van Veelen, Matthijs; Allen, Benjamin; Hoffman, Moshe; Simon, Burton; Veller, Carl

Working Paper

Inclusive Fitness

Tinbergen Institute Discussion Paper, No. 16-055/I

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: van Veelen, Matthijs; Allen, Benjamin; Hoffman, Moshe; Simon, Burton; Veller, Carl (2016) : Inclusive Fitness, Tinbergen Institute Discussion Paper, No. 16-055/I, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/145362

TI 2016-055/I
Tinbergen Institute Discussion Paper

Inclusive Fitness

Matthijs van Veelen[1]
Benjamin Allen[2]
Moshe Hoffman[3]
Burton Simon[4]
Carl Veller[5]

[1] *Faculty of Economics and Business, University of Amsterdam, and Tinbergen Institute, the Netherlands;*
[2] *Emmanuel College, Boston, United States;*
[3] *Rady School of Management, UC San Diego, United States;*
[4] *University of Colorado, Denver, United States;*
[5] *Harvard University, Cambridge, United States.*

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Inclusive Fitness

Matthijs van Veelen[1,2,3,*], Benjamin Allen[3,4,5],
Moshe Hoffman[3,6,7], Burton Simon[8], and Carl Veller[3,9].

This paper reviews and addresses a variety of issues relating to inclusive fitness. The main question is: are there limits to the generality of inclusive fitness, and if so, what are the perimeters of the domain within which inclusive fitness works? This question is addressed using two well known tools from evolutionary theory: the replicator dynamics, and adaptive dynamics. Both are combined with population structure. How generally Hamilton's rule applies depends on how costs and benefits are defined. We therefore consider costs and benefits following from Karlin & Matessi's (1983) "counterfactual method", and costs and benefits as defined by the "regression method" (Gardner et al., 2011). With the latter definition of costs and benefits, Hamilton's rule always indicates the direction of selection correctly, and with the former it does not. How these two definitions can meaningfully be interpreted is also discussed. We also consider cases where the qualitative claim that relatedness fosters cooperation holds, even if Hamilton's rule as a quantitative prediction does not.

We furthermore find out what the relation is between Hamilton's rule and Fisher's Fundamental Theorem of Natural Selection. We also consider cancellation effects – which is the most important deepening of our understanding of when altruism is selected for – and we discuss preference evolution. Finally we also explore the remarkable (im)possibilities for empirical testing with either definition of costs and benefits in Hamilton's rule.

[1]Department of Economics and Business, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. [2]Tinbergen Institute, The Netherlands. [3]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA. [4]Department of Mathematics, Emmanuel College, MA 02115, USA. [5]Center for Mathematical Sciences and Applications, Harvard University, Cambridge, MA 02138, USA. [6]Rady School of Management, UC San Diego, La Jolla, CA 92093, USA. [7]Department of Computer Science and Engineering, UC San Diego, La Jolla, CA 92093. [8]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80202, USA. [9]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. *Corresponding author: c.m.vanveelen@uva.nl.

1 Introduction

In 1964 Hamilton introduced the most famous rule in evolutionary biology. In two back to back papers, he formulated a model, and derived a rule from it that we now know as Hamilton's rule. That rule states that selection will favour altruistic behaviour if the benefits to the recipient times the relatedness between actor and recipient outweigh the costs to the actor. This captured both a qualitative insight – genes can make the individuals that they are in do things that are bad for that particular individual, but good for copies of that gene in other individuals – and an elegant, intuitive, and simple quantification of that phenomenon. Both Hamilton's rule, and the notion of "inclusive fitness", which the rule suggests is maximized by evolution, have since become standard material for both theoretically and empirically inclined biologists. As is natural for a landmark paper, it came with indications that also signal to outsiders that this is an important result. The paper has about half as many citations as Darwin's "*On the Origin of Species*", and is one of the core ingredients in Richard Dawkins' "*The Selfish Gene*", which is one of the must-read books for anyone with a general interest in science.

Besides being a monumental breakthrough, Hamilton's rule is also the topic of a controversy. In the early '80s Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) already suggested that not all evolutionary scenarios lead to maximization of inclusive fitness, but those papers did not receive enough attention to make it to the collective memory of evolutionary biology. In the last 7 years a renewed criticism of the generality of inclusive fitness has appeared, the most notable of which was voiced in a paper by Martin Nowak, Corina Tarnita and E.O. Wilson. The recent exchange concerning the generality of inclusive fitness does not yet show any signs of convergence, and positions range all the way from "*Hamilton's rule almost never holds*" (Nowak et al., 2010) to "*Inclusive fitness is as general as the genetical theory of natural selection itself*" (Abbot et al., 2010).

In this paper, we will review and address a variety of issues relating to inclusive fitness. We will for instance consider the relation between Hamilton's rule and Fisher's Fundamental Theorem of Natural Selection (Section 2), discuss cancellation effects, which is the most important refinement of our understanding of how individuals sharing genes matters for evolution (Section 7), and consider the question how Hamilton's rule can be tested empirically (Section 9). The recurrent theme, however, will be the central question in this controversy, which is: are there limits to the generality of inclusive fitness, and if so, what are the perimeters of the domain within which inclusive fitness works? In order to shed light on this in a simple and accessible way, we chose to consider two very well known dynamics from evolutionary theory: the replicator dynamics (Section 3) and adaptive dynamics (Section 6) – besides, of course, Hamilton's own dynamical model, which is discussed in Section 2.[1]

[1] Many papers in the domain of inclusive fitness consider different models – such as for instance Wright's Islands model. This paper is not meant to be a review that encompasses all of inclusive fitness theory. It

We will argue that the difference in opinions on the generality of inclusive fitness stems from a disagreement on how to define the costs and benefits in Hamilton's rule – as suggested by Birch (2014). In this paper, we will follow Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) and define costs and benefits by comparing actual fitnesses with counterfactuals. This is not one of the options considered in Birch (2014), but the "counterfactual method" is a classic possibility that is worth exploring. Alternatively, one can define costs and benefits using the "regression method" (for example Gardner et al., 2011, West & Gardner, 2013, and Marshall, 2011). In cases that are uncontroversial, these different definitions lead to the same b and c. In cases that are subject to debate, they lead to different costs and benefits. If costs and benefits are defined according to the regression method, inclusive fitness always matches the direction of selection for any given linear specification (the fact that it does so *for any given linear specification* is more than a detail; the model specification issue turns out to be central to the interpretation). If costs and benefits are defined using counterfactuals, this is not the case. Even then, though, we can still stake out a sizable set of models where inclusive fitness works.

Knowing that it makes a difference how costs and benefits are defined helps understand why points of view concerning the generality of Hamilton's rule are so different. But while it helps understand the divergence of opinions, it still allows for disagreement on which choice for b and c is better. Both methods will therefore be discussed in some detail.

In Section 3 we will consider Karlin & Matessi's original counterfactual method for defining costs and benefits, and conclude that these definitions have an undesirable property. With their definitions, the cost of cooperating rather than defecting is not necessarily minus the cost of defecting instead of cooperating. Inclusive fitness therefore was bound to only work in special cases, since there is the possibility that both the inclusive fitness of cooperation vs. defection and the inclusive fitness of defection vs. cooperation are positive, or that both are negative. An alternative, and perhaps – with hindsight – also more natural version of the counterfactual method does not allow for such inconsistencies. It does, however, still allow for inclusive fitness to not agree with the direction of selection.

In Section 4 we will discuss the regression method. We give a derivation of the result that, with costs and benefits defined by the regression method, Hamilton's rule always holds. The starting point for this result, however, is that we already have a specification for the regression, and that this specification is linear. That implies that Hamilton's rule holds just as much for costs and benefits that follow from one linear specification as it does for costs and benefits that follow from another. For the regression method to be well-defined for all possible models (or datasets), it would therefore need to be combined with a method for choosing between different specifications. Subsequently, whatever criterion one would use for choosing one linear specification over another should also be used to choose between linear and non-linear specifications, or between different non-linear ones. In other

mostly aims at understanding and illustrating the reasons for the controversy using relatively simple and well known models.

words, if we have a way to decide whether or not a *linear* variable should be included in the specification, then we immediately also have a criterion that should be used to decide whether or not a *non-linear* variable is to be included. This follows from the fact that least squares regressions treat linear and non-linear variables exactly the same. We therefore argue that Hamilton's rule, using the regression method, cannot both always be uniquely defined, and generally valid. The general validity depends crucially on the specification being linear, while any criterion that one could use for choosing between different linear specifications will immediately imply that there will also be models (or datasets) where the same criterion will rule in favour of a non-linear specification. Non-linearity therefore remains a problem for inclusive fitness.

The different topics related to Hamilton's rule will be discussed in the following order. In Section 2 we will revisit Hamilton's paper itself. There we will also discuss how his result relates to the literature at the time, and to Fisher's Fundamental Theorem of Natural Selection, both in the interpretation of Ewens (1989) and in the interpretation of Lessard (1997). It turns out that Hamilton's rule is the social generalization of Fisher's FTNS in neither of the two interpretations, while it does generalize results by Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b) to a setting with social traits.

Hamilton (1964b) conjectured that his rule would also be valid outside the confines of his model, and in the remainder of the paper we will look at other model settings. In Section 3 we consider the replicator dynamics with population structure. While Hamilton's model setup assumes a diploid species and considers difference equations, the replicator dynamics imply a switch to a haploid setting with differential equations. Here we find that in order for Hamilton's rule – with costs and benefits defined using the counterfactual method – to agree with the direction of selection, we need "equal gains from switching".

In Section 4 we discuss the regression method.

In Section 5 we look at comparative statics for the replicator dynamics. Comparative statics capture qualitative results, that may hold, even if Hamilton's rule – which is a quantitative prediction – does not apply. We find that there are indeed model settings in which a higher relatedness unambiguously fosters cooperation, even though Hamilton's rule, with costs and benefits according to the counterfactual method, does not hold.

In Section 6 we discuss how inclusive fitness describes what happens under adaptive dynamics, and what its limitations are there. Adaptive dynamics considers a continuous space of phenotypes, and assumes a monomorphic population. Here we find that for Hamilton's rule to hold – again, with costs and benefits according to the counterfactual method – it is enough if fitnesses are linear *locally*, and if populations do indeed remain close to being monomorphic.

In Section 7 we look at cancellation effects, which occur when not only opportunities for cooperation are local, but competition is local too. The insight that these two opposite effects occur (Wilson, Pollock and Dugatkin, 1992; Taylor, 1992a,b) is the most important deepening of our understanding of kin selection. For social behaviour to evolve, it is not

enough that interactants are related. What is needed is that there is a discrepancy between the two effects. Those that get the opportunity to cooperate, or that seek each other out for cooperation, need to be more related than those that they compete with.

Section 8 then goes on to discuss recent advances in the evolution of human preferences, which relates to the economics literature.

Section 9 discusses how inclusive fitness can be tested empirically, and revisits the replicator dynamics from Section 3, the adaptive dynamics from Section 6, and the examples that illustrate cancellation effects from Section 7. Observing violations of Hamilton's rule empirically is by definition impossible when costs and benefits are defined according to the regression method. But also with the counterfactual method, not just any violation of Hamilton's rule lends itself to observation by measuring costs and benefits of those phenotypes that survived selection (as opposed to as selection takes place). What is required for that to work is that different phenotypes coexist in equilibrium. The empirical literature nonetheless shows surprisingly many violations, also in cases where we would not expect those to be observed, and we will explain what causes these "false negatives".

Section 10 concludes.

2 Hamilton's rule and Fisher's Fundamental Theorem of Natural Selection

We revisit the central result in Hamilton's (1964) milestone paper and discuss how it relates to the literature at the time, and to Fisher's Fundamental Theorem of Natural Selection, both in the interpretation of Ewens (1989) and in the interpretation of Lessard (1997). It turns out that Hamilton's rule is the social generalization of Fisher's FTNS in neither of the two interpretations, while it does generalize results by Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b) to a setting with social traits.

2.1 Hamilton's rule

The point of departure in Hamilton (1964) is a single locus and a set of alleles $G_1, ..., G_n$. These give rise to genotypes G_iG_j, $1 \leq i, j \leq n$. Before we go to the central claim in Hamilton's two papers, we should perhaps first look at the typical question in the literature at that point, so that we understand why Hamilton chose his setup. In the few papers in Hamilton's reference list (Mulholland & Smith, 1959, Scheuer & Mandel, 1959, Kingman, 1961a,b), such genotypes G_iG_j always concerned properties that only affected the carrier itself, and not its relatives. The core question that was addressed in those papers was whether or not average fitness will always increase. This turns out to be a deep question in some settings, and trivially true, or trivially untrue, in others.

