tau\}].\end{aligned}$$

We directly compare variances for the AMW, IPW, and outcome regression estimators. In the case where the outcome model is correctly specified, the efficiency of $\hat{\tau}^{\theta^*}_{\text{AMW}}$ may be lower than that of $\hat{\tau}^{\theta^*}_{\text{reg}}$. Moreover, if the outcome model is incorrectly specified, $\hat{\tau}^{\theta^*}_{\text{AMW}}$ maybe less efficient than the IPW estimator. However, the single-robust outcome regression and IPW estimators are biased if the corresponding parametric models are misspecified. In contrast, the linear expression for AMW derived in the supplementary material is similar to AIPW, indicating that AMW possesses the advantage of double robustness in the large sample setting. Therefore, AMW consistently estimates the ATE if either the propensity score or outcome model is correctly specified. Specifically, if both nuisance models are correct, the variance of the AMW estimator can be expressed as:

$$\mathbb{E}\left[\frac{\mathbb{V}\{Y(0) \mid X\}}{\pi_0(e^*)} + \frac{\mathbb{V}\{Y(1)|X\}}{\pi_1(e^*)} + \{u_1(X,\beta^*_1) - u_0(X,\beta^*_0) - \tau\}^2\right].$$

This variance achieves the semiparametric efficiency bound as the variance of AIPW (Robins et al. 1994; Hahn 1998; Tsiatis 2006). Hence, in the context of a large sample, the efficiency of the AMW estimator is comparable to that of the AIPW estimator.

Theorem 2 *Under Assumptions 1-2, as $n_1 \to \infty$, if either the propensity score model or the outcome model is correctly specified, let $p = \mathbb{E}(A)$, we have*

$$n^{1/2}\left(\hat{\tau}^{t,\theta^*}_{\text{AMW}} - \tau^t\right) \to N\left(0, \Sigma^{t,\theta^*}_{\tau}\right), \tag{2.15}$$

$$\begin{aligned}\Sigma^{t,\theta^*}_{\tau} = &\frac{1}{p^2}\mathbb{E}\left[\left\{\frac{1-\pi_0(e^*)}{\pi_0(e*)}\right\}^2 \times \mathbb{V}\{(1-A)R^*(0)|X\} + \mathbb{V}\{AR^*(1)|X\}\right] \\ &+ \frac{1}{p^2}\mathbb{E}[\mathbb{V}\{YA - u_0\left(X, \beta^*_0\right)A - \tau^t\}].\end{aligned} \tag{2.16}$$

2.5.2 Asymptotic distribution based on estimated θ

We acknowledge that θ is usually unknown and should be estimated in practical applications. Therefore, this section will examine the effect of estimating the nuisance parameters on the AMW estimators within the framework of Abadie and Imbens (2016) and Yang and Zhang (2020). To investigate the limiting distributions of the AMW estimators with estimated $\hat{\theta}$, we implement M-estimation to obtain $\hat{\theta}$. The nuisance parameter $\hat{\theta}$ in candidate propensity score and outcome models relies on the following estimating equations:

$$\Psi(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}\psi(A_i, X_i, Y_i; \theta) = \frac{1}{n}\sum_{i=1}^{n}\begin{Bmatrix} \psi_1(A_i, X_i; \alpha) \\ \psi_2\left(A_i, X_i, Y_i; \beta_0\right) \\ \psi_3\left(A_i, X_i, Y_i; \beta_1\right) \end{Bmatrix},$$

$\psi_1(A_i, X_i, Y_i; \alpha) = \frac{\partial e(X;\alpha)}{\partial \alpha}\frac{A-e(X,\alpha)}{e(X;\alpha)\{1-e(X,\alpha)\}}$, $\psi_2(A_i, X_i, Y_i; \beta_0) = (1-A)\frac{\partial u_0\left(X;\beta_0\right)}{\partial \beta_0}\left\{Y - u_0\left(X, \beta_0\right)\right\}$, and $\psi_3(A_i, X_i, Y_i; \beta_1) = A\frac{\partial u_1\left(X;\beta_1\right)}{\partial \beta_1}\left\{Y - u_1\left(X, \beta_1\right)\right\}$.

Based on the Locally Asymptotically Normal (LAN) model, we define the local parameter $\theta_n = \theta^* + h/\sqrt{n}$, where θ^* is the true parameter and h is a constant. To invoke Le Cam's third lemma (Le Cam and Yang 1990; van der Vaart 2000) for LAN model, we first derive the limiting distribution between $n^{1/2}(\hat{\tau}^{\theta_n}_{\text{AMW}} - \tau^{\theta_n})$, $n^{1/2}(\theta_n - \theta)$, and $\log(p^{\theta^*}/p^{\theta_n})$ under

p^{θ_n}, where p^{θ_n} is the probability measure with θ_n and p^{θ^*} is the true probability measure of the random variables. Next, we obtain the limiting distribution $n^{1/2}(\hat{\tau}^{\theta_n}_{\text{AMW}} - \tau)$ by Le Cam's third lemma and derive the corresponding coefficients by utilizing martingale theory (Andreou and Werker 2012). Lastly, the asymptotic distribution for $n^{1/2}(\hat{\tau}^{\hat{\theta}}_{\text{AMW}} - \tau)$ can be approximated by replacing θ_n with $\hat{\theta}$.

Theorem 3 *Under Assumptions 1-2, and regularity conditions specified in the supplementary material, if either the propensity score model or the outcome model is correctly specified, we have*

$$n^{1/2}\left(\hat{\tau}^{\hat{\theta}}_{\text{AMW}} - \tau\right) \to N\left(0, \Sigma^{\hat{\theta}}_{\tau}\right), \tag{2.17}$$

where $\Sigma^{\hat{\theta}}_{\tau} = \Sigma^{\theta^*}_{\tau} - C_1^{\top} I^{-1}_{\theta^*} C_1 + C_2^{\top} \Sigma_{\theta^*} C_2$, $I^{-1}_{\theta^*} = \mathbb{E}\left\{\psi(A, X, Y; \theta^*)\psi(A, X, Y; \theta^*)^{\top}\right\}^{-1}$, $\Sigma_{\theta^*} = \varphi^{-}_{\theta^*} I_{\theta^*} (\varphi^{-}_{\theta^*})^{\top}$ *with* $\varphi^{-}_{\theta^*} = \mathbb{E}\{\partial \psi(A, X, Y; \theta^*)/\partial \theta\}^{-}$, *coefficients* C_1 *and* C_2 *are illustrated in the supplementary material.*

The variance $\Sigma^{\hat{\theta}}_{\tau}$ incorporates two additional terms, namely $-C_1^{\top} I^{-1}_{\theta^*} C_1$ and $C_2^{\top} \Sigma_{\theta^*} C_2$, compared to $\Sigma^{\theta^*}_{\tau}$ in (2.14). The term $-C_1^{\top} I^{-1}_{\theta^*} C_1$ is analogous to the reduction term obtained by Abadie and Imbens (2016), which demonstrates the correlation between the score function for θ in parametric models and the matching estimator, leading to a decrease in estimated variance. Besides, the term $C_2^{\top} \Sigma_{\theta^*} C_2$ is similar to the variance inflation term obtained by Yang and Zhang (2020), where τ relies on the nuisance parameters by equation $\mathbb{E}[u_1(X, \beta_1) - u_0(X, \beta_0) + AR/e(X, \alpha) - (1-A)R/\{1 - e(X, \alpha)\}]$ with the misspecification of either the propensity score or outcome model. The difference between $\Sigma^{\theta^*}_{\tau}$ in (2.14) and $\Sigma^{\hat{\theta}}_{\tau}$ in (2.17) is unknown, since the sum of the variance reduction and variance inflation terms is unknown. Therefore, we cannot assert that the estimated score can always improve estimation, and it might even reduce estimation efficiency.

Theorem 4 *Under Assumptions 1-2, and regularity conditions specified in the supplementary material, for* $r \geq 1$, $0 < n_1^r/n < \infty$, *if either the propensity score model or the outcome*

model is correctly specified, we have

$$n^{1/2}\left(\hat{\tau}^{t,\hat{\theta}}_{\mathrm{AMW}} - \tau^t\right) \to N\left(0, \Sigma^{t,\hat{\theta}}_{\tau}\right), \tag{2.18}$$

where $\Sigma^{t,\hat{\theta}}_{\tau} = \Sigma^{t,\theta^*}_{\tau} - C_1^{t,\top} I^{-1}_{\theta^*} C_1^t + C_2^{t,\top} \Sigma_{\theta^*} C_2^t$, C_1^t *and* C_2^t *are illustrated in the supplementary material.*

The extension of the discussion on large sample distribution properties from the AMW estimator for ATE to that for ATT is straightforward.

2.6 Simulation study

To examine the finite sample properties of the proposed AMW estimator $\hat{\tau}_{\mathrm{AMW}}$ for estimating the average treatment effect (ATE), as well as other existing methods such as weighting and matching, two experiments are conducted. One experiment uses the extreme propensity score distribution, while the other uses the standard distribution. The study aims to verify the AMW estimator's stable property, indicating that it maintains a small mean squared error even when propensity scores are close to zero or one. Additionally, the double robustness property of the AMW estimator is to determine its consistency when either the propensity score or the outcome model is correctly specified. Lastly, the validity of the variance estimated by standard bootstrap is also investigated. The simulations are conducted on 1000 Monte Carlo simulated datasets for each scenario to obtain solid results.

2.6.1 Simulation setting

This simulation compares three types of estimators: 1) weighting type estimators such as IPW and AIPW, 2) matching type estimator PSM, and 3) the AMW estimator with unfixed K

and the AMWF estimator, which is the AMW estimator with fixed $K = 1$. Parameter K is selected using cross-validation, repeated 25 times for robust bias estimation.

Set sample size to $n = 1000$, and generate variables $Z_j \overset{iid}{\sim} U[1-\sqrt{3}, 1+\sqrt{3}]$, where $j = 1, .., 12$. Transformations are applied to incorporate nonlinearity and correlation into the confounder model: $X_1 = \exp(Z_1)$, $X_2 = \exp(Z_2)$, $X_3 = \log(Z_3+1)^2$, $X_4 = \log(Z_4+1)^2$, $X_5 = \sin(Z_5 - Z_6)$, $X_6 = \cos(Z_5 + Z_6)$, $X_7 = \sin(Z_7)$, $X_8 = \cos(Z_7 - 1)$, $X_9 = (Z_8 > 0.4)$, $X_{10} = (Z_8 > -0.4)$, $X_{11} = (Z_9 > 0.3)$, $X_{12} = (Z_{10} > -0.3)$. Standardize the transformed X by $\{X_j - \mathbb{E}(X_j)\}/\sqrt{\{\mathbb{V}(X_j)\}}$ in case some covariates may significantly influence the results. Potential outcome models are generated as $Y(0) = \beta^\top X + \epsilon(0)$ and $Y(1) = \beta^\top X + \epsilon(1)$, where $\epsilon(0) \sim N(0,4)$, $\epsilon(1) \sim N(0,1)$ and $\beta^\top = (1,1,1,1,-1,-1,-1,-1,1,-1,1,-1) \times 0.2$. Meanwhile, the probability of receiving treatment $A = 1$ is $\text{logit}^{-1}(\alpha_k^\top X)$ with $\alpha_k^\top = (1,1,1,1,-1,-1,-1,-1,1,-1,1,-1) \times c_k$, when $c_1 = 1$ induces a extreme propensity score distribution with a heavy tail, and $c_2 = 0.3$ results in a standard propensity score distribution with fewer extreme values.

Two model specifications are used to estimate the propensity scores: 1). a correctly specified logistic model $e(X, \alpha^1) = \text{logit}^{-1}(\alpha^{1,\top} X)$; 2). a misspecified logistic model based on the original variables $e(X, \alpha^0) = \text{logit}^{-1}(\alpha^{0,\top} Z)$. Similarly, we consider two model specifications for the outcomes: 1). a correctly specified linear model $u_a(X, \beta_a^1) = \beta_a^{1,\top} X$; 2). a misspecified linear model based on the original variables $u_a(X, \beta_a^0) = \beta_a^{0,\top} Z$. Therefore, we can use four combinations to fit the propensity score and outcome models. To make it easier, each estimator's name is associated with two numbers that indicate the choice of fitting models. The first number indicates the selection for the propensity score model, while the second number represents the choice for the outcome model. Here, we use "1" to denote the correctly specified model and "0" to denote the misspecified model. For instance, "AMW01" represents the AMW estimator that combines a misspecified propensity score model and a correctly specified outcome model.

Standard nonparametric bootstrap approach is utilized to obtain the variances of all

estimators. Specifically, 500 bootstrap replicates are conducted in the non-stable case with the correct propensity score model to minimize the impact of extreme propensity scores. In other cases, 100 bootstrap replicates are generated. Five summary statistics are computed to evaluate the performance of the estimators:

- The sample mean value of 1000 simulated estimators ("mean"),
- The sample standard deviation of 1000 simulated estimators ("sd"),
- The average value of 1000 standard deviations from bootstrap replicates ("bootsd"),
- The MSE of 1000 simulated estimators ("mse"),
- The coverage rates obtained by 95% bootstrap quantiles ("cr").

2.6.2 Results

Table 1 summarizes the performance of different estimators based on five summary statistics, while Figure 2.1 displays the box plots for each estimator. The results indicate that the AMW estimators consistently achieve the minimum MSE or near-minimum MSE among all the estimators. When the propensity score distribution is not extreme and the population is adequate, the difference between the AMW estimator and the locally efficient AIPW estimator is negligible. However, when the propensity scores are extreme, the AMW estimators outperform the AIPW estimator, as the inverse of the extreme weights in the weighting type estimators leads to unstable estimates. We remove the outliers beyond the range of $(-5, 5)$ on the y-axis in Figure 2.1 to provide a more harmonized presentation. The PSM estimator consistently has a larger MSE than the AMW estimators, which justifies the efficiency gain from the augmented terms in the AMW estimators.

Table 2.1: Simulation results.

Name	Extreme propensity score distribution					Non-extreme propensity score distribution				
type	mean	sd	bootsd	mse	cr	mean	sd	bootsd	mse	cr
AMW11	0.01	0.427	0.401	0.182	0.946	-0.002	0.23	0.231	0.053	0.938
IPW11	-0.044	1.248	0.768	1.558	0.846	-0.008	0.233	0.222	0.055	0.927
AIPW11	-0.051	1.129	0.704	1.277	0.93	-0.006	0.229	0.219	0.053	0.919
PSM11	0.087	0.7	0.598	0.497	0.94	0.005	0.271	0.257	0.073	0.967
AMWF11	0.003	0.319	0.6	0.101	0.952	-0.004	0.214	0.256	0.046	0.963
AMW01	-0.009	0.302	0.309	0.091	0.95	0	0.22	0.224	0.048	0.943
IPW01	0.682	0.231	0.235	0.518	0.17	0.314	0.2	0.2	0.139	0.634
AIPW01	-0.001	0.293	0.296	0.086	0.938	-0.001	0.214	0.21	0.046	0.915
PSM01	0.674	0.3	0.296	0.544	0.315	0.315	0.234	0.236	0.154	0.719
AMWF01	-0.006	0.279	0.339	0.078	0.956	0.002	0.212	0.245	0.045	0.96
AMW10	0.156	0.432	0.392	0.21	0.926	0.023	0.224	0.229	0.051	0.957
IPW10	0.071	1.388	0.732	1.929	0.871	0	0.224	0.222	0.05	0.943
AIPW10	0.085	1.364	0.735	1.867	0.891	0.005	0.223	0.221	0.05	0.948
PSM10	0.103	0.705	0.602	0.507	0.961	0.004	0.265	0.257	0.07	0.973
AMWF10	0.591	0.289	0.609	0.433	0.973	0.28	0.209	0.259	0.122	0.975
AMW00	0.682	0.248	0.26	0.527	0.232	0.32	0.203	0.214	0.144	0.66
IPW00	0.692	0.239	0.236	0.536	0.155	0.321	0.199	0.199	0.143	0.602
AIPW00	0.691	0.244	0.241	0.537	0.163	0.322	0.199	0.199	0.143	0.609
PSM00	0.693	0.305	0.299	0.574	0.307	0.317	0.234	0.237	0.155	0.72
AMWF00	0.66	0.22	0.3	0.484	0.325	0.318	0.198	0.235	0.14	0.724

The results obtained from both scenarios indicate that the AMW-type estimators are doubly robust, which is a significant advantage over singly robust estimators like the IPW and PSM estimators. Additionally, we observe that the bias of the AMWF10 estimator is not negligible in both settings, which supports the use of unfixed K in the AMW estimators. Specifically, the AMWF10 estimator can be considered as the PSM estimator for $K = 1$ with an incorrect outcome term, which introduces a new bias for AMWF.

Furthermore, we find that the standard deviation estimated by bootstrap for the AMWF estimator is consistently higher than the simulated standard deviation. In contrast, the smoothed AMW estimator with unfixed K can provide a more reliable and consistent standard deviation estimation from the standard bootstrap. As a result, the AMW estimator with unfixed K shows a more promising coverage rate compared to other matching type estimators.

2.7 Real data analysis

In this section, we utilize AMW estimators to study the causal effects in three publicly accessible datasets (Loh and Vansteelandt 2021). To generate the AMW estimators, we first posit propensity score and outcome models. Although the outcome model selections can vary across studies, logistic regression is used to estimate all propensity scores. Additionally, all covariates are scaled before being incorporated into the models to prevent certain variables from dominating the results. Once the AMW estimators are obtained, we apply standard bootstrap with 500 repetitions to investigate the standard errors. Since researchers cannot verify the actual causal effects, we assess the matching performances using standardized differences of covariates (Abadie and Imbens 2011) between the treatment and control groups. Reduced differences after matching indicate a successful balance between the two groups. This paper includes plots of the absolute standardized differences, with blue

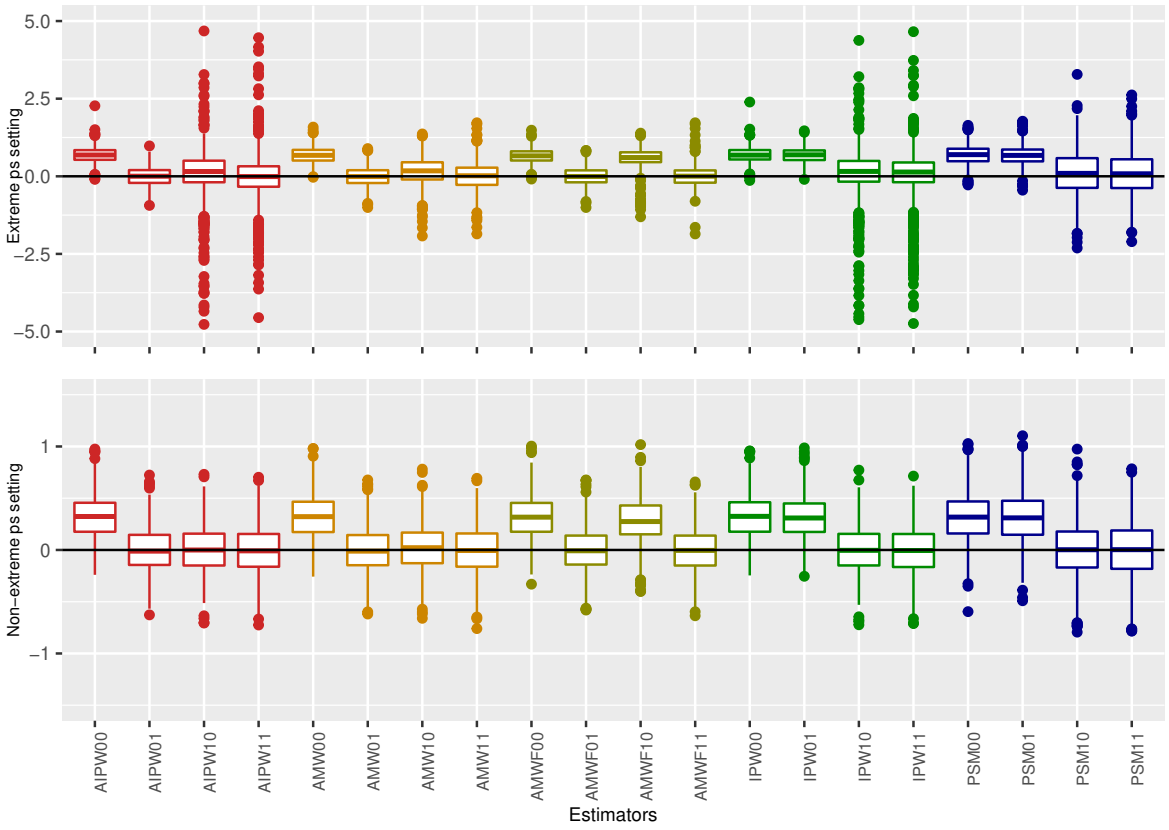


Figure 2.1: Box plot for ATE.

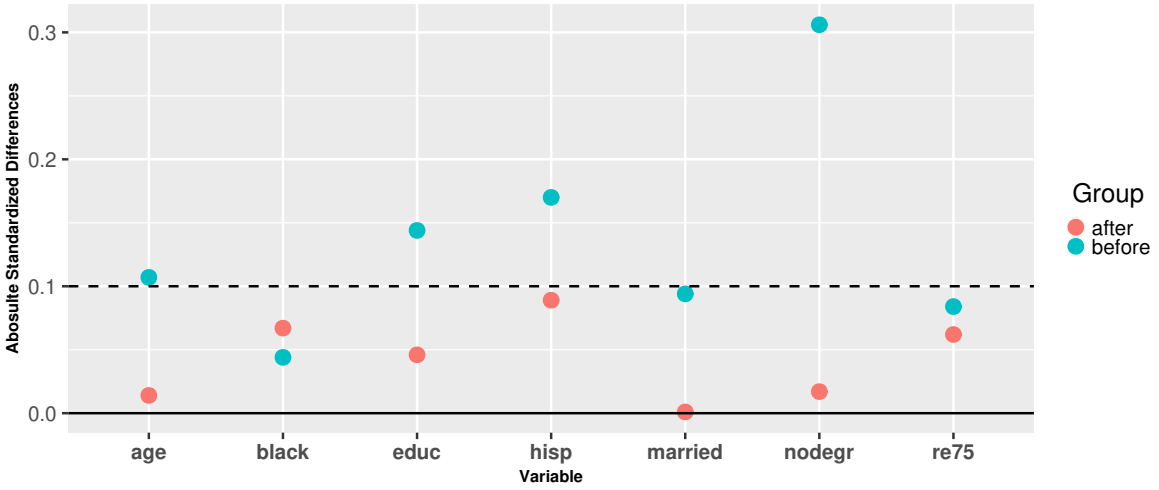


Figure 2.2: Standardized differences plot for Labor Training Program data.

dots representing differences before matching and red dots representing differences after matching for corresponding covariates. Tables displaying standardized differences are presented in the supplementary material.

2.7.1 Labor training program

LaLonde (1986) and Dehejia and Wahba (1999) analyzed the National Supported Work (NSW-DW) dataset, which is a subset of a labor training program. The dataset, available in the "Matching" package in R, aims to examine the causal effect of a job training program. There are a total of 445 individuals in the NSW-DW dataset, with 185 individuals in the treatment group. The outcome is represented by the actual earnings of each person in 1978, and seven covariates are taken into account in the linear outcome model. Specifically, the second-order terms of two numeric covariates, namely age and actual earnings in 1975, are included in both the propensity score and outcome models (Dehejia and Wahba 1999).

To provide a more comprehensive evaluation of this program, we analyze the ATT for participating in the job training program, as the ATE may not be sufficient. Figure 2.2

displays the standardized differences, illustrating that the AMW estimator achieves greater balance of covariates between the two groups. The ATT, which measures the average earning increase for the treated in one year, is 1872.53 dollars, with a standard error of 692.70. The empirical 95% confidence interval, which reflects the uncertainty surrounding the ATT, is (514.84, 3230.22). These results suggest that the job training program has a significant impact on the average earning of the treated group, consistent with the NSW-DW dataset's estimated ATT of 1794.34 (Otsu and Rai 2017).

2.7.2 AIDS clinical trials group study

Hammer et al. (1996) designed a double-blind study to assess the efficacy of AIDS treatments, comparing the effects of a single treatment of either zidovudine or didanosine to the combination treatment of zidovudine plus didanosine or zalcitabine. The dataset used for this study, AIDS Clinical Trials Group Study 175 (ACTG175), is available in the R package "speff2trial." As HIV can attack the immune system and reduce CD4 cell counts, the primary outcome of interest in our design is the difference between CD4 cell counts after 96 weeks and the baseline. Units with missing outcomes are excluded to maintain the validity of the modeling process, resulting in a reduced model with 1342 individuals. While the original study was designed as a randomized experiment, missing values posed challenges when comparing the two therapy groups directly. Thus, it is necessary to employ the AMW estimator to eliminate confounding bias and obtain reliable effects. The outcome model utilizes a linear model with all 14 covariates.

Figure 2.3 displays the standardized differences for covariates before and after matching, which demonstrates that the AMW estimator effectively enhances the balance between the two groups. The average treatment effect (ATE) is calculated to be 40.852 with a standard error of 8.220, yielding a 95% empirical lower bound for ATE of 24.740. This result, which is

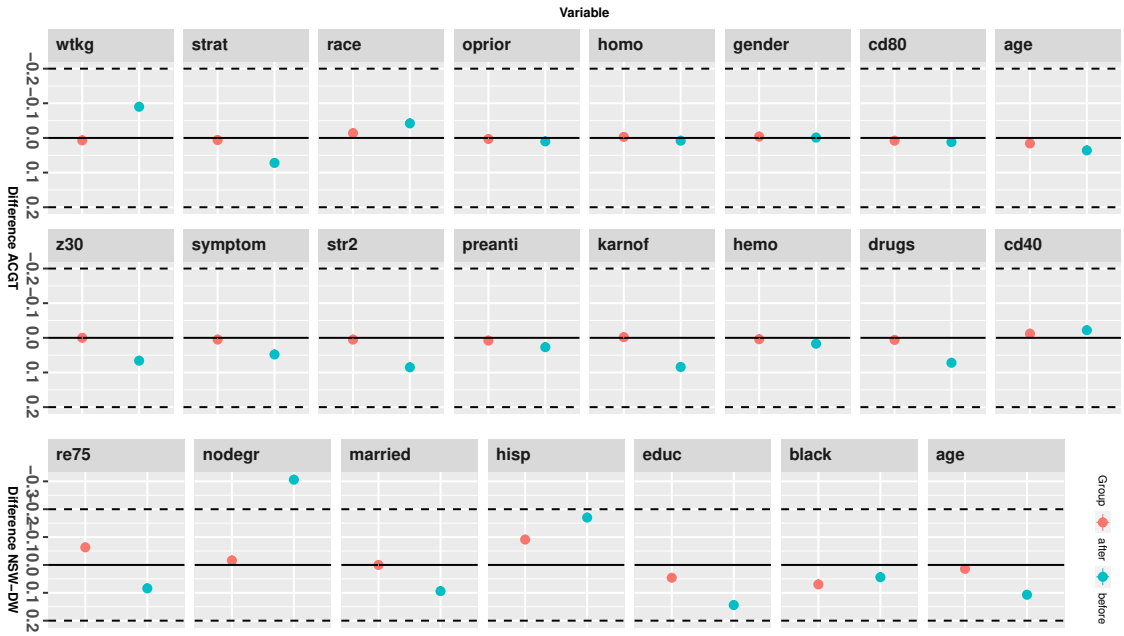


Figure 2.3: Standardized differences plot for ACTG data.

greater than zero, suggests a significant improvement in CD4 cell counts resulting from combination therapy.