One setting in which it is trivially true, is if we 1) assume that these fitnesses are growth rates in a differential equation, and if we moreover 2) assume that these fitnesses are not frequency dependent – that is: the fitness of genotype G_iG_j does not depend on the distribution of genotypes in the population that G_iG_j lives in. In this case it is relatively straightforward that average fitness will go up. Another setting in which this is trivially true is if we assume that all alleles can be ranked from unambiguously bad – G_1 – to unambiguously good – G_n. In other words, if one can order the alleles such that $j > i$ implies that the fitness of G_jG_k is larger than the fitness of G_iG_k for all k, then the frequency of G_j is always increasing relative to the frequency of G_i, both in difference equations (i.e. in discrete time) and in differential equations (continuous time). If fitnesses are furthermore not frequency dependent, then this implies that average fitness increases. In discrete time, the fitness of a genotype G_iG_j is then defined as the mean number of offspring produced by individuals of that genotype. Everything that happens within a generation is collapsed in this number – in Hamilton's model organisms reproduce "once and for all at the end of a fixed period" – so this can incorporate both differences in viability and differences in fecundity.[2]

[2] Some papers explicitly look only at differences in viability. In simple models, these also translate linearly into offspring, so nothing is lost if we subsume viabilities in expected numbers of offspring. One convention is to have every successful gamete counts for half an offspring, which is what we adopt here.

Whether or not average fitness will increase – still in the standard, non-social setting – becomes a more difficult question if update steps are discrete – that is, if we have a difference equation, and not a differential equation – and if there are pairs of alleles that cannot unambiguously be ranked. Alleles G_i and G_j can not be ranked unambiguously if there are alleles G_k and G_l such that the fitness of G_iG_k is larger than the fitness of G_jG_k, but the fitness of G_iG_l is smaller than the fitness of G_jG_l. In such a setting, one could imagine that, when not already in equilibrium, the update step overshoots the equilibrium values in such a way that average fitness would go down. This is a far from trivial question, and it is the question that Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b) address. Their answer is positive: also with difference equations, and allowing for alleles that cannot be unambiguously ranked, average fitness will go up every step of the way. We do still have to assume that those fitnesses are not frequency dependent though.

Because Hamilton's result is sometimes also described as a social version of Fisher's Fundamental Theorem of Natural Selection (FTNS), it is worth emphasizing that, first of all, Hamilton does not present it as such – there is no reference to the FTNS in the papers. The papers that he does cite are only sideways related to the FTNS, and in no way proof for it, although it should be said that the relation between the results in those papers and the FTNS was, at the time, not well understood (see Price, 1972b, Ewens, 1989, Lessard, 1997, and Section 2.2 below).

The big difference between Hamilton (1964) and the previous literature is of course that in Hamilton's Part I the genotypes come with social effects; they do not only imply fitness effects on the carrier itself, but also on its relatives, and it is explicitly allowed for this to include different effects on different relatives, all at the same time. A genotype G_iG_j therefore comes with a vector $(\delta a_1, ..., \delta a_m)_{ij}$ of effects on itself – δa_1 – and on the fitnesses of $m-1$ relatives – $\delta a_2, ..., \delta a_m$ – which have relatednesses $r_2, ..., r_m$ to the focal individual. Since individual number 1 is the focal individual itself, $r_1 = 1$. Other than that, the setting is the same as in Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b); we are 1) looking at a difference equation, 2) there is no frequency dependence, and 3) the fitness effects could be anything. This is the setting for which the question whether or not average fitness increases for non-social traits was non-trivial. Also Hamilton assumes that the frequency of (ordered) genotype G_iG_j is p_ip_j, where p_i and p_j are the frequencies of allele G_i and G_j. This reflects random mating in a population with non-overlapping generations, and is in line with Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b), but not with Fisher's setup (see Section 2.2 below).

The question whether or not average fitness will always increase now turns into a different one, which is if perhaps it is average *inclusive* fitness that will always increase here. The

Not all papers are equally explicit about this, but switching to counting every successful gamete as one full offspring would amount to a different normalization of fitness, leaving the results intact.

inclusive fitness $R^{\bullet}_{ij}$ of genotype G_iG_j is defined as baseline fitness 1 plus the weighted sum of the fitness effects, with relatednesses as weights: $R^{\bullet}_{ij} = 1 + \sum_{k=1}^{m} r_k \ (\delta a_k)_{ij}$. Hamilton denotes *average* inclusive fitness by $R^{\bullet}_{..}$, which is short for $\sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j R^{\bullet}_{ij}$.

The central result in Hamilton (1964a) states that a sufficient condition for average inclusive fitness to not decrease is that the average *diluting effect* is nonnegative. The diluting effect can be seen as the complement of the inclusive fitness effect. If a social trait has fitness effects $(\delta a_1, ..., \delta a_m)_{ij}$, then those effects are divided, and subsequently aggregated, into the *inclusive fitness effect* $\delta R^{\bullet}_{ij} = \sum_{k=1}^{m} r_k \ (\delta a_k)_{ij}$ and the *diluting effect* $\delta S^{\bullet}_{ij} = \sum_{k=1}^{m} (1 - r_k) \ (\delta a_k)_{ij}$; every effect on fitnesses is weighted by r_k for its contribution to inclusive fitness, and by the remaining $(1 - r_k)$ for its contribution to the dilution term. If we then further aggregate all those dilution terms over all genotypes, we get the (overall, average) dilution term $\delta S_{..} = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \delta S^{\bullet}_{ij}$. If this dilution term is nonnegative, that is, if $\delta S_{..} \geq 0$, then the result is that average inclusive fitness will not decrease (page 7, Hamilton 1964a).

There are some minor mathematical points that one can make. The first is that $\delta S_{..} \geq 0$ is indeed a sufficient condition for the change in average inclusive fitness not to be negative. Hamilton does however not give an example of a case where this condition is violated – that is, where $\delta S_{..} < 0$ – and where average inclusive fitness actually goes down. In other words, from this paper, we do not know if this condition is important to consider, or if it is really a redundant requirement, and a stronger claim – that average inclusive fitness will *never* decrease – will perhaps also hold.

A second minor issue is that Hamilton concludes from this result that

> It follows that $R^{\bullet}_{..}$ certainly maximizes (in the sense of reaching a local maximum of $R^{\bullet}_{..}$) if it never occurs in the course of selective changes that $\delta S_{..} < 0$.

For this to follow, we need not just that $\delta S_{..}$ is never negative, but that it is strictly positive everywhere other than at the optimum. With some additional math, one can show that too (see Theorem 1 in Van Veelen, 2007).

This result, which is about *average* inclusive fitness being maximized, does not imply that *individuals* will behave as if they all maximize their inclusive fitness. With a few simple counterexamples one can show that this is not the case. Assume, for instance, that there are two alleles, and that the heterozygote has higher inclusive fitness than both homozygotes. In this case *average* inclusive fitness is maximized in a population state that has positive shares of both alleles. This implies that there will always also be homozygotes, and therefore there will still be an unavoidable share of individuals that have an inclusive fitness that is not the highest that it could be. Another counterexample is that with heterozygotes that have an inclusive fitness that is *lower* than both homozygotes, the population dynamics can get stuck in a local optimum, while a trait with a higher inclusive fitness is still available.

If we want to make the maximization of inclusive fitness by *individuals* follow from Hamilton's result, we will have to make an extra assumption. This extra assumption is

that inclusive fitnesses of all heterozygotes G_iG_j must lie between the inclusive fitnesses of their homozygote counterparts G_iG_i and G_jG_j (see Theorem 2 in Van Veelen, 2007).

It is worthwhile to realize that this is not an innocuous assumption. As touched upon above, the question whether *average* fitness increases in the case of non-social traits is only non-trivial if 1) we have a difference equation, 2) there is no frequency dependence, and 3) there are no restrictions on the fitness effects we consider. If we rule out heterozygote over- or underdominance, in terms of their inclusive fitnesses, then that means we can rule out coexistence of different alleles[3] – unless they are indistinguishable in terms of their inclusive fitnesses. In other words, if we make the extra assumption needed for maximization of inclusive fitness by *individuals*, then really all that matters is which homozygote has the largest inclusive fitness. And if we do look at homozygote fitnesses anyway, we might just as well drop the diploid model, and consider a haploid version.

From Section 3 onwards we will refer to Hamilton's rule in this individual sense. Hamilton's rule then states that behaviour with higher inclusive fitness will be selected for, at the expense of behaviour with lower inclusive fitness. If there are only two behaviours present in the population – one benchmark behaviour, with fitness effects 0, and one alternative behaviour – and the behaviour only affects the fitnesses of the actor and one typical other agent (such as, for instance, a sibling), then Hamilton's rule is reduced to its simplest, best known form: the alternative behaviour is selected when $rb > c$, or $rb - c > 0$. As we will see in sections 3 and 4, there are different ways to define the costs and benefits in Hamilton's rule. When we step outside the confines of his model, in which fitness transfers are independent of the genotype of the recipient, these different definitions start to diverge, and this is the source of disagreements on the validity of Hamilton's rule. Following Karlin & Matessi (1983) we will also consider Hamilton's rule to be qualitatively valid if higher relatednesses are more conducive to cooperation, in a sense that will be made more precise in Section 5.

2.2 Fisher's Fundamental Theorem of Natural Selection

It is sometimes suggested that Hamilton's model extends or generalizes Fisher's Fundamental Theorem of Natural Selection (FTNS) into the domain of social behaviours, or that it was at least inspired by it (see for instance Grafen, 2004). For the longest time it has been somewhat unclear what the FTNS actually claimed, which made it hard to judge whether or not Hamilton's result was in fact a generalization. Now, with the advantage of later papers on the FTNS (Price, 1972b, Ewens, 1989, 1992, Lessard & Castilloux, 1995, Lessard, 1997), we can see that Hamilton's result is not a generalization of Fisher's Fundamental

[3] If we rank the alleles such that $R_{ii} < R_{jj}$ implies that $i < j$, then under the assumption that there is no over- or underdominance, this implies that there is no fixed point of the dynamics in which alleles with unequal homozygote inclusive fitnesses coexist. In order to show that, assume that they do, and compare the allele with the largest homozygote inclusive fitness with the one with the smallest. The smallest has positive frequency, but is dominated by the largest.

Theorem, but of a different result that can be found in the papers that Hamilton (1964a) cites, and for which it was presumably unclear at the time how they related to Fisher's Fundamental Theorem.

In order to explain the differences, we first follow the presentation of the FTNS in Ewens (1989). The fitnesses and frequencies of the (ordered) genotypes G_iG_j are denoted by w_{ij} and P_{ij}, $1 \leq i, j \leq n$.[4] Fitness w_{ij} can be interpreted as the mean number of offspring produced by individuals whose genotype is ij, from the beginning to the end of the current generation, where every successful gamete counts for half an offspring. Ewens (1989) interprets w_{ij} as a measure of viability, while Lessard (1997) suggests a broader definition that also encompasses fecundity differences. The frequency of allele G_i is $p_i = \sum_j P_{ij}$, but it is *not* assumed that $P_{ij} = p_i p_j$. The frequency of allele G_i in the new generation is $p'_i = \sum_j P_{ij} w_{ij} / \overline{w}$, where $\overline{w} = \sum_{i,j} P_{ij} w_{ij}$ is the mean fitness of the population.

Rather than looking at the change in mean fitness $\overline{w}$, the FTNS, as interpreted by Ewens (1989), looks at something else. Suppose one were to choose $\alpha_1, ..., \alpha_n$ such that they minimize $\sum_{i,j} P_{ij} \left(w_{ij} - \overline{w} - \alpha_i - \alpha_j\right)^2$. This may look a bit like a statistical exercise, where the 'true' fitnesses $\overline{w}+\alpha_i+\alpha_j$ are estimated by treating w_{ij} as noisy data, and treating the differences between $\overline{w} + \alpha_i + \alpha_j$ and w_{ij} as i.i.d. draws from a random distribution with expectation 0. What it really does, however, is assign a number α_i to each allele A_i that best represents that allele's contribution to fitnesses in the current population; note that w_{ij} is a fixed quantity, which is assumed to stay the same over generations, and *not* a noisy observation, which would change with every draw. Being joined with allele A_i might be good news for allele A_k, and bad news for A_l, but on average, given the current type frequencies, the effect of A_i is quantified by α_i. Obviously, which $\alpha_1, ..., \alpha_n$ minimizes $\sum_{i,j} P_{ij} \left(w_{ij} - \overline{w} - \alpha_i - \alpha_j\right)^2$ depends on the P_{ij}'s.

Fisher's FTNS, as interpreted by Ewens (1989), states that if we evaluate the change in frequencies using $\overline{w} + \alpha_i + \alpha_j$ – and not w_{ij} – then this change equals the "additive genetic variance", divided by the mean fitness in the population. The additive genetic variance is defined as $\sigma_A^2 = \sum_{i,j} P_{ij} \left(\alpha_i + \alpha_j\right)^2$, and this is obviously non-negative, and only 0 in equilibrium. In other words,

$$\sum_{i,j} \left(P'_{ij} - P_{ij}\right) \left(\overline{w} + \alpha_i + \alpha_j\right) = \frac{\sigma_A^2}{\overline{w}} \geq 0. \tag{2.1}$$

No assumption is made about how the alleles in the new generation are matched, as long as the frequencies of genotypes in the new generation are consistent with the frequencies of the alleles in the new generation, that is, as long as $\sum_j P'_{ij} = p'_i$.

Ewens (1989) and Price (1972b) convincingly argue that the claim is correct, but also that the quantity that is shown to be larger than 0 is perhaps not that interesting to look

[4] In Ewens (1989) the number of alleles is denoted by m. Since Hamilton already uses m for the number of individuals affected by the social trait, we stick to using n for the number of alleles, as in Hamilton (1964a). Also, alleles are denoted with A's in Ewens (1989), but G's in Hamilton (1964).

at, because in the new generation, the old α's no longer apply; if we repeat the minimizing exercise for the new generation, we typically get a different set $\alpha'_1, ..., \alpha'_n$.

The interpretation of Lessard (1997) features two minimizations. The first one is the same minimization as in Ewens (1989), which concerns fitnesses w_{ij}. The second minimization concerns growth rates W_{ij}, which are defined as $W_{ij} = \left(\frac{P'_{ij}}{P_{ij}}\right)\overline{w}$. One of the results in Lessard (1997) is that minimizing $\sum_{i,j} P_{ij} \left(w_{ij} - \overline{w} - \alpha_i - \alpha_j\right)^2$ gives the same values for $\alpha_1, ..., \alpha_n$ as minimizing $\sum_{i,j} P_{ij} \left(W_{ij} - \overline{W} - \alpha_i - \alpha_j\right)^2$. Because W_{ij} and w_{ij} can very well be unequal, the "residual addends" may also differ. With a mix of the notation in Lessard (1997) and Ewens (1989), the residual addends are defined as $\delta_{ij} = w_{ij} - \overline{w} - \alpha_i - \alpha_j$ and $\varepsilon_{ij} = W_{ij} - \overline{W} - \alpha_i - \alpha_j$, both for all i, j (Lessard, 1997, immediately looks at the multi-locus case, but since Hamilton, 1964, is a single locus model, here we translate back to the single locus setup that Ewens, 1989, uses in his first two sections).

Fisher's FTNS, as interpreted by Lessard (1997), does concern the change in frequencies using w_{ij} – and not, as in Ewens' (1989) interpretation, $\overline{w} + \alpha_i + \alpha_j$. Lessard (1997) gives a decomposition of the change in average fitness that also allows for changes in w_{ij} (see equation 38 on page 127 in Lessard, 1997). The vectors of effects in Hamilton (1964), however, are constant, and therefore we will also consider constant w_{ij}'s, as also Ewens (1989) does. This implies that Lessard's decomposition has two remaining non-zero terms:

$$\sum_{i,j} \left(P'_{ij} - P_{ij}\right) w_{ij} = \frac{\sigma_A^2}{\overline{w}} + \frac{\sum_{i,j} P_{ij}\varepsilon_{ij}\delta_{ij}}{\overline{w}}. \tag{2.2}$$

The first term on the right hand side of this equation is the same as the one term on the right hand side of Equation (2.1), but now this reflects the change in average fitness due to changes in frequencies, insofar as they can be accounted for by the effects of genotypes as described by the parameters of the linear model. The FTNS is now interpreted as a statement about this first term only. The total change in average fitness can of course still be negative, if the second term on the right hand side is negative, and outweighs the first (the second term is shortened to $cov\left(\varepsilon, \delta\right)/\overline{w}$ in Lessard, 1997, which is justified by the observation that $\sum_{i,j} P_{ij}\delta_{ij} = \sum_{i,j} P_{ij}\varepsilon_{ij} = 0$, and by interpreting the frequencies as probabilities in a random draw from the parent population).

The setup in Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b) is different. Here (ordered) genotypes at this locus are assumed to be in Hardy-Weinberg proportions – it is assumed that $P_{ij} = p_i p_j$ and $P'_{ij} = p'_i p'_j$ – and these papers show that

$$\sum_{i,j} \left(p'_i p'_j - p_i p_j\right) w_{ij} \geq 0. \tag{2.3}$$

Hamilton's result is the social version of this latter result. If we take $\left(\delta a_1, ..., \delta a_m\right)_{ij}$ such that $\left(\delta a_1, ..., \delta a_m\right)_{ij} = \left(\delta a_1, 0, ..., 0\right)_{ij}$ for all i and j – that is, if it reflects a trait with no

effect on others – then $R^{\bullet}_{ij}$ reverts to being individual fitness, and can be interpreted as w_{ij} in Mulholland & Smith (1959), Scheuer & Mandel (1959) and Kingman (1961a,b).

2.3 Fitness, reproductive value, and topics not covered

We restricted attention to the basic, discrete-time, multi-allele, single-locus model with nonoverlapping generations. Ewens (1989) also includes a continuous time version, and a discrete time, multi-locus version. Lessard (1997) is a multilocus model from the beginning, and also includes both continuous and discrete time versions, both with and without overlapping generations. The reason why we restrict attention to the more basic version is that this setup matches Hamilton's (1964). It also makes the definition of fitness – the success in leaving progeny (Darwin, 1956, p. 64) – uncomplicated; the fitness w_{ij} is the mean number of offspring produced by individuals of genotype ij, one generation down the road. This definition incorporates viability as well as fecundity differences, if we assume that mating and reproduction do not change gene frequencies from the current generation at the time of maturity to the next generation at the time of conception (for viability selection) and if mating does not change gene frequencies in the current generation from the time of conception to the time of reproduction, and if neither meiotic drive nor gametic selection takes place (for fecundity selection); see Ewens (1989, 1992), Castilloux and Lessard (1995), Lessard and Castilloux (1995), and Lessard (1997).

With more complicated, or more detailed models, the definition of fitness may require more than just counting offspring. With haplodiploid organisms, males and females are not the same in their expected future contribution to the population (Price, 1970, Oster et al, 1977, Benford, 1978, Pamilo & Crozier, 1982, Frank, 1986, Grafen, 1986, Taylor, 1988). Helping someone get an extra offspring in the further away future may not be the same as helping someone get extra offspring now (Fisher, 1930, Leslie, 1948, Charlesworth, 1980). An offspring on one node in a network may not contribute to future generations in the same way as an offspring on another node in a heterogeneous network does (Maciejewski, 2014, Taylor & Maciejewski, 2014). All of these examples can be encompassed by defining different classes of individuals (by sex, age, or position in the network, for instance) and by using this class-structured populations to define class-specific reproductive values to replace fitnesses (Taylor, 1990, Grafen, 2006, Barton & Etheridge, 2011).

In the remainder of this paper we will consider models that are all symmetric, and for which there is a degenerate class structure, with one class only. Therefore we cannot benefit from the richness that using richer class structures would allow for. Also there is no need to distinguish between fitness and reproductive value. Such a simple setup comes with restrictions on the species and phenomena that can be modeled. One of the most interesting phenomena in social evolution is eusociality. Symmetric models like the ones we will see in the following sections are hardly appropriate to approach the question when and why eusociality will evolve and be maintained. Also the question which sex ratios to expect

requires different models. In this paper we will therefore not discuss some of the most interesting topics from the literature in the last 50 years. This implies that we will also not discuss the eusociality part of Nowak, Tarnita & Wilson (2010), and only pay attention to how costs and benefits are defined in part A of their Supplementary Information, which contains a model setup that is different from the model of eusociality in part C of their Supplementary Information. Our setup therefore sidesteps the question whether or not inclusive fitness helps understand eusociality. The symmetric setup nonetheless leaves us with more than enough to explore, and allows us to answer interesting questions concerning the generality of Hamilton's rule.

3 Replicator dynamics

In this section we consider the replicator dynamics combined with population structure. We also discuss the "counterfactual method", which defines costs and benefits by comparing fitnesses under one behaviour to what they would have been under the alternative behaviour. This approach was suggested in Karlin & Matessi (1983). With their method, the inclusive fitness of cooperation is not necessarily minus the inclusive fitness of defection, but using a natural, improved version of their approach, consistency is restored. With costs and benefits according to the counterfactual method, we find that in order for Hamilton's rule to agree with the direction of selection, the fitness effects need to satisfy "equal gains from switching".

Hamilton (1964b) conjectured that his rule would also be valid outside the confines of his model. In this section we will look at the replicator dynamics as an alternative model. The replicator dynamics are haploid, while Hamilton's model was diploid, but this choice nonetheless connects relatively naturally with what we found in Section 2. There we have seen that in order to make Hamilton's central result imply that *individuals* will behave as if they all maximize their inclusive fitness – and not just that *average* inclusive fitness is maximized – we need to make extra assumptions. These extra assumptions restrict heterozygote inclusive fitnesses, and they imply that all that matters for the outcome of the dynamics is how homozygote inclusive fitnesses compare. The outcome of the dynamics under these extra assumptions therefore is not sensitive to a change from a diploid to a haploid model, where the genotypes are the homozygotes. A considerable share of the inclusive fitness literature on cooperation moreover also uses haploid models.

3.1 2-player games

Hamilton describes his intuition in a 1963 prequel in the *American Naturalist* as follows:

> As a simple but admittedly crude model we may imagine a pair of genes g and G such that G tends to cause some kind of altruistic behaviour while g is null. Despite the principle of 'survival of the fittest' the ultimate criterion which determines whether G will spread or not is not whether it is to the benefit of the behaver but whether or not it is to the benefit of the gene G; and this will be the case if the average net result of the behavior is to add to the gene-pool a handful of genes containing G in higher concentration than does the gene-pool itself. With altruism this will happen only if the affected individual is a relative of the altruist, therefore having an increased chance of carrying the gene, and if the advantage conferred is large enough [...].

The setup in Hamilton (1964a) is one that follows this intuition, and therefore he formulates the problem in terms of what economists would call an "individual choice problem". The gene G causes its bearer to give up c in order for its relative (sibling, nephew, niece) to

gain b. All that matters is what happens, on average, to the frequency of copies of G, and all we need to compare is the loss to the donor and the relatedness-weighted benefit to the recipient.

The model setup in Hamilton (1964a, page 2) is "*particularly adapted to deal with interactions between relatives of the same generation*". Obviously, between any pair of same-generation relatives, both of them have both roles; both are a possible donor as well as a possible receiver. It is therefore very natural to think of a population of pairs, in which every pair is playing a game. With both of them choosing between giving and not giving, and with benefits being larger than costs, the game they are playing becomes a prisoners' dilemma. Because there are more possibilities in games between pairs of individuals than just prisoners' dilemmas, we will mention all typical cases. We will also use a way of picturing games that is more common in economics than it is in biology.

In Hamilton (1964a), costs and benefits are additions and subtractions to a basic fitness of 1. That implies that the game between two possible donors is given by the following matrix.

$$\begin{array}{cc} & g \qquad\quad G \\ g & 1 \qquad 1+b \\ G & 1-c \quad 1+b-c \end{array}$$

The numbers in this matrix are the fitnesses of an individual that has the genotype that is indicated in the column to the left of the matrix, when facing the genotype that is indicated in the row above the matrix. In the replicator dynamics, fitnesses are rates of increase (or decrease). Following the custom in classical (not evolutionary) game theory, where games are typically *not* assumed to be symmetric, we could complement the entries in the matrix by also indicating what the opponent gets. With the assumption of symmetry, this is redundant information – we could infer that from the first matrix already – but it will useful in rendering the game graphically.

$$\begin{array}{cc} & g \qquad\qquad G \\ g & 1,1 \qquad\quad 1+b,1-c \\ G & 1-c,1+b \quad 1+b-c,1+b-c \end{array}$$

In Fig. 1 below, we chose $c = 1$ and $b = 3$, and plotted all four payoff combinations from the payoff matrix. Any two points between which only one player's choice is different are furthermore joined by a line; for instance the points $(1, 1)$ and $(1 + b, 1 - c)$ are joined, because the first corresponds to (g, g) and the second to (g, G).

In this section, and in Sections 5 and 6, we will use the terms "fitness" and "payoff" interchangeably. As we will see in Section 7, that is not always OK; payoffs from a game may translate to fitness effects in intricate ways. The replicator dynamics, however, assume

that, even if interactions are not taking place in a well-mixed population, competition is totally symmetric, and everyone competes with everyone else equally intensely. In Section 7 we will consider examples where the local interaction structure makes both competition and cooperation local affairs. Here we assume that structure only affects who has the opportunity to cooperate with whom, while competition is a global affair, in which everyone's payoffs directly translate into fitnesses. This fits a situation with kin recognition relatively well, where competition with those that are recognized as kin may very well be equally intense as with those that are not recognized as kin.