2.7.3 Right heart catheterization

Connors et al. (1996) and Hirano and Imbens (2001) analyzed the RHC dataset to investigate the safety and effectiveness of the Right Heart Catheterization procedure for severely ill patients in ICUs. The RHC procedure was previously believed to increase patient survival rates, but subsequent reanalysis by Hirano and Imbens (2001) suggested otherwise. We aim to use the AMW estimator to analyze the RHC dataset and evaluate its effectiveness on this challenging dataset. The RHC dataset, which contains 5735 units with 63 variables, is available at https://hbiostat.org/data/. For our analysis, we select 50 variables and convert categorical variables into dummy variables for both the propensity score and outcome models. We model the outcome variable, which is a binary indicator of whether the patient

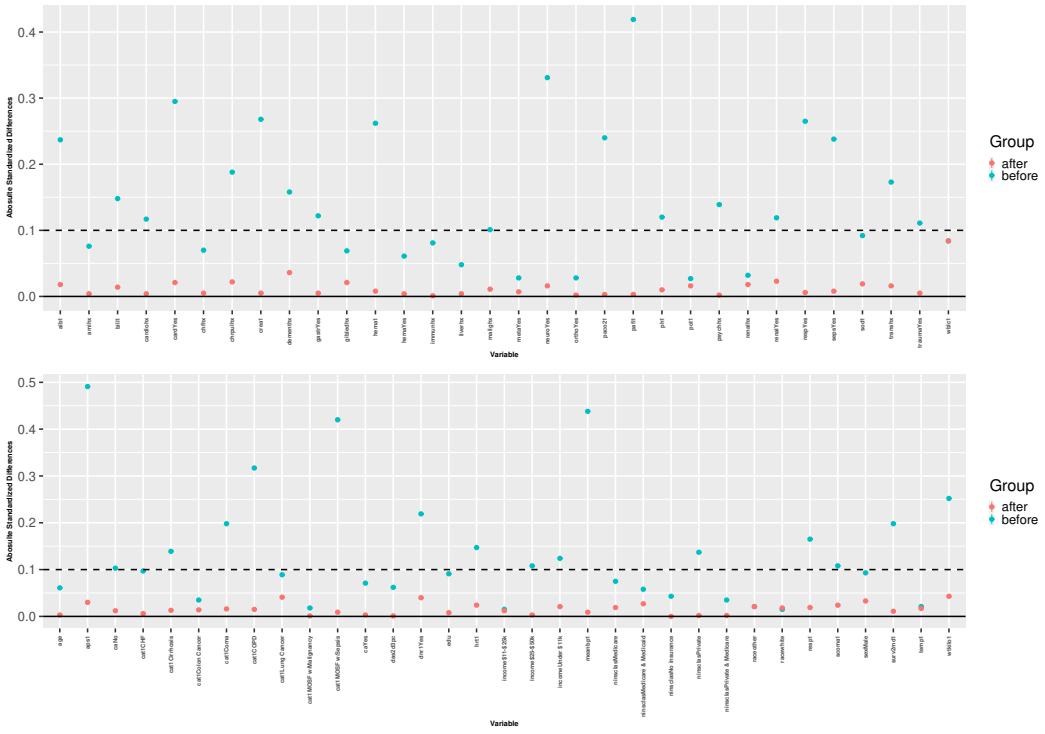


Figure 2.4: Standardized differences plot for RHC data.

survives for 30 days (the survival patient denoted as "1", and the dead patient denoted as "0"), using logistic regression. Our goal is to estimate the causal risk difference by comparing the survival probabilities between the two experimental groups.

Figure 2.4 demonstrates that the use of the AMW estimator results in reduced differences between the observed covariates in the two groups before and after matching. Our estimated ATT of -0.0641 is consistent with the effects obtained by overlap and optimal matching estimators, as reported by Li et al. (2017). However, the standard error for our estimator, 0.0241, is higher than those obtained by the other two estimators. Our findings suggest that the RHC procedure is associated with increased risk for patients, which contrasts with the effects estimated by traditional approaches that do not account for confounding bias.

2.8 Discussion

We propose a new approach called Augmented Match Weighted (AMW) estimator for estimating general causal estimands. In comparison with PSM, AMW achieves the semi-parametric efficiency bound for unfixed K and exhibits double robustness due to the inclusion of additional augmentation terms. Unlike the AIPW approach, AMW avoids the need to directly invert the propensity scores, thereby increasing its stability. Our simulation results demonstrate that the naive bootstrap method can be used to obtain variance estimates for AMW. Overall, the AMW estimator is a practical and powerful alternative to existing methods for causal inference. However, to address the issue of computationally tractable problems, we typically fix the value of K to estimate the variance through bootstrap in CV. While this approach does not pose any significant trouble in practice, a more standard method is to allow K to vary independently in each bootstrap iteration to achieve a more precise selection.

Assumption 1 requires all confounders should be measured and included to construct the propensity score and outcome models so that people may include all variables in models to guarantee the assumption hold. However, Brookhart et al. (2006) emphasized that instrumental variables may not improve estimate treatment effects for the propensity score model since they are uncorrelated with outcomes. The variables related to outcomes can provide better estimation even under model misspecifications. Hence, Zhang et al. (2021) recommended a simple yet effective strategy that employs lasso (Hastie et al. (2009)) for variable selection in the outcome model, followed by using the chosen features with appropriately tuned propensity scores to improve the efficiency of matching. This approach can significantly reduce bias and variance in PSM. Likewise, one can investigate how various variables might affect AMW, where AMW could be less susceptible to instrumental variables. Multiple variable selection techniques, such as lasso and Adaptive lasso, can be combined

with AMW to enhance its efficiency.

Meanwhile, Chernozhukov et al. (2018) investigated the application of debiased/double machine learning (DML) methods for estimating high-dimensional nuisance parameters, and established broad theoretical foundations with weak assumptions. This led to the integration of the DML framework into AMW, allowing for the use of modern nonparametric techniques (such as xgboost and random forest) in estimating causal effects.

On the other hand, Yang et al. (2016) proposed a framework that employs PSM to estimate treatment effects in scenarios with more than two treatment options, and AMW could be similarly adapted to handle such cases. Additionally, it is interesting to extend the AMW estimator to longitudinal (Buzkova and Lumley (2007)) or survival analysis (Choi and O'Malley (2017)).

2.9 Acknowledgements

T.X. was supported by N.S.F. DEB-1754142. S.Y. was partially supported by NIH 1R01AG066883 and 1R01ES031651. T.X. served as the first author of this paper, contributing significantly to the study's approach, theory, coding, writing, and data analysis.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76:1537–1557.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29:1–11.

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *J Am Stat Assoc*, 107:833–843.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84:781–807.

Andreou, E. and Werker, B. J. (2012). An alternative asymptotic analysis of residual-based statistics. *Rev Econ Stat*, 94:88–99.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *Am J Epidemiol*, 163:1149–1156.

Brookhart, M. A. and Van Der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational statistics & data analysis*, 50(2):475–498.

Buzkova, P. and Lumley, T. (2007). Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics*, 35:485–500.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:1–68.

Choi, J. and O'Malley, A. J. (2017). Estimating the causal effect of treatment in observational studies with survival time end points and unmeasured confounding. *Journal of the Royal Statistical Society: Series C*, 66:159–185.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897.

Cui, Y. and Tchetgen, E. T. (2019). Selective machine learning of doubly robust functionals. *arXiv preprint arXiv:1911.02029*.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J Am Stat Assoc*, 94:1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Rev Econ Stat*, 84:151–161.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Rev Econ Stat*, 86:77–90.

Guo, S. and Fraser, M. W. (2014). *Propensity Score Analysis: Statistical Methods and Applications*, volume 11. Thousand Oaks, CA: SAGE.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100:417–430.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*, volume 2. Springer.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64:605–654.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2:259–278.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge UK.

Ju, C., Schwab, J., and van der Laan, M. J. (2019). On adaptive propensity score truncation in causal inference. *Statistical methods in medical research*, 28(6):1741–1760.

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

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76:604–620.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer: Berlin.

Leacy, F. P. and Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat Med*, 33:3488–3508.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2017). Balancing covariates via propensity score weighting. *J Am Stat Assoc*, page doi:10.1080/01621459.2016.1260466.

Loh, W. W. and Vansteelandt, S. (2021). Confounder selection strategies targeting stable treatment effect estimators. *Statistics in Medicine*, 40(3):607–630.

Mack, Y. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9:1–15.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *J Am Stat Assoc*, 112:1720–1732.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*, 89:846–866.

Rolling, C. A. and Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):749–769.

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

Rothenhäusler, D. (2020). Model selection for estimation of causal parameters. *arXiv preprint arXiv:2008.12892*.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29:159–183.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, England.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25:1–21.

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

van der Vaart (2000). *Asymptotic Statistics*, volume 3. Cambridge university press, Cambridge: Cambridge University Press.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72:1055–1065.

Yang, S. and Kim, J. K. (2017). A semiparametric inference to regression analysis with missing covariates in survey data. *Statistica Sinica*, 27:261–285. Accepted for publication.

Yang, S. and Zhang, Y. (2020). Multiply robust matching estimators of average and quantile treatment effects. *arXiv preprint arXiv:2001.06049*.

Zhang, Y., Yang, S., Ye, W., Faries, D. E., Lipkovich, I., and Kadziola, Z. (2021). Practical recommendations on double score matching for estimating causal effects. *Statistics in medicine*.

CHAPTER

3

INTERLOCUS GENE CONVERSION, NATURAL SELECTION, AND PARALOG HOMOGENIZATION

3.1 Abstract

Following a duplication, the resulting paralogs tend to diverge. While mutation and natural selection can accelerate this process, they can also slow it. Here, we quantify the paralog homogenization that is caused by point mutations and interlocus gene conversion (IGC).

Among 164 duplicated teleost genes, the median percentage of post-duplication codon substitutions that arise from interlocus gene conversion rather than point mutation is estimated to be between 7% and 8%. By differentiating between the nonsynonymous codon substitutions that homogenize the protein sequences of paralogs and the non-homogenizing nonsynonymous substitutions, we estimate the homogenizing nonsynonymous rates to be higher for 163 of the 164 teleost data sets as well as for all 14 data sets of duplicated yeast ribosomal protein-coding genes that we consider. For all 14 yeast data sets, the estimated homogenizing nonsynonymous rates exceed the synonymous rates.

3.2 Introduction

Due to both natural selection and mutation, it has long been appreciated that duplicated regions of the genome may not evolve independently of one another (e.g., Hood et al. 1975; Tartof 1975). One kind of mutation responsible for this lack of independence is known as interlocus gene conversion (IGC) and also as non-allelic gene conversion. An IGC mutation results in a stretch of DNA sequence from one paralog being replaced by the sequence in the corresponding region from another paralog. The evolutionary impact of IGC is difficult to assess because studying it necessitates jointly considering paralogs.

As a result of this difficulty, the possibility of IGC has conventionally been ignored when probabilistic models of sequence change have been employed to characterize molecular evolution or infer phylogenies. One way to avoid consideration of IGC has been to exclusively focus on molecular evolution in single-copy genomic regions. This narrow focus is unfortunate because genomes are often rich in duplicated regions.

While some hypothesized ancient whole-genome duplications are controversial (e.g., Abbasi 2010), it is clear that taxonomically-important lineages including angiosperms, teleosts, and yeast experienced them (e.g., Wolfe and Shields 1997; Van de Peer et al. 2009;

Van de Peer et al. 2010; Vanneste et al. 2014). For these lineages, genes that are now single-copy may have had ancestral paralogs. Therefore, IGC-induced dependence between ancestral paralogs may have shaped the DNA of genes that are single-copy in extant genomes.

Rather than ignoring IGC, we have been developing a model-based phylogenetic approach for studying it (Ji et al. 2016). It considers sequence changes that result from both point and IGC mutations. For 14 data sets of ribosomal protein-coding genes that resulted from a historical genome-wide duplication event in the yeast lineage, the estimated percentage of changes attributable to IGC ranges from 20% to 38% (Ji et al. 2016). These values are especially high when one considers that a homogenizing substitution due to IGC cannot occur unless corresponding codons differ due to a point mutation prior to the IGC event. Therefore, the percentages from IGC cannot exceed 50%. Because yeast ribosomal protein-coding genes are unusually conserved (Kellis et al. 2004) and had been previously recognized for their IGC (Evangelisti and Conant 2010), their high IGC percentages are probably unrepresentative of other duplicated genes.

In contrast, Harpak et al. (2017) examined mammalian intronic regions that were the result of recent duplications. They adopted an inference procedure with similarities to Ji et al. (2016, see also Ji 2017) and concluded that IGC did not have a substantial long-term impact on paralog divergence for their intronic loci. Because so few IGC studies have been done, the overall evolutionary influence of IGC remains uncertain, as do the consequences of ignoring IGC when inferring phylogenies or estimating divergence times.

Motivated by the desire to assess IGC in additional data sets so as to broaden the understanding of how generally relevant IGC is to molecular evolution, we apply and extend our inference procedure to study IGC in 164 protein-coding data sets from teleosts. We conclude that substitutions due to IGC rather than point mutation are responsible for a non-negligible proportion of all post-duplication sequence changes in these data sets. This casts doubt on the usual practice in molecular evolutionary analyses of ignoring IGC

following gene duplication.

Beyond establishing the presence of IGC in the teleost data, we also carefully examine the patterns of nonsynonymous change that they have experienced. By analyzing both the teleost data and ribosomal protein-coding genes from yeast, we come to the biologically plausible conclusion that the amino acid that is encoded at a protein position in one paralog is a useful predictor of nonsynonymous rates at the corresponding codon triplet of the other paralog. We conclude by discussing the implications of these nonsynonymous rate patterns, weaknesses of our approach, and promising directions for IGC-related research.

3.3 New approaches

3.3.1 The basic IGC model

Ji et al. (2016) extended conventional codon-based substitution models to reflect IGC, but a simpler version of the same approach can add IGC to the HKY (Hasegawa et al. 1985) or other 4-state nucleotide substitution models. We first review the Ji et al. (2016) approach to facilitate presentation of a generalization that is introduced here. While the IGC model incorporates dependence between corresponding codons of paralogs, it assumes independent evolution among the codons within a paralogous gene. Because IGC mutations homogenize tracts of adjacent sequence positions, actual genome evolution violates this independence assumption. However, the degree of this violation is unknown because it depends on the level of intralocus recombination. When recombination rates are high, the independence assumption may be reasonable because recombination will lessen the dependence in fixation probabilities between sequence sites that comprise an IGC tract. For sufficiently high recombination rates, the IGC inference procedure is effectively a maximum likelihood procedure whereas the approach is more accurately classified as a maximum

composite likelihood procedure when recombination rates are low (Ji et al. 2016).

The IGC model describes both codon substitutions that originate with point mutation and those that originate with IGC. Consider corresponding codons of two paralogs, with the first paralog being occupied by a codon triplet denoted i and the second being occupied by a triplet denoted i'. For codon substitutions that originate with point mutations, the instantaneous rate of change from triplet state i (i') to j (j') in the first (second) paralog will be denoted Q_{ij} ($Q_{i'j'}$). To specify Q_{ij}, any conventional model of nucleotide or codon substitution can be adopted. All analyses conducted for this study have reduced the number of free parameters to be estimated by setting $Q_{ij} = Q_{i'j'}$ when $i = i'$ and $j = j'$.

Because $Q_{i,j}$ represents the rates of codon substitutions that arise from point mutation, $Q_{i,j} = 0$ for all triplets i and j that differ at more than one codon position. As is conventional for codon models, we do not model substitutions that involve stop codons. For i and j that differ exactly at one position where j has nucleotide type h, the rates are

$$Q_{i,j} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ u\pi_h\kappa & \text{for a synonymous transition} \\ u\pi_h\omega & \text{for a nonsynonymous transversion} \\ u\pi_h\kappa\omega & \text{for a nonsynonymous transition,} \end{cases} \tag{3.1}$$

with $\pi_A + \pi_C + \pi_G + \pi_T = 1$ and $0 \leq \pi_h \leq 1$ for $h \in \{A, C, G, T\}$. The u in Equation 3.1 is a normalization constant that makes the expected rate per codon of substitutions that arise from point mutation equal to one. In the notation of the PAML software (Yang 2007), this parameterization of Q_{ij} is referred to as the $F1\times4MG + \kappa + \omega$ model.

For a genetic code with 61 possible codon states, joint consideration of the two paralogs requires specification of the instantaneous rate of change $Q_{(i,i'),(j,j')}$ from the 61^2 possible combinations of i and i' to each of the 61^2 possible combinations of j and j'. Our joint substitution model with IGC has:

$$Q_{(i,i'),(j,j')} = \begin{cases} 0 & i \neq j,\ i' \neq j' \\ Q_{i,j} & i \neq j,\ i' = j',\ j \neq j' \\ Q_{i',j'} & i = j,\ i' \neq j',\ j \neq j' \\ Q_{i,j} + \nu & i \neq j,\ i' = j',\ j = j' \\ Q_{i',j'} + \nu & i = j,\ i' \neq j',\ j = j', \end{cases} \tag{3.2}$$

where $\nu = \tau$ if the change is synonymous and where $\nu = \omega\tau$ if the change is nonsynonymous. The parameter τ controls the amount of IGC. Our practice is to normalize rates to make the expected rate per codon equal to one for substitutions that arise from point mutation. This normalization is performed for the case where IGC is absent (i.e., $\tau = 0$) and we do not renormalize to make the normalization dependent on the value of τ. We note that IGC events simultaneously change multiple nucleotides when a codon differs at more than one position in the two paralogs.

3.3.2 The ω_H/ω_N IGC model

For the Ji et al. (2016) model, excess homogenization of amino acid types at corresponding positions in two paralogs is explained by IGC. An alternative – but not mutually exclusive

– explanation for excess homogenization could be an excess of nonsynonymous point mutations that become fixed. These competing explanations can be evaluated by fitting models with or without IGC and with or without special treatment of homogenizing nonsynonymous substitutions.

We therefore generalize the codon substitution rates of Equation 3.1 by having the amino acid encoded by a codon in one paralog influence the nonsynonymous rates of the corresponding codon triplet in the other paralog. Specifically, we replace the ω parameter by ω_H when the nonsynonymous change homogenizes amino acids that are encoded by corresponding codons in the two paralogs. Similarly, we replace ω by ω_N when the nonsynonymous change is non-homogenizing.

As with the Ji et al. (2016) model, this model has synonymous changes due to IGC occur at rate τ. Because all IGC events homogenize paralogs, this model has rate $\omega_H\tau$ for all nonsynonymous IGC change. The rates of codon substitution that originate with a point mutation become

$$Q_{i,j} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ u\pi_h\kappa & \text{for a synonymous transition} \\ u\pi_h\omega_H & \text{for a nonsynonymous homogenizing transversion} \\ u\pi_h\omega_N & \text{for a nonsynonymous non-homogenizing transversion} \\ u\pi_h\kappa\omega_H & \text{for a nonsynonymous homogenizing transition} \\ u\pi_h\kappa\omega_N & \text{for a nonsynonymous non-homogenizing transition.} \end{cases} \tag{3.3}$$

In Equation 3.3, note that ω_H is used for nonsynonymous changes that cause the nucleotide triplets in corresponding codon positions of the two paralogs to encode the same amino acid even when the triplets differ at the nucleotide level. The value of u is again set so that the expected rate of substitution per codon equals one in the absence of IGC (i.e., when all substitutions are attributable to point mutation). However, the rate of change for this model at a codon position in one paralog depends on the encoded amino acid at the corresponding position in the other paralog. For the standard genetic code, normalizing to find the value of u that makes the expected rate equal to one involves considering a rate matrix with 61^2 rows and columns.

When considering codon substitutions that originate with a point mutation, we denote the rates of Equation 3.3 as the ω_H/ω_N model. The special case of $\omega_H = \omega_N$ that yields the rates of Equation 3.1 is denoted as the ω model. Because these two treatments of point mutation can each be adopted with or without IGC, the four model possibilities are termed: $\omega_H/\omega_N + IGC$, $\omega_H/\omega_N - IGC$, $\omega + IGC$, and $\omega - IGC$.

3.4 Results

3.4.1 Duplicated teleost genes

Most teleost species are descended from a whole-genome duplication that occurred between 300 and 400 million years ago (e.g., Christoffels et al. 2004; Vandepoele et al. 2004; Crow et al. 2006). Conant (2020) inferred syntenic relationships between the genomes of 8 post-duplication teleosts and an outgroup. The species tree relating the taxa from that study is depicted in Figure 3.1.

Here, we characterize IGC that has occurred subsequent to this teleost genome duplication (TGD). We relied upon the orthology/paralogy inferences of Conant (2020) to

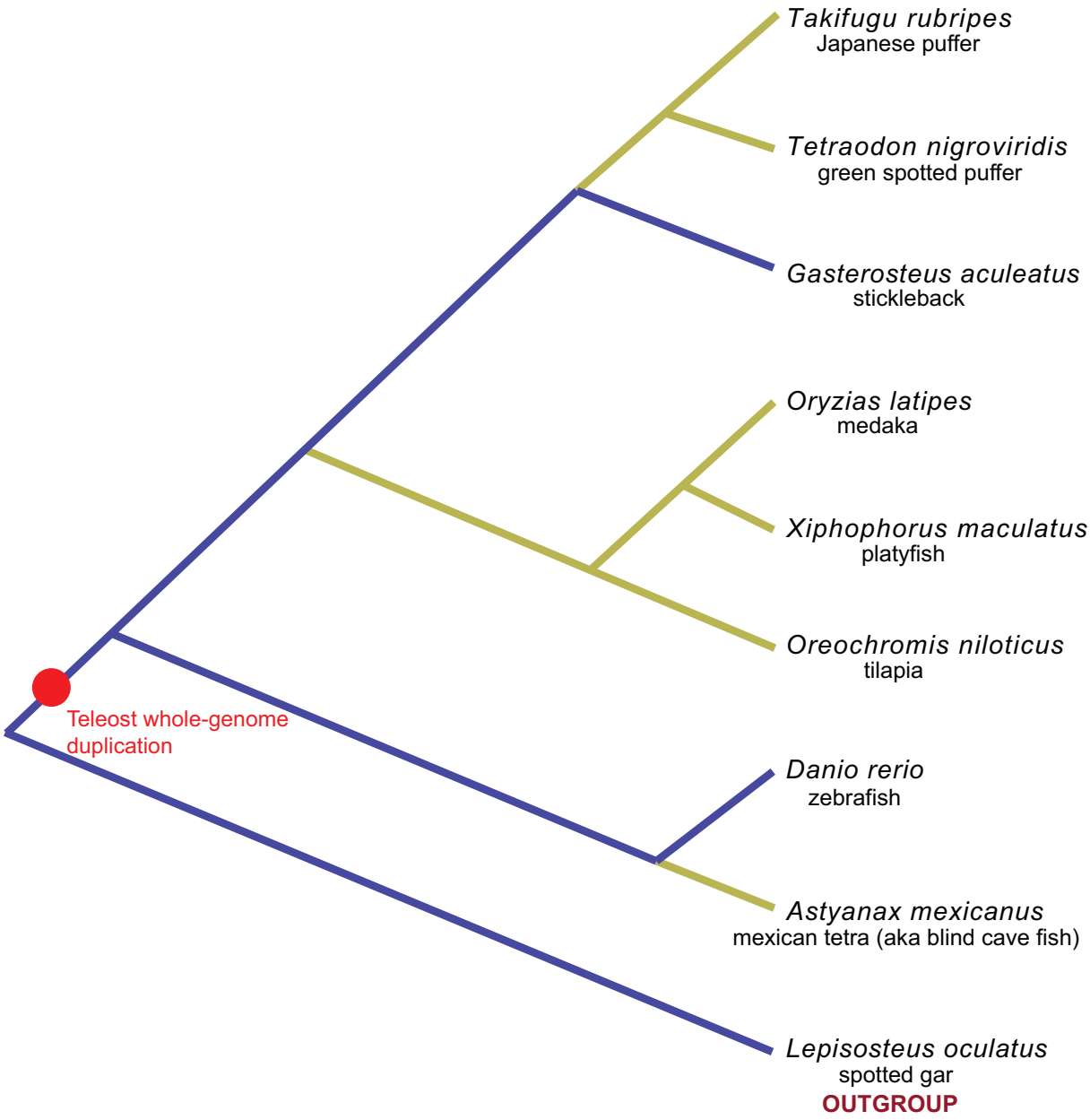


Figure 3.1: Species tree of the teleosts in this study. For the subtree shown in blue, all 164 data sets have representative sequences (i.e., two paralogs each from zebrafish and stickleback plus one sequence from the outgroup gar). For the post-duplication taxa that are connected to the subtree via branches colored yellow, some data sets include two paralogs and others do not include any.

collect 164 data sets of duplicated and aligned protein-coding genes. All 164 data sets have a sequence from the outgroup taxon. For 37 of the 164 data sets, all 8 post-TGD taxa have both paralog sequences. The remaining 127 data sets do not include both paralogs from all 8 post-TGD taxa, presumably because of deletion of one or both paralogs subsequent to the TGD event. These 127 data sets were selected based on two paralogs being available from both the stickleback and zebrafish genomes. For each of the other 6 post-TGD taxa considered by Conant (2020), sequences from that species are included in a data set only when two paralog sequences are available. Further details and justification regarding our procedure for collecting the data sets can be found in **Materials and Methods**.

3.4.2 Duplicated yeast genes

Ji et al. (2016) quantified the evolutionary impact of IGC in 14 groups of yeast ribosomal protein-coding genes that resulted from an ancient genome duplication. These data sets each consist of a single sequence from an outgroup taxon as well as two paralogs from each of 6 yeast species that are descended from the ancient whole-genome duplication event. Here, we analyze the 14 data sets from Ji et al. (2016) with the ω_H/ω_N model. We contrast the yeast results to those from the teleosts.

3.4.3 Model comparison

We can assess four different models with the most flexible being the $\omega_H/\omega_N + IGC$ model and the simplest being the $\omega - IGC$ model. The $\omega_H/\omega_N - IGC$ and $\omega + IGC$ models represent intermediates between these extremes. Likelihood ratio tests that use twice the difference between maximum log-likelihood values as a test statistic are available for comparing different pairs of these models.

To assess a null hypothesis that homogenizing and non-homogenizing nonsynonymous

rates are equal (i.e., $\omega_H = \omega_N$) versus an alternative without the equality constraint, the null distribution of the likelihood ratio test statistic is approximately (i.e., asymptotically) a χ^2_1 random variable when codon positions are assumed to independently evolve. For the case of testing the $\omega + IGC$ model as the null hypothesis and the $\omega_H/\omega_N + IGC$ model as the alternative, the χ^2_1 test statistic would not be appropriate when recombination is low enough to mean that our inferences under the null hypothesis need to be considered as maximum composite likelihood estimates rather than maximum likelihood estimates. Here, we assume that recombination rates are sufficiently high.

To assess the null hypothesis that $\tau = 0$ versus the alternative that τ is free to vary, the fact that $\tau = 0$ is at the boundary of the parameter space should be considered. In this case, the null distribution of the test statistic is approximately an equiprobable mixture of χ^2_0 and χ^2_1 distributions where the χ^2_0 distribution has value 0 with probability 1 (Self and Liang 1987; Goldman and Whelan 2000; Ota et al. 2000). To compare the null hypothesis that $\tau = 0$ and also $\omega_H = \omega_N$ versus an alternative that has neither of these constraints, the appropriate test statistic is an equiprobable mixture of χ^2_1 and χ^2_2 random variables.

One way to decompose the test statistic for comparing the $\omega - IGC$ and $\omega_H/\omega_N + IGC$ models is as the sum of the test statistic for comparing the $\omega - IGC$ and $\omega + IGC$ models and the test statistic for comparing the $\omega + IGC$ and $\omega_H/\omega_N + IGC$ models. This decomposition represents a path from the simplest ($\omega - IGC$) model to the most flexible model ($\omega_H/\omega_N + IGC$) by first adding IGC to the simplest model to get the intermediate $\omega + IGC$ model and then distinguishing between homogenizing and non-homogenizing nonsynonymous change to obtain the $\omega_H/\omega_N + IGC$ model. An alternative decomposition for comparing the $\omega - IGC$ and $\omega_H/\omega_N + IGC$ models is as the sum of the test statistic for comparing the $\omega - IGC$ and $\omega_H/\omega_N - IGC$ models and the test statistic for comparing the $\omega_H/\omega_N - IGC$ and $\omega_H/\omega_N + IGC$ models. This alternative decomposition represents a path from the simplest to the most flexible model by first distinguishing between the two

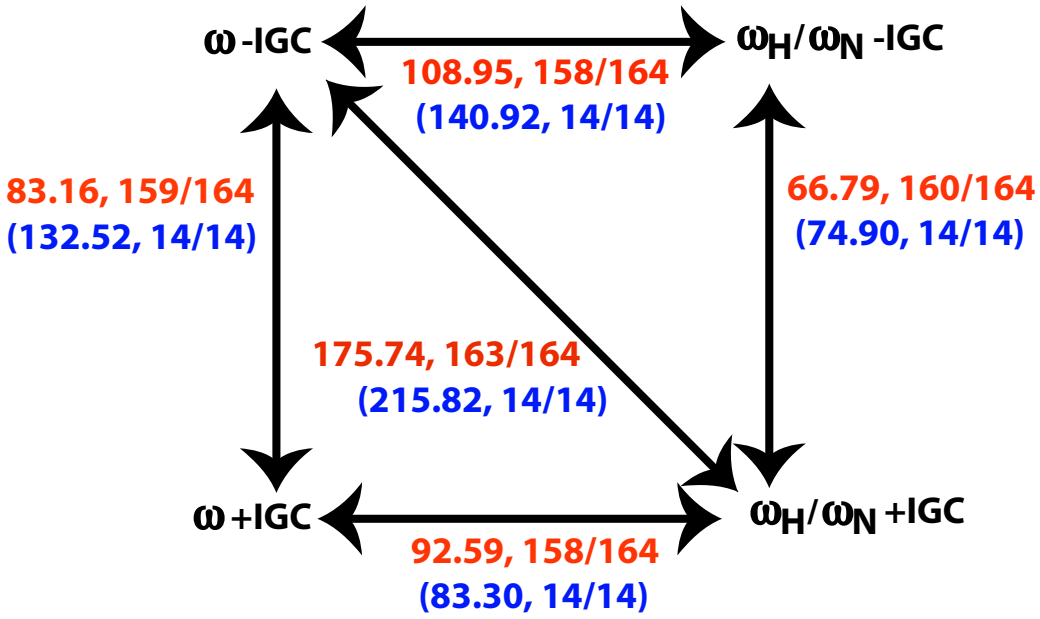


Figure 3.2: Bidirectional arrows indicate model comparisons that were performed. For each comparison, the top line summarizes results from the 164 teleost data sets and the parenthesized bottom line has results from the 14 yeast ribosomal protein-coding data sets. Each line contains the sample mean among the data sets of the test statistic (i.e., twice the log-likelihood differences between models) followed by the proportion of the data sets for which the null hypothesis was rejected at a significance level of 0.05.

sorts of nonsynonymous change to yield the intermediate $\omega_H/\omega_N - IGC$ model and then adding IGC to obtain the $\omega_H/\omega_N + IGC$ model.