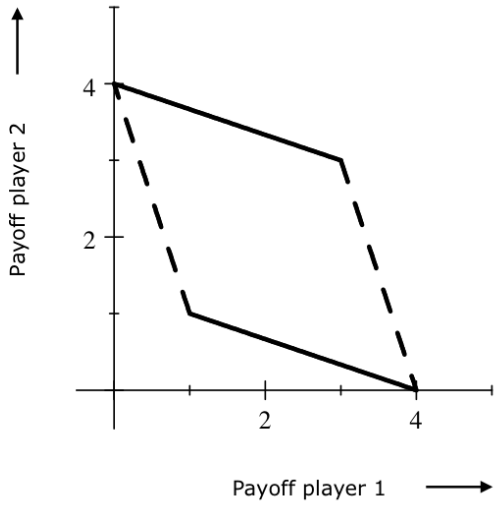


Figure 1. Prisoners' dilemma with equal gains from switching. The four corners of the lozenge reflect the payoffs to the players for the four possible combinations of strategies; (g, g) for the left/down corner, (g, G) for the right/down corner, (G, g) for the left/up corner, and (G, G) for the right/up corner. The solid lines connect outcomes in which player 1 is always of the same type, and player 2 switches. The dashed lines connect outcomes in which player 2 is always of the same type, and player 1 switches. Equal gains from switching is sometimes also referred to as fitness effects being additive, and for the picture this implies that the solid lines are parallel and equally long, and the dashed ones too.

The game above is described in Nowak & Sigmund (1990) as a prisoners' dilemma with "equal gains from switching" (see also Wild & Traulsen, 2007). Equal gains from switching means that the effect of switching between strategies on one's own fitness as well as the effect on the other's fitness is independent of what the other individual does. Sometimes this is also referred to as "additive fitness effects". This is obviously the case in the above game, because of the way it is constructed in the first place; it is a combination of two mirrored individual choice problems. However, not every prisoners' dilemma has this property. Consider the following matrix of fitnesses.

	g	G
g	1.5	4
G	0	3

In this game, the costs to oneself of switching from g to G are 1.5 if one is paired with a g, and 1 if one is paired with a G. The benefits to the other are 2.5 when paired with a g and 3 when paired with a G. This game therefore does not have equal gains from switching (see Fig. 2).

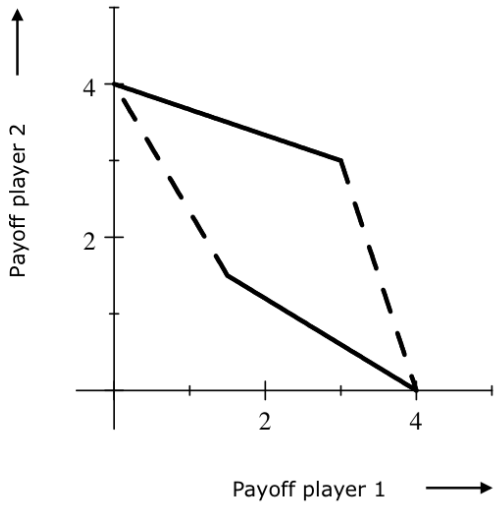


Figure 2. Prisoners' dilemma with unequal gains from switching, or non-additive fitness effects. The four corners again reflect the payoffs to the players for the four possible combinations of strategies; (g, g) for the left/down corner, (g, G) for the right/down corner, (G, g) for the left/up corner, and (G, G) for the right/up corner.

Altogether different games are also possible between two players with two actions. The following payoff matrix defines a stag hunt game.

	g	G
g	4	4
G	1	7

While the games above have only one pure Nash equilibrium, the stag hunt game has two pure Nash equilibria – (g, g) and (G, G) – and one mixed equilibrium (see Fig. 3).

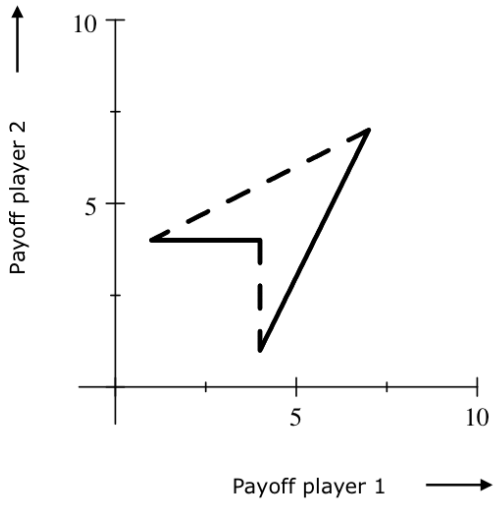


Figure 3. Stag hunt game. The right/up corner represents the payoffs that players get at (G, G), the lower point on the diagonal represents payoffs at (g, g), the bottom corner represents (g, G), and the left corner (G, g).

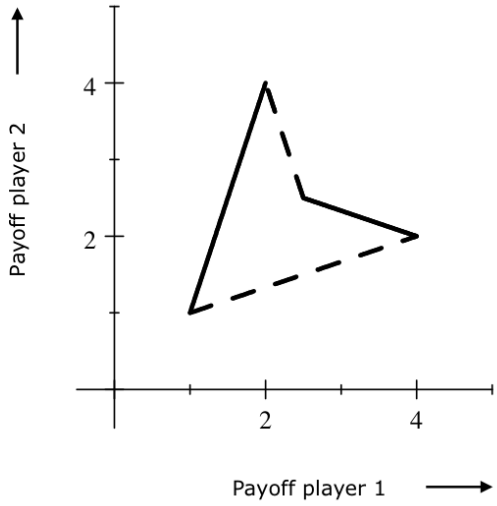


Figure 4. Hawk dove game. The left/down corner represents the payoffs that players get at (g, g), the higher up point on the diagonal represents payoffs at (G, G), the right corner represents (g, G), and the up corner (G, g).

The last type of game is the hawk-dove game, a.k.a. the snowdrift game.

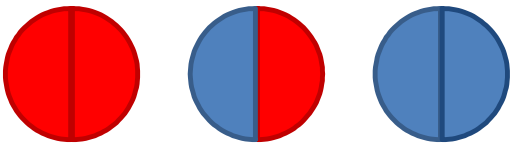
	g	G
g	1	4
G	2	2.5

This game has no pure symmetric equilibria – only pure asymmetric ones. The only symmetric equilibrium is a mixed one (see Fig. 4).

3.2 Forming pairs

In order to illustrate how relatedness can feature in a natural way, we begin with a totally unrealistic, but nonetheless instructive genetical system. Assume that parent pairs are randomly formed from a very large pool in which the frequency of cooperators is p and the frequency of defectors is $(1-p)$ – we switch here from g and G to the more standard notation of D for defect and C for cooperate. That means that a fraction p^2 of the parent pairs are both C, $2p(1-p)$ of the parent pairs are $\{C,D\}$ pairs, and $(1-p)^2$ of the parent pairs are all D pairs. Assume that both parents are haploid, and that at reproduction, every offspring has a 50% chance of inheriting the genotype of either parent. We assume that all parent pairs produce two offspring, and we will consider those pairs of siblings. Obviously all $\{C,C\}$ parent pairs produce only C offspring, and all $\{D,D\}$ parent pairs produce only D offspring. A $\{C,D\}$ parent pair has a 25% chance of producing two C's, a 25% chance of producing two D's and a 50% chance of producing one C and one D. That means that all offspring pairs together occur in the following frequencies.

$$\begin{aligned}
\{D,D\}: &\quad (1-p)^2 + \tfrac{1}{4}\cdot 2p(1-p) = \tfrac{1}{2}(1-p)^2 + \tfrac{1}{2}(1-p)\\
\{C,D\}: &\quad \tfrac{1}{2}\cdot 2p(1-p) = \tfrac{1}{2}\cdot 2p(1-p) + \tfrac{1}{2}\cdot 0\\
\{C,C\}: &\quad p^2 + \tfrac{1}{4}\cdot 2p(1-p) = \tfrac{1}{2}p^2 + \tfrac{1}{2}p
\end{aligned}$$



Well mixed (r = 0)	25%	50%	25%
Full sibs (r = 0.5)	37.5%	25%	37.5%
Clones (r = 1)	50%	0%	50%

Figure 5. Three population structures. In all populations the overall frequency of both defectors (red) and cooperators (blue) is 50%, but relatednesses are different.

If pairs were the result of random matching – as the parent pairs are – then the frequencies would be $(1-p)^2$, $2p(1-p)$, and p^2, respectively. If the pairs were pairs of clones, on the other hand, then the frequencies would be $1-p$, 0, and p, respectively. Because the frequencies of the sibling pairs are exactly halfway between those, it makes perfect sense that relatedness in this case should also be halfway between 0 and 1. It also makes perfect sense to generalize the frequencies of pair types, using an assortment parameter α:

$$\begin{aligned}
\{D,D\}: &\quad (1-\alpha)(1-p)^2+\alpha(1-p)\\
\{C,D\}: &\quad (1-\alpha)\cdot 2p(1-p)\\
\{C,C\}: &\quad (1-\alpha)p^2+\alpha p
\end{aligned}$$

At $\alpha=0$ we get the pair frequencies that random matching would give, at $\alpha=1$ the pair frequencies of clonal pairs, both with a frequency of C's that is p. In other words, for a given frequency p, the higher α is, the fewer $\{C,D\}$ pairs, and the more $\{D,D\}$ and $\{C,C\}$ pairs.

The above description gives pair types as a function of frequency p and assortment parameter α. We would however also like to be able to start at the other end, with pair frequencies, and work our way back to p and a measure of assortment or relatedness. This can be done in an easy and intuitive way too. Denote the frequency of $\{D,D\}$-groups, that is, groups with 0 cooperators, by f_0, denote the frequency of $\{C,D\}$-groups by f_1, and denote the frequency of $\{C,C\}$-groups by f_2. The frequency p of C's in the overall population is then recovered in an obvious way; $p=\frac{1}{2}f_1+f_2$. Let $\mathbb{P}(C|C)$ furthermore denote the probability of facing a C if you are a C yourself. That probability equals the share of cooperators in the population that are facing another cooperator, divided by the share of all cooperators in the population. In other words,

$$\mathbb{P}(C|C)=\frac{2f_2}{2f_2+f_1}=\frac{f_2}{p}$$

Let $\mathbb{P}(C|D)$ denote the probability of being paired with a C if you are a D yourself. That probability equals the share of defectors in the population that are facing a cooperator, divided by the share of all defectors in the population;

$$\mathbb{P}(C|D)=\frac{f_1}{f_1+2f_0}=\frac{f_1}{2(1-p)}$$

If we define relatedness as the difference between those two conditional probabilities, then we recover the assortment parameter α; $r=\mathbb{P}(C|C)-\mathbb{P}(C|D)=\frac{f_2}{p}-\frac{f_1}{2(1-p)}=\frac{2\left((1-\alpha)p^2+\alpha p\right)}{2p}-\frac{(1-\alpha)\cdot 2p(1-p)}{2(1-p)}=(1-\alpha)p+\alpha-(1-\alpha)p=\alpha$. This property of a population structure is the definition of relatedness that we will use throughout the paper, although at some points we will use equivalent definitions to compute it.

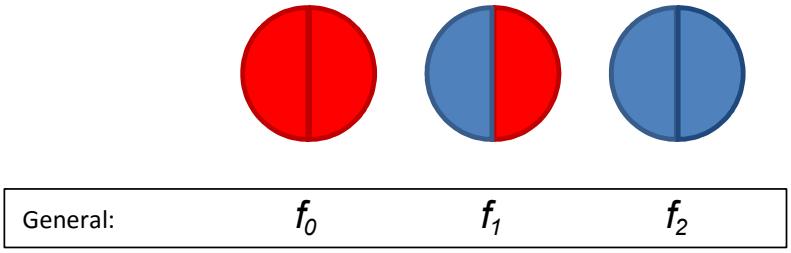


Figure 6. General population structure.

In order to illustrate that all combinations of frequencies f_0, f_1, and f_2 define one combination of p and r and vice versa, we draw a simplex (see Fig. 7). The proportions f_0, f_1, and f_2 have to add up to 1, and every point on the simplex represents a vector (f_0, f_1, f_2) with $f_0 \geq 0, f_1 \geq 0, f_2 \geq 0$ and $f_0 + f_1 + f_2 = 1$. The corners of the simplex are $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. Note that on this simplex, straight vertical lines are population states with constant frequency p, and the curves from the left down corner to the right down corner are lines of constant relatedness. Any such line, with constant relatedness $r \in [0, 1]$, has to go through the corners where $p = 0$ and $p = 1$. The straight line on the bottom side of the simplex reflects $r = 1$, as any population on that line has no mixed groups, and only groups with two D's or two C's – so $\mathbb{P}(C|C) = 1$ and $\mathbb{P}(C|D) = 0$. The higher up the curve is, the more mixed groups there are, and the lower relatedness is. The curve for $r = 0$ follows the shares of the group types that the binomial distribution with probability p would give.

In the setting of Hamilton's paper, it is natural to assume that relatedness does not change with frequency p. The production of pairs of full siblings for instance simply imposes that relatedness is 0.5, whatever the frequency p of a gene is (the example at the very beginning of Section 3.2 indicates how that works). One could imagine that perhaps there are population structures for which this may not be the case, and where r varies with p, but here we will stick to population structures with a fixed and constant r, which fits Hamilton's setup perfectly.

For a given r we know for every frequency p how many groups of the three types there are. With those, and the game payoffs, we can compute the average payoffs of both strategies. The replicator dynamics (Taylor & Jonker, 1978) is a natural way to translate that into a differential equation; the time-derivative of the frequency of cooperators p is p times the difference between the average payoff of cooperators and the average payoff in the population as a whole:

$$\begin{aligned} \dot{p} &= p\left(\overline{\pi}_C - \overline{\overline{\pi}}\right) \\ &= p\,(1-p)\,(\overline{\pi}_C - \overline{\pi}_D) \end{aligned} \tag{3.1}$$

It turns out that with relatedness r and payoff matrix A, the change in frequency is the

same as it is in the replicator dynamics without population structure, but then with a transformed payoff matrix; $A' = rB + (1-r)\,A$, where $[B]_{ij} = [A]_{ii}$, that is, B is a matrix where all elements on row i are the same as the i'th diagonal element of the matrix A (Van Veelen, 2011b).

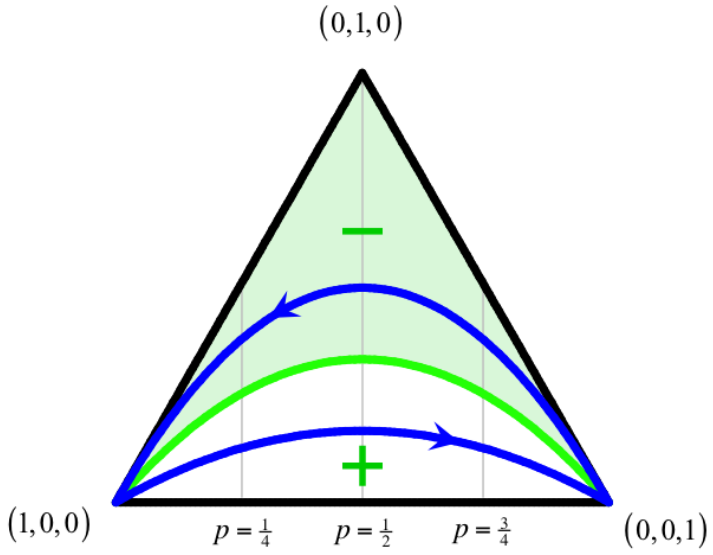


Figure 7. The simplex with equal gains from switching. The threshold (green line) for the game from Figure 1 is at $r = \frac{1}{3}$. The blue lines are trajectories for $r = 0$ (top, well mixed population), and $r = \frac{2}{3}$ (down). The bottom of the simplex reflects population states with $r = 1$ (clones). The corners represent population states with only $\{D, D\}$-groups (left bottom), only $\{C, C\}$-groups (right bottom) and only $\{C, D\}$-groups (top).

Fig. 7 illustrates how Hamilton's rule now shows up nicely on the simplex. For any game, we can divide the simplex up into two parts; one where the frequency of cooperators increases, and one where it decreases. For a given prisoners' dilemma with equal gains from switching, the line that separates the plus-region from the minus-region has the exact same shape as a line with constant r. This implies that with a given r, either the population state is always below this line – that is: it has a *higher* relatedness than the threshold relatedness – or it is always above it. Therefore, if cooperation is selected for one frequency, it is selected for all; trajectories go to one corner of the simplex, irrespective of the starting point on the constant r-curve. Which corner that is, is given by Hamilton's rule.

Without equal gains from switching, the shape of the line that separates the plus-region from the minus-region no longer has the same shape as a line with constant r, and therefore they may intersect. That implies that it is possible that either the dynamics do not converge to a corner, or that they do, but that it depends on the starting point which corner that is.

In other words, we could get stable coexistence – as is typical in hawk-dove games already at $r = 0$ – or bistability – which is typical in stag hunt or coordination games already at $r = 0$.

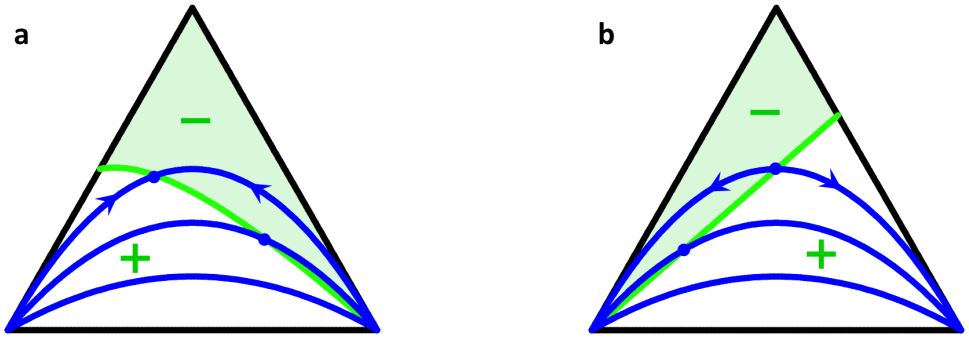


Figure 8. (a) Hawk dove and (b) stag hunt. The green lines separate the plus-regions, where cooperator frequencies increase, from the minus-regions, where cooperator frequencies decrease, for the games from Figure 4 and 3, respectively. Since the green line does not have the same shape as the constant r-arcs (the blue lines) there is a range of r's for which we either get coexistence (a) or bistability (b).

In Hamilton's original setup, as well as in the case of equal gains, costs, benefits, and relatedness are fixed quantities, and they combine to a rule that predicts the direction of selection for *any* current frequency. That is no longer possible, if the direction of selection changes at the intersection of the threshold (the green curves in Fig. 8) and a line with constant r (the blue curves in Fig. 8).

In their evaluation of Hamilton's rule – or, as they called it "the Hamilton rule" – Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) require costs and benefits to be independent of the "kin-group structure", which in their case implies that it must also be independent of the current frequency of cooperators. We can, however, choose to allow benefits b and costs c to change with the population state, and treat Hamilton's rule not as a global prediction for the success of a cooperative mutant, but as a local criterion, that may depend on the frequency p of cooperators.

3.2.1 Costs, benefits, and counterfactuals

As we will see, whether or not Hamilton's rule applies, depends on how b and c are defined. Below we consider the definition according to the counterfactual method, as used by Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986). In Section 4 we will describe the definition according to the regression method, as used by Gardner et al. (2011), along with

their application to the same example. We will use standard notation for entries in the payoff matrix.

$$\begin{array}{cc} & \begin{array}{cc} D & C \end{array} \\ \begin{array}{c} D \\ C \end{array} & \begin{array}{cc} P & T \\ S & R \end{array} \end{array}$$

Cooperators only In order to compute the costs and benefits of cooperation, we can go over all cooperators in the population and compare their current fitness to what their fitness would have been, had they defected. For those that are matched with another cooperator, this difference is $T-R$. For those that are matched with a defector, this difference is $P-S$. If we weigh those differences with how many cooperators are matched with cooperators and how many with defectors, we find the average cost of cooperating instead of defecting to be

$$c = (r + (1-r)\,p)\,(T-R) + (1-r)\,(1-p)\,(P-S)\,.$$

Similarly, if we go over all cooperators again, and now compare the fitness of their interaction partner to what their interaction partner's fitnesses would have been, had they themselves defected, we find the average benefits to their interaction partner to be

$$b = (r + (1-r)\,p)\,(R-S) + (1-r)\,(1-p)\,(T-P)\,.$$

The criterion for C-players to win at frequency p is $\overline{\pi}_C - \overline{\pi}_D > 0$. If we rewrite that, we find:

$$\begin{aligned} (r + (1-r)\,p)\,R + (1-r)\,(1-p)\,S &> (r + (1-r)\,(1-p))\,P + (1-r)\,pT \\ r\,((1-p)\,(R-S)) + pR + (1-p)\,S &> rp\,(P-T) + (1-p)\,P + pT \\ r\,((1-p)\,(R-S) + p\,(T-P)) &> p\,(T-R) + (1-p)\,(P-S)\,. \end{aligned} \tag{3.2}$$

This criterion is not the same as $rb > c$ if we use the b and c as we just computed them, unless $P + R = T + S$, that is, unless the game satisfies equal gains from switching.

One property that costs and benefits should have, however, is that the cost of cooperating should be minus the cost of defecting, and the benefits of cooperating should be minus the benefits of defecting. In other words, it should not matter whether we take the benchmark to be cooperation or defection. This, however, is not the case for the definition used by Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986), already for this simple example. If we switch to having cooperation as the benchmark, and compute the costs and benefits of defection, we find $c = (1-r)\,p\,(R-T) + (r + (1-r)\,(1-p))\,(S-P)$ and $b = (1-r)\,p\,(S-R) + (r + (1-r)\,(1-p))\,(P-T)$, which are not minus the c and b

we found when defection was the benchmark. The following alternative definition is not sensitive to the benchmark, and therefore this is the one we will use in the remainder of the paper.

Cooperators and defectors In their computation of costs and benefits, Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) only consider those that actually cooperate. Alternatively, one could consider not only the cooperators, but all individuals in the population, since every individual had the opportunity either to cooperate, or to defect. For all individuals facing a cooperator, the difference in their own fitness between cooperation and defection is $T - R$. For all individuals facing a defector, this difference is $P - S$. The difference to their interaction partner's fitness is $R - S$, if their partner is a cooperator, and $T - P$ if their partner is a defector. Since p is the share of cooperators, p is also the share of individuals that is matched with one. Average costs and benefits therefore are:

$$\begin{aligned} c &= p\,(T - R) + (1 - p)\,(P - S) \\ b &= p\,(R - S) + (1 - p)\,(T - P)\,. \end{aligned} \tag{3.3}$$

Also with this b and c, Hamilton's rule does not match the criterion for the frequency of cooperators to increase, again unless we have equal gains from switching. If indeed $P + R = T + S$, then these two definitions using counterfactuals, as well as the regression method definition from Gardner et al. (2011), all coincide.

3.3 Interactions between more than two individuals

Besides dyadic interactions, there are also interactions that take place between more than two individuals. In humans, there are football teams and orchestras, corporations and armies. Also eusocial insects and cells cooperate in large to astronomical numbers.

With more than two players, it could still be that the costs and benefits of one individual's possible cooperation are independent from what the others do, so that a generalized version of equal gains from switching holds. In this case, average costs and benefits are not frequency dependent, and all three definitions of b and c coincide. One would expect that inclusive fitness will therefore agree with the prediction again – and it does. Theorem 5 in Van Veelen (2011b) claims even more. With "generalized equal gains from switching" not only the *sign* of inclusive fitness matters, but also the *absolute value*; inclusive fitness becomes a parameter in the replicator equation, and not only determines the *direction* of selection, but also the *speed*. A simple example of an "n-player game" with equal gains from switching involves n individuals living in a group, where their living in a group implies that every so often they get paired with another individual from the same group to play a prisoners' dilemma with equal gains from switching. Or, even simpler, every so often one gets the opportunity to give the other benefit b at cost c to itself.

Many games of cooperation do not have (generalized) equal gains from switching though. In football teams and orchestras team performance might be as good as the weakest link. When going to war with a neighbouring tribe, whether or not I contribute may not make much of a difference if everybody else already does – in which case we will win anyway – or if nobody does – in which case we lose anyway. Only if there is a fair chance that I might tip the balance, are there benefits to be gained from my switching from defection to cooperation.

A general, symmetric n-player game with 2 strategies is determined by its $2n$ payoffs; symmetry implies that no one individual is special, so payoffs only depend on how many cooperators there are in a group, and whether or not one is a cooperator or a defector oneself. Payoffs therefore can be denoted by $\pi_{i,C}$ for $i = 1, ..., n$ and $\pi_{i,D}$ for $i = 0, ..., n-1$. The first – $\pi_{i,C}$ – is the payoff of a cooperator in a group with i cooperators, including itself. The second – $\pi_{i,D}$ – is the payoff of a defector in a group with i cooperators. Together this amounts to $2n$ parameters that can be chosen freely.