Figure 3.2 summarizes the results of hypothesis tests between our models for the teleost and yeast data. The mean log-likelihood improvement among data sets is substantial for each step in each of the two paths for decomposing the comparison of the $\omega - IGC$ and $\omega_H/\omega_N + IGC$ models via an intermediate model (Figure 3.2). The substantial improvements arise both when IGC is first added to the simplest model and when the differentiation between nonsynonymous substitution types is first added. These results suggest that IGC and differentiation of nonsynonymous substitution types are not redundant and that both features are important.

While Figure 3.2 summarizes the mean behavior among data sets of test statistics, Figure 3.3 displays the behavior for individual data sets. When IGC is added to the $\omega - IGC$ model to produce the intermediate $\omega + IGC$ model, the improvement in log-likelihood tends to be big but tends to represent only a moderate proportion of the total improvement in log-likelihood between the $\omega - IGC$ and $\omega_H/\omega_N + IGC$ models (Figure 3.3A). Similarly, when differentiating between the two types of nonsynonymous change is added to the $\omega - IGC$ model to produce the intermediate $\omega_H/\omega_N - IGC$ model, the improvement in log-likelihood tends to be big but again tends to represent only a moderate proportion of the total improvement in log-likelihood between the $\omega - IGC$ and $\omega_H/\omega_N + IGC$ models (Figure 3.3B).

3.4.4 IGC proportions

As explained in Ji et al. (2016), our phylogenetic approach has difficulty in extracting IGC information from branches separating a duplication event from the first post-duplication speciation. It also has difficulty estimating the expected number of codon substitutions

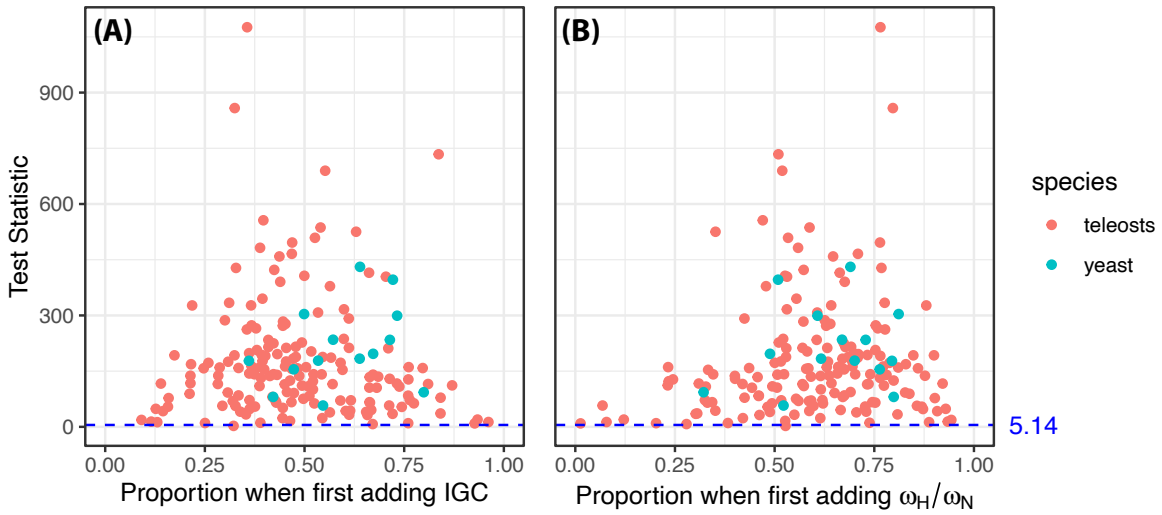


Figure 3.3: Likelihood ratio test statistics for the teleost and yeast data sets when comparing the ω - IGC model to intermediate models versus when comparing the ω - IGC model to the ω_H/ω_N + IGC model. The x-axes represent ratios of likelihood ratio test statistics (i.e., twice the difference between the maximum log-likelihood of the alternative and null hypotheses). The numerator of these ratios is the test statistic when the null hypothesis is the ω - IGC model and the alternative is an intermediate model with one additional free parameter. The denominator of these ratios is the test statistic when the null hypothesis is the ω - IGC model and the alternative is the ω_H/ω_N + IGC model that has two additional free parameters. The y-axes represent the test statistic in the denominator of the ratio (i.e., the null hypothesis is the ω - IGC model and the alternative is the ω_H/ω_N + IGC model). For the test statistics displayed on the y-axes, horizontal lines indicate the critical value of $y = 5.14$ at the 0.05 significance level. (A) The intermediate model is $\omega + IGC$. (B) The intermediate model is ω_H/ω_N - IGC.

that occurred on these branches. As demonstrated by simulation, evolutionary inferences regarding these initial post-duplication branches tend to be overly sensitive to model assumptions. When we estimate the proportion of codon substitutions that arise via IGC rather than point mutation, we therefore only consider the branches on the species tree that occur subsequent to the first post-duplication speciation.

As shown in Ji et al. (2016), these proportions can be inferred via the "matrix of exponentials" technique of Tataru and Hobolth (2011). Using the Tataru-Hobolth technique on each of our 164 teleost data sets, the estimated IGC proportions when assuming the $\omega+IGC$ model range among data sets from slightly less than 0.02 to slightly more than 0.22 with a median value of about 0.08. Figure 3.4 depicts the distribution of these estimated proportions. The estimated proportion of codon substitutions due to IGC rather than point mutation is relatively insensitive to whether the $\omega+IGC$ model or the ω_H/ω_N+IGC model is assumed. For example, the median difference in these proportions among the 164 teleost data sets has the proportion from the ω_H/ω_N+IGC model being about 0.005 lower than from the $\omega+IGC$ model. For the teleost genes when assuming the ω_H/ω_N+IGC model, the estimated proportions range from slightly more than 0.01 to slightly less than 0.22 with a median value between 0.07 and 0.08.

3.4.5 Estimates of ω_H and ω_N

Figure 3.5 plots estimates of ω_N versus ω_H from the 164 teleost data sets when assuming the ω_H/ω_N+IGC model. The ω_N estimates were all between approximately 0.01 and 0.24 with a median of about 0.08. For 159 of the 164 teleost data sets, the ω_H and ω_N estimates had $0<\omega_N<\omega_H<1$. One data set violated this pattern in that the estimated ω_N was small and the estimated ω_H was even closer to 0. The remaining four data sets violated the pattern in that each yielded an ω_H estimate that was above 1.0, with all 4 estimates being

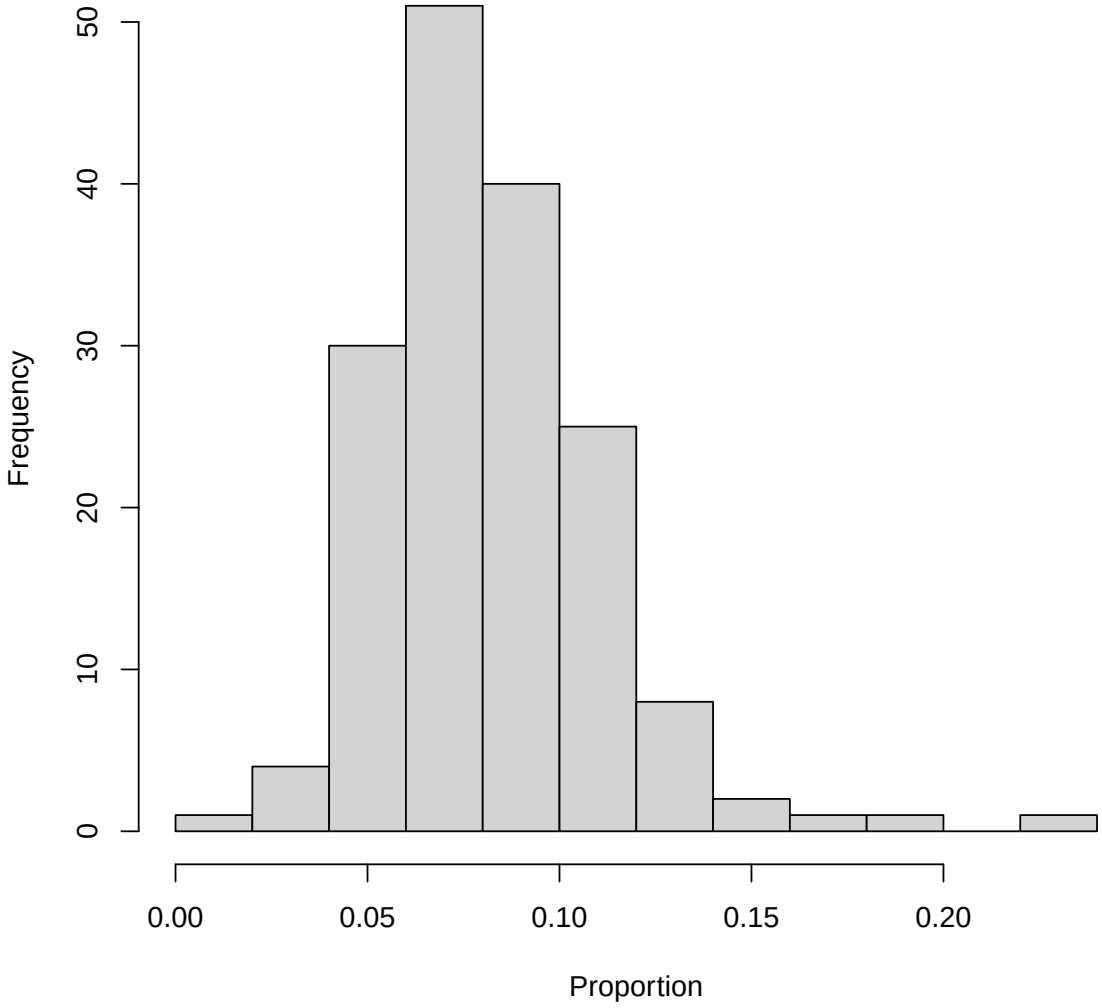


Figure 3.4: A histogram of the estimated proportions of codon substitutions that are due to IGC for the 164 teleost data sets when the $\omega + IGC$ model is assumed.

between 1.0 and 1.4.

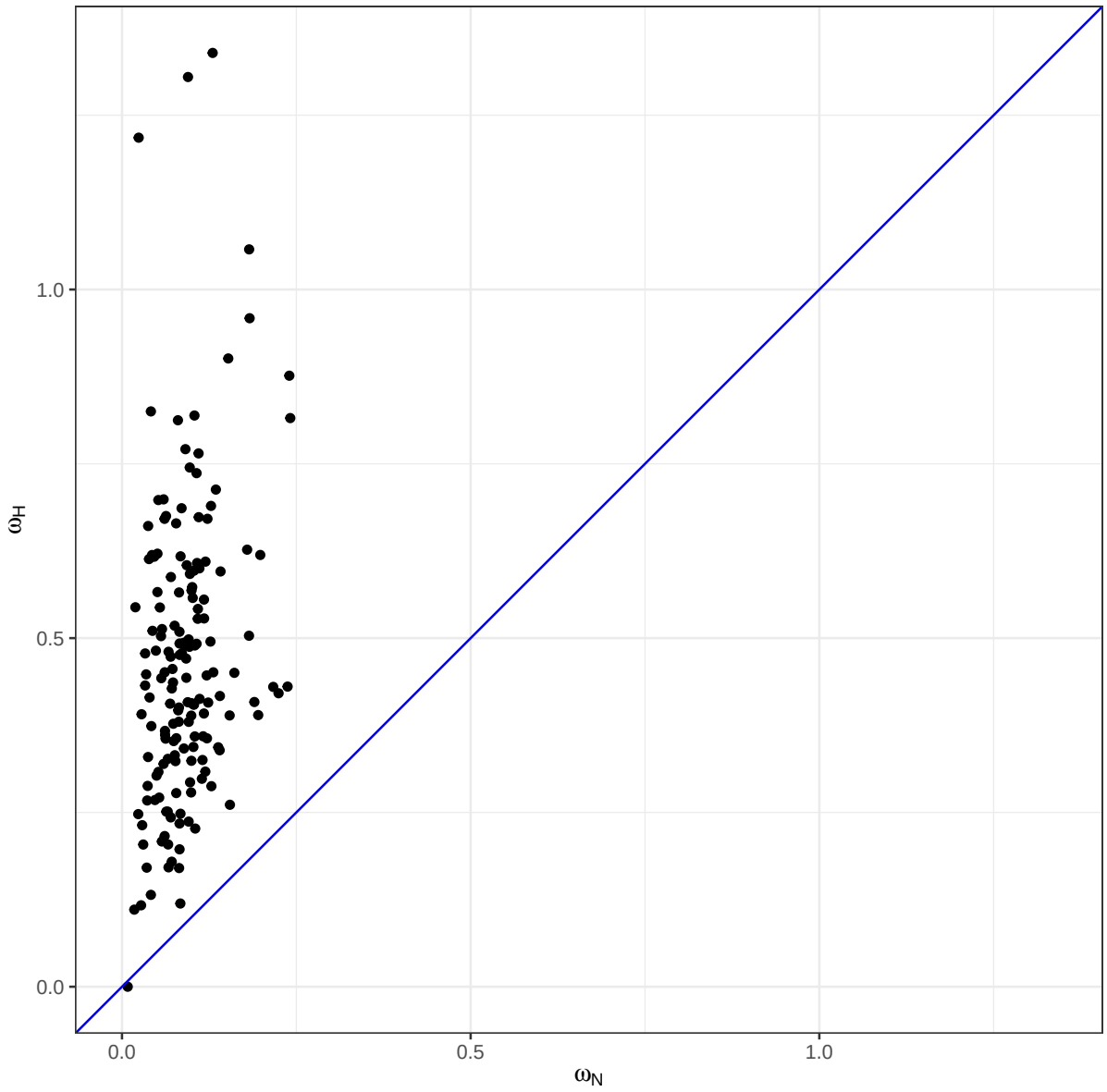


Figure 3.5: A plot of the estimated ω_N (x-axis) and ω_H (y-axis) values for the 164 teleost data sets when assuming the $\omega_H/\omega_N + IGC$ model. The diagonal line represents $y = x$.

In contrast, the 14 yeast data sets all have estimates with the $\omega_H/\omega_N + IGC$ model that

satisfy $0 < \omega_N < 1 < \omega_H$. The yeast ω_N estimates range from approximately 0.02 to 0.17 with a median of 0.08. The smallest of the yeast ω_H estimates is 1.36, a value that slightly exceeds the largest of the 164 ω_H estimates from teleosts. The median ω_H estimate from yeast is 5.70. Two of the 14 ω_H estimates are ∞. For these two data sets, the encoded amino acids of the two paralogs are identical within species but different between species. However, paralogs for these two data sets do have some within-species synonymous variation.

The ω_N estimates are not greatly affected by whether the $\omega_H/\omega_N + IGC$ or the $\omega_H/\omega_N - IGC$ model is assumed, with the yeast data sets showing more sensitivity than the teleost data sets. For the teleosts, subtracting the ω_N estimate of the $\omega_H/\omega_N - IGC$ model from that of the $\omega_H/\omega_N + IGC$ model yields differences that range from approximately -0.007 to 0.014 with a median that is very close to 0. For the yeast data sets, 13 of the 14 ω_N differences are positive but all are still relatively close to 0. Specifically, the ω_N differences for yeast range from about -0.015 to 0.030 with a median of 0.008. Similarly, when the ω_N estimate from the $\omega_H/\omega_N - IGC$ model is divided by the ω_N estimate from the $\omega_H/\omega_N + IGC$ model, the median ratio is approximately 0.997 among the teleost data sets and 0.864 among the yeast data sets.

The ω_H estimates are quite sensitive to whether the $\omega_H/\omega_N + IGC$ or the $\omega_H/\omega_N - IGC$ model is assumed. For the teleost data sets, Figure 3.6 plots the ω_H estimates from the $\omega_H/\omega_N - IGC$ model versus the estimates from the $\omega_H/\omega_N + IGC$ model. In 158 of the 164 cases, the ω_H estimates are larger from the $\omega_H/\omega_N - IGC$ model. Figure 3.6 shows that the disparity in ω_H estimates grows as the estimated values of ω_H from the $\omega_H/\omega_N - IGC$ model get larger. For the yeast data sets, the ω_H estimates from the $\omega_H/\omega_N - IGC$ model are especially large and are presumably too large to be stable estimates. For the $\omega_H/\omega_N - IGC$ case, our results had all 14 yeast ω_H values exceeding 5 and had 12 of the 14 estimates exceeding 20.

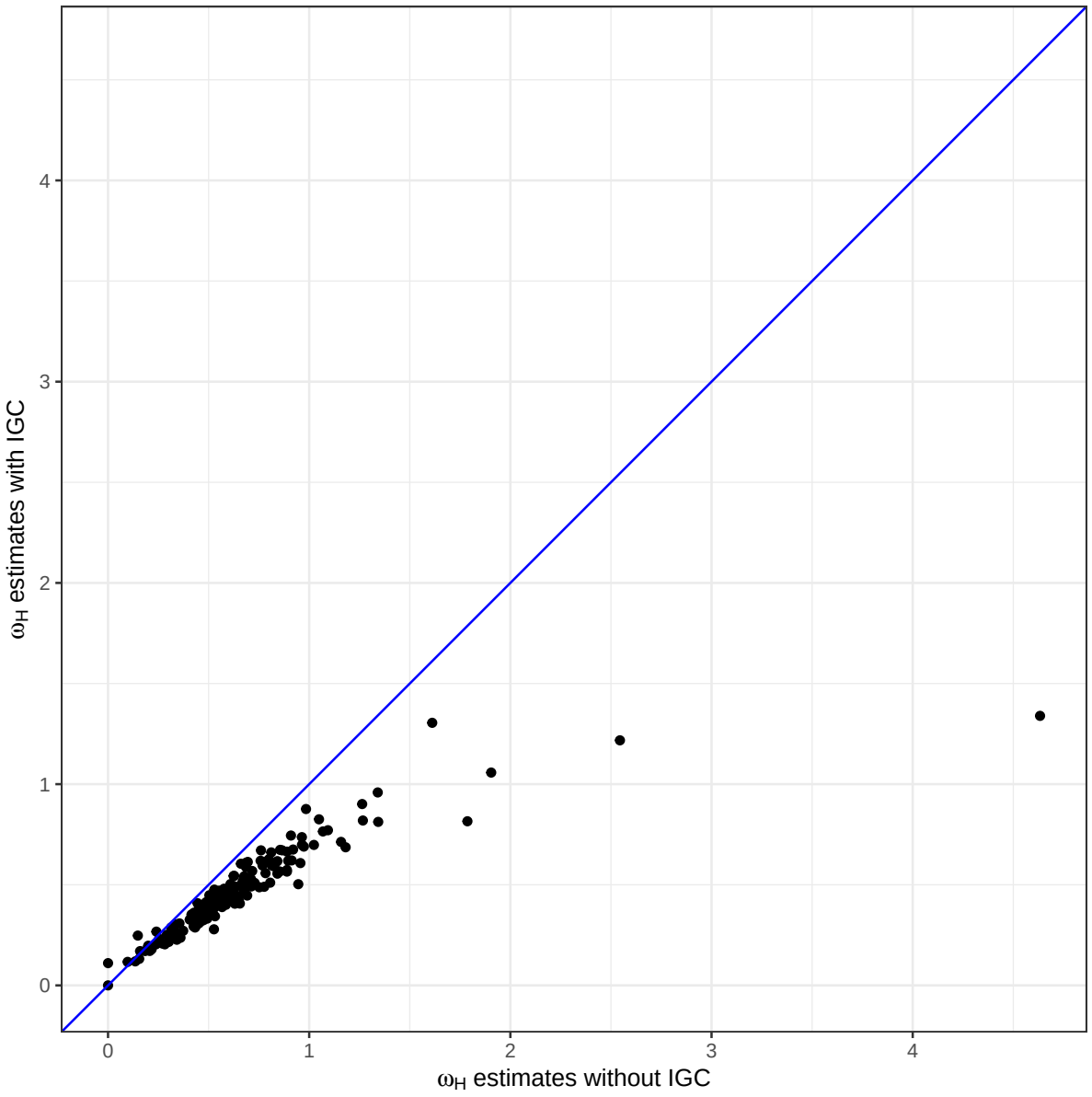


Figure 3.6: The estimates of ω_H from the 164 teleost data sets when assuming the $\omega_H/\omega_N - IGC$ model (x-axis) versus the estimates of ω_H when assuming the $\omega_H/\omega_N + IGC$ model (y-axis).

3.5 Discussion

3.5.1 Estimates of IGC and nonsynonymous rates

There is strong evidence for IGC in nearly all teleost data sets. This evidence is compelling whether or not homogenizing and non-homogenizing nonsynonymous changes are separately treated. While it is unclear whether teleost genes experienced enough interlocus gene conversion to have an important influence on phylogeny inference or divergence time estimation, we view the quantification of interlocus gene conversion as being fundamentally necessary for the study of molecular evolution.

The estimated proportion of codon substitutions from IGC is relatively robust to how we model nonsynonymous rates. We believe this robustness is explained by much of the information about IGC coming from homogenizing synonymous substitutions. However, the estimates of ω_H are sensitive to whether or not IGC is allowed. Presumably, this sensitivity arises because IGC is responsible for some of the homogenizing nonsynonymous changes and those can only be explained by a large value of ω_H when IGC is ignored.

Except for 5 of the 164 teleost data sets, our parameter estimates have $0 < \omega_N < \omega_H < 1$. This means nonsynonymous rates are less than synonymous rates whether or not they are homogenizing or non-homogenizing. The fact that ω_N and ω_H differ establishes that the evolution of one paralog can illuminate the evolution of another.

For yeast ribosomal protein-coding genes, all 14 data sets yielded $0 < \omega_N < 1 < \omega_H$. This could imply that there is strong selection for the yeast paralogs to converge at the protein sequence level. However, the relative values of ω_H and ω_N should be cautiously interpreted. Multiple explanations for inferring $\omega_H > \omega_N$ are possible. For example, the ω_H/ω_N model does not explicitly incorporate variation of preferred amino acid types among protein positions. The CAT model does explicitly incorporate such variation and can substantially

improve model fit (Lartillot and Philippe 2004). As an alternative explanation, a finding that $\omega_H > \omega_N$ would be consistent with there being a correlation between corresponding paralog positions of preferred amino acid types.

These alternative explanations are not mutually exclusive. Although less parametric strategies would also be attractive, a direct way to disentangle these possibilities would be to explicitly incorporate variation of preferred amino acids among protein positions as was done in the pioneering work of Lartillot and Philippe (2004). One can envision developing model-based strategies that distinguish between homogenizing and non-homogenizing nonsynonymous changes when attempting to characterize paralog evolution with respect to non-functionalization, neofunctionalization, subfunctionalization, or functional stasis. The basic idea would be that existing and widely-used strategies for parameterizing nonsynonymous rates could be supplemented by explicitly including homogenizing nonsynonymous substitution rates that vary among sites and/or lineages.

3.5.2 Tetrasomic inheritance versus IGC

An evolutionary genomic study by Parey et al. (2022) persuasively suggests that the whole genome duplication in teleosts was an autopolyploid event that was followed by a period of tetrasomic inheritance and then rediploidization to yield the paralogs that are found in extant genomes. The Parey et al. (2022) work further indicates that this rediploidization occurred at different times in different portions of ancestral teleost genomes. Importantly, the Parey et al. (2022) evidence for tetrasomic inheritance does not extend as late as the first post-duplication speciation event relating the taxa that we studied. Because our phylogenetic approach for quantifying IGC is only able to infer IGC that occurs subsequent to the first post-duplication speciation among the taxa in a sample (Ji et al. 2016), the tetrasomic inheritance detected by Parey et al. (2022) does not explain our results. A hypothesis that

has tetrasomic inheritance extended beyond the periods detected by Parey et al. (2022) and beyond the first post-duplication speciation among our taxa would face the challenge of explaining why there seems to be a non-negligible IGC impact for nearly all of the 164 teleost data sets, even though these data sets represent diverse genomic regions.

3.5.3 Future directions for studying IGC

Techniques for studying the intersection of IGC and evolution remain primitive. The approach used here can be classified as a composite likelihood procedure unless recombination rates are sufficiently high. Ideally, the approach would not ignore the fact that IGC mutations affect stretches of consecutive sequence positions. Although the approach yielded reasonable parameter estimates even when data were simulated by having IGC events affect sequence tracts (Ji et al. 2016), it can be challenging to measure the uncertainty of parameters estimated with composite likelihood approaches (e.g., Varin et al. 2011). Better handling of IGC tracts is needed.

Furthermore, the IGC inference approach is computationally challenging because the joint consideration of paralogs greatly increases the state space associated with matrices of evolutionary rates. This is the reason why the analyses performed here have been restricted to two paralogs per genome. Monte Carlo data augmentation strategies are one possibility for studying molecular evolution when more than two paralogs per genome are considered.

Another attractive research direction would account for the relationship between IGC rates and paralog divergence. Gene conversion rates decrease as paralog divergence increases (e.g., Chen et al. 2007), but a refined characterization of this relationship is unavailable and should be pursued. Finally, it is unclear whether the evolutionary impacts of IGC have taxonomic-specific patterns and whether the impact is different for segmental and whole genome duplications.

3.6 Materials and methods

3.6.1 Teleost data set collection

Conant (2020) employed the POInT software (Conant and Wolfe 2008; Emery et al. 2018) to estimate orthology/paralogy relationships among protein-coding genes in the genomes of 8 representative species that are descended from the TGD. Conant (2020) linked these paralog sets to genes from the spotted gar (*Lepisosteus oculatus*), a representative of an outgroup lineage that diverged from other teleosts prior to the whole-genome duplication. The POInT software assigns confidence scores to its best estimate of orthology/paralogy relationships at homologous genomic loci that are descended from a whole genome duplication. For our teleost analyses, we considered only protein-coding loci with a confidence score that equalled or exceeded 95%.

The implementation of our IGC model does not account for the possibility of paralog deletion following duplication. Proper handling would include the possible influence via IGC of deleted paralogs on surviving ones. For the data sets that we analyzed, we therefore only included the subset of post-duplication taxa for which both paralogs were present.

We further restricted our investigations to cases with a single identifiable homolog in the outgroup gar genome and with two paralogs in both the post-duplication zebrafish and the post-duplication three–spined stickleback. This requirement was instituted because zebrafish and stickleback are representatives of the earliest post-duplication speciation event on the phylogeny relating our taxa (see Figure 3.1). Because our phylogenetic approach relies on evolution subsequent to the earliest post-duplication speciation event to characterize IGC, we did not want to include data sets where the earliest post-duplication speciation was later than the zebrafish-stickleback split.

Each protein-coding DNA data set was further processed by translating the sequences

and aligning the resulting amino acid sequences with Version 13.45.0.4846264 of the T-Coffee software (Notredame et al. 2000). Amino acids in the aligned data were then replaced with the corresponding codon triplets. All alignment columns with at least one gapped position were removed prior to subsequent analysis.

We eliminated potentially problematic data sets by establishing two additional criteria. Both criteria involved maximum likelihood estimation with the $\omega - IGC$ model. The first criterion eliminated data sets with at least one overly long estimated branch length. Our motivation was that long branch length estimates could stem from sequencing or assembly error, alignment mistakes, and/or paralogy/orthology misidentification. Specifically, a data set was removed from further consideration if any estimated branch lengths exceeded 1.5 codon substitutions per codon.

The second criterion only involved 5 sequences from each of the 164 teleost data sets. Because all data sets include 2 paralogs from both stickleback and zebrafish as well as one sequence from gar, we compared the maximum log-likelihood value for these 5 sequences being related via the paralogy/ortholog relationships inferred by Conant (2020) with the maximum log-likelihood value when the orthology/paralogy relationships between the zebrafish and stickleback paralog pairs are switched from those inferred by Conant (2020). Because it suggests problematic orthology/paralogy relationships, data sets were eliminated if the latter maximum likelihood value exceeded the former.

3.6.2 The ω_H/ω_N model for lineages that are not post-duplication

When analyzing the yeast and teleost data sets with the ω_H/ω_N model, a choice has to be made about how to model the nonsynonymous rates prior to the duplication and also on the lineage to the outgroup. While other choices are possible, we chose to have ω_N be the nonsynonymous rate factor for these evolutionary intervals. Also, we normalized the rates

for these evolutionary intervals to have one expected codon substitution arising from point mutation per time unit.

3.7 Data availability

Software and instructions for performing the IGC analyses are available at: https://github.com/xji3/IGCexpansion. Teleost and yeast data sets as well as output from IGC analyses are available at: https://github.com/Yixuan39/IGC-fish.

3.8 Acknowledgments

Y.Y., T.X., and J.L.T. were supported by N.S.F. DEB-1754142. G.C. was supported by N.S.F. DEB-2241312. H.K. was supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research 22K11950. X.J. acknowledges support by the NVIDIA Corporation. T.X. served as the second author of this paper, contributing to the study's approach, coding, and data analysis alongside the first author.

References

Abbasi, A. A. (2010). Piecemeal or big bangs: correlating the vertebrate evolution with proposed models of gene expansion events. *Nature Reviews Genetics*, 11(2):166.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775.

Christoffels, A., Koh, E. G., Chia, J.-m., Brenner, S., Aparicio, S., and Venkatesh, B. (2004). *Fugu* genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Molecular Biology and Evolution*, 21(6):1146–1151.

Conant, G. C. (2020). The lasting after-effects of an ancient polyploidy on the genomes of teleosts. *PLoS One*, 15(4):e0231356.

Conant, G. C. and Wolfe, K. H. (2008). Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics*, 179(3):1681–1692.

Crow, K. D., Stadler, P. F., Lynch, V. J., Amemiya, C., and Wagner, G. P. (2006). The "fish-specific" hox cluster duplication is coincident with the origin of teleosts. *Molecular Biology and Evolution*, 23(1):121–136.

Emery, M., Willis, M. M. S., Hao, Y., Barry, K., Oakgrove, K., Peng, Y., Schmutz, J., Lyons, E., Pires, J. C., Edger, P. P., et al. (2018). Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genetics*, 14(3):e1007267.

Evangelisti, A. M. and Conant, G. C. (2010). Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biology and Evolution*, 2:826–834.

Goldman, N. and Whelan, S. (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978.

Harpak, A., Lan, X., Gao, Z., and Pritchard, J. K. (2017). Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proceedings of the National Academy of Sciences*, 114(48):12779–12784.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22(2):160–174.

Hood, L., Campbell, J., and Elgin, S. (1975). The organization, expression, and evolution of antibody genes and other multigene families. *Annual Review of Genetics*, 9:305–353.

Ji, X. (2017). *Phylogenetic approaches for quantifying interlocus gene conversion.* PhD thesis, North Carolina State University.

Ji, X., Griffing, A., and Thorne, J. L. (2016). A phylogenetic approach finds abundant interlocus gene conversion in yeast. *Molecular Biology and Evolution*, 33(9):2469–2476.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H., and Kishino, H. (2000). Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution*, 17(5):798–803.

Parey, E., Louis, A., Montfort, J., Guiguen, Y., Crollius, H. R., and Berthelot, C. (2022). An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication. *Genome Research*, 32(9):1685–1697.

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

Tartof, K. D. (1975). Redundant genes. *Annual Review of Genetics*, 9:355–385.

Tataru, P. and Hobolth, A. (2011). Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time markov chains. *BMC Bioinformatics*, 12:465.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10(10):725–732.

Van de Peer, Y., Maere, S., and Meyer, A. (2010). 2r or not 2r is not the question anymore. *Nature Reviews Genetics*, 11(2):166–166.

Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A., and Van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences*, 101(6):1638–1643.

Vanneste, K., Maere, S., and Van de Peer, Y. (2014). Tangled up in two: a burst of genome duplications at the end of the cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1648):20130353.

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

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

CHAPTER

4

THE RELATIONSHIP BETWEEN INTERLOCUS GENE CONVERSION AND PARALOG DIVERGENCE

4.1 Abstract

4.2 Introduction

Interlocus gene conversion (IGC) is a type of mutation that homogenizes paralogs by replacing a contiguous subsequence in one paralog with the corresponding stretch from another. Careful consideration has been devoted to how gene function is shaped by duplication mutations and subsequent point mutations (e.g., Ohno 1970; Force et al. 1999; Birchler and Veitia 2012). However, less consideration has been given to the role of IGC mutations. Variation between paralogs that has been introduced via point mutation can be erased by IGC and can also spread to additional gene family members by IGC. This interplay between duplication and IGC has often been ignored when the effects of duplication on the evolution of gene function have been studied. IGC can introduce genetic variation into one paralog that has already been exposed to natural selection when in the other paralog. Because fitness effects at corresponding sequence positions are likely to be correlated between paralogs, the distribution of fitness consequences of IGC mutations is likely to differ from the distribution of fitness consequences of new point mutations. Sometimes, IGC might introduce co-adapted positions from one paralog into another. These features make IGC mutations particularly interesting from an evolutionary perspective.

There does not appear to be a single molecular mechanism that can explain all gene conversion mutations (Yin et al. 2017), but it is clear that the rate of IGC mutations is negatively correlated with the divergence between paralogs (e.g., Mansai et al. 2011). The detailed nature of this negative correlation between paralog divergence and IGC mutation rate has not been well characterized. Similarly, the relationship between paralog divergence and the rate of IGC mutations that become fixed is understudied.

As a result, the evolutionary importance of IGC mutations is poorly understood. Nu-

cleotide substitutions that arise via point mutation tend to be modeled as occurring independently following the duplication event that gives rise to paralogs. Typically, nucleotide substitutions that originate with point mutations are assumed to be independent among positions within a paralog and independent between paralogs. A challenge in studying the evolutionary impact of IGC partly stems from the fact that characterizing IGC requires joint consideration of paralog sequences. When there are two paralogs that consist of L positions, the computational demand of allowing dependence amongst all positions is daunting because the the number of rows and columns in evolutionary rate matrices is 4^{2L} or 61^{2L}, depending on whether sequence positions are modelled as being occupied by one of 4 possible nucleotide types or by one of 61 possible codon triplets.

Ji et al. (2016) introduced a phylogenetic approach for modeling paralog evolution following duplication. With this approach, two paralogs that result from duplication initially have identical DNA sequences. These paralogs subsequently evolve via sequence changes that arise from point mutations and IGC events. Nucleotide or codon substitutions that originate with point mutations are modeled with conventional strategies that have independent changes among sequence positions and that have independent changes among paralogs. Ji et al. (2016) incorporate IGC by jointly considering the states at corresponding sequence sites in the paralogs, but they ignore the dependence that is attributable to single events simultaneously homogenizing multiple positions between the two paralogs.

According to the study by Harpak et al. (2017), as paralogs become more diverged, the IGC rate quickly approaches zero. Consequently, the researchers suggested that the quick decline in the IGC rate as paralogs diverge would minimize the evolutionary impact of IGC. However, the quantitative analysis of the relationship between IGC rate and paralog divergence level is not well-developed.

Here, we investigate the relationship between IGC rate and paralog divergence with both analytical and likelihood approaches. We begin by incorporating IGC and paralog

divergence into ordinary differential equations, using the Jukes-Cantor model as our basic setting (Jukes and Cantor 1969). We then solve the corresponding equations to examine some consequences of a connection between IGC and paralog divergence.

Next, we revisit the likelihood approach introduced by Ji et al. (2016) to estimate branch-specific IGC rates. We apply this approach to analyze a set of 14 yeast ribosomal protein-coding genes and 37 teleost protein-coding genes (Chapter 3). By doing so, we obtain simple estimates for the branch-specific IGC rates. These estimates are used in a simple nonparametric approach for assessing whether IGC and paralog divergence are related.

Due to limitations of the nonparametric approach, we then modify and apply a generalization of the Ji et al. (2016) model to IGC inference. Our analyses suggest that IGC rates quickly decrease as paralog identity decreases, but they also indicate that the evolutionary impact of IGC can be too substantial to ignore. We conclude this chapter with a brief discussion of how our approach could be improved and extended.

4.3 Behavior of a simple IGC model

The Jukes-Cantor (JC) model (Jukes and Cantor 1969) is the simplest model for nucleotide substitutions that originate with point mutation. It assumes all sites evolve independently and identically. When the expected substitution rate per site is scaled to 1, the JC model has the rate of change from nucleotide type i to nucleotide type j ($j \neq i$) be 1/3. When there are two paralogs and the JC model is employed to describe substitutions that originate with point mutations, a variety of treatments of IGC can be envisioned to jointly model the evolution of the two paralogs.

For a site where one paralog has nucleotide type i and the other site has j with $i \neq j$, the simple IGC treatment of Ji et al. (2016) would have IGC events homogenize both paralogs to i at rate τ and would have IGC events homogenize both paralogs to j also at rate τ.

Combining the JC model for substitutions that originate with point mutation with this simple IGC treatment, the joint behavior of a site for a pair of paralogs can then be studied with a 16-state continuous-time Markov process. When the nucleotide types at the site are identical, they become different at rate 2 because substitutions that originate with a point mutation occur at rate 1 in each paralog. When the nucleotide types at the site differ, they become homogenized at a rate $2\tau + 2/3$ because either of the paralogs can be an IGC donor at rate τ and because $1/3$ of substitutions that originate with a point mutation will homogenize according to the JC model.

We treat time 0 as the time of occurrence of the duplication event that forms the paralogs. We measure times $t > 0$ as the expected number of substitutions that originate with a point mutation per site per paralog. With the simple IGC treatment, the expected proportion $I(t)$ of sites with identical nucleotide types becomes

$$\frac{dI(t)}{dt} = 2\{1 - I(t)\}\tau - 2I(t) + \frac{2}{3}\{1 - I(t)\}. \tag{4.1}$$

At stationarity, this simple IGC model has

$$I(\infty) = \frac{1+3\tau}{4+3\tau}.$$

Figure 4.1 illustrates the relationship between paralog identities and time that results from this simple IGC treatment. When there is no IGC, the JC model has a stationary distribution with an expected proportion of $1/4$ of all sites being identical between the two paralogs. When τ approaches ∞, the proportion of identical sites between the paralogs approaches 1, and gene duplicates do not diverge. With this simple IGC treatment, the homogenization effect of IGC increases the expected paralog identity.

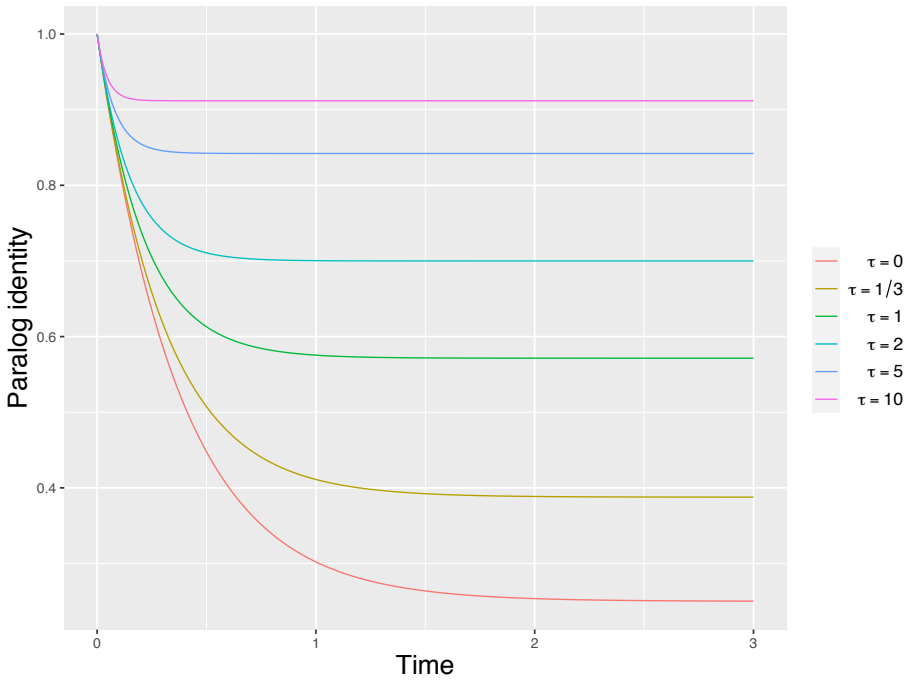


Figure 4.1: The expected proportion of sites that differ between paralogs is plotted relative to time (i.e., the expected number per paralog position of post-duplication nucleotide substitutions that originated with point mutation). The Jukes-Cantor model is employed to describe the substitutions that originate with point mutations and different colors represent different values of τ.

Harpak et al. (2017) carried out simulations to investigate the relationship between paralog divergence and IGC rate. They considered two types of paralog-identity thresholds H $(0 \leq H \leq 1)$. One type of threshold is a "global" threshold where IGC occurs at a fixed expected rate if the paralog identity exceeds H and where no IGC occurs if the paralog identity is below the threshold. If we modify the previous simple Jukes-Cantor treatment of IGC to incorporate this global threshold scenario, the stationary state is

$$I(\infty) = \begin{cases} \frac{1}{4} & H > \frac{1+3\tau}{4+3\tau,} \\ \\ \frac{1+3\tau}{4+3\tau} & \text{otherwise.} \end{cases} \tag{4.2}$$

The discreteness of the stationary state arises from the threshold's cutoff level.

4.3.1 Divergence model

A threshold treatment represents a rather extreme way to model the relationship between paralog divergence and IGC rates. Instead of the threshold model, a smoother and decreasing function of paralog identity can model IGC rates. One model has the rate at which a site with nucleotide types i and j ($i \neq j$) become homogenized being

$$2\tau_D I(t)^K, \tag{4.3}$$

where the factor of 2 is attributable to symmetrical handling of the two paralogs with respect to IGC. Specifically, the factor corresponds to an assumption that the two paralogs are equally likely to serve as IGC donor and IGC recipient (i.e., IGC is equally likely to homogenize a site where one paralog has nucleotide type i and the other has type j to yield both paralogs with type i as it is to yield both paralogs with type j). The case of $K = 0$ corresponds to the simple model that has IGC rates be independent of paralog divergence. Biologically plausible values of K would be positive, resulting in IGC rates decreasing as paralog identity decreases. The biological intuition for $K > 0$ is that IGC mutations are thought to involve a DNA strand from one paralog displacing the corresponding DNA strand from another paralog (e.g., Chen et al. 2007). The chance of strand displacement increases with the amount of nucleotide identity between the paralogs.

With this model, a Jukes-Cantor treatment of point mutation yields

$$\frac{dI(t)}{dt} = 2\{1-I(t)\}\tau_D I(t)^K - 2I(t) + \frac{2}{3}\{1-I(t)\}, \tag{4.4}$$

and the stationary proportion satisfies

$$I(\infty) = \frac{1 + 3\tau_D I(\infty)^K}{4 + 3\tau_D I(\infty)^K}.$$

If $I(t) = 1$, the right side of Equation 4.4 is less than zero. In contrast, the right side of Equation 4.4 is greater than zero if $I(t) = 0$. Therefore, a continuous function has at least one root for $I(\infty)$ in the interval $(0, 1)$, although there may be multiple roots for the corresponding equation. However, only one root corresponds to the equilibrium state $I(\infty)$, which is the first root after $I(0) = 1$. Figure 4.2 depicts the distribution of $I(\infty)$ for various τ and K values. As τ increases or K decreases, $I(\infty)$ approaches one.

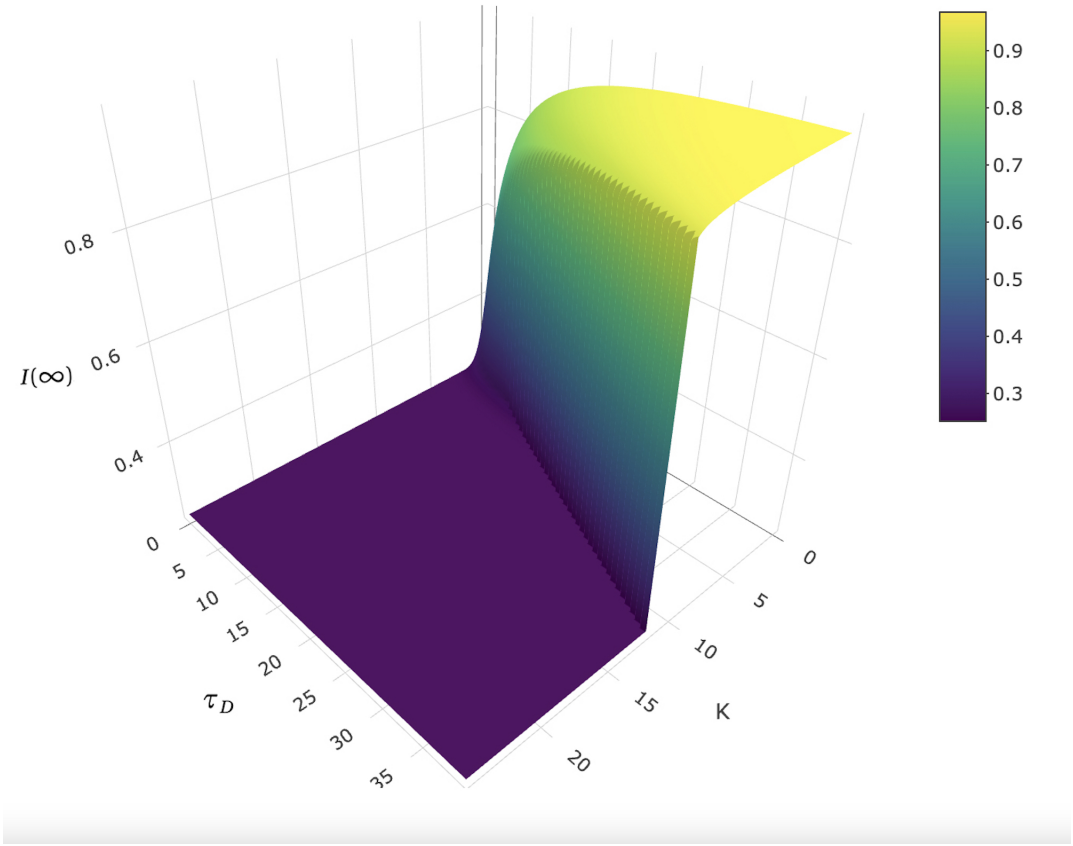


Figure 4.2: Equilibrium distribution for paralog identities at stationarity (i.e., $I(\infty)$). The x-axis represents the K value, the y-axis represents the τ_D value, and the z-axis represents the $I(\infty)$ value.

Sequence changes can be categorized as either point mutations or IGC mutations. If the proportion of change from IGC at time t is significant, then IGC is important. The relationship between the parameter $I(t)$ and its corresponding proportion $C(t)$ is described for $\tau_D = 10$ in Figure 4.3, where

$$C(t) = \frac{\tau_D I(t)^K}{\tau_D I(t)^K + 1}.$$

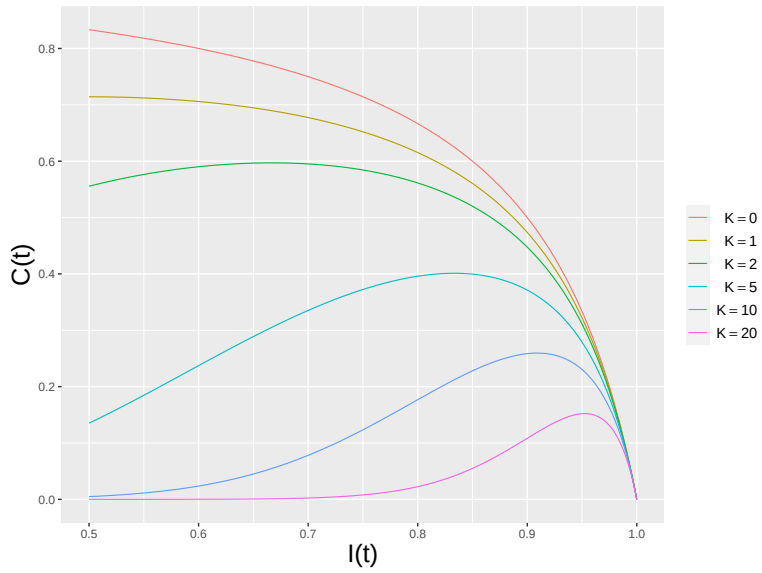


Figure 4.3: The proportion of change due to IGC when $\tau_D = 10$ for different values of K. The x-axis represents the paralog identity (i.e., $I(t)$), and the y-axis represents the proportion of change due to IGC (i.e., $C(t)$).

4.4 Inference with a more realistic IGC model

The appropriateness of the JC model for IGC inference is limited to ideal circumstances with strong assumptions. In Section 4.4.1, we will revisit the basic likelihood approach of Ji et al. (2016) because it facilitates investigating IGC in more biologically-plausible situations. We will then consider ways to characterize the relationship between IGC rates and paralog divergence.

4.4.1 Quantitative treatment of IGC

To separately consider synonymous and nonsynonymous change, Muse and Gaut (1994) describe a codon substitution model for sequence changes that originate with a point mutation. The codon substitution model that we adopt has an instantaneous rate $Q_{i,j}$ of

change from codon i to j and utilizes the parameter ω to differentiate between synonymous and nonsynonymous change. The model employs the parameter κ to distinguish transversions and transitions. The rates below correspond to codon substitutions that result in a nucleotide of type h ($h \in \{A, C, G, T\}$) being introduced into one of the three codon positions. The π_h parameters denote what would be the stationary frequency of nucleotide type h ($\pi_A + \pi_C + \pi_G + \pi_T = 1$) if there were no natural selection. The model has rates of codon substitution from triplet i to j ($i \neq j$) being

$$Q_{i,j} = \begin{cases} 0, \text{ at least two positions are different} \\ \kappa\pi_h, \text{ synonymous transition,} \\ \pi_h, \text{ synonymous transversion,} \\ \omega\kappa\pi_h, \text{ nonsynonymous transition,} \\ \omega\pi_h, \text{ nonsynonymous transversion.} \end{cases} \tag{4.5}$$

The independently-evolving paralog (IND) model is a straightforward extension to two paralogs of the above variation of the model of Muse and Gaut (1994). Because the IND model describes the evolution of two paralogs that change only by substitutions that originate with point mutation, the IND model has a state space with 61^2 possible states. The IND model specifies rate $Q_{(i,i'),(j,j')}$ of change from the joint states (i, i') to (j, j') with i and j being the codon states of the first paralog and i' and j' being the codon states of the second paralog. Because the IND model does not allow two separate substitutions to occur simultaneously at the two paralogs, the only possible positive rates occur when exactly one of the two paralogs changes codon state

$$Q_{(i,i'),(j,j')} = \begin{cases} 0 & i \neq j, i' \neq j', \\ Q_{i,j} & i \neq j, i' = j', \\ Q_{i',j'} & i = j, i' \neq j'. \end{cases} \tag{4.6}$$

The Ji et al. (2016) model builds upon the IND model by allowing change via IGC mutation as well as point mutation. Because IGC induces a dependence between paralogs through changes that homogenize the codon states, the Ji et al. (2016) model starts with the rates from the IND model but adds an amount ν when IGC could also cause a change. The Ji et al. (2016) model rates therefore are

$$Q_{(i,i'),(j,j')} = \begin{cases} 0 & i \neq j, i' \neq j', \\ Q_{i,j} & i \neq j, i' = j', j \neq j', \\ Q_{i',j'} & i = j, i' \neq j', j \neq j', \\ Q_{i,j} + \nu & i \neq j, i' = j', j = j', \\ Q_{i',j'} + \nu & i = j, i' \neq j', j = j', \end{cases} \tag{4.7}$$

where $\nu = \tau$ for synonymous change and $\nu = \omega\tau$ for nonsynonymous change.

As is conventional for evolutionary inference, the pruning algorithm (Felsenstein 1981) can be used to compute the likelihood with the Ji et al. (2016) model on a phylogenetic tree. Because the state space can become large when jointly considering paralogs, the computational efficiency of obtaining the likelihood can be improved by Al-Mohy and Higham (2011) algorithm. We also note that a very similar modeling strategy can be used to combine IGC with conventional 4-state nucleotide substitution models. The use of 4-state models may be desirable when protein-coding regions are not being analyzed and/or when the 61 states of codon models is computationally prohibitive.

4.4.2 Simple estimator of branch-specific IGC rates

The Ji et al. (2016) model assumes that the IGC rate at a site where paralogs differ is independent of overall paralog divergence. This assumption can be investigated. To do so, we start by finding the maximum likelihood estimates (MLE) of parameters from the Ji

et al. (2016) model for a particular data set. We then apply the algorithm of Tataru and Hobolth (2011) to calculate three different posterior expectations for each sequence on each branch of the phylogeny. These posterior expectations are each conditional upon the tree topology, the parameter estimates, and the data. The values per sequence of these three quantities represent sums of posterior expectations per site over all sites. The first of the three quantities is the expected number of changes from IGC per paralog sequence pair per branch. For a branch B, the posterior expectation for the number of IGC changes will be denoted $\widehat{N}_B$. For a pair of paralogs, $\widehat{N}_B$ includes both the substitutions where the first paralog was IGC donor and the second was IGC recipient and also the IGC substitutions where the donor and recipient roles are reversed.

The second and third of the posterior expectations are referred to as dwell times. Specifically, the second quantity is a synonymous dwell time whereas the third quantity is termed a nonsynonymous dwell time. The estimated synonymous dwell time for Branch B will be denoted $\widehat{t}_{s,B}$. It represents the sum over all codon sites of the expected amount of time on each branch that the two paralogs had a synonymous difference. Because there is a corresponding pair of sites for each of the two paralogs, each instant where there is a synonymous difference between the corresponding sites gets counted twice (i.e., once for each paralog). The estimated nonsynonymous dwell time for branch B is similar and will be denoted $\widehat{t}_{n,B}$. It represents the sum over all sequence sites of the expected amount of time that the two paralogs had a nonsynonymous difference.

According to the Ji et al. (2016) model, paralog sites that differ by a synonymous difference experience IGC events with the first paralog being IGC donor at rate τ and they also experience IGC events with the second paralog being IGC donor at rate τ. Because the dwell times count each instant on a branch once for each sequence in a paralog pair, the expected number of synonymous IGC events on a branch should be the product of τ and the synonymous dwell time $t_{s,B}$. Because nonsynonymous IGC rates are $\omega\tau$ rather

than τ according to the Ji et al. (2016) model, the expected number of nonsynonymous IGC events on a branch should be the product of $\omega\tau$ and the nonsynonymous dwell time $t_{n,B}$. If we represent the branch-specific IGC rate for Branch B by τ_B, we can therefore obtain an estimate

$$\widehat{\tau}_B = \frac{\widehat{N}_B}{\widehat{t}_{s,B} + \widehat{\omega}\widehat{t}_{n,B}}. \tag{4.8}$$

Via Equation 4.8, we obtained branch-specific IGC rate estimates for the 14 data sets of yeast ribosomal protein-coding genes that are considered Ji et al. (2016). We also obtained branch-specific IGC rate estimates for the teleost data sets that are described in Chapter 3. We specifically considered only the 37 teleost data sets of Chapter 3 where all post-duplication taxa have retained 2 paralogs.

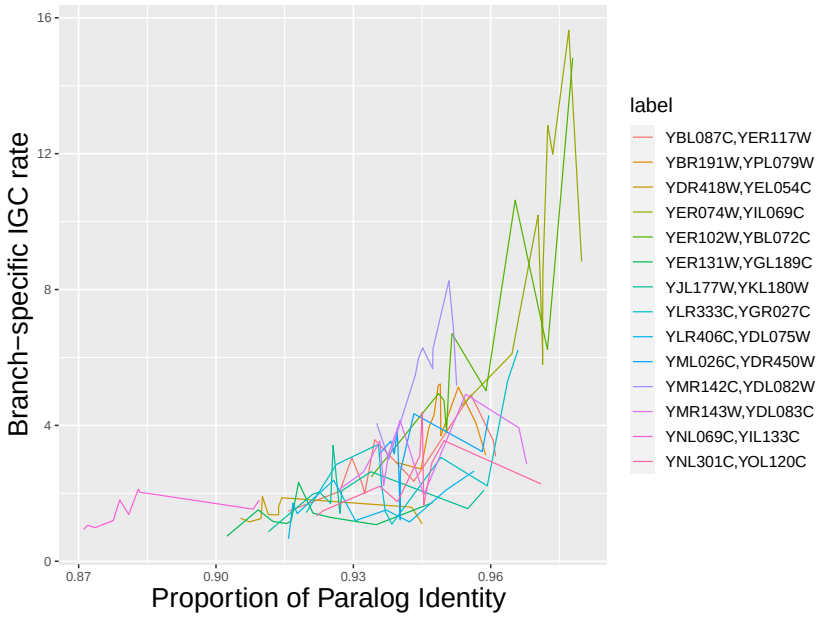
For each branch in each data set, we compared the branch-specific estimates of IGC rates to the approximate proportion on the branch of nucleotide identity between paralogs. Because the phylogenetic approach for IGC inference is not able to effectively separate substitutions from point mutation and IGC on the branch between the duplication and the first post-duplication speciation (Ji et al. 2016), we only consider branches that are subsequent to the first post-duplication speciation.