The population structure is determined by $f_0, ..., f_n$, where f_i is the frequency of groups with i cooperators and $n - i$ defectors. Because these have to add up to 1, there are only n, and not $n + 1$ degrees of freedom. With $n = 2$ these group frequencies are uniquely determined by r and p; any choice for r and p comes with one unique combination of f_0, f_1 and f_2, and any combination of f_0, f_1 and f_2 that adds up to 1 implies one unique combination of r and p (see Section 3.2). With $n > 2$ that is no longer true. Because the space of population states has dimension n, there is a multitude of possible population states that are consistent with the same value for r and p (see Fig. 9) and in some the average payoff of cooperators might be higher than that of defectors, and in others the defectors might have a higher average payoff.

If we just look at what one could describe as the most basic criterion – whether cooperators have a higher fitness than defectors – then all of those parameters enter there. The average payoff to a cooperator is

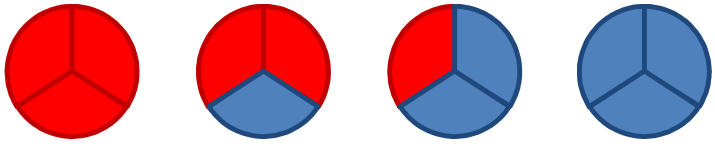
$$\overline{\pi}_C = \frac{\sum_{i=1}^{n} i \cdot f_i \cdot \pi_{i,C}}{np}.$$

Similarly, the average payoff to a defector is

$$\overline{\pi}_D = \frac{\sum_{i=0}^{n-1} (n-i) \cdot f_i \cdot \pi_{i,D}}{n\,(1-p)}.$$

The fully general criterion $\overline{\pi}_C > \overline{\pi}_D$ will therefore always involve all of the f_i's. Generalized equal gains from switching puts a restriction on the admissible games. Suppose all payoffs are defined as follows: $\pi_{i,C} = 1 + ib - c$ and $\pi_{i,D} = 1 + ib$. Now all payoffs are functions of

two parameters only, and the game satisfies equal gains from switching. Substituting the payoffs into the above equations, we see that $\overline{\pi}_C > \overline{\pi}_D$ if and only if $r\mathbf{b} > \mathbf{c}$, where net aggregate benefits $\mathbf{b}$ are given by $\mathbf{b} = (n-1)\,b$ and net costs $\mathbf{c}$ by $\mathbf{c} = c - b$.



Well mixed (r = 0)	1/8	3/8	3/8	1/8
Not well mixed, but also r = 0	0	3/4	0	1/4
Full sibs (r = 0.5)	5/16	3/16	3/16	5/16
Not full sibs, but also r = 0.5	1/4	3/8	0	3/8

Figure 9. Population structures with $n = 3$. The table gives values for f_0, f_1, f_2 and f_3.

Without the restriction to games that satisfy equal gains from switching, knowing relatedness r may not be enough to determine whether cooperation gets selected. The Rock Band game illustrates this. Suppose a band only sounds good if all three players have rehearsed. If the band sounds good, then all players get a payoff of 2 from it. Rehearsing comes at a personal cost to the individual of 1. Payoffs to each band member therefore are $\pi_{i,C} = -1$ if $i = 1$ or 2, $\pi_{i,C} = +1$ if $i = 3$, and $\pi_{i,D} = 0$ for all i. For this game, the criterion $\overline{\pi}_C > \overline{\pi}_D$ can be rewritten as $\frac{f_3}{p} > \frac{1}{2}$. From Fig. 9 we can conclude that in the case of siblings, cooperators will be selected for at $p = 0.5$, because at that frequency $\frac{f_3}{p} = \frac{5}{8} > \frac{1}{2}$. For the last population state in Fig. 9, on the other hand, cooperation is selected against – because $\frac{f_3}{p} = \frac{3}{8} < \frac{1}{2}$ – even though it has both also a relatedness of $r = 0.5$ and also a frequency of cooperators of $p = 0.5$.

If we use the Karlin and Matessi counterfactual method of calculating costs and benefits, and only consider cooperators, we arrive at average costs of $(f_1 + 2f_2 - 3f_3)/3p$ and average benefits of $2f_3/p$. If, instead, we consider all individuals, average costs are $f_0 + f_1 + \frac{1}{3}f_2 - f_3$ and average benefits are $2f_3 + \frac{2}{3}f_2$. Neither of these choices make Hamilton's rule work here.

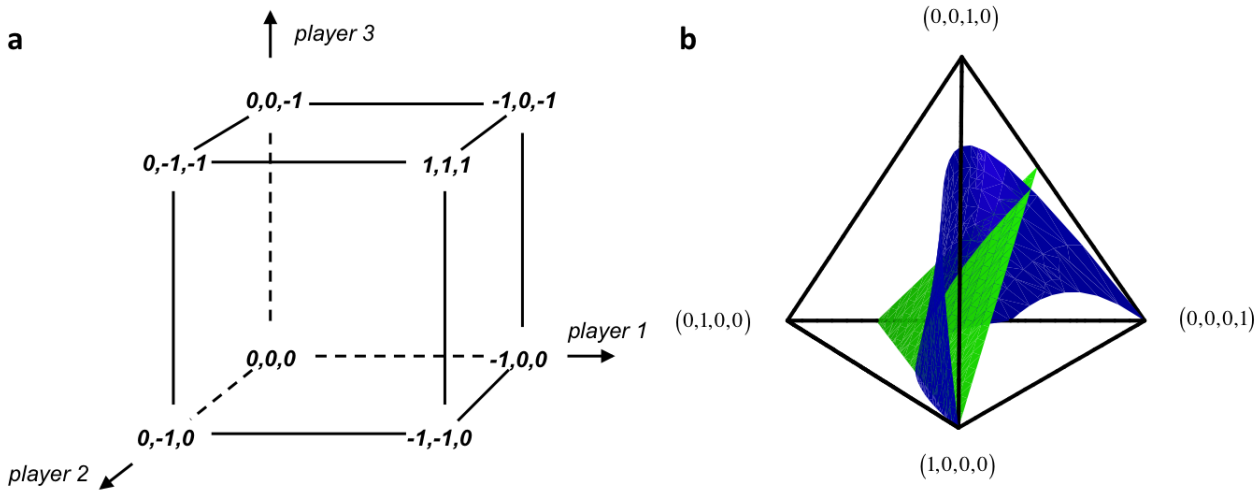


Figure 10. The rock band game. (a) The payoffs of the game and (b) the 3D simplex for the game. A point in this simplex represents a population state (f_0, f_1, f_2, f_3), with $f_0 + f_1 + f_2 + f_3 = 1$ and $0 \leq f_i \leq 1$ for $i = 0, 1, 2, 3$. The vertex closest to us is $f_0 = 1$, the rightmost vertex is $f_3 = 1$, the leftmost vertex is $f_1 = 1$ and the top vertex is $f_2 = 1$. The surface that separates the plus-region from the minus region (green) does not have the same shape as a constant-r surface (blue; $r = \frac{1}{4}$). All constant-r surfaces stretch from the vertex $(1, 0, 0, 0)$, where the frequency of cooperators is 0, to the vertex $(0, 0, 0, 1)$, where the frequency of cooperators is 1.

Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) find that in order for Hamilton's rule to be qualitatively valid, the fitness functions for cooperators and defectors need to be parallel linear functions, which they find is only true if the game satisfies what we call (generalized) equal gains from switching (Matessi & Karlin, 1986, page 697). Their interpretation of Hamilton's rule is that it is a global criterion, and independent of the current frequency p. Both the 2-player and the 3-player cases that we discussed show that the same restriction applies also if we allow Hamilton's rule to be a local criterion, where costs and benefits are allowed to vary with p, provided that we stick to the definition of costs and benefits using counterfactuals. At the end of Section 4, we will revisit these examples, this time using the definition of costs and benefits that follow from the regression method (Gardner et al., 2011).

In the literature, there are many papers about n-player games that do not satisfy equal gains from switching; for example Zheng et al., 2007, Milinski et al., 2008, Kurokawa & Ihara, 2009, Pacheco et al., 2009, Souza et al., 2009, Wang et al., 2009, van Veelen, 2009, 2011a,b, Archetti & Scheuring, 2010, 2011, Gokhale & Traulsen, 2010, 2011, Santos & Pacheco, 2011, van Veelen & Nowak, 2012.

4 The regression method

In this section we will discuss the "regression method" for defining costs and benefits, and we derive the result that, with costs and benefits defined according to this method, Hamilton's rule *always* holds. The starting point for this result, however, is that we already have chosen a specification for the regression, and that this specification is linear. What the costs and benefits are therefore is not necessarily uniquely defined, and may, for one and the same model (or dataset) differ across alternative linear specifications. For the regression method to be well-defined for all possible models (or datasets) it would therefore need to be combined with a method for choosing between different specifications. Subsequently, whatever criterion one would use for choosing one linear specification over another should also be used to choose between linear and non-linear specifications, or between different non-linear ones. We therefore argue that Hamilton's rule, using the regression method, cannot both always be uniquely defined, and generally valid. The general validity depends crucially on the specification being linear, while any criterion that one would use for choosing between different linear specifications will immediately imply that there will also be models, or datasets, where the same criterion will rule in favour of a non-linear specification.

The regression method, as employed in Queller (1992a,b), Gardner et al. (2011), West & Gardner (2013), Marshall (2011, 2015) and Rousset (2015) is the basis for the result that Hamilton's rule *always* holds, provided that we interpret the regression coefficients that the method implies as the benefits and costs in Hamilton's rule. The difference in opinion on the generality of Hamilton's rule results from a disagreement on whether interpreting the regression coefficients as costs and benefits is justified (Okasha, 2016, Okasha & Martens, 2016a,b, see also Birch, 2014, and Birch & Okasha, 2014). Because the regression method is central to the claim of generality, and because the interpretation of the regression coefficients as costs and benefits is central to the disagreement (see also Allen et al. 2013), it is worthwhile to discuss the regression method in detail. Moreover, whether one agrees or disagrees with that interpretation, in any case it is useful to have a formal derivation of the result itself at hand.

The name "regression method" suggests a link with statistics, and the computation of the regression coefficients is indeed the same as in standard statistical exercises, or at least very similar. There is however a significant difference. More often than not, the regression method is applied to *theoretical* models, computing benefits and costs for a given fitness function and, possibly, a given state of the population. This implies that the true model is known. Also the statement that Hamilton's rule always holds is a claim in the theory domain; it states that whatever the true theoretical model is, the regression method will always return costs and benefits such that Hamilton's rule agrees with the direction of selection. That is an exercise that is fundamentally different from statistics, where regressions are applied to *data* in order to uncover an *unknown model* that generated those data. In statistics, regressions therefore are inevitably combined with statistical tests.

The result that is the basis for the use of regressions in statistics is the Gauss-Markov theorem. This theorem states that the parameter estimates (which is how the regression coefficients are interpreted here) that result from applying an (ordinary least squares) regression have desirable properties, such as being unbiased, and having minimum variance among all unbiased estimators. The trueness of the theorem however does depend on assumptions concerning the distribution of the noise term, and on the assumption that the model specification is correct. If the true model is different from the specification chosen, then the regression coefficients are typically no longer unbiased estimates of the parameters they are meant to estimate. Statistical tests are concerned with the model specification part, and they try to find out if the data are not at odds with the specification that is chosen, and with the assumptions about the noise term. A statistical exercise therefore combines finding the correct model specification on the one hand with regressions that estimate model parameters given that specification on the other.

The "regression method" has taken the recipe for estimating parameters out of its statistical domain, and applies it, mostly, to theoretical models. That means that the rationale for using regressions in statistics (the Gauss-Markov theorem) no longer applies. Actually, the regression method is typically applied when the known model is *different* from the specification chosen for the regression method, for instance because the true model is non-linear, while the specification for the regression method is linear. The interpretation of the regression coefficients therefore cannot be the same as their interpretation in statistics.

The regression method is silent about the choice of a specification, and we will see that this presents us with a problem. Below we will derive the claim that Hamilton's rule always holds. After the formal derivation, we will point to the fact that the claim of generality is true, *whichever linear specification is chosen* (and provided that the regression does not lead to an underdetermined system of equations; see Allen & Nowak, 2015). The result therefore implies that Hamilton's rule holds just as much for costs and benefits that follow from one linear specification as it does for costs and benefits that follow from another. This in turn implies that there is a specification issue that needs resolving in order for Hamilton's rule to be uniquely defined; if we do not solve the specification issue, we can have multiple Hamilton's rules, with differing costs and benefits. We will argue that any sensible criteria that one would use for choosing one linear specification over the other immediately imply that there are also cases where the same criteria will decide in favour of non-linear specifications and against linear ones. This then undermines the general validity of Hamilton's rule, which requires the specification to be linear.