To approximate the proportion of nucleotide identity on each branch following the first post-duplication speciation, we used the Tataru and Hobolth (2011) algorithm to estimate the expected proportions of time on a branch where the average codon differs between paralogs at exactly 0, exactly 1, exactly 2, or exactly 3 codon positions. These four expected proportions are each ratio of dwell times and branch lengths, with the dwell times being calculated by the Tataru and Hobolth (2011) algorithm. With the estimated expected proportions of times that the average codon differs at 0, 1, 2, or 3 nucleotide positions, it is straightforward to infer the average proportion of nucleotide difference (or identity) on the

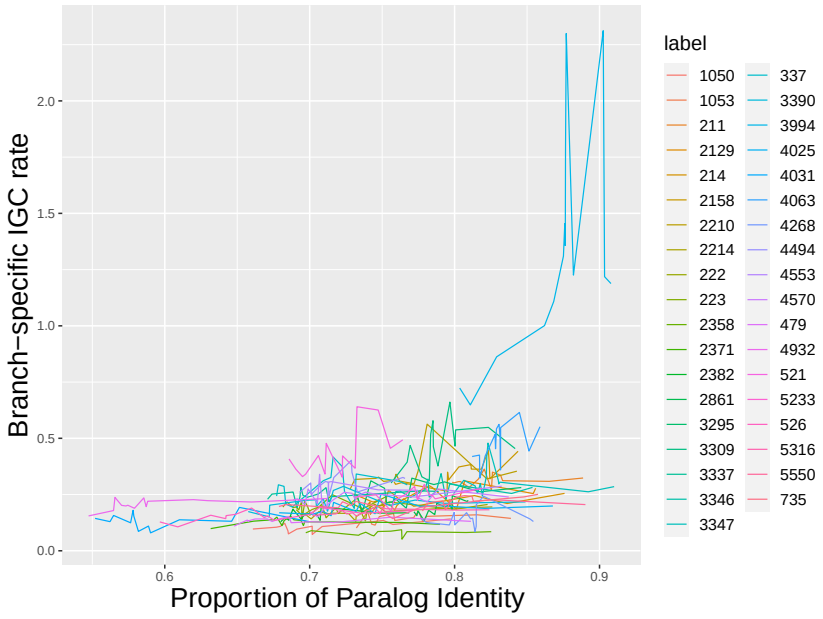
branch.

For all branches that are subsequent to the first post-duplication speciation, Figure 4.4 plots the branch-specific IGC rate estimates from the 14 yeast data sets versus the proportion of nucleotide sites that are identical between paralogs on the branch. Figure 4.4b is a similar plot except that results from the 37 teleost protein-coding genes of Chapter 3 where all post-duplication species have retained both paralogs. Because the expected rate of change that originates with point mutation is set to 1 per site per paralog, a branch-specific IGC rate that exceeds 1 represents a case where paralogs are homogenized by IGC at a higher rate than they experience change from point mutation. Figure 4.4 shows that many of the branch-specific IGC rates from yeast genes exceed 1. Although few of the branch-specific IGC rates from teleost genes exceed 1, the teleost rates tend to be far enough above 0 that it would seem problematic to make the conventional assumption that IGC can be ignored.

Figure 4.4 has an appearance that is consistent with the expectation that IGC rates and paralog identity are positively correlated (e.g., see also Chen et al. 2007; Harpak et al. 2017). However, the estimated branch-specific IGC rates that are displayed in Figure 4.4 are all determined by first analyzing data with the Ji et al. (2016) model that has the IGC parameter τ be independent of paralog divergence. Different estimates of τ will lead to different estimates of branch-specific IGC rates because the formula for $\widehat{\tau}_B$ in Equation 4.8 uses terms (e.g., $\widehat{N}_B$, $\widehat{t}_{s,B}$, and $\widehat{t}_{n,B}$) that are conditional upon the estimate of τ. This circularity in relying upon a divergence-independent IGC rate model to derive branch-specific IGC rate estimates could be eliminated by instead relying on an IGC model where paralog divergence affects IGC rates.



(a) Yeast genes



(b) Teleost genes

Figure 4.4: Estimated branch-specific IGC rates versus corresponding paralog identities for 14 Yeast genes and 37 Teleost genes. Branch-specific rates and paralog identities were estimated via the "simple" estimator of Section 4.4.2. Different data sets are represented with different colors, and an estimate is plotted for each branch that is subsequent to the first post-duplication speciation. Teleost data set labels refer to the "Pillar" identification system employed for the data sets at: https://github.com/Yixuan39/IGC-fish.

4.4.3 Nonparametric approach for comparing IGC rates to paralog divergence

The relationship between paralog divergence and branch-specific IGC rate can also be investigated via a nonparametric approach. The idea is that paralogs are identical in sequence at the time that they are formed via duplication and then the paralogs will subsequently diverge. This means that paralog divergence on a parent branch of an evolutionary tree will typically be lower than paralog divergence of a child branch. If branch-specific IGC rates decrease as paralog divergence increases, we can therefore expect that branch-specific IGC rates on parent branches will be higher than the rates on child branches.

Specifically, the branch-specific IGC rate estimate from a parent branch can be compared to the branch-specific rate estimates of each of its child branches. For a null hypothesis of paralog divergence being independent of IGC rate, one might expect the branch-specific IGC rate estimate from a child branch to be equally likely to be higher or lower than the estimate from its parent. In contrast, the biologically-plausible alternative hypothesis would be that branch-specific IGC rates should decrease as paralog divergence increases. This one-sided alternative hypothesis would be consistent with a pattern where branch-specific IGC rate estimates on child branches would tend to be lower than the estimates from the parent branches.

Therefore, we can consider whether the branch-specific IGC rate estimate from a child branch is lower than the estimate from the parent branch for each pair of parent and child branches. We do these for all pairs of branches where both the parent and the child are subsequent to the first post-duplication speciation. For the null hypothesis of independence, the number of pairs where the child estimate is lower can be approximated as having a binomial distribution with the number of trials being the number of parent-child pairs and with the probability of each outcome being 0.5. The alternative hypothesis would be that

the probability should be greater than 0.5 of a child branch having a lower branch-specific IGC rate estimate than its parent.

The binomial approximation is not technically correct when all parent-child branch pairs are used on a phylogeny for the simple reason that each internal branch on a rooted bifurcating tree can be associated with up to three parent-child pairs where two of the pairs have the branch being a parent and the remaining pair has the internal branch being a child. This violates the binomial assumption of independent trials that each have the same outcome probabilities. For example, even if the branch-specific rates of parent and child are independent of one another, knowing that an internal branch has a higher rate than one of its children will mean that it is also likely to have a higher rate than its other child branch. However, even if branch-specific rates are independent of one another, knowing that a parent branch has a higher rate than one of its children also suggests that the parent branch is likely to have a higher rate than the internal branch of which it is a child. When the null hypothesis of independence of rates among branches is correct, this correlation of rates among different parent-child branch pairs on a tree that is due to individual branches being represented in multiple pairs would serve to make the variance of our test statistic smaller than that of a binomial distribution. A strategy that permutes rates among branches would yield a more suitable null distribution than a binomial, but we do not pursue that direction here.

We applied the nonparametric test to the 14 yeast ribosomal protein-coding data sets and the 37 teleost data sets. For the yeast case, each data set had 8 parent-child pairs that are subsequent to the first post-duplication speciation so that the total number of parent-child pairs was $112 = 8 \times 14$. Of these 112 pairs, the branch-specific IGC rate estimate from the child branch was lower than the estimate from the parent branch for 63 pairs and was higher for 49 pairs. For the yeast data sets, we therefore obtain a P-Value of approximately 0.11. In contrast, when we apply the nonparametric test to the 12 parent-child pairs of

branches from each of the 37 teleost data sets, we find that 275 of the $444 = 12 \times 37$ pairs yield a lower branch-specific IGC rate on the child branch than the parent branch. In this case, the null hypothesis of paralog divergence being independent of the IGC rate yields a P-Value of less than 10^{-5}.

We attribute the rejection of the null hypothesis for the teleost data sets and the failure to reject for the yeast data sets to the high IGC rates that the yeast protein-coding ribosomal genes have experienced. Because the high IGC rates in these yeast genes have caused so much homogenization of paralog sequences, paralog divergence can be similar between parent and child branches. For example, the 112 parent-child pairs for the yeast data had 21 cases where the estimated paralog divergence of the child was lower than the estimated divergence of the parent. In contrast, only 7 of the 444 parent-child pairs from the yeast data have an estimated paralog divergence of the child being lower than for the parent. As a consequence of paralog divergence being similar in parent and child branches for the yeast data sets, branch-specific IGC rates are likely to be similar in parent and child branches even if IGC rates are a function of paralog divergence. This similarity between parent and child branches in IGC rates reduces the power to reject the null hypothesis for the yeast data.

4.4.4 Divergence model implementation

The suggestion of a relationship between paralog identity and IGC rate motivates development of a model that lacks some of the circularity of the Section 4.4.2 estimator of branch-specific IGC rates. We choose to model the relationship by adding an additional parameter denoted K ($\infty > K > -\infty$) to the Ji et al. (2016) model. Rather than having the rate of an IGC change at codons that differ by a synonymous substitution be τ, our model has the IGC rate be $\tau_D I^K$ where I represents the proportion of identical nucleotide types

between the two paralogs and where the IGC rate for a synonymous difference approaches τ_D as I approaches 1. If corresponding codons of two paralogs encode different amino acids, we have the rate of a specific IGC change be $\omega\tau_D I^K$. We refer to this parameterization as the Divergence model. The Ji et al. (2016) model can be considered as the special case of the Divergence model that has $K = 0$. Although negative values of K are technically possible, biologically plausible values of K would be positive and would yield the expected result that IGC rates decrease as paralog divergence increases.

An ideal implementation of the Divergence model would have IGC rates at corresponding codons change as the paralog identity changes. However, we choose a more computationally-tractable implementation that has IGC rates be piecewise homogeneous. Specifically, paralog identities and IGC rates on different branches of a phylogeny are allowed to differ but paralog identities and IGC rates are treated as being constant along each branch. This piecewise-homogeneous treatment facilitates likelihood calculation via the pruning algorithm of Felsenstein (1981).

To explain the details of our piecewise-homogeneous implementation, we let X represent the observed molecular sequence data and we have θ denote the parameters of the Divergence model (i.e., K, τ_D, parameters that specify the rate matrix for sequence changes that originate with point mutation, and branch lengths). We will use I_D to represent a vector that specifies the nucleotide-level paralog sequence identities for each of the post-duplication branches on the phylogeny. For a branch with a paralog identity level I and for a site where the two paralogs differ in sequence state, the Divergence model has the rate at which one paralog serves as IGC donor and the other paralog serves as IGC recipient be equal to $\tau_D I^K$. The Divergence model assumes that this branch-specific IGC rate of $\tau_D I^K$ applies both when the first paralog serves as IGC donor and when the second paralog serves as IGC donor.

Our iterative algorithm for estimating the parameters θ of the piecewise-homogeneous

divergence model can be outlined as follows:

- **Step 1:** Set $j = 0$.
- **Step 2:** Find the maximum likelihood estimates of all parameters from the Ji et al. (2016) model. Using these parameter estimates and the Ji et al. (2016) model, apply the Tataru and Hobolth (2011) algorithm as described in Section 4.4.2 to estimate the average nucleotide-level paralog identity on each post-duplication branch of the phylogeny. These paralog identities will serve as the initial guess $I_D^{(0)}$ for the branch-specific paralog identities of the divergence model.
- **Step 3:** Set $\theta^{(j+1)}$ to the θ from the Divergence model that maximize $P(X|\theta, I_D^{(j)})$. The $P(X|\theta, I_D^{(j)})$ quantities can be calculated because the branch-specific IGC rate for each post-duplication branch can be determined for the piecewise-homogeneous Divergence model with the paralog identity vector together with the parameters τ_D and K.
- **Step 4:** Using the branch-specific IGC rates calculated in Step 3 along with the parameters $\theta^{(j+1)}$ that were estimated in Step 3, apply the Tataru and Hobolth (2011) algorithm to obtain new paralog identities that will serve as $I_D^{(j+1)}$.
- **Step 5:** Determine whether the algorithm has converged. Our implementation assesses how $\theta^{(j)}$ changes as the value of j changes. However, the $P(X|\theta^{(j+1)}, I_D^{(j)})$ or $I_D^{(j+1)}$ could also be examined for how they change as j changes. If the algorithm is determined to have converged, the final parameter estimates $\widehat{\theta}$ are set to $\theta^{(j+1)}$, the final estimates of paralog identities $\widehat{I}_D$ are set to $I_D^{(j+1)}$, and the algorithm can be terminated. Otherwise, set $j = j + 1$ and go to Step 3.

We view $\widehat{\theta}$ from our piecewise-homogeneous implementation to be an approximation of the maximum likelihood estimates that would result from the Divergence model when

paralog identities are allowed to change within a branch. When $K = 0$, the approximation would be exact. The approximation would presumably be relatively good when K is close to 0. We also expect the piecewise-homogeneous approximation to be better when a phylogeny consists of relatively short branches than when a phylogeny has branches that are long enough for paralog identities to be likely to substantially change along the branches. Similarly, we consider $P(X|\widehat{\theta}, \widehat{I}_D)$ to be an approximation of the maximum likelihood value that could be achieved by the Divergence model when paralog identities are allowed to change within a branch.

4.5 Results

4.5.1 Paralog divergence in yeast

The 14 yeast data sets were analyzed with the piecewise-homogeneous Divergence model and a summary of the results is shown in Table 4.1. The piecewise-homogeneous implementation yielded positive estimates of K for 13 of the 14 data sets. All 13 of the positive K estimates exceeded 1 and 12 of the 14 had the estimate $\widehat{K} > 10$. As would be expected for the 13 data sets where $\widehat{K} > 1$, the estimated τ_D values all exceeded the maximum likelihood estimates of the τ parameter from the Ji et al. (2016) model that correspond to τ_D when $K = 0$. Figure 4.5 displays the range of branch-specific IGC rates that are estimated from the piecewise-homogeneous Divergence implementation. The large ranges for several of these genes are attributable to the high values that are estimated for K. More details concerning individual branch-specific rate estimates from the yeast genes are shown in Figure 4.7 in the Supporting information.

If we consider $P(X|\widehat{\theta}, \widehat{I}_D)$ to be the maximum likelihood value that would arise if paralog identities are allowed to change within a branch and if we ignore the fact that single IGC

Table 4.1: Results from analyzing the yeast ribosomal protein-coding data sets. The rows of this table summarize results from individual data sets. The "len" column shows the length in codons of the data sets. The "ll" column displays the maximum likelihood value from analysis of the data set by the Ji et al. (2016) model. The "Diff" column shows the difference between the $\log P(X|\widehat{\theta},\widehat{I}_D)$ from the piecewise-homogeneous implementation and the maximum likelihood value from the Ji et al. (2016) model. The "P" column represents the P-Value associated with the likelihood ratio test of the null hypothesis that $K = 0$ (see text for accompanying assumptions). P-Values that are less than 0.005 are rounded to 0. The "Proportion" column shows the estimated proportion of changes that originate from IGC rather than point mutation according to the piecewise-homogeneous analysis. The "$\widehat{\tau}_{SK,D}$" column shows gene-specific estimates of τ_D when the 14 yeast data sets are jointly estimated with the piecewise-homogeneous implementation and where the data sets are constrained to share a common value of the K parameter that is estimated to be $K = 30.05$ (see Section 4.5.4).

Data Set	len	ll	Diff	P	$\widehat{\tau}$	$\widehat{\tau}_D$	$\widehat{K}$	Proportion	$\widehat{\tau}_{(SK,D)}$
YBL087C_YER117W	136	-1367.68	6.69	0	2.81	22.35	42.55	0.19	13.46
YBR191W_YPL079W	159	-1467.29	0.06	0.73	3.83	6.31	10.37	0.28	14.16
YDR418W_YEL054C	163	-1739.18	0.04	0.78	1.41	1.14	-2.76	0.20	10.44
YER131W_YGL189C	118	-1205.18	0.09	0.67	1.36	2.72	9.26	0.18	20.65
YER074W_YIL069C	133	-1251.96	8.59	0	7.47	62.76	64.27	0.28	20.76
YER102W_YBL072C	198	-2058.96	10.98	0	4.87	117.85	68.53	0.15	10.32
YJL177W_YKL180W	183	-1837.06	0.70	0.24	1.77	4.08	13.27	0.17	10.45
YLR406C_YDL075W	112	-1178.10	2.64	0.02	1.65	14.44	34.65	0.16	20.07
YLR333C_YGR027C	107	-1262.00	6.59	0	3.28	33.93	40.52	0.19	11.37
YMR142C_YDL082W	197	-2054.05	5.51	0	5.71	97.85	57.53	0.25	15.00
YMR143W_YDL083C	134	-1209.75	2.42	0.03	3.16	12.3	27.97	0.25	26.97
YML026C_YDR450W	140	-1377.25	0.45	0.34	3.64	8.24	16.44	0.26	13.61
YNL301C_YOL120C	185	-2139.31	6.11	0	2.48	21.76	39.64	0.20	25.20
YNL069C_YIL133C	197	-2322.83	1.42	0.09	1.46	5.55	12.53	0.19	13.60

events might homogenize multiple sites, we can perform a likelihood ratio test of the null hypothesis that $K = 0$. In other words, the null hypothesis is the Ji et al. (2016) model that does not have IGC rates depending on paralog divergence. For a significance level of 0.05, this null hypothesis could be rejected for 8 of the 14 data sets.

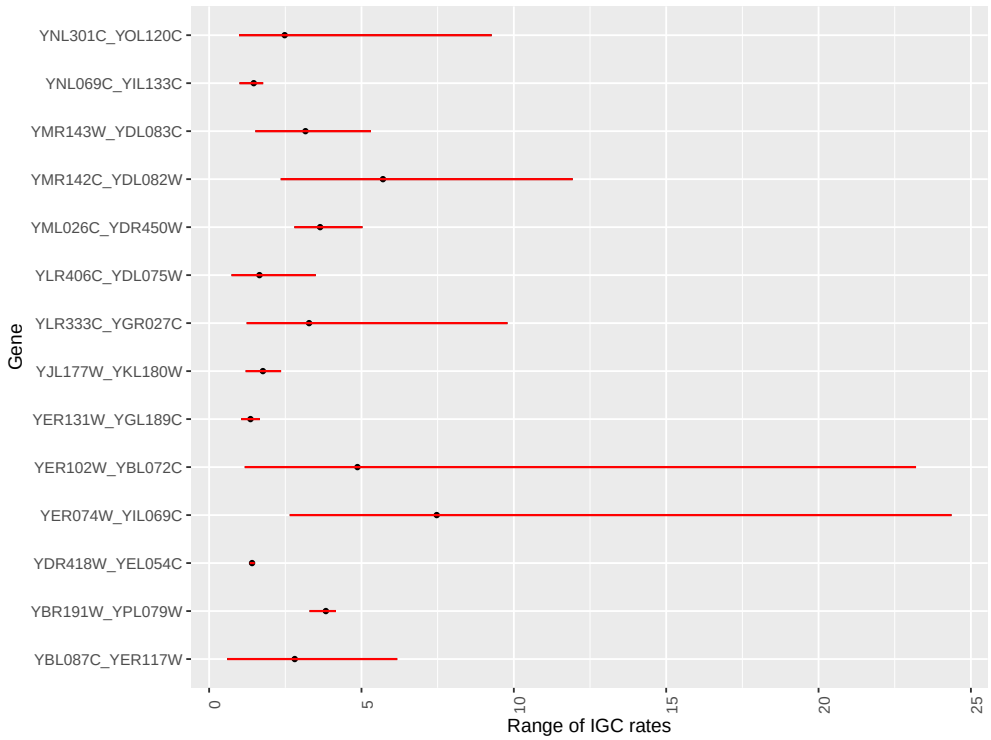


Figure 4.5: The range of estimated branch-specfic IGC rates from the Divergence model versus the IGC rates estimated from the Ji et al. (2016) model for 14 Yeast data sets. For each data set, the black point is the τ estimate from the Ji et al. (2016) model, and the red line shows the range of the branch-specific IGC rates from the Divergence model.

4.5.2 Analysis of teleosts

Table 4.2 displays results from the 37 teleost data sets of Chapter 3 that have two paralogs for each post-duplication taxon. The teleost data sets tend to yield lower IGC rate estimates than do the yeast genes, but most teleost data sets still exhibit substantial IGC evidence. For the divergence model, some data sets yield high estimated values for τ_D, but the paralog divergence and the positive K estimates still yield generally lower branch-specific IGC rate estimates than for the yeast data. For a significance level of 0.05, the (approximate) likelihood ratio tests indicate a significant improvement in the log-likelihoods for 26 out of 37 genes. Specifically, Gene 3309 is hard to get robust estimation due to the short length.

Table 4.2: Results from analyzing the teleost data sets. Column labels are as in Table 4.1.

Data Set	len	ll	Diff	P	$\widehat{\tau}$	$\widehat{\tau_D}$	$\widehat{K}$	Proportion
211	560	-12408.76	0.36	0.4	0.31	0.67	3.93	0.09
214	581	-14134.16	9.87	0	0.21	3.96	12.6	0.06
222	590	-16785.38	2.06	0.04	0.19	0.87	5.04	0.06
223	379	-10253.74	2.23	0.03	0.21	1.99	6.82	0.05
337	239	-7550.08	1.92	0.05	0.29	1.5	4.9	0.08
479	775	-24481.33	0.93	0.17	0.14	0.49	3.54	0.04
521	334	-10402.19	4.66	0	0.44	7.66	9.28	0.06
526	304	-10683.17	3.73	0.01	0.16	1.42	5.46	0.04
735	1490	-33364.44	20.82	0	0.22	5.54	15.6	0.06
1050	320	-8753.76	10.45	0	0.11	33.16	22.41	0.00
1053	408	-10245.91	2.23	0.03	0.14	1.67	9.62	0.04
2129	500	-14905.77	4.89	0	0.24	2.45	8.88	0.06
2158	837	-20915.64	22.22	0	0.33	56	26.35	0.11
2210	395	-11831.41	1.77	0.06	0.33	1.24	4.9	0.09
2214	641	-17658.15	7.86	0	0.18	4.10	10.87	0.05
2358	285	-7828.6	0.08	0.69	0.08	0.25	3.88	0.03
2371	296	-9081.5	0.00	1	0.13	0.12	-0.07	0.04
2382	526	-14950.46	4.71	0	0.16	4.59	10.29	0.05
2861	204	-4299.52	5.56	0	0.18	53.66	29.09	0.11
3295	229	-5844.69	2.28	0.03	0.26	1.96	7.86	0.07
3309	134	-3450.64	4.91	0	0.44	68.23	22.34	0.42
3337	566	-16481.07	4.12	0	0.25	2.73	7.22	0.07
3346	786	-21649.26	2.00	0.05	0.26	0.91	5.12	0.07
3347	466	-10382.02	0.25	0.48	0.3	0.16	-3.19	0.10
3390	253	-7734.06	2.07	0.04	0.2	1.12	5.16	0.06
3994	281	-5778.83	25.82	0	1.09	7.72	14	0.10
4025	1087	-32137.02	7.11	0	0.18	0.93	5.41	0.06
4031	174	-6400.75	0.06	0.73	0.15	0.23	0.85	0.05
4063	331	-6663.05	3.45	0.01	0.51	3.11	10.47	0.12
4268	175	-3468.79	0.07	0.71	0.15	0.35	3.97	0.06
4494	245	-7112.81	2.55	0.02	0.3	3.1	7.47	0.06
4553	314	-9377.87	1.05	0.15	0.25	0.87	3.87	0.07
4570	491	-13044.53	12.28	0	0.23	4.5	10.89	0.06
4932	495	-19556.84	1.66	0.07	0.21	10.38	7.85	0.03
5233	761	-16742.42	6.22	0	0.25	20.48	19.24	0.10
5316	1886	-52780.28	8.12	0	0.17	1.3	6.92	0.11
5550	255	-7438.96	0.22	0.51	0.21	0.42	2.43	0.06

4.5.3 Average IGC rate

Based on Tables 4.1 and 4.2, there is a strong positive correlation between the τ_D and K estimates. The standard deviations of τ_D and K estimates also appear to be large (see Tables 4.4 and 4.5). Because the yeast and teleost data sets are related via a phylogenetic tree with a small number of branches, it is difficult to obtain disentangle τ_D and K in our analyses. Even though it is challenging to separate these two parameters from a single data set, there may be some information concerning the relationship between IGC rate and paralog divergence if each data set is employed to calculate a single average paralog divergence over branches and a single average IGC rate over all branches. These averages could then be compared across data sets. This is the motivation for the following approach.

Consider the Divergence model and let τ_{lm} be the IGC rate on branch m for gene l. After log transformation,

$$log(\tau_{lm}) = log(\tau_{Dl}) + K_l * log(I_{lm}).$$

Averaging over all branches yields,

$$\overline{log(\tau_l)} = log(\tau_{Dl}) + K_l * \overline{log(I_l)},$$

where $\overline{log(\tau_l)}$ denotes the average logarithm of branch-specific IGC rates and $\overline{log(I_l)}$ denotes for the average logarithm of the paralog identity.

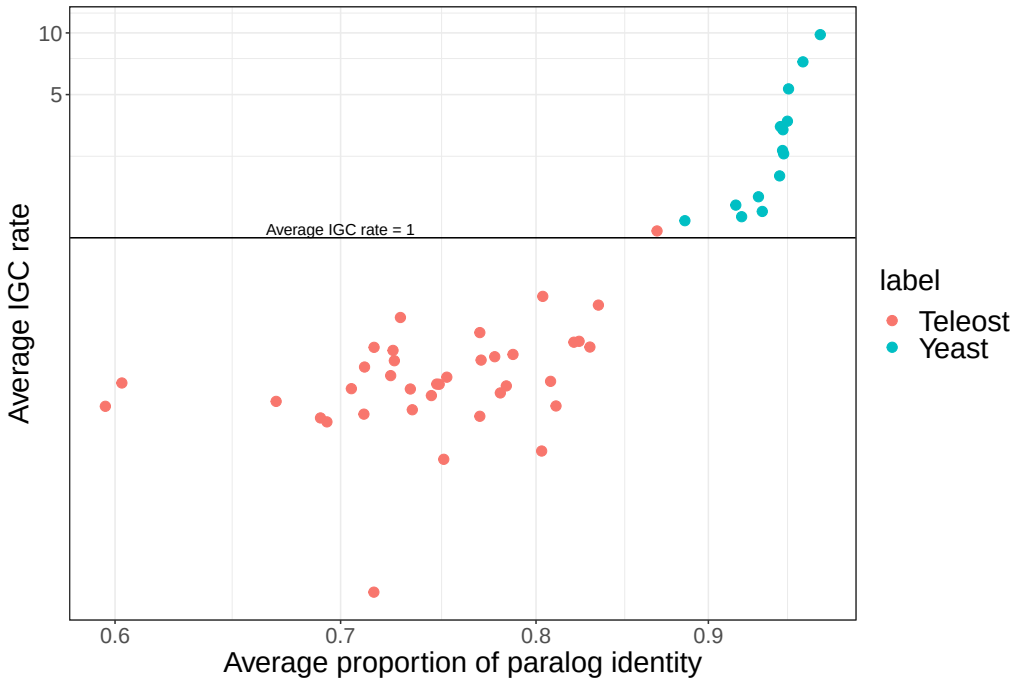


Figure 4.6: Average over branches of logarithms of branch-specific IGC rate estimates versus average over branches of logarithms of estimated paralog identities. The 14 yeast data sets and the 37 teleost data sets are each represented by a single point. The plot is made on a log-log scale but the units on the x and y axes are shown as rates and paralog identities rather than logarithm of rates and logarithm of identities.

For the 14 yeast and 37 teleost data sets, Figure 4.6 plots the average logarithm of the branch-specific IGC rate estimates versus the average logarithm of the paralog identity. If we separately consider the yeast and the teleost data sets, it appears that the yeast data sets have a strong relationship between IGC rate and paralog identity. While the yeast relationship does not appear to be linear on the log-log scale that is plotted, it does appear that the average logarithm of IGC rates for the yeast data can be well-described by a monotonically increasing and simple function of the average logarithm of paralog identity. The teleost data sets seem to have a positive correlation on this log-log scale, but the relationship seems weaker than for the yeast data sets and the teleost relationship seems to have a different nature than that exhibited by the yeast data sets.

4.5.4 Joint likelihood analysis

Section 4.5.3 demonstrates IGC rate and paralog identity relationships based on single-gene statistics. Here, we employ more comprehensive joint analyses to share information across yeast data sets. Specifically, we contrast the situation where all yeast data sets are allowed to have separate estimates of all parameters for our implementation of the Divergence model to three special cases that impose sharing of parameter values across data sets. The most restrictive "concatenated" case has all parameters shared across all 14 yeast data sets. In this situation, the parameters concerning changes that originate with point mutation and the IGC parameters τ_D and K and the branch lengths are all shared by the 14 data sets. The next most restrictive situation shares K and τ_D across data sets but does not constrain the other parameters. We also examine a case that constrains all 14 data sets to have the same K value but allows different τ_D values for different data sets. Computational concerns did not permit completion of these shared-parameter analyses for the 37 teleost data sets.

Table 4.3: Maximum log-likelihood values and parameter estimates from the 14 yeast ribosomal protein-coding sets for different IGC treatments. The "No IGC", "Basic IGC", and "Divergence" columns show log-likelihood differences. Each log-likelihood difference represents a maximum log-likelihood value for a specific analysis minus the maximum log-likelihood value of −23663.05 that was obtained for the simplest and least parameter-rich model. The simplest model does not permit IGC (i.e., $\tau_D = 0$) and shares all parameter values amongst all data sets (i.e., "Concatenated" case). The "No IGC" column represents analyses where $\tau_D = 0$. The Basic IGC model is the Ji et al. (2016) model where the value of $\tau_D = \tau$ is inferred, but $K = 0$. The "Divergence" column represents analyses where both τ_D and K are inferred. The "$\widehat{\tau}$" column shows estimates of $\tau_D = \tau$ when $K = 0$. The "$\widehat{\tau}_D$" and "$\widehat{K}$" columns show estimates when τ_D and K are jointly inferred. The "Concatenated" row represents causes where all 14 data sets are forced to share the same parameter values. The "Shared K and shared τ_D" row constrains K and τ_D parameters to have the same values for all data sets but allows all other model parameters (e.g., branch lengths) to differ between data sets. The "Shared K but different τ_D" row constrains only K parameters to have the same values for all data sets. The "Different K and different τ_D" row separately estimates the IGC parameters for each data set.

Case	No IGC	Basic IGC	Divergence	$\widehat{\tau}$	$\widehat{\tau}_D$	$\widehat{K}$
Concatenated	0	902.17	916.62	2.85	10.32	21.76
Shared K and shared τ_D	264.81	1129.55	1205.25	2.78	15.67	29.87
Shared K but different τ_D	——	1192.42	1229.12	Table 4.1	Table 4.1	30.05
Different K and different τ_D	——	——	1244.74	Table 4.1	Table 4.1	Table 4.1

Table 4.3 clearly demonstrates the large improvement in model fit when allowing IGC. A substantial improvement in log-likelihood also appears when the K parameter is freely estimated rather than being constrained to 0 as in the Ji et al. (2016) model. For example, the Divergence model with "Shared K and shared τ_D" yields a log-likelihood improvement of more than 75 log-likelihood units when only a single free parameter (the freely estimated K value) is added to the Basic IGC model with "Shared K and shared τ_D". In contrast, the Basic IGC model with "Shared K but different τ_D" yields a log-likelihood improvement of only about 63 log-likelihood units when 13 freely estimated parameters (i.e., the τ_D parameters for different genes) are added to the Basic IGC model with "Shared K and shared τ_D." While

statistically significant improvements are likely to result from allowing different data sets to have different τ_D values and from further allowing different data sets to have different K values, these improvements are less sizeable than the aforementioned ones.

4.6 Discussion

The evolutionary relationship between IGC rate and paralog divergence is difficult to characterize because the history of sequence changes over time is not directly observed. Therefore, it tends to be unclear when and whether IGC mutations occurred and also what was the paralog divergence at the time of the IGC mutations. In addition, evolutionary sequence changes due to IGC represent mutations that occur and then eventually survive (i.e., are "fixed" in population genetic jargon). A challenge with IGC mutations is that a single mutational event can influence multiple positions in a sequence tract, but subsequent recombination events can influence which of the affected positions are eventually fixed. However, fixation of mutations can be influenced by natural selection and natural selection itself can be problematic to study and model.

Despite these obstacles, this chapter introduces two lines of evidence (the piecewise-homogeneous implementation of the Divergence model and the nonparametric approach) that collectively suggest a tendency for IGC rates to decrease as paralog divergence increases. While this tendency is both biologically plausible and generally accepted (e.g., Harpak et al. 2017; Dumont 2015), it has not previously been sufficiently examined. Our careful examination seems warranted because IGC seems to be responsible for a non-negligible proportion of sequence change in duplicated genes (see Chapter 3) and because it is typically ignored by researchers who study molecular evolution and/or infer evolutionary history. The overlapping but non-identical lines of evidence that are introduced in this chapter better establish the negative correlation between IGC rates and paralog divergence.

The Divergence model has parameters τ_D and K and specifies that the per-site IGC rate is $\tau_D I^K$ when the paralog identity proportion is I. While we feel confident that the value of K should exceed 0, we expect that our point estimates of K are inaccurate. Unfortunately, we cannot reliably report the uncertainty of our estimates. Via an inverted Hessian matrix when $\log P(X|\widehat{\theta}, \widehat{I_D})$ is treated as the log-likelihood and all parameters but K and τ_D are fixed at their estimated values, we tried to estimate standard deviations associated with the K estimates from our piecewise-homogeneous implementation of the Divergence models. The values that we obtained for individual genes are in Supplementary Tables 4.4 and 4.5. However, given the high estimated values for the K parameter, the estimated standard deviations seem too small to be plausible.

A thorough simulation study of our piecewise-homogeneous estimation procedure will be necessary, but substantial changes to the procedure may also be worth exploring. For example, the procedure assumes that each IGC event only affects a single codon. The effects of this assumption will depend both on the length distribution of IGC mutations and on the amount of subsequent recombination that separates the fates on different sites that are affected by a single IGC mutational tract. The Divergence model assumes the tract length is one, violating the fact that IGC tract length can be as long as $4kb$ (Judd and Petes 1988; Chen et al. 2007). The high recombination rates (Lian et al. 2022) will reduce the tract length, which may be favorable for the Divergence model. Tract lengths can be incorporated into IGC inference (Ji and Thorne 2019), but the extra required computation may limit the ability to add other biologically important features to IGC inference techniques.

Another feature of our inference procedure that may be problematic is its piecewise-homogeneous nature. Paralog identities change on a branch. Our piecewise-homogeneous procedure ignores this fact and also ignores the fact that estimated branch-specific paralog identities have associated uncertainty. There is no obvious reason to expect substantial bias in the average paralog identities that are used by our implementation and that are

calculated by the Tataru-Hobolth algorithm. However, our piecewise-homogeneous model has the IGC rate on a branch being $\tau_D I^K$ with I being the paralog identity. Even if the estimate $\widehat{I}$ is an unbiased estimate of the average paralog identity on a branch, it could be the case that $\widehat{I}^K$ is a biased estimate of I^K on the branch. One unexplored option would be to change **Step 4** of our piecewise-homogeneous implementation to estimate the set of I^K values on different branches of the tree rather than to first estimate the different paralog identities on the branches of the tree and then raise each branch's value to the power of K.

There is some reason for cautious optimism concerning the quest to better understand the evolutionary impact of IGC. Most importantly, low-cost and high-throughput DNA sequencing may eventually make it feasible to collect direct evidence of IGC mutations from trios of parent-offspring genomes. Recently, Vollger et al. (2023) employed high-fidelity long-read DNA sequencing to identify the genetic consequences of putative IGC mutations that are segregating in human populations. While the work of Vollger et al. (2023) was not a study that examined the influence of IGC mutation on interspecific genetic variation, a better characterization of IGC mutational events will inevitably eventually lead to an improved of the long-term evolutionary consequences of IGC.

4.7 Acknowledgments

Y.Y., T.X., and J.L.T. were supported by N.S.F. DEB-1754142. G.C. was supported by N.S.F. DEB-2241312. H.K. was supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research 22K11950. X.J. acknowledges support by the NVIDIA Corporation. T.X. served as the first author of this paper, contributing significantly to the study's approach, theory, coding, writing, and data analysis.

References

Al-Mohy, A. H. and Higham, N. J. (2011). Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM journal on scientific computing*, 33(2):488–511.

Birchler, J. A. and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746–14753.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775.

Dumont, B. L. (2015). Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. *BMC genomics*, 16(1):456.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545.

Harpak, A., Lan, X., Gao, Z., and Pritchard, J. K. (2017). Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proceedings of the National Academy of Sciences*, 114(48):12779–12784.

Ji, X., Griffing, A., and Thorne, J. L. (2016). A phylogenetic approach finds abundant interlocus gene conversion in yeast. *Molecular biology and evolution*, 33(9):2469–2476.

Ji, X. and Thorne, J. L. (2019). A phylogenetic approach disentangles interlocus gene conversion tract length and initiation rate. *arXiv preprint arXiv:1908.08608*.

Judd, S. R. and Petes, T. (1988). Physical lengths of meiotic and mitotic gene conversion tracts in saccharomyces cerevisiae. *Genetics*, 118(3):401–410.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism. Academic Press, New York*, pages 21–123.

Lian, Q., Maestroni, L., Gaudin, M., Llorente, B., and Mercier, R. (2022). Remarkably high rate of meiotic recombination in the fission yeast schizosaccharomyces pombe. *bioRxiv*, pages 2022–12.

Mansai, S. P., Kado, T., and Innan, H. (2011). The rate and tract length of gene conversion between duplicated genes. *Genes*, 2(2):313–331.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724.

Ohno, S. (1970). *Evolution by gene duplication*. Allen and Unwin, London.

Tataru, P. and Hobolth, A. (2011). Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time markov chains. *BMC bioinformatics*, 12(1):465.

Vollger, M. R., Dishuck, P. C., Harvey, W. T., DeWitt, W. S., Guitart, X., Goldberg, M. E., Rozanski, A. N., Lucas, J., Asri, M., et al. (2023). Increased mutation and gene conversion within human segmental duplications. *Nature*, 617(7960):325–334.

Yin, Y., Dominska, M., Yim, E., and Petes, T. D. (2017). High-resolution mapping of heteroduplex dna formed during uv-induced and spontaneous mitotic recombination events in yeast. *eLife*, 6:e28069.

4.8 Supporting information

Table 4.4: Estimated standard deviations for the 14 Yeast genes. The "$s(\widehat{\tau_D})$" and "$s(\widehat{K})$" columns show estimated standard deviations from the piecewise-homogeneous implementation from an inverted Hessian matrix when $\log P(X|\widehat{\theta},\widehat{I}_D)$ is treated as the log-likelihood and all parameters but K and τ_D are fixed at their estimated values.

Data Set	$\widehat{\tau_D}$	$\widehat{K}$	$s(\widehat{\tau_D})$	$s(\widehat{K})$	Data Set	$\widehat{\tau_D}$	$\widehat{K}$	$s(\widehat{\tau_D})$	$s(\widehat{K})$
YBL087C_YER117W	22.35	42.55	12.27	12.88	YLR406C_YDL075W	14.44	34.65	11.54	14.3
YBR191W_YPL079W	6.31	10.37	6.22	19.94	YLR333C_YGR027C	33.93	40.52	6.09	4.19
YDR418W_YEL054C	1.14	-2.76	1.1	11.42	YMR142C_YDL082W	97.85	57.53	13.78	3.82
YER131W_YGL189C	2.72	9.26	3.25	15.55	YMR143W_YDL083C	12.3	27.97	6.05	10.4
YER074W_YIL069C	62.76	64.27	23.61	9.6	YML026C_YDR450W	8.24	16.44	4.77	11.23
YER102W_YBL072C	117.85	68.53	8.37	2.99	YNL301C_YOL120C	21.76	39.64	7.08	6.64
YJL177W_YKL180W	4.08	13.27	2.71	10.22	YNL069C_YIL133C	5.55	12.53	3.59	6.2

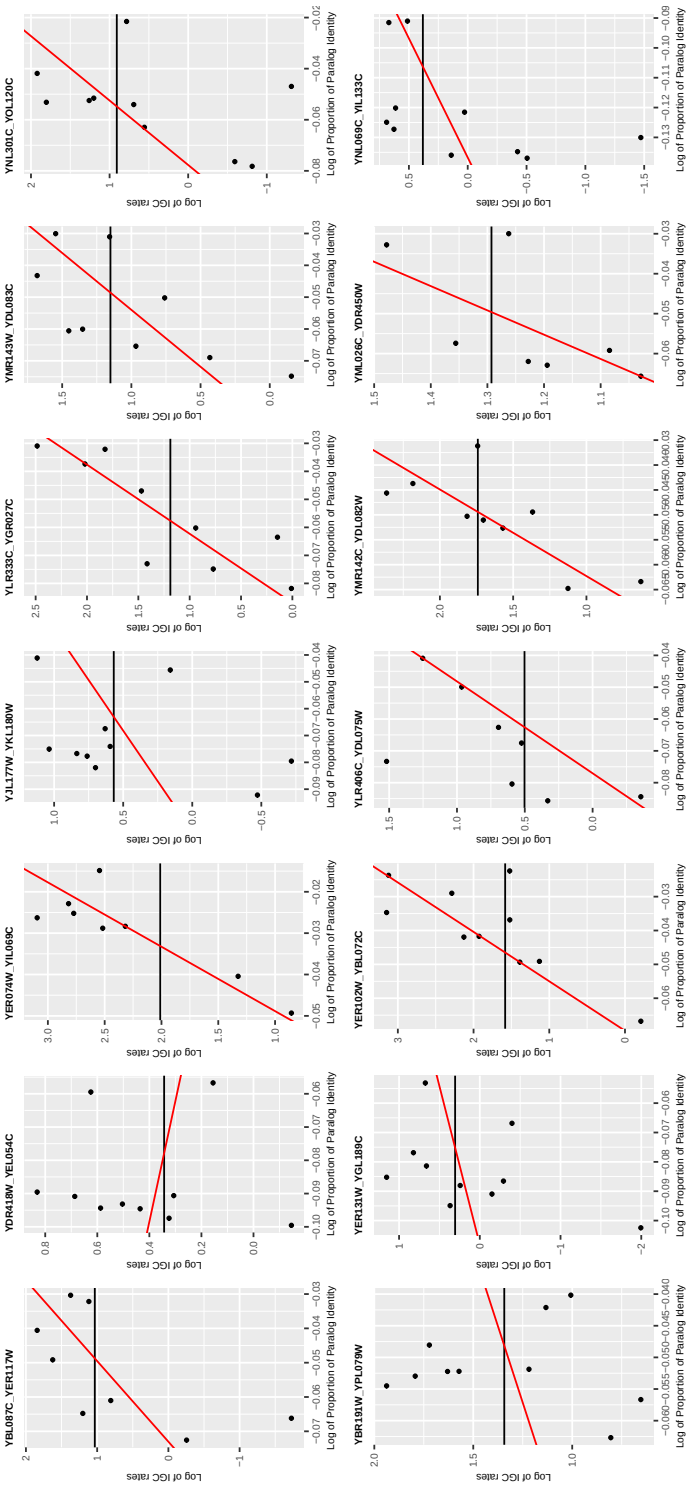


Figure 4.7: The log of IGC rate vs log paralog identity. In each plot, the red line represents the parametric form of the Divergence model; the black line represents the $\log(\tau)$ as estimated from the Ji et al. (2016) model; the dots represent branch-specific IGC rate estimates.

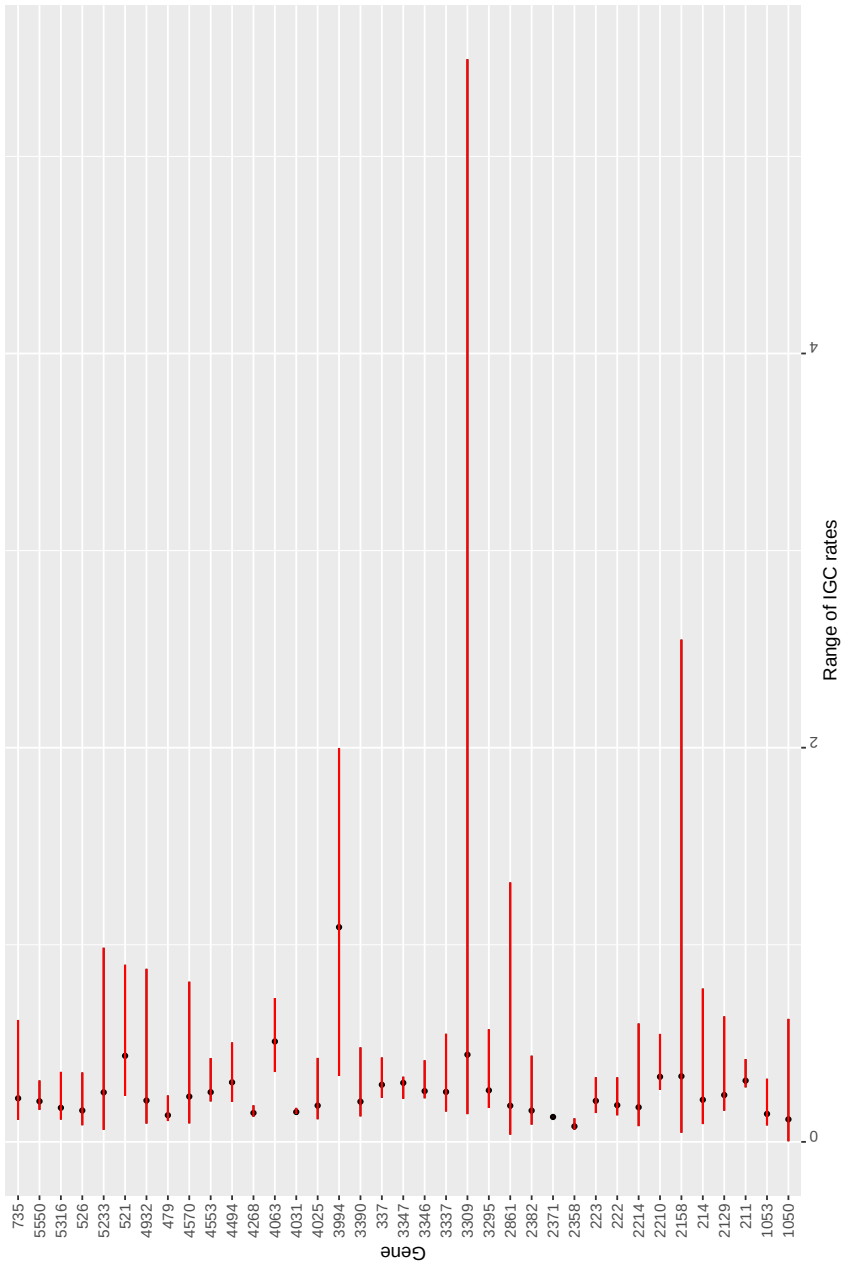


Figure 4.8: The range of estimated branch-specfic IGC rates from the Divergence model versus the IGC rates estimated from the Ji et al. (2016) model for 37 Teleost data sets.

Table 4.5: Estimated standard deviations for the 37 Teleost genes. The "$s(\widehat{\tau_D})$" and "$s(\widehat{K})$" columns show estimated standard deviations from the piecewise-homogeneous implementation from an inverted Hessian matrix when $\log P(X|\widehat{\theta}, \widehat{I}_D)$ is treated as the log-likelihood and all parameters but K and τ_D are fixed at their estimated values.

Data Set	$\widehat{\tau_D}$	$\widehat{K}$	$s(\widehat{\tau_D})$	$s(\widehat{K})$	Data Set	$\widehat{\tau_D}$	$\widehat{K}$	$s(\widehat{\tau_D})$	$s(\widehat{K})$
211	0.67	3.93	0.46	3.54	3295	1.96	7.86	1.23	2.63
214	3.96	12.6	1.75	2.4	3309	68.23	22.34	4.1	0.55
222	0.87	5.04	0.49	1.97	3337	2.73	7.22	1.25	1.55
223	1.99	6.82	1.29	2.13	3346	0.91	5.12	0.49	2.24
337	1.5	4.9	0.92	1.92	3347	0.16	-3.19	0.14	4.41
479	0.49	3.54	0.36	2.13	3390	1.12	5.16	0.74	2.19
521	7.66	9.28	3.65	1.71	3994	7.72	14	1.98	2.12
526	1.42	5.46	0.81	1.56	4025	0.93	5.41	0.33	1.29
735	5.54	15.6	1.26	1.38	4031	0.23	0.85	0.26	2.29
1050	33.16	22.41	1.2	0.2	4063	3.11	10.47	1.29	2.54
1053	1.67	9.62	1.19	3.19	4268	0.35	3.97	0.77	10.49
2129	2.45	8.88	1.19	2.11	4494	3.1	7.47	1.67	1.89
2158	56	26.35	1.79	0.36	4553	0.87	3.87	0.57	2.13
2210	1.24	4.9	0.75	2.31	4570	4.5	10.89	1.61	1.6
2214	4.1	10.87	1.75	1.82	4932	10.38	7.85	1.47	0.4
2358	0.25	3.88	0.61	8.44	5233	20.48	19.24	2.41	0.47
2371	0.12	-0.07	0.14	3.29	5316	1.3	6.92	0.45	1.28
2382	4.59	10.29	3.39	2.57	5550	0.42	2.43	0.36	3.06
2861	53.66	29.09	16.08	2.81					

CHAPTER

5

FUTURE WORK

5.1 Causal inference

Causal inference has been widely used in clinical trials (Lipkovich et al. 2020) and AB tests (Syrgkanis et al. 2021) to study the effects of measurement, policy and treatment. However, its application in financial statistics is not common. Specifically, investigating how pairs trading can benefit from causal inference by identifying relationships and executing trades based on historical data is less quantitatively developed. Traditionally, researchers relied on statistical models, such as copula (Trivedi et al. 2007), to explore the relationship between target stock prices and paired factors, such as prices of other assets (e.g., stocks,

commodities, bonds) or economic metrics (e.g., index, interest rate). By doing so, traders have been able to manage their portfolios through long/short positions, ultimately seeking to generate profits.

Observational studies that rely only on associations can be less effective than studies using causal inference. Thus, the use of a causal inference approach in pairs trading can be more beneficial. Let Y_i be the prices for target stock, and X_i be the financial factors during the time period $i = 1, \ldots, n$, and let Z be the other factors affecting these prices. The objective is to investigate the parameter τ in the equation

$$Y = \tau X + f(Z)$$

.

If considering Y_i as a categorical factor, the AMW can be implemented to estimate the effect τ. The AMW estimator described in Chapter 2 is initially derived using parametric nuisance models. However, the financial data sets often contain complex covariates that demand the utilization of robust machine learning models. The double/debiased machine learning approach (Chernozhukov et al. 2018) offers frameworks with great flexibility for utilizing machine learning methods to deal with nonlinear, autocorrelated, and high-dimensional covariates. This approach allows for the integration of powerful machine learning techniques and also establishes theoretical foundations for employing machine learning models within the given framework. The asymptotic properties of the AMW estimator may be established based on nonparametric nuisance models. However, it is important to choose appropriate models and ensure interoperability when employing AMW carefully. Furthermore, the approach discussed in Chapter 2 for determining tuning parameters can also be adapted for machine learning applications.

On the other hand, if Y_i is a continuous factor, the double/debiased machine learning

approach offers a viable method to investigate causal relationships. The approach involves obtaining residuals ϵ_X and ϵ_Y by conditioning on the models f_x and f_y respectively:

$$\epsilon_X = f_x(Z) - X,\ \epsilon_Y = f_y(Z) - Y.$$

Hence, τ is obtained by linear regression as:

$$\epsilon_Y = \tau \epsilon_X + \epsilon.$$

The other ad-hoc aspect of the model pertains to the sampling of prices in the target stocks, which is complicated by the correlation among the data points. To ensure the model is validated and useful for real trading, distinct sampling strategies should be implemented for high-frequency, middle-frequency, and low-frequency trading.

5.2 Interlocus gene conversion

Segmental duplications result in there being additional copies of a portion of a genome. Immediately after a segmental duplication, the resulting copies are identical in sequence. With a segmental duplication, homogenization due to IGC is a possibility as soon as the first point mutation occurs. Both the Ji et al. (2016) parameterization and the Divergence model of Chapter 4 are suitable for segmental duplications because they allow homogenization due to IGC as soon as the first point mutation occurs.

The duplicated genes from yeast and teleosts that are analyzed in this thesis are the result of whole-genome duplications rather than segmental duplications. Whole-genome duplications are typically categorized as either having an autopolyploid or an allopolyploid origin. There has been careful thought about what constitutes autopolyploidy versus al-

lopolyploidy, and the distinction is not always clear (e.g., see Doyle and Sherman-Broyles 2017). For our purposes, allopolyploidy can be said to be the cases where the doubling in genome size is preceded by an interspecific hybridization. The yeast whole-genome duplication may have been an allopolyploid event (e.g., see Marcet-Houben and Gabaldón 2015) whereas the teleost duplication may have been an autopolyploid event (Parey et al. 2022).

When paralogs within a genome arise because interspecific hybridization leads to allopolyploidy, the paralogs from the different species can already differ in sequence at the time of the duplication event. These differences would not have been subject to homogenization until the time of the duplication event. However, the Ji et al. (2016) model and the Divergence model both assume that IGC can homogenize paralogs as soon as the first point mutations occur following the most recent common ancestor of the paralogs. A better treatment of IGC for allopolyploids might separate the time of most recent common ancestry from the time when the duplication event occurs. While a refinement in this direction seems worthy of future research, it will also be challenging in light of the estimation difficulties in studying IGC events that occur prior to the first post-duplication speciation event. Success in characterizing IGC following allopolyploidy may be highly dependent on the availability of evolutionary lineages that diverged soon before and soon after a whole-genome duplication.

5.2.1 Augmenting sequence histories

The Divergence model utilizes a piecewise-homogeneous implementation to approximate a nonhomogeneous Poisson process. This approach assumes a constant IGC rate throughout the corresponding branch, regardless of overall paralog divergence levels. We would like to relax this flawed assumption by instead basing IGC inference on (augmented) sequence

histories that are consistent with the observed sequence data at the tips of an evolutionary tree.

Each sequence history specifies exactly which DNA sequences existed during all periods represented by an evolutionary tree. This is accomplished by specifying the sequence at the root of the tree and also the instants and the branches where each specific sequence change occurred. With an augmented history, IGC rates could be influenced by all sites in two paralogs and could be adjusted for each sequence change.

While it is difficult with complicated models of sequence change to sum over all possible sequence histories that are consistent with observed data at tips of a tree and it is therefore difficult to calculate a (full) likelihood, it is less challenging for many models to calculate the probability density of a specific sequence history. An augmented sequence history needs to be consistent with the observed data so that an inference procedure would sample possible sequence histories according to their relative probability densities.

To sample augmented evolutionary histories, we developed an endpoint-conditioned sampling approach, such as described in work by Hobolth and Stone (2009). In our case, the endpoint-conditioned histories are simulated according to the Ji et al. (2016) model. The procedure is to change one endpoint-conditioned site history at a time. If the sampled ending states at a specific sequence location do not match the given ending states at that location, the sampled histories are discarded and regenerated until the sampled ending states align with the observed ending states.

Our basic approach is to sample histories at individual sequence locations according to a simple model that assumes all sequence locations (e.g., codons or nucleotide positions) to change independently. We then add a second step that involves accepting or rejecting the sampled history. Histories at a site are proposed independently of the histories at other sites, but histories are accepted or rejected according to the model of interest (e.g., the Divergence model). Joint consideration of all sequence locations allows models that have

dependence between sequence locations

The second part of the simulation focuses on the augmented history. In this stage, we augment the generated evolutionary histories using a parametric model (e.g., IGC rates that are $\tau_D I^K$). Applying the augmentation approach for sampling histories and matching them with the given ending states often results in rejection of proposed histories.