4.1 One variable

Because the regression method is a general approach, that can be applied to populations with a discrete distribution of traits as well as populations with a continuous distribution of traits, we will use probability measures to describe population states. This subsumes

continuous probability densities as well as discrete probability measures. A probability measure μ will reflect the distribution of trait values at time 0, or in the parent population, and gives probabilities with which a randomly drawn individual from that population has certain trait values. In the case that the trait distribution takes on a discrete set of values $x_1, ..., x_n$, with corresponding frequencies $p_1, ..., p_n$, the probability measure μ will be a point measure, where $\mu(x_i) = p_i$. In particular, if $x_1, ..., x_N$ are distinct trait values in the parent population, each occuring once, then $\mu(x_i) = \frac{1}{N}$ for each i. In most applications of the regression method, this represents a population state in a theoretical model. It is however also possible to have the probability distribution represent the parent population in a dataset, in which case the probability measure will automatically be discrete.

Besides a probability measure μ, we have a function $f : S \to \mathbb{R}_0^+$. This S is the set of trait values, and f is integrable with respect to μ. This function typically reflects how reproduction depends on x in a model, but it could also reflect realized reproduction in a dataset. The population state after reproduction can be written as a new probability measure λ, which just reflects what one round of reproduction according to fitness function f does to the distribution of trait values in the population starting at μ; $\lambda(T) = \int_T f(x)\, d\mu / \int f d\mu$ for any measurable set T. Dividing by $\int f d\mu$ normalizes the new probability measure, so that it integrates to 1 again. Sometimes a function f is constructed so that $\int f d\mu = 1$ by definition, but normalizing has the same effect.

Together, μ and f contain all the relevant information about a transition from one generation to the other. In a theory model f is a fitness function, that determines what the next generation will be like, if the current is μ. In an empirical exercise, μ and f together represent a dataset, where μ represents the parent generation, and f and μ together make λ, which represents the offspring generation. Even though μ and f are perfectly informative, one might still want to replace f with a polynomial, without affecting certain characteristics of the transition. For this polynomial with degree n we write $g_n(x) = a_0 + a_1 x + ... + a_n x^n$. Suppose furthermore that the coefficients $a_0, ..., a_n$ are chosen so that they minimize the squared difference between f and g_n:

$$\min_{a_0,...,a_n} \int (f - g_n)^2 \, d\mu.$$

The first order conditions – setting the derivatives w.r.t. $a_i, i = 0, ..., n$ equal to 0 – imply that:

$$\int x^i f d\mu = \int x^i g_n d\mu, \quad i = 0, ..., n. \tag{4.1}$$

The first two of these $n + 1$ conditions – the ones for $i = 0$ and $i = 1$ – imply that we can replace f with g_n without affecting the change in average x – of course assuming that $n \geq 1$

$$\frac{\int x f d\mu}{\int f d\mu} - \int x d\mu = \frac{\int x g_n d\mu}{\int g_n d\mu} - \int x d\mu. \tag{4.2}$$

In other words, one can replace f by any polynomial of degree 1 or higher, without affecting the change in average x, if we choose the polynomial coefficients so that they minimize the squared difference. If one would want to replace f by a polynomial of the *lowest* possible degree and preserve this property, it is enough to take $n = 1$.

If we do indeed choose $n = 1$, then that implies

$$\int f d\mu = a_0 + a_1 \int x d\mu \qquad \Rightarrow \quad a_0 = \int f d\mu - a_1 \int x d\mu \tag{4.3}$$

and

$$\begin{aligned} \int x f d\mu &= a_0 \int x d\mu + a_1 \int x^2 d\mu \qquad (4.4) \\ \Rightarrow \quad \int x f d\mu &= \left(\int f d\mu - a_1 \int x d\mu \right) \int x d\mu + a_1 \int x^2 d\mu \\ \Rightarrow \quad a_1 &= \frac{Cov\left(X, f\right)}{Var\left(X\right)}. \end{aligned}$$

Moreover, we can use (4.4) and (4.3) to rewrite the condition for the change in average x to be positive[5]:

$$\frac{\int x f d\mu}{\int f d\mu} - \int x d\mu > 0 \qquad \Leftrightarrow \qquad a_1 > 0 \tag{4.5}$$

Summarizing, we found that one can replace $f\left(x\right)$ by $g_1\left(x\right) = a_0 + a_1 x$ without consequences for the change in average x, provided that we choose a_0 and a_1 such that they minimize $\int \left(f - g_1\right)^2 d\mu$. Moreover, the average x goes up if and only if $a_1 > 0$. It might be useful to also mention that the function g_1 is *not* a local linearization of f.

4.2 Two variables

For considering cases with two relevant trait values, let μ be a probability measure on $\mathbb{R}^2$. These two quantities can be thought of as values for two different traits, which will be useful as a reference. They can also be interpreted as values of the same trait, the first representing the trait value that the agent itself has, the second representing the trait

[5] One can use (4.4) to rewrite $\frac{\int x f d\mu_1}{\int f d\mu_1} - \int x d\mu_1 > 0$ as $\frac{a_0 \int x d\mu_1 + a_1 \int x^2 d\mu_1}{\int f d\mu_1} - \int x d\mu_1 > 0$. With (4.3), this can be rewritten as $\frac{\left(\int f d\mu_1 - a_1 \int x d\mu_1\right) \int x d\mu_1 + a_1 \int x^2 d\mu_1}{\int f d\mu_1} - \int x d\mu_1 > 0$, or $\frac{a_1 \left(\int x^2 d\mu_1 - \left(\int x d\mu_1\right)^2\right)}{\int f d\mu_1} > 0$. This is true if and only if $a_1 > 0$.

value of its interaction partner. The second interpretation will lead to the first version of Hamilton's rule.

Let f be a function on $\mathbb{R}^2$. In the first interpretation, where the two variables represent different traits in the same individual, this induces a probability measure λ, representing the distribution of the two traits in the next generation in the same way as it did with one variable; $\lambda(T) = \int_T f(x,y)\, d\mu / \int f d\mu$ for any measurable set T in $\mathbb{R}^2$. In the second interpretation, y does not represent a different trait within the same individual, but the value of the same trait in another individual (the interaction partner), which is not heritable. This fitness function f therefore only informs us about the distribution of x in the next generation. We can use λ_x to denote the marginal probability measure that represents this distribution; $\lambda_x(T) = \int_{T\times\mathbb{R}} f(x,y)\, d\mu / \int f d\mu$ for any measurable set T in $\mathbb{R}$. If we would like to arrive at a complete description of the new generation, then more information is required, or more assumptions need to be made. One possibility is that the transition as a whole tracks a model, the assumptions of which imply a fitness function f as well as how individuals are matched in every new generation. Another possibility is that more straightforward assumptions about matching are made, which define, for every distribution of traits x, what the according joint distribution of x and y is, for instance reflecting interactions between siblings. For derivations of Hamilton's rule, however, it is enough to have the marginal probability distribution. Hamilton's rule can perfectly well pertain to one transition only, in which case the matchings in the next generation do not matter.

Together, μ and f again contain all the relevant information about a transition from one generation to the other, and again we will replace f with a function of degree 1, which this time uses 2 variables: $g(x,y) = a_{00} + a_{10}x + a_{01}y$. Suppose that we minimize the squared difference between f and g:

$$\min_{a_{00},a_{10},a_{01}} \int (f-g)^2 \, d\mu$$

The first order conditions imply that

$$\int f d\mu = a_{00} + a_{10}\int x d\mu + a_{01}\int y d\mu \tag{4.6}$$

$$\int x f d\mu = a_{00}\int x d\mu + a_{10}\int x^2 d\mu + a_{01}\int xy d\mu \tag{4.7}$$

$$\int y f d\mu = a_{00}\int y d\mu + a_{10}\int xy d\mu + a_{01}\int y^2 d\mu \tag{4.8}$$

It is possible that this system does not have a unique solution. That happens if the distribution μ is such that any individual's y value follows linearly from their x value. If μ puts positive probabilities only on individuals with $y = Ax + B$, for constants A and B, then there are infinitely many combinations a_{00}, a_{10} and a_{01} that would produce one and the

same function $g(x,y)$. Therefore, if a given $g(x,y)$ minimizes $\int (f-g)^2 d\mu$, then so do all equivalent choices for a_{00}, a_{10} and a_{01}. This naturally shows up in the first order conditions; in this case (4.8) is equal to A times (4.7) plus B times (4.6). Equation (4.8) is then redundant, leaving us with a system with 2 equations and 3 unknowns. One possible way to arrive at such a situation is if the parent population consists of two (x,y)-combinations only. We will return to this possibility in Section 4.5.6.

In the typical, and more interesting case where there is a unique solution to this system of equations, we can use (4.7) and (4.6) to rewrite the change in average x

$$\begin{aligned}
\frac{\int xfd\mu}{\int fd\mu} - \int xd\mu &= \frac{a_{00}\int xd\mu + a_{10}\int x^2 d\mu + a_{01}\int xyd\mu}{\int fd\mu} - \int xd\mu \qquad (4.9)\\
&= \frac{\left(\int fd\mu - a_{10}\int xd\mu - a_{01}\int yd\mu\right)\int xd\mu + a_{10}\int x^2 d\mu + a_{01}\int xyd\mu}{\int fd\mu} - \int xd\mu\\
&= \frac{a_{10}\left(\int x^2 d\mu - \left(\int xd\mu\right)^2\right) + a_{01}\left(\int xyd\mu - \int yd\mu\int xd\mu\right)}{\int fd\mu}\\
&= \left(a_{10} + \frac{Cov(X,Y)}{Var(X)}a_{01}\right)\frac{Var(X)}{\int fd\mu}.
\end{aligned}$$

If x and y represent different traits, then this equation captures the possibility that higher values of x can be selected for, not because having a high value of x is fitness enhancing per se, but because the covariance between the traits is high enough, and having a high trait value of y is good for fitness. If we have a model where x is the genotype of the agent, and y is the genotype of its interaction partner – sometimes also denoted as x' rather than y – then $\frac{Cov(X,Y)}{Var(X)}$ can be interpreted as the relatedness between interaction partners. If we moreover interpret a_{10} and a_{01} as costs and benefits ($c=-a_{10}$ and $b=a_{01}$) then it follows that Hamilton's rule *always* holds. In other words, $\int xfd\mu/\int fd\mu - \int xd\mu > 0$ if and only if $a_{10} + \frac{Cov(X,Y)}{Var(X)}a_{01} > 0$.[6]

In a model with two binary traits, where $X=1$ if the agent is a cooperator, and $Y=1$ if its interaction partner is, $\frac{Cov(X,Y)}{Var(X)} = \mathbb{P}(C|C) - \mathbb{P}(C|D) = r$.[7] If μ and f would represent data rather than a model, this quantity would be the *sample* covariance over

[6] One can also include the normalization in the b and c, and define $c=-a_{10}/\int fd\mu_1$ and $b=a_{01}/\int fd\mu_1$. This is also what one gets if the normalization is done at the construction of f, that is, if we use $\widehat{f} = f/\int fd\mu_1$ instead of f. This is done in, amongst others, Gardner et al. (2011).

[7] $\mathbb{P}(C|C) - \mathbb{P}(C|D) = \frac{\mathbb{E}[XY]}{\mathbb{E}[X]} - \frac{\mathbb{E}[Y]-\mathbb{E}[XY]}{1-\mathbb{E}[X]} = \frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X]-\mathbb{E}^2[X]}$. With binary variables, $\mathbb{E}[X] = \mathbb{E}\left[X^2\right]$, so this equals $\frac{Cov(X,Y)}{Var(X)}$.

the *sample* variance, which would amount to being an estimate of relatedness, rather than being relatedness itself.

Notice that (4.8) is not actually used in the derivation. That implies that, even though the regression method prescribes that we do minimize $\int (f-g)^2 \, d\mu$ with respect to all three variables, we could actually have chosen *any* value for a_{01}, minimized $\int (f-g)^2 \, d\mu$ only with respect to a_{00} and a_{10}, and have arrived at a Hamilton's rule with the exact same derivation. An alternative choice for a_{01} will typically lead to a alternative values for a_{00} and a_{10} too, and different choices for a_{01} would therefore come with different Hamilton's rules, with different b's and c's, that all still correctly reflect the direction of selection. One particular alternative choice for a_{01} would be $a_{01} = 0$. In that case the model specification would revert to the case with one variable, where we wrote a_0 for a_{00} and a_1 for a_{10}.