Our experience is that our augmentation approach can have satisfactory computational requirements for nucleotide-based models of sequence change that yield $4^2 = 16$ possible joint states. However, when treating protein-coding evolution with 61-state codon models, our experience is that the resulting 61^2 joint states make our implementation of the augmentation procedure become unsatisfactory in the amount of required computational time for accurate parameter inference. Future efforts to accelerate the computation seem warranted.

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Doyle, J. J. and Sherman-Broyles, S. (2017). Double trouble: taxonomy and definitions of polyploidy. *New Phytologist*, 213(2):487–493.

Hobolth, A. and Stone, E. A. (2009). Simulation from endpoint-conditioned, continuous-time markov chains on a finite state space, with applications to molecular evolution. *The annals of applied statistics*, 3(3):1204.

Ji, X., Griffing, A., and Thorne, J. L. (2016). A phylogenetic approach finds abundant inter-locus gene conversion in yeast. *Molecular biology and evolution*, 33(9):2469–2476.

Lipkovich, I., Ratitch, B., and Mallinckrodt, C. H. (2020). Causal inference and estimands in clinical trials. *Statistics in Biopharmaceutical Research*, 12(1):54–67.

Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS biology*, 13(8):e1002220.

Parey, E., Louis, A., Montfort, J., Guiguen, Y., Crollius, H. R., and Berthelot, C. (2022). An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication. *Genome Research*, 32(9):1685–1697.

Syrgkanis, V., Lewis, G., Oprescu, M., Hei, M., Battocchi, K., Dillon, E., Pan, J., Wu, Y., Lo, P., Chen, H., et al. (2021). Causal inference and machine learning in practice with econml and causalml: Industrial use cases at microsoft, tripadvisor, uber. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 4072–4073.

Trivedi, P. K., Zimmer, D. M., et al. (2007). Copula modeling: an introduction for practitioners. *Foundations and Trends® in Econometrics*, 1(1):1–111.

APPENDIX

APPENDIX

A

SUPPORTING INFORMATION FOR "AUGMENTED MATCH WEIGHTED ESTIMATORS FOR AVERAGE TREATMENT EFFECTS" IN CH2

A.1 Proof of lemma 2

We divide AMW estimator into two parts as a function $B(K)$ of K and constant C, where

$$\begin{aligned} B(K) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{M_{e(X),i}}{K}A_i\hat{R}_i + \frac{M_{e(X),i}}{K}(1-A_i)\hat{R}_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{m_{A_i}(e_i^*)(1-A_i) - m_{A_i}(e_i^*)A_i\right\}, \end{aligned}$$

$$C = \hat{\tau}_{\text{reg}} + \frac{1}{n}\sum_{i=1}^{n}\left[A_i\hat{R}_i - (1-A_i)\hat{R}_i\right].$$

Since both $R_i^*(A_i)$ and e_i^* are given, we can apply the theorems from Mack and Rosenblatt (1979) to illustrate the bias and variance of $m_{A_i}(e_i^*)$:

$$\frac{24f^3(e^*)}{K^2} \times Bias\left\{\hat{m}_{A_i}(e_i^*)\right\} = \begin{cases} \frac{(m_1 f)''(e^*) - m_1(e^*)f''(e^*)}{n_1^2} & \text{if } A_i = 1, \\ \frac{(m_0 f)''(e^*) - m_0(e^*)f''(e^*)}{n_0^2} & \text{if } A_i = 0. \end{cases}$$

$$Var\left\{\hat{m}_{A_i}(e_i^*)\right\} = \begin{cases} \mathbb{V}\{AR^*(1)|e^*\} & \text{if } A_i = 1, \\ \mathbb{V}\{(1-A)R^*(0)|e^*\} & \text{if } A_i = 0. \end{cases}$$

Since each unit is iid, add the single bias and variance to get the bias and variance of $B(K)$:

$$n^2 \times Var\{B(K)\} = \frac{n_0\mathbb{V}\{AR^*(1)|e^*\} + n_1\mathbb{V}\{(1-A)R^*(0)|e^*\}}{K},$$

$$\begin{aligned} \frac{24nf^3(e^*)}{K^2} \times Bias\{B(K)\} &= \left\{(m_1 f)''(e^*) - m_1(e^*)f''(e^*)\right\}\frac{n_0}{n_1^2} \\ &\quad - \left\{(m_0 f)''(e^*) - m_0(e^*)f''(e^*)\right\}\frac{n_1}{n_0^2}. \end{aligned}$$

A.2 Proof of theorem 1

A.2.1 Rewrite the estimator with KNN

For simplicity, $\hat{m}_{A_i}(e_i^*)$ is denoted as:

$$\hat{m}_{A_i}(e_i^*) = \frac{\sum_{j\in\mathscr{J}_K(i)} D_{R_{e_i}}\left(e_i^* - e_j^*\right) R_j^*(A_i) I_{(A_j=A_i)}}{\sum_{j\in\mathscr{J}_K(i)} D_{R_{e_i}}\left(e_i^* - e_j^*\right) I_{(A_j=A_i)}} = \frac{h_n(e_i^*)}{f_n(e_i^*)}.$$

Let $\hat{\tau}^{\theta^*}_{\text{AMW}}$ be the estimator with given scores, and rewrite our estimator into three parts:

$$\hat{\tau}^{\theta^*}_{\text{AMW}} = \hat{\tau}^{\theta^*}_{\text{reg}} + \hat{\tau}^{\theta^*}_1 - \hat{\tau}^{\theta^*}_0,$$

where

$$\hat{\tau}^{\theta^*}_1 = \frac{1}{n}\sum_{i=1}^{n}\left\{R_i^*(1)A_i + \hat{m}_{A_i}(e_i^*)(1-A_i)\right\},$$

$$\hat{\tau}^{\theta^*}_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_{A_i}(e_i^*)A_i + R_i^*(0)(1-A_i)\right\}.$$

Let's make expectation that $m_{A_i}(e_i^*) = \mathbb{E}\left\{R_i^*(A_i)|e_i^*\right\}$, and rewrite

$$\hat{\tau}^{\theta^*}_1 = \frac{1}{n}\sum_{i=1}^{n}\left[R_i^*(1)A_i + m_1(e_i^*)(1-A_i) + \left\{\hat{m}_1(e_i^*) - m_1(e_i^*)\right\}(1-A_i)\right],$$

$$\hat{\tau}^{\theta^*}_0 = \frac{1}{n}\sum_{i=1}^{n}\left[R_i^*(0)(1-A_i) + m_0(e_i^*)A_i + \left\{\hat{m}_0(e_i^*) - m_0(e_i^*)\right\}A_i\right].$$

Detect the relationship between $\hat{m}_1(e_i^*) - m_1(e_i^*)$ by lemma from Mack and Rosenblatt (1979):

$$h_n(\mathbf{x}) = f(\mathbf{x})\pi(\mathbf{x})m(\mathbf{x}) + O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right),$$

$$f_n(\mathbf{x}) = f(\mathbf{x})\pi(\mathbf{x}) + O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right).$$

Denote $\pi_0(e_i^*) = \mathbb{P}\left\{I_{(A_i=0)} \mid e_i^*\right\}$, $\pi_1(e_i^*) = \mathbb{P}\left\{I_{(A_i=1)} \mid e_i^*\right\}$. Hence by Taylor expansion:

$$\hat{m}_1(e_i^*) - m_1(e_i^*) = \frac{1}{f(e_i^*)\pi_1(e_i^*)}\left\{h_n(e_i^*) - f_n(e_i^*)m_1(e_i^*)\right\} + O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right).$$

We express the similar form for $\hat{m}_0(e_i^*) - m_0(e_i^*)$.

A.2.2 U stat

Basing on the definition, we know $m_{Ai}(e_i^*) = \mathbb{E}\left\{R_i^*(A_i)|e_i^*\right\} = 0$,

$$\begin{aligned}
\hat{\tau}_{\text{AMW}}^{\theta^*} = & \frac{1}{n}\sum_{i=1}^{n}\left\{u_1(X_i,\beta_1^*) - u_0(X_i,\beta_0^*) + R_i^*(1)A_i - R_i^*(0)(1-A_i)\right\} \\
& - \sum_{i=1}^{n}\left(A_i\frac{1}{n}\frac{1}{f(e_i^*)\pi_0(e_i^*)}\sum_{j=1}^{n}\left[1_{(A_j=0)}D_{R_{e_i^*}}\left(e_i^* - e_j^*\right)\left\{R_j^*(0) - \hat{m}_0(e_i^*)\right\}\right]\right) \\
& + \sum_{i=1}^{n}\left((1-A_i)\frac{1}{n}\frac{1}{f(e_i^*)\pi_1(e_i^*)}\sum_{j=1}^{n}\left[1_{(A_j=1)}D_{R_{e_i^*}}\left(e_i^* - e_j^*\right)\left\{R_j^*(1) - \hat{m}_1(e_i^*)\right\}\right]\right) \\
& + O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right).
\end{aligned}$$

Define

$$U_1 = \frac{1}{n^2}\sum_{i=1}^{n}A_i\frac{1}{f(e_i^*)}\pi_0(e_i^*)\sum_{j=1}^{n}1_{(A_j=0)}D_{R_{e_i^*}}\left\{f(e_i^*) - e_j^*\right\}\left\{R_j^*(0) - \hat{m}_0(e_i^*)\right\}.$$

Using U statistics (van der Vaart 2000):

$$U_1 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}h\left(Z_i, Z_j\right),$$

where $Z_i = (R_i^*(0), A_i, e_i^*)$, and

$$\begin{aligned} h(Z_i, Z_j) = & \frac{1}{2}[A_i 1_{(A_j=0)} \frac{1}{f(e_i^*)\pi_0(e_i^*)} D_{R_{e_i^*}}(e_i^* - e_j^*)\{R_j^*(0) - \hat{m}_0(e_i^*)\} \\ & + A_j 1_{(A_i=0)} \frac{1}{f(e_j^*)\pi_0(e_j^*)} D_{R_{e_i^*}}(e_j^* - e_i^*)\{R_i^*(0) - \hat{m}_0(e_j^*)\}] \\ = & \frac{1}{2}(\zeta_{ij} + \zeta_{ji}). \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\zeta_{ij}|Z_i) = & \mathbb{E}\left(A_i \frac{1}{f(e_i^*)\pi_0(e_i^*)} \mathbb{E}\left[1_{(A_j=0)} D_{R_{e_i^*}}(e_i^* - e_j^*)\{R_j^*(0) - \hat{m}_0(e_i^*)\} | Z_i \right] \right) \\ = & \mathbb{E}\left[A_i \frac{1}{f(e_i^*)\pi_0(e_i^*)} f(e_j^*)\pi_0(e_j^*)\{m_0(e_i^*) - \hat{m}_0(e_i^*)\} \right] \\ = & O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right). \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\zeta_{ji}|Z_i) = & \mathbb{E}\left[A_j \frac{1}{f(e_j^*)\pi_0(e_j^*)} 1_{(A_i=0)} D_{R_{e_j^*}}(e_j^* - e_i^*)\{R_i^*(0) - \hat{m}_0(e_i^*)\} | Z_i \right] \\ = & (1-A_i)\mathbb{E}\left(\mathbb{E}\left[A_j \frac{1}{f(e_j^*)\pi_0(e_j^*)} D_{R_{e_j^*}}(e_j^* - e_i^*)\{R_i^*(0) - \hat{m}_0(e_i^*)\} | Z_i, R_{e_j^*} \right] | Z_i \right) \\ = & (1-A_i)\frac{1-\pi_0(e_i^*)}{\pi_0(e_i^*)}\{R_i^*(0) - \hat{m}_0(e_i^*)\} + O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right). \end{aligned}$$

Hence,

$$\begin{aligned} \hat{\tau}_{\mathrm{AMW}}^{\theta^*} - \tau = & \frac{1}{n}\sum_{i=1}^{n}\{Y_i A_i - Y_i(1-A_i) - u_0(X_i, \beta_0^*)A_i + u_1(X_i, \beta_1^*)(1-A_i)\} - \tau \\ & - \frac{1}{n}\sum_{i=1}^{n}(1-A_i)\frac{1-\pi_0(e_i^*)}{\pi_0(e_i^*)} R_i^*(0) + \frac{1}{n}\sum_{i=1}^{n} A_i \frac{1-\pi_1(e_i^*)}{\pi_1(e_i^*)} R_i^*(1) \\ & + O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right). \end{aligned}$$

$$\hat{\tau}^{\theta^*}_{\mathrm{AMW}} - \tau = \frac{1}{n}\sum_{i=1}^{n}\left\{ Y_i A_i - Y_i(1-A_i) - u_0(X_i,\beta_0^*)A_i + u_1(X_i,\beta_1^*)(1-A_i)\right\} - \tau$$
$$+ \frac{1}{n}\sum_{i=1}^{n}(1-A_i)R_i^*(0) - \frac{1}{n}\sum_{i=1}^{n}A_i R_i^*(1)$$
$$- \frac{1}{n}\sum_{i=1}^{n}(1-A_i)\frac{1}{\pi_0(e_i^*)}R_i^*(0) + \frac{1}{n}\sum_{i=1}^{n}A_i\frac{1}{\pi_1(e_i^*)}R_i^*(1)$$
$$+ O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right).$$

$$\hat{\tau}^{\theta^*}_{\mathrm{AMW}} - \tau = \frac{1}{n}\sum_{i=1}^{n}\left\{ u_1\left(X_i,\beta_1^*\right) - u_0\left(X_i,\beta_0^*\right)\right\} - \tau$$
$$- \frac{1}{n}\sum_{i=1}^{n}(1-A_i)\frac{1}{\pi_0(e_i^*)}R_i^*(0) + \frac{1}{n}\sum_{i=1}^{n}A_i\frac{1}{\pi_1(e_i^*)}R_i^*(1)$$
$$+ O\left(\frac{K}{n}\right)^2 + O\left(\frac{1}{K}\right).$$

This linear form is identical to that of AIPW estimators. Hence, the AMW estimator shares a similar property with AIPW estimators.

A.2.3 The double robust property

The decomposition for the linear form is obtained as τ_1^* and τ_0^*, where

$$\tau_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[u_1\left(X_i,\beta_1^*\right) + A\frac{1}{\pi_1(e_i^*)}\{Y(1) - u_1\left(X_i,\beta_1^*\right)\}\right],$$

$$\tau_0^* = \frac{1}{n}\sum_{i=1}^{n} -\left[u_0\left(X_i,\beta_0^*\right) + (1-A_i)\frac{1}{\pi_0(e_i^*)}\{Y(0) - u_0\left(X_i,\beta_0^*\right)\right].$$

And the expectations of τ_1^* and τ_0^* are:

$$\mathbb{E}(\tau_1^*) = \mathbb{E}\{Y(1)\} + \mathbb{E}\left[\frac{\{A-\pi_1(e^*)\}}{\pi_1(e*)}\{Y(1) - u_1(X,\beta_1^*)\}\right],$$

$$-\mathbb{E}(\tau_0^*) = \mathbb{E}\{Y(0)\} + \mathbb{E}\left[\frac{\{1-A-\pi_0(e^*)\}}{\pi_0(e^*)}\{Y(0)-u_0(X,\beta_0^*)\}\right].$$

If only the outcome model is correct, we obtain $\mathbb{E}\{Y(1)|X\} = u_{1,true}(X,\beta_1)$, so

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\{A-\pi_1(e^*)\}}{\pi_1(e^*)}\{Y(1)-u_1(X,\beta_1^*)\}|X\right]\right) = 0.$$

We also obtain $\mathbb{E}\{Y(0)|X\} = u_{0,true}(X,\beta_0)$, so

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\{1-A-\pi_0(e^*)\}}{\pi_0(e^*)}\{Y(0)-u_0(X,\beta_0^*)\}|X\right]\right) = 0.$$

Therefore, the AMW estimator is unbiased when only the outcome model is correct.

If only the propensity score model is correct, we obtain

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\{A-\pi_1(e^*)\}}{\pi_1(e^*)}|X\right]\right) = \mathbb{E}\left\{\frac{\pi_{1,true}(e^*)-\pi_1(e^*)}{\pi_1(e^*)}\right\},$$

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\{1-A-\pi_0(e^*)\}}{\pi_0(e^*)}|X\right]\right) = \mathbb{E}\left\{\frac{\pi_{0,true}(e^*)-\pi_0(e^*)}{\pi_0(e^*)}\right\},$$

so

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\{A-\pi_1(e^*)\}}{\pi_1(e^*)}\{Y(1)-u_1(X,\beta_1^*)\}|X\right]\right) = 0,$$

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\{1-A-\pi_0(e^*)\}}{\pi_0(e^*)}\{Y(0)-u_0(X,\beta_0^*)\}|X\right]\right) = 0.$$

Therefore, the AMW estimator is unbiased when only the correct propensity score model.

The variance for the linear form is:

$$\Sigma_\tau^{\theta^*} = \mathbb{E}\left[\left\{\frac{1}{\pi_0(e^*)}\right\}^2 \mathbb{V}\{(1-A)R^*(0)\mid X\} + \left\{\frac{1}{\pi_1(e^*)}\right\}^2 \mathbb{V}\{AR^*(1)|X\}\right]$$
$$+\mathbb{E}\left[\mathbb{V}\{u_1(X,\beta_1^*)-u_0(X,\beta_0^*)-\tau\}\right].$$

A.3 Proof of theorem 2

The idea is identical to the proof for Theorem 1, and we keep the central step for obtaining Theorem 2.

A.3.1 Rewrite the estimator with KNN

Let $\hat{\tau}^{t,\theta*}_{\text{AMW}}$ be the estimator with given scores, and rewrite our estimator into three parts:

$$\hat{\tau}^{t,\theta^*}_{\text{AMW}} = \hat{\tau}^{t,\theta^*}_{\text{reg}} + \hat{\tau}^{t,\theta^*}_{1} - \hat{\tau}^{t,\theta^*}_{0},$$

where

$$\hat{\tau}^{t,\theta^*}_{1} = \frac{1}{n_1}\sum_{i=1}^{n} R^*_i(1)A_i,$$

$$\hat{\tau}^{t,\theta^*}_{0} = \frac{1}{n_1}\sum_{i=1}^{n} \hat{m}_{A_i}(e^*_i)A_i.$$

A.3.2 KNN transform

Let's make expectation that $m_{A_i}(e^*_i) = \mathbb{E}(R^*_i(A_i)|e^*_i)$, and rewrite

$$\hat{\tau}^{t,\theta^*}_{0} = \frac{1}{n_1}\sum_{i=1}^{n}\left[m_0(e^*_i)A_i + \left\{\hat{m}_0(e^*_i) - m_0(e^*_i)\right\}A_i\right].$$

Set $\pi_0(e^*_i) = \mathbb{P}\left\{I_{(A_i=0)} \mid e^*_i\right\}$, hence by Taylor expansion:

$$\hat{m}_0(e^*_i) - m_0(e^*_i) = \frac{1}{f(e^*_i)\pi_0(e^*_i)}\left\{h_n(e^*_i) - f_n(e^*_i)m_0(e^*_i)\right\} + O\left(\frac{K}{n_1}\right)^2 + O\left(\frac{1}{K}\right).$$

A.3.3 U stat

We rewrite the estimator:

$$\begin{aligned}\hat{\tau}^{t,\theta^*}_{\text{AMW}} = &\frac{1}{n_1}\sum_{i=1}^{n}\left\{A_i Y_i - A_i u_0(X_i,\beta_0^*) + A_i R_i^*(1) - m_0(e_i^*)A_i\right\}\\ &-A_i\frac{1}{n_1}\frac{1}{f(e_i^*)\pi_0(e_i^*)}\sum_{j=1}^{n}\left[1_{(A_j=0)}D_{R_{e_i^*}}\left(e_i^*-e_j^*\right)\left\{R_j^*(0)-\hat{m}_0(e_i^*)\right\}\right]\\ &+O\left(\frac{K}{n_1}\right)^2+O\left(\frac{1}{K}\right).\end{aligned}$$

By the U stat (van der Vaart 2000):

$$\begin{aligned}\mathbb{E}(\zeta_{ji}|Z_i) =&\mathbb{E}\left[A_j\frac{1}{f\left(e_j^*\right)\pi_0(e_j^*)}1_{(A_i=0)}D_{R_{e_j^*}}\left(e_j^*-e^*{}_i\right)\left\{R_i^*(0)-\hat{m}_0(e_i^*)\right\}|Z_i\right]\\ =&(1-A_i)\mathbb{E}\left(\mathbb{E}\left[A_j\frac{1}{f\left(e_j^*\right)\pi_0(e_j^*)}D_{R_{e_j^*}}\left(e_j^*-e_i^*\right)\left\{R_i^*(0)-\hat{m}_0(e_i^*)\right\}|Z_i,R_{e_j^*}\right]|Z_i\right)\\ =&(1-A_i)\frac{1-\pi_0(e_i^*)}{\pi_0(e_i^*)}\left\{R_i^*(0)-\hat{m}_0(e_i^*)\right\}+O\left(\frac{K}{n_1}\right)^2+O\left(\frac{1}{K}\right).\end{aligned}$$

$$\begin{aligned}\hat{\tau}^{t,\theta^*}_{\text{AMW}}-\tau^t = &\frac{1}{n_1}\sum_{i=1}^{n}\left\{A_i Y_i - A_i u_0(X_{i,},\beta_0^*) + A_i R_i^*(1) - m_0(e_i^*)A_i\right\}-\tau^t\\ &-\frac{1}{n_1}\sum_{i=1}^{n}(1-A_i)\frac{1-\pi_0(e_i^*)}{\pi_0(e_i^*)}\left\{R_i(0)-\hat{m}_0(e_i^*)\right\}\\ &+O\left(\frac{K}{n_1}\right)^2+O\left(\frac{1}{K}\right).\end{aligned}$$

$$\begin{aligned}\hat{\tau}^{t,\theta^*}_{\text{AMW}}-\tau^t = &\frac{1}{n_1}\sum_{i=1}^{n}\left\{A_i Y_i - A_i u_0(X_{i,},\beta_0^*) + A_i R_i^*(1)\right\}-\tau^t\\ &-\frac{1}{n_1}\sum_{i=1}^{n}(1-A_i)\frac{1-\pi_0(e_i^*)}{\pi_0(e_i^*)}R_i^*(0)\\ &+O\left(\frac{K}{n_1}\right)^2+O\left(\frac{1}{K}\right).\end{aligned}$$

A.4 Proof of theorem 3

Suppose parameters are estimated, we can abstract the basic frame from (Abadie and Imbens 2016) and van der Vaart (2000) to do inference with $\theta^\top = (\alpha^\top, \beta_0^\top, \beta_1^\top)$. Under the true models, define $\theta^{*,\top} = (\alpha^{*,\top}, \beta_0^{*,\top}, \beta_1^{*,\top})$ satisfy $\mathbb{E}\left\{\psi(A, X, Y; \theta^*)\right\} = 0$, which can be obtain by M-estimation:

$$\Psi(\theta^*) \equiv \mathbb{E}\left\{\psi(A, X, Y; \theta^*)\right\} = \mathbb{E}\left\{\begin{array}{c} \psi_1(A, X, \alpha^*) \\ \psi_2\left(A, X, Y, \beta_0^*\right) \\ \psi_3\left(A, X, Y, \beta_1^*\right) \end{array}\right\} = 0.$$

[Assumption 3] 1. Estimator $\hat{\theta}$ is a zero of Ψ_n converge in probability to θ^* a zero of Ψ; 2. $\varphi(\theta)$ is score function as $E(\partial\psi(A, X, Y; \theta)/\partial\theta)$, which is nonsingular around θ^*.

By Taylor expansion

$$\sqrt{n}(\hat{\theta} - \theta^*) = -\varphi_{\theta^*}^- * \Psi(\theta^*) + o_p(1),$$

and

$$\Psi(\theta^*) \xrightarrow{d} N(0, I_{\theta^*}),$$

apply delta method, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \varphi_{\theta^*}^- I_{\theta^*} (\varphi_{\theta^*}^-)^\top),$$

where I_{θ^*} is $\mathbb{E}\left\{\psi(\theta^*)\psi(\theta^*)^\top\right\}$, and denote $\varphi_{\theta^*}^- I_{\theta^*} (\varphi_{\theta^*}^-)^\top = \Sigma_{\theta^*}$

[Assumption 4] Regularity Conditions for Local Normality Define any θ_n around fixed θ^*, where $\theta_n = \theta^* + h/\sqrt{n}$, which can be treated as local shift from θ^*.

1. $S_i^{\theta_n}$ are scored based on θ_n, which have a compact and convex supporting set and are

continuous and bounded away from zero.

2. $u(X_i, \theta_n)$ and $\sigma^2(\theta_n)$ are satisfying Lipschitz continuity conditions.

3. $E\left\{R(A_i)^3|\theta_n\right\}$ is uniformly bounded.

Gain the likelihood ratio based on θ^* and θ_n. Let p^{θ^*} be a density from $\mathbb{P}^n_{\theta^*}$, and take $l^{\theta^*} = log(p^{\theta^*})$. Apply the asymptotic local normal theory (van der Vaart 2000) and log density from (Yang and Zhang 2020):

$$log\left(\frac{p^{\theta^*}}{p^{\theta_n}}\right) = -h^T \varphi(\theta^*) * I^-_{\theta^*} * \Psi(\theta^*) - \frac{1}{2} h^T \Sigma_{\theta^*} h + o_p(1).$$

A.4.1 Asymptotic joint distribution

Under distribution of θ_n, the joint distribution can be derived as:

$$\left\{ \begin{array}{c} n^{1/2}\left(\hat{\tau}^{\theta_n}_{\text{AMW}} - \tau^{\theta_n}\right) \\ n^{1/2}\left(\hat{\theta} - \theta_n\right) \\ \log\left(p^{\theta^*}/p^{\theta_n}\right) \end{array} \right\} \xrightarrow{d} \mathcal{N}\left\{ \left(\begin{array}{c} 0 \\ 0 \\ -\frac{1}{2} h^\top \Sigma^{-1}_{\theta^*} h \end{array} \right), AV \right\},$$

$$AV = \left(\begin{array}{ccc} \Sigma_\tau & C_1^\top \varphi^{-1}_{\theta^*} & -C_1^\top I^{-1}_{\theta^*} \varphi_{\theta^*} h \\ \varphi^{-1}_{\theta^*} C_1 & \Sigma_{\theta^*} & -h \\ -h^\top \varphi_{\theta^*} I^{-1}_{\theta^*} C_1 & -h & h^\top \Sigma^{-1}_{\theta^*} h \end{array} \right),$$

where $C_1 = Cov\left\{\Psi(\theta_n), n^{1/2}\left(\hat{\tau}^{\theta_n}_{\text{AMW}} - \tau^{\theta_n}\right)\right\}$.

A.4.2 Martingale theory

Define D_n be a linear form for

$$Cov\left\{\Psi(\theta_n), n^{1/2}\left(\hat{\tau}^{\theta_n}_{\text{AMW}} - \tau^{\theta_n}\right)\right\} = w_1^\top \Psi(\theta_n) + w_2\left\{n^{1/2}\left(\hat{\tau}^{\theta_n}_{\text{AMW}} - \tau^{\theta_n}\right)\right\}$$

with weight $w_1 = (w_{11}^\top, w_{12}^\top, w_{13}^\top)^\top$, w_2:

$$\begin{aligned}
D_n = {} & w_{11}^\top n^{-1/2} \sum_{i=1}^{n}\left[\frac{\partial e(X_i,\alpha_n)}{\partial \alpha}\frac{A_i - e(X_i,\alpha_n)}{e(X_i,\alpha_n)\{1-e(X_i,\alpha_n)\}}\right] \\
& + w_{12}^\top n^{-1/2} \sum_{i=1}^{n}\left[(1-A_i)\frac{\partial u_0\left(X_i,\beta_{0,n}\right)}{\partial \beta_0}\left\{Y_i - u_0\left(X_i,\beta_{0,n}\right)\right\}\right] \\
& + w_{13}^\top n^{-1/2} \sum_{i=1}^{n}\left[A_i\frac{\partial u_1\left(X_i,\beta_{1,n}\right)}{\partial \beta_1}\left\{Y_i - u_1\left(X_i,\beta_{1,n}\right)\right\}\right] \\
& + w_2 n^{-1/2} \sum_{i=1}^{n}\left[\left\{u_1\left(X_i,\beta_{1,n}\right) - u_0\left(X_i,\beta_{0,n}\right)\right\} - \tau^{\theta_n}\right] \\
& - w_2 n^{-1/2} \sum_{i=1}^{n}\left[(1-A_i)\frac{1}{\pi_0(e_{i,n})}\left\{Y_i - u_0\left(X_i,\beta_{0,n}\right)\right\}\right] \\
& + w_2 n^{-1/2} \sum_{i=1}^{n}\left[A_i\frac{1}{\pi_1(e_{i,n})}\left\{Y_i - u_1\left(X_i,\beta_{1,n}\right)\right\}\right] + o_P(1).
\end{aligned}$$

Hence, rewrite $D_n = \Sigma_{i'=1}\phi_{n,i'}$ into martingale presentation to implement martingale central limit theorem:

For $0 \le i' \le n$,

$$\phi^1_{n,i'} = w_{11}^\top n^{-1/2}\left[\frac{\partial e(X_{i'},\alpha_n)}{\partial \alpha}\frac{e(X_{i'}) - e(X_{i'},\alpha_n)}{e(X_{i'},\alpha_n)\{1-e(X_{i'},\alpha_n)\}}\right],$$

$$\phi^2_{n,i'} = w_{11}^\top n^{-1/2}\left[\frac{\partial e(X_{i'},\alpha_n)}{\partial \alpha}\frac{A_{i'} - e(X_{i'})}{e(X_{i'},\alpha_n)\{1-e(X_{i'},\alpha_n)\}}\right],$$

$$\phi^{3}_{n,i'} = w_{12}^{\top} n^{-1/2} \{1 - e(X_{i'})\} \frac{\partial u_0\left(X_{i'}, \beta_{0,n}\right)}{\partial \beta_0} \left\{u_0(X_{i'}) - u_0\left(X_{i'}, \beta_{0,n}\right)\right\},$$

$$\phi^{4}_{n,i'} = -w_{12}^{\top} n^{-1/2} \{A_{i'} - e(X_{i'})\} \frac{\partial u_0\left(X_{i'}, \beta_{0,n}\right)}{\partial \beta_0} \left\{u_0(X_{i'}) - u_0\left(X_{i'}, \beta_{0,n}\right)\right\},$$

$$\phi^{5}_{n,i'} = w_{13}^{\top} n^{-1/2} e(X_{i'}) \frac{\partial u_1\left(X_{i'}, \beta_{1,n}\right)}{\partial \beta_1} \left\{u_1(X_{i'}) - u_1\left(X_{i'}, \beta_{1,n}\right)\right\},$$

$$\phi^{6}_{n,i'} = w_{13}^{\top} n^{-1/2} \{A_{i'} - e(X_{i'})\} \frac{\partial u_1\left(X_{i'}, \beta_{1,n}\right)}{\partial \beta_1} \left\{u_1(X_{i'}) - u_1\left(X_{i'}, \beta_{1,n}\right)\right\},$$

$$\phi^{7}_{n,i'} = w_2 n^{-1/2} \left[\left\{u_1\left(X_{i'}, \beta_{1,n}\right) - u_0\left(X_{i'}, \beta_{0,n}\right)\right\} - \tau^{\theta_n}\right],$$

$$\phi^{8}_{n,i'} = -w_2 n^{-1/2} (1 - A_{i'}) \frac{1}{\pi_0(e_{i',n})} \left\{u_0(X_{i'}) - u_0\left(X_{i'}, \beta_{0,n}\right)\right\},$$

$$\phi^{9}_{n,i'} = w_2 n^{-1/2} A_{i'} \frac{1}{\pi_1(e_{i',n})} \left\{u_1(X_{i'}) - u_1\left(X_{i'}, \beta_{1,n}\right)\right\},$$

with σ field like $\sigma_{i'} = \sigma(A_1, \ldots, A_{i'}, X_1, \ldots, X_{i'})$.

For $n < i' \leq 2n$,

$$\phi^{10}_{n,i'} = w_{12}^{\top} n^{-1/2} \left\{1 - A_{(i'-n)}\right\} \frac{\partial u_0\left(X_{i'}, \beta_{0,n}\right)}{\partial \beta_0} \left[Y_{(i'-n)} - u_0\left\{X_{(i'-n)}\right\}\right],$$

$$\phi^{11}_{n,i'} = w_{13}^{\top} n^{-1/2} A_{(i'-n)} \frac{\partial u_1\left(X_{i'}, \beta_{1,n}\right)}{\partial \beta_1} \left[Y_{(i'-n)} - u_1\left\{X_{(i'-n)}\right\}\right],$$

$$\phi^{12}_{n,i'} = -w_2 n^{-1/2} \left\{1 - A_{(i'-n)}\right\} \frac{1}{\pi_0\left\{e_{(i'-n),n}\right\}} \left[Y_{(i'-n)} - u_0\left\{X_{(i'-n)}\right\}\right],$$

$$\phi^{13}_{n,i'} = w_2 n^{-1/2} A_{(i'-n)} \frac{1}{\pi_1\left\{e_{(i'-n),n}\right\}} \left[Y_{(i'-n)} - u_1\left\{X_{(i'-n)}\right\}\right],$$

with σ field like $\sigma_{i'} = \sigma(A_1, \ldots, A_n, X_1, \ldots, X_n, Y_{i'-1}, \ldots, Y_{i'-n})$.

It is easy to verify the martingale property based on given fields, then we can apply the central limitation theory as $D_n \sim N(0,V)$, where $V = \lim \Sigma_{i'=1}^{2n} \mathbb{E}(\phi_{n,i'}^2 | \sigma_{i'})$. We derive the form of C_1, and $C_1 = \left(c_1^\top, c_2^\top, c_3^\top\right)^\top$ with

$$
\begin{aligned}
c_1 = & \mathbb{E}\left[\left\{u_1\left(X,\beta_1^*\right) - u_0\left(X,\beta_0^*\right) - \tau^{\theta^*}\right\} \frac{\partial e\left(X,\alpha^*\right)}{\partial \alpha} \frac{A - e\left(X,\alpha^*\right)}{e\left(X,\alpha^*\right)\{1 - e\left(X,\alpha^*\right)\}}\right] \\
& + \mathbb{E}\left[\frac{1}{\pi_0(e^*)}\left\{u_0(X) - u_0\left(X,\beta_0^*\right)\right\} \frac{\partial e\left(X,\alpha^*\right)}{\partial \alpha} \frac{e\left(X,\alpha^*\right)}{e\left(X,\alpha^*\right)\{1 - e\left(X,\alpha^*\right)\}}\right] \\
& + \mathbb{E}\left[\frac{1}{\pi_1(e^*)}\left\{u_1(X) - u_1\left(X,\beta_1^*\right)\right\} \frac{\partial e\left(X,\alpha^*\right)}{\partial \alpha} \frac{1 - e\left(X,\alpha^*\right)}{e\left(X,\alpha^*\right)\{1 - e\left(X,\alpha^*\right)\}}\right],
\end{aligned}
$$

$$
\begin{aligned}
c_2 = & \mathbb{E}\left[\left\{u_1\left(X,\beta_1^*\right) - u_0\left(X,\beta_0^*\right) - \tau^{\theta^*}\right\}(1 - A) \frac{\partial u_0\left(X,\beta_0^*\right)}{\partial \beta_0}\left\{u_0(X) - u_0\left(X,\beta_0^*\right)\right\}\right] \\
& - \mathbb{E}\left[\frac{\partial u_0\left(X,\beta_0^*\right)}{\partial \beta_0}\left\{u_0(X) - u_0\left(X,\beta_0^*\right)\right\}^2 \frac{1}{\pi_0(e^*)}\right] \\
& - \mathbb{E}\left\{\frac{\partial u_0\left(X,\beta_0^*\right)}{\partial \beta_0} \sigma_0^2 \frac{1}{\pi_0(e^*)}\right\},
\end{aligned}
$$

$$
\begin{aligned}
c_3 = & \mathbb{E}\left[\left\{u_1\left(X,\beta_1^*\right) - u_0\left(X,\beta_0^*\right) - \tau^{\theta^*}\right\} A \frac{\partial u_1\left(X,\beta_1^*\right)}{\partial \beta_1}\left\{u_1(X) - u_1\left(X,\beta_1^*\right)\right\}\right] \\
& + \mathbb{E}\left[\frac{\partial u_1\left(X,\beta_1^*\right)}{\partial \beta_1}\left\{u_1(X) - u_1\left(X,\beta_1^*\right)\right\}^2 \frac{1}{\pi_1(e^*)}\right] \\
& + \mathbb{E}\left\{\frac{\partial u_1\left(X,\beta_1^*\right)}{\partial \beta_1} \sigma_1^2 \frac{1}{\pi_1(e^*)}\right\}.
\end{aligned}
$$

A.4.3 Le Cam's third theory

Abstract Le Cam's third theory from :

If

$$
\left(X_n, \log \frac{dQ_n}{dP_n}\right) \xrightarrow{P_n} N_{k+1}\left\{\begin{pmatrix} u \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right\},
$$

then

$$X_n \xrightarrow{Q_n} N_k(u+\tau,\Sigma).$$

Le Cam's third theory is applied to get asymptotic distribution for the AWM estimator. Firstly, the specific form of τ^{θ_n} is provided with the AIPW estimator since AIPW is related to both propensity scores and prognostic scores:

$$\tau^{\theta_n} = \mathbb{E}\left[u_1\left(X,\beta_{1,n}\right) - u_0\left(X,\beta_{0,n}\right) + \left\{\frac{AR_n}{e(X,\alpha_n)} - \frac{(1-A)R_n}{1-e(X,\alpha_n)}\right\}\right].$$

Then, do Taylor expansion:

$$\tau^{\theta_n} = \tau^{\theta^*} + \frac{\partial \tau^{\theta}}{\partial \theta}|_{\theta=\theta^*}(\theta^* - \theta_n) + o(n^{-1/2}).$$

What's more, denote $C_2 = (\frac{\partial \tau^{\theta}}{\partial \theta}|_{\theta=\theta^*})$, under distribution of θ^* and apply Le Cam' third lemma to get:

$$\begin{Bmatrix} n^{1/2}\left(\hat{\tau}^{\theta_n}_{\mathrm{AMW}} - \tau\right) \\ n^{1/2}\left(\hat{\theta} - \theta_n\right) \end{Bmatrix} \rightarrow \mathcal{N}\left\{\begin{pmatrix} -h^{\top}\varphi_{\theta^*} I^{-1}_{\theta^*} C_1 - C_2^{\top} h \\ -h \end{pmatrix}, \begin{pmatrix} \Sigma_{\tau} & C_1^{\top}\varphi^{-1}_{\theta^*} \\ \varphi^{-1}_{\theta^*} C_1 & \Sigma_{\theta^*} \end{pmatrix}\right\}.$$

By replacing θ_n with θ^*,

$$\begin{Bmatrix} n^{1/2}\left(\hat{\tau}^{\theta^*+h/\sqrt{n}}_{\mathrm{AMW}} - \tau\right) \\ n^{1/2}\left(\hat{\theta} - \theta^*\right) \end{Bmatrix} \rightarrow \mathcal{N}\left\{\begin{pmatrix} -h^{\top}\varphi_{\theta^*} I^{-1}_{\theta^*} C_1 - C_2^{\top} h \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{\tau} & C_1^{\top}\varphi^{-1}_{\theta^*} \\ \varphi^{-1}_{\theta^*} C_1 & \Sigma_{\theta^*} \end{pmatrix}\right\}.$$

Apply the conditional normal distribution:

$$n^{1/2}\left(\hat{\tau}^{\hat{\theta}}_{\mathrm{AMW}} - \tau\right) | n^{1/2}\left(\hat{\theta} - \theta^*\right) = h \rightarrow N\left(C_2^{\top} h, \Sigma_{\tau} - C_1^{\top} I^{-1}_{\theta^*} C_1\right).$$

Marginalize over the h to get unconditional distribution:

$$n^{1/2}\left(\hat{\tau}_{\text{AMW}}^{\theta^*+h/\sqrt{n}}-\tau\right)\rightarrow N\left(0,\Sigma_\tau-C_1^\top I_{\theta^*}^{-1}C_1+C_2^\top\Sigma_{\theta^*}C_2\right).$$

A.5 Proof of theorem 4

The idea is identical to the proof for Theorem 3, and we just give the main steps to derive Theorem 4.

A.5.1 Asymptotic joint distribution

Under distribution of θ_n, the joint distribution can be derived as:

$$\left\{\begin{array}{c} n^{1/2}\left(\hat{\tau}_{\text{AMW}}^{t,\theta_n}-\tau^{t,\theta_n}\right) \\ n^{1/2}\left(\hat{\theta}-\theta_n\right) \\ \log\left(p^{\theta^*}/p^{\theta_n}\right) \end{array}\right\}\xrightarrow{d}\mathcal{N}\left\{\left(\begin{array}{c} 0 \\ 0 \\ -\frac{1}{2}h^\top\Sigma_{\theta^*}^{-1}h \end{array}\right),AV^t\right\},$$

$$AV^t=\left(\begin{array}{ccc} \Sigma_\tau^t & C_1^{t,\top}\varphi_{\theta^*}^{-1} & -C_1^{t,\top}I_{\theta^*}^{-1}\varphi_{\theta^*}h \\ \varphi_{\theta^*}^{-1}C_1^t & \Sigma_{\theta^*} & -h \\ -h^\top\varphi_{\theta^*}I_{\theta^*}^{-1}C_1^t & -h & h^\top\Sigma_{\theta^*}^{-1}h \end{array}\right),$$

where $C_1^t=Cov\left\{\Psi(\theta_n),n^{1/2}\left(\hat{\tau}_{\text{AMW}}^{t,\theta_n}-\tau^{t,\theta_n}\right)\right\}$.

A.5.2 Martingale theory

Define D_n^t be a linear form:

$$Cov\left\{\Psi(\theta_n),n^{1/2}\left(\hat{\tau}_{\text{AMW}}^{t,\theta_n}-\tau^{t,\theta_n}\right)\right\}=w_1^{t,\top}\Psi(\theta_n)+w_2^t\left\{n^{1/2}\left(\hat{\tau}_{\text{AMW}}^{t,\theta_n}-\tau^{t,\theta_n}\right)\right\},$$

with weight $w_1^t = (w_{11}^{t,\top}, w_{12}^{t,\top}, w_{13}^{t,\top})^\top, w_2^t$:

$$
\begin{aligned}
D_n^t = {} & w_{11}^{t,\top} n_1^{-1/2} \sum_{i=1}^{n} \left[\frac{\partial e(X_i, \alpha_n)}{\partial \alpha} \frac{A_i - e(X_i, \alpha_n)}{e(X_i, \alpha_n)\{1 - e(X_i, \alpha_n)\}} \right] \\
& + w_{12}^{t,\top} n_1^{-1/2} \sum_{i=1}^{n} \left[(1 - A_i) \frac{\partial u_0\left(X_i, \beta_{0,n}\right)}{\partial \beta_0} \left\{ Y_i - u_0\left(X_i, \beta_{0,n}\right) \right\} \right] \\
& + w_{13}^{t,\top} n_1^{-1/2} \sum_{i=1}^{n} \left[A_i \frac{\partial u_1\left(X_i, \beta_{1,n}\right)}{\partial \beta_1} \left\{ Y_i - u_1\left(X_i, \beta_{1,n}\right) \right\} \right] \\
& + w_2^t n_1^{-1/2} \sum_{i=1}^{n} \left[\left\{ Y_i - u_0\left(X_i, \beta_{0,n}\right) \right\} A_i - \tau^{t,\theta_n} \right] \\
& - w_2^t n_1^{-1/2} \sum_{i=1}^{n} \left[(1 - A_i) \frac{1}{\pi_0(e_{i,n})} \left\{ Y_i - u_0\left(X_i, \beta_{0,n}\right) \right\} \right] \\
& + w_2^t n_1^{-1/2} \sum_{i=1}^{n} \left[A_i \left\{ Y_i - u_1\left(X_i, \beta_{1,n}\right) \right\} \right] + o_P(1).
\end{aligned}
$$

Apply Martingale theory (Andreou and Werker 2012), for $0 \le i' \le n$,

$$
\phi_{n,i'}^{t,1} = w_{11}^{t,\top} n_1^{-1/2} \left[\frac{\partial e(X_{i'}, \alpha_n)}{\partial \alpha} \frac{e(X_{i'}) - e(X_{i'}, \alpha_n)}{e(X_{i'}, \alpha_n)\{1 - e(X_{i'}, \alpha_n)\}} \right],
$$

$$
\phi_{n,i'}^{t,2} = w_{11}^{t,\top} n_1^{-1/2} \left[\frac{\partial e(X_{i'}, \alpha_n)}{\partial \alpha} \frac{A_{i'} - e(X_{i'})}{e(X_{i'}, \alpha_n)\{1 - e(X_{i'}, \alpha_n)\}} \right],
$$

$$
\phi_{n,i'}^{t,3} = w_{12}^{t,\top} n_1^{-1/2} \{1 - e(X_{i'})\} \frac{\partial u_0\left(X_{i'}, \beta_{0,n}\right)}{\partial \beta_0} \left\{ u_0(X_{i'}) - u_0\left(X_{i'}, \beta_{0,n}\right) \right\},
$$

$$
\phi_{n,i'}^{t,4} = -w_{12}^{t,\top} n_1^{-1/2} \{A_{i'} - e(X_{i'})\} \frac{\partial u_0\left(X_{i'}, \beta_{0,n}\right)}{\partial \beta_0} \left\{ u_0(X_{i'}) - u_0\left(X_{i'}, \beta_{0,n}\right) \right\},
$$

$$
\phi_{n,i'}^{t,5} = w_{13}^{t,\top} n_1^{-1/2} e(X_{i'}) \frac{\partial u_1\left(X_{i'}, \beta_{1,n}\right)}{\partial \beta_1} \left\{ u_1(X_{i'}) - u_1\left(X_{i'}, \beta_{1,n}\right) \right\},
$$

$$
\phi_{n,i'}^{t,6} = w_{13}^{t,\top} n_1^{-1/2} \{A_{i'} - e(X_{i'})\} \frac{\partial u_1\left(X_{i'}, \beta_{1,n}\right)}{\partial \beta_1} \left\{ u_1(X_{i'}) - u_1\left(X_{i'}, \beta_{1,n}\right) \right\},
$$

$$
\phi_{n,i'}^{t,7} = w_2^t n_1^{-1/2} \left[\left\{ u\left(X_{i'}, \beta_{1,n}\right) - u\left(X_{i'}, \beta_{0,n}\right) \right\} A_{i'} - \tau^{t,\theta_n} \right],
$$

$$\phi_{n,i'}^{t,8} = w_2^t n_1^{-1/2} A_{i'} \left\{ u_1(X_{i'}) - u_1\left(X_{i'}, \beta_{1,n}\right) \right\},$$

$$\phi_{n,i'}^{t,9} = w_2^t n_1^{-1/2} \left\{ u_1\left(X_{i'}, \beta_{1,n}\right) - u_0\left(X_{i'}, \beta_{0,n}\right) \right\} A_{i'} - \tau^{\theta_n},$$

with σ field like $\sigma_{i'} = \sigma(A_1, \ldots, A_{i'}, X_1, \ldots, X_{i'})$.

For $n < i' \leq 2n$,

$$\phi_{n,i'}^{t,10} = w_2^t n_1^{-1/2} \left[\left\{ Y_{(i'-n)} - u_1\left(X_{(i'-n)}, \beta_{1,n}\right) \right\} A_{(i'-n)} \right],$$

$$\phi_{n,i'}^{t,11} = w_{12}^{t,\top} n_1^{-1/2} \left\{ 1 - A_{(i'-n)} \right\} \frac{\partial u_0\left(X_{(i'-n)}, \beta_{0,n}\right)}{\partial \beta_0} \left[Y_{(i'-n)} - u_0\left\{ X_{(i'-n)} \right\} \right],$$

$$\phi_{n,i'}^{t,12} = w_{13}^{t,\top} n_1^{-1/2} A_{(i'-n)} \frac{\partial u_1\left(X_{(i'-n)}, \beta_{1,n}\right)}{\partial \beta_1} \left[Y_{(i'-n)} - u_1\left\{ X_{(i'-n)} \right\} \right],$$

$$\phi_{n,i'}^{t,13} = -w_2^t n_1^{-1/2} \left\{ 1 - A_{(i'-n)} \right\} \frac{1}{\pi_0\left\{ e_{(i'-n),n} \right\}} \left[Y_{(i'-n)} - u_0\left\{ X_{(i'-n)} \right\} \right],$$

$$\phi_{n,i'}^{t,14} = w_2^t n_1^{-1/2} A_{(i'-n)} \left[Y_{(i'-n)} - u_1\left\{ X_{(i'-n)} \right\} \right],$$

with σ field like $\sigma_{i'} = \sigma(A_1, \ldots, A_n, X_1, \ldots, X_n, Y_{i'-1}, \ldots, Y_{i'-n})$.

And we obtain C_1^t with:

$$\begin{aligned} c_1^t = {} & \mathbb{E}\left[\left\{ u(X, \beta_1^*) A - u\left(X, \beta_0^*\right) A - \tau^{t,\theta^*} \right\} \frac{\partial e(X, \alpha^*)}{\partial \alpha} \frac{A - e(X, \alpha^*)}{e(X, \alpha^*)\{1 - e(X, \alpha^*)\}} \right] \\ & + \mathbb{E}\left[\frac{1}{\pi_0(e^*)} \left\{ u_0(X) - u_0\left(X, \beta_0^*\right) \right\} \frac{\partial e(X, \alpha^*)}{\partial \alpha} \frac{e(X, \alpha^*)}{e(X, \alpha^*)\{1 - e(X, \alpha^*)\}} \right] \\ & + \mathbb{E}\left[\left\{ u_1(X) - u_1\left(X, \beta_1^*\right) \right\} \frac{\partial e(X, \alpha^*)}{\partial \alpha} \frac{1 - e(X, \alpha^*)}{e(X, \alpha^*)\{1 - e(X, \alpha^*)\}} \right], \end{aligned}$$

$$\begin{aligned} c_2^t = {} & \mathbb{E}\left[\left\{ u\left(X, \beta_1^*\right) A - u\left(X, \beta_0^*\right) A - \tau^{t,\theta^*} \right\} (1 - A) \frac{\partial u_0\left(X, \beta_0^*\right)}{\partial \beta_0} \left\{ u_0(X) - u_0\left(X, \beta_0^*\right) \right\} \right] \\ & - \mathbb{E}\left[\frac{\partial u_0\left(X, \beta_0^*\right)}{\partial \beta_0} \left\{ u_0(X) - u_0\left(X, \beta_0^*\right) \right\}^2 \frac{1}{\pi_0(e^*)} \right] \\ & - \mathbb{E}\left\{ \frac{\partial u_0\left(X, \beta_0^*\right)}{\partial \beta_0} \sigma_0^2 \frac{1}{\pi_0(e^*)} \right\}, \end{aligned}$$

$$
\begin{aligned}
c_3^t =& \mathbb{E}\left[\left\{u\left(X,\beta_1^*\right)A - u\left(X,\beta_0^*\right)A - \tau^{t,\theta^*}\right\}A\frac{\partial\, u_1\left(X,\beta_1^*\right)}{\partial\,\beta_1}\left\{u_1(X) - u_1\left(X,\beta_1^*\right)\right\}\right] \\
&+\mathbb{E}\left[\frac{\partial\, u_1\left(X,\beta_1^*\right)}{\partial\,\beta_1}\left\{u_1(X) - u_1\left(X,\beta_1^*\right)\right\}^2\right] \\
&+\mathbb{E}\left[\left\{Y - u_1(X)\right\}\frac{\partial\, u_1\left(X,\beta_1^*\right)}{\partial\,\beta_1}\left\{Y - u_1\left(X,\beta_1^*\right)\right\}\right] \\
&+\mathbb{E}\left\{\frac{\partial\, u_1\left(X,\beta_1^*\right)}{\partial\,\beta_1}\sigma_1^2\right\}.
\end{aligned}
$$

A.5.3 Le Cam's third theory

$$
\tau^{t,\theta_n} = \mathbb{E}\left[YA - u_0\left(X,\beta_{0,n}\right)A + \left\{AR_n - \frac{(1-A)e(X,\alpha_n)R_n}{1-e(X,\alpha_n)}\right\}\right].
$$

By Taylor expansion:

$$
\tau^{t,\theta_n} = \tau^{t,\theta^*} + \frac{\partial\,\tau^{t,\theta}}{\partial\,\theta}|_{\theta=\theta^*}(\theta^* - \theta_n) + o(n^{-1/2}).
$$

Marginalize over the h to get unconditional distribution:

$$
n^{1/2}\left(\hat{\tau}_{\text{AMW}}^{t,\theta^*+h/\sqrt{n}} - \tau^t\right) \to N\left(0, \Sigma_\tau^t - C_1^{t,\top} I_{\theta^*}^{-1} C_1^t + C_2^{t,\top}\Sigma_{\theta^*}C_2^t\right).
$$

A.6 Tables

Table A.1: Standard difference for NSW-DS data regarding ATE.

covariate	age	educ	black	hisp	married	nodegr	re75
before	0.107	0.144	0.044	-0.170	0.094	-0.306	0.084
after	0.014	0.046	0.069	-0.090	-0.001	-0.016	-0.062

Table A.2: Standard difference for NSW-DS data regarding ATT.

covariate	age	educ	black	hisp	married	nodegr	re75
before	0.107	0.144	0.044	-0.170	0.094	-0.306	0.084
after	0.013	0.046	0.069	-0.092	-0.003	-0.016	-0.062

Table A.3: Standard difference for ACGT175 data regarding ATE.

covariate	age	wtkg	hemo	homo	drugs	karnof	oprior	z30
before	0.036	-0.090	0.017	0.008	0.072	0.084	0.010	0.066
after	0.016	0.004	0.007	-0.003	0.006	-0.002	0.003	0.000
covariate	preanti	race	gender	str2	strat	symptom	cd40	cd80
before	0.027	-0.042	-0.001	0.085	0.072	0.048	-0.022	0.012
after	0.008	-0.014	-0.004	0.005	0.006	0.005	-0.012	0.008

Table A.4: Standard difference for RHC data regarding ATT (1).

covariate	age	sexMale	raceother	racewhite	edu
before	-0.061	0.093	0.021	0.015	0.091
after	0.003	0.033	-0.021	0.018	-0.008
covariate	income	income	income	ninsclas	nonsclasMedicare
	$11-$25k	$25-$50k	Under $11K	Medicare	&Medicaid
before	0.015	0.108	-0.124	-0.075	-0.058
after	-0.012	-0.003	0.021	-0.019	0.027
covariate	ninsclas No	ninsclas	ninsclas	cat1CHF	cat1Cirrhosis
	insurance	Private	Private&Med		
before	0.043	0.137	0.035	0.097	-0.139
after	0	0.002	-0.002	0.006	-0.013
covariate	cat1Colon	cat1Coma	cat1COPD	cat1Lung	cat1MOSF
	Cancer			Cancer	w/Malignancy
before	-0.035	-0.198	-0.317	-0.089	0.018
after	-0.014	-0.016	-0.015	-0.041	0.001
covariate	cat1MOSF	das2d3pc	dnr1Yes	caNo	caYes
	w/Sepsis				
before	0.42	0.062	-0.219	0.103	-0.071
after	0.009	0.001	-0.04	0.012	-0.003
covariate	surv2md1	aps1	scoma1	wtkilo1	temp1
before	-0.198	0.491	-0.108	0.252	-0.021
after	-0.011	0.03	-0.024	0.043	-0.017
covariate	meanbp1	resp1	hrt1	pafi1	paco21
before	-0.438	-0.165	0.147	-0.419	-0.24
after	0.009	0.019	0.024	-0.003	-0.003

Table A.5: Standard difference for RHC data regarding ATT (2).

covariate	ph1	wblc1	hema1	sod1	pot1
before	-0.12	0.084	-0.262	-0.092	-0.027
after	0.01	0.084	-0.008	-0.019	-0.016
covariate	crea1	bili1	alb1	respYes	cardYes
before	0.268	0.148	-0.237	-0.265	0.295
after	0.005	-0.014	0.018	-0.006	0.021
covariate	neuroYes	gastrYes	renalYes	metaYes	hemaYes
before	-0.331	0.122	0.119	-0.028	-0.061
after	-0.016	-0.005	0.023	0.007	0.004
covariate	sepsYes	traumaYes	orthoYes	cardiohx	chfhx
before	0.238	0.111	0.028	0.117	0.07
after	0.008	0.005	0.002	0.004	0.005
covariate	dementhx	psychhx	chrpulhx	renalhx	liverhx
before	-0.158	-0.139	-0.188	0.032	-0.048
after	-0.036	0.002	0.022	0.018	-0.004
covariate	gibledhx	malighx	immunhx	transhx	amihx
before	-0.069	-0.101	0.081	0.173	0.076
after	-0.021	-0.011	0.001	0.016	-0.004

References

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84:781–807.

Andreou, E. and Werker, B. J. (2012). An alternative asymptotic analysis of residual-based statistics. *Rev Econ Stat*, 94:88–99.

Mack, Y. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9:1–15.

van der Vaart (2000). *Asymptotic Statistics*, volume 3. Cambridge university press, Cambridge: Cambridge University Press.

Yang, S. and Zhang, Y. (2020). Multiply robust matching estimators of average and quantile treatment effects. *arXiv preprint arXiv:2001.06049*.