Because specification issues are a recurrent theme, it may help to stress that if the true model f would be a linear function of both x and y, a specification with x only would not stop us from getting the direction of selection right, but will make us draw incorrect conclusions. Suppose that $f(x,y) = d_{00} + d_{10}x + d_{01}y$ with $d_{10} < 0$ and $d_{01} > 0$, and suppose furthermore that f and μ combine in such a way that the average value of the first trait increases; $\int x f d\mu / \int f d\mu - \int x d\mu > 0$. If we then choose the specification from the previous subsection, with one variable and $g_1(x) = a_0 + a_1 x$, then minimizing the squared difference between f and g would have to return a value $a_1 > 0$, because the average trait value has increased. One would – incorrectly – conclude from this choice of g that having a high trait value itself is a good thing. If instead we use the specification from this section, with $g(x,y) = a_{00} + a_{10}x + a_{01}y$, we bring this integral all the way down to 0, and find $a_{00} = d_{00}$, $a_{10} = d_{10}$, and $a_{01} = d_{01}$ – provided that μ has sufficiently rich support, to avoid trivial cases. We would then conclude that having a higher trait value for x is in fact not good ($a_{10} < 0$) but that it is selected anyway, because it covaries sufficiently much with y, and higher values of y are good ($a_{01} > 0$). Both specifications come with Hamilton's rules that indicate the direction of selection correctly; the first specification has one with $b = a_1 > 0$ and $c = 0$, and the second specification has one with $b = a_{10} < 0$ and $c = a_{01} > 0$. Given that the second specification matches f perfectly, it makes perfect sense to suggest that the first Hamilton's rule is not the right one, even though it also matches the direction of selection correctly, and that the second Hamilton's rule is in fact the right one. Here, with only two specifications to choose from, that is an obvious point to make. In Sections 4.3 and 4.4 we will have a richer set of specifications to choose from. There we will make the same point, which then may seem less immediately obvious.

The setup with a function f only allows for a fixed number of offspring to go with every combination of x and y. This implies that in a theory model, this restricts attention to models where the number of offspring is deterministic. Also, if the probability measure represents a dataset, then this rules out the possibility of having multiple observations with the same value of (x,y), but different numbers of offspring. In Appendix A we show that

one can relax this restriction without changing the results. The arguments there also justify using functions f that are not restricted to return integers.

4.2.1 More than one equally related interaction partner

The above conclusions hold also if there is not just one interaction partner, but if individuals interact in groups of size $k > 2$. In that case we can take y to be the sum of trait values of all $k-1$ interaction partners. If we do, then $\frac{Cov(X,Y)}{Var(X)}$ amounts to $k-1$ times the relatedness between two individuals in the same group. One can still interpret a_{01} as the per partner benefits. If we multiply this by the number of interaction partners – which is $k-1$ – then we could interpret that product as total benefits, and we again get Hamilton's rule; the change in average x is positive if $rb - c > 0$, where $c = -a_{10}$, $b = (k-1)\, a_{01}$, and $r = \frac{1}{k-1}\frac{Cov(X,Y)}{Var(X)}$.

4.3 More than two variables, more than one Hamilton's rule

The starting point in Hamilton (1964) is an array of fitness effects on individuals with different relatednesses to the agent. Thereby Hamilton's model not only allows for behaviour with effects on siblings *or* on cousins, but also on siblings *and* cousins at the same time. Also elsewhere in the literature, behaviours are considered that simultaneously affect different individuals that have distinct degrees of relatedness to the agent (see for instance Grafen, 2007b). This can be encompassed by allowing for more than two independent variables. The first will then pertain to the agent, the second to the first type of interaction partner, the third to the second type of interaction partner, and so on. We will therefore allow for f to depend on m independent variables, and also consider functions g to do the same, but moreover are linear: $g(x_1, ..., x_m) = a_{0,...,0} + a_{1,0,...,0}x_1 + ... + a_{0,...,0,1}x_m$. Suppose again that we minimize the squared difference between f and g:

$$\min_{a_{0,\ldots,0},\, a_{1,0,\ldots 0},\, \ldots\, ,\, a_{0,\ldots,0,1}} \int (f-g)^2 \, d\mu$$

The first order conditions imply that

$$\frac{\int x_1 f d\mu}{\int f d\mu} - \int x_1 d\mu = \left(a_{1,0,\ldots,0} + \frac{Cov(X_1, X_2)}{Var(X_1)} a_{0,1,0,\ldots,0} + ... + \frac{Cov(X_1, X_m)}{Var(X_1)} a_{0,,\ldots,0,1} \right) \frac{Var(X_1)}{\int f d\mu} \tag{4.10}$$

If we now interpret $-a_{1,0,\ldots,0}$ as the costs c of the behaviour to the agent, $a_{0,1,0,\ldots,0}$ up to $a_{0,\ldots,0,1}$ as the benefits b_2 to b_m to differently related individuals, and $\frac{Cov(X_1,X_2)}{Var(X_1)}$ up to $\frac{Cov(X_1,X_m)}{Var(X_1)}$ as relatednesses r_2 to r_m to these different types of individuals, then, also in this more general setup, it follows that Hamilton's rule always holds; $\int x_1 f d\mu / \int f d\mu - \int x_1 d\mu >$

0 if and only if $-c + r_2 b_2 + ... + r_m b_2 > 0$. This nicely mirrors Hamilton's original setup (see also Section 2).

The derivation is a straightforward generalization of the derivation with two variables in Section 4.2. Also here, it only uses the first order conditions that pertain to the derivatives with respect to $a_{0,...,0}$ and $a_{1,0,...,0}$. Again that implies that there is scope for multiple specifications, leading to multiple Hamilton's rules, all of which indicate the direction of selection correctly. We can choose to minimize $\int (f-g)^2 \, d\mu$ using all coefficients, but we can also set some of them to 0, and only minimize with respect to the others. As long as $a_{0,...,0}$ and $a_{1,0,...,0}$ are not set to 0, these will all result in Hamilton's rules. Examples in Section 9.3 illustrate that. Picking the right one is again a relatively straightforward task if f has the same general form as g; if $f(x_1, ..., x_m) = d_{0,...,0} + d_{1,0,...,0} x_1 + ... + d_{0,...,0,1} x_m$, with all coefficients non-zero, then the squared difference between f and g is reduced to 0 if we choose the specification that includes all coefficients.

4.4 More than two variables with higher order terms

Although the regression method as a way to determine benefits and costs is restricted to linear terms (see Gardner et al., 2011, Box 4), we would also like to allow for a more general setup, where the difference between f and a polynomial $g_{\mathcal{J}}$ is minimized. This requires a little notation. The set of all coefficients that are included in this polynomial is $\mathcal{J}$. This is a finite subset of $\mathbb{N}_0^m$, and an element of $\mathcal{J}$ is a vector, elements of which indicate the exponents of the variables in the term they are a (possibly non-zero) coefficient for. In other words, $g_{\mathcal{J}} = \sum_{j \in \mathcal{J}} a_j x_1^{j_1} x_2^{j_2} ... x_m^{j_m}$. The minimization then becomes

$$\min_{j \in \mathcal{J}} \int (f - g_{\mathcal{J}})^2 \, d\mu.$$

The first order conditions of this minimization imply a more general form of the identities in Sections 4.1, 4.2 and 4.3, all of which are special cases of the general version. For brevity, we write $X^{(j)} = X_1^{j_1} X_2^{j_2} ... X_m^{j_m}$.

$$\frac{\int x_1 f d\mu}{\int f d\mu} - \int x_1 d\mu = \left(\sum_{j \in \mathcal{J} \setminus \{a_{0,...,0}\}} \frac{Cov\left(X_1, X^{(j)}\right)}{Var\left(X_1\right)} a_j \right) \frac{Var\left(X_1\right)}{\int f d\mu} \tag{4.11}$$

This implies that the change in average trait value $\int x_1 f d\mu / \int f d\mu - \int x_1 d\mu$ is larger than 0 if and only if $\sum_{j \in \mathcal{J} \setminus \{a_{0,...,0}\}} \frac{Cov\left(X_1, X^{(j)}\right)}{Var(X_1)} a_j > 0$.

The derivation is again a straightforward generalization of the derivations in Sections 4.2 and 4.3, where only the first order conditions that pertain to $(0, ..., 0)$ and $(1, 0, ..., 0)$ are used. This implies that this identity holds, whatever set of coefficients $\mathcal{J}$ we allow to be non-zero, as long as $(0, ..., 0)$ and $(1, 0, ..., 0)$ are included.

This leaves us with a possible multitude of rules. The choice of $g_{\mathcal{J}}$ – or in other words: the choice which coefficients to include in $\mathcal{J}$ – is the specification. For a given f and μ, different specifications may lead to different values for the coefficients that are included in both, and all specifications produce rules that indicate the direction of selection correctly. Some of those are Hamilton's rules. If the regression contains only linear terms and a fixed term – in other words, $\sum_{k=1}^{m} j_k \leq 1$ for all a_j with $j \in \mathcal{J}$ – then we are back in the situation described in Section 4.3. Others, that do include coefficients for non-linear terms, qualify as proper generalizations of Hamilton's rule, but are not Hamilton's rules themselves. The specification problem now amounts to finding criteria for choosing the right $\mathcal{J}$.

For functions f that are polynomials themselves, one can imagine that the recipe for finding the right specification involves starting with a fixed term, and a coefficient for x_1, and then adding ever more coefficients. What one will typically find is that, as coefficients are added to $\mathcal{J}$, the values of the coefficients that are already in there will keep changing, until the point where all coefficients that are non-zero in f are included in $g_{\mathcal{J}}$, at which point $\int (f - g_{\mathcal{J}})^2 \, d\mu = 0$. After this, every coefficient that is added will get the value 0 in the minimization, and the coefficients already in there will stop changing. At this point we have found the right specification, because $f = g_{\mathcal{J}}$.

If $f(x_1, ..., x_m) = 1 - x_1 + x_2 + x_3$ – where x_1 might represent the agent's own trait value, x_2 the trait value of the agent's sibling, and x_3 the trait value of the agent's cousin – such a recipe will typically choose the specification $g(x_1, ..., x_m) = a_{0,...,0} + a_{1,0,...,0}x_1 + a_{0,1,0,...,0}x_2 + a_{0,0,1,0,...,0}x_3$ over $g(x_1, ..., x_m) = a_{0,...,0} + a_{1,0,...,0}x_1$ and over $g(x_1, ..., x_m) = a_{0,...,0} + a_{1,0,...,0}x_1 + a_{0,1,0,...,0}x_2$, even though those other two also come with Hamilton's rules that get the direction of selection right. If $f(x_1, ..., x_m) = 1 + x_1x_2$, then this recipe would choose $g(x_1, ..., x_m) = a_{0,...,0} + a_{1,1,0,...,0}x_1x_2$, which comes with a rule, but not a Hamilton's rule. The examples below, as well as in Section 9.3, illustrate that further. Of course this recipe would need to be augmented when applied to functions f that are not themselves polynomials, but for functions f that are, there is no reason to treat the decision whether or not to include coefficient $a_{0,0,1}$ any different from the decision whether or not to include coefficient $a_{1,1,0}$.

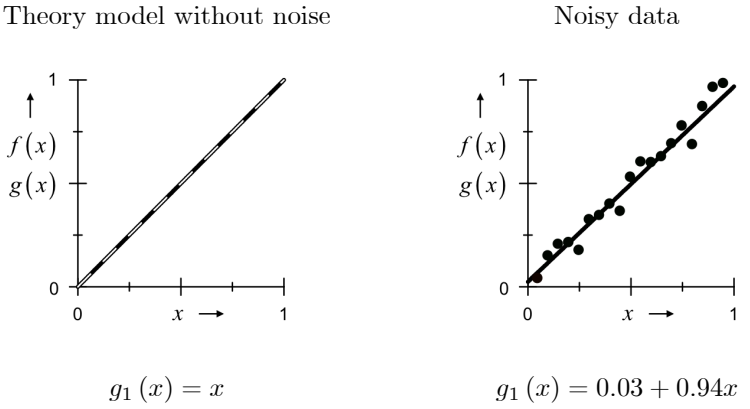
4.5 Coefficients, models and statistics

Our preferred definition of costs and benefits follows the "counterfactual method". This approach compares fitnesses with their counterfactuals, and has the differences between them determine the costs and benefits of cooperation. For us, therefore, there is no need to replace fitness functions f, that we can use directly to derive model implications, with other functions g, and minimize the squared difference. The regression method is claimed to be a method for defining costs and benefits, and, for a given linear specification, it is. What we argue, however, is that the right specification does not fall from the sky, but has to be chosen too. The regression method, without a recipe how to chose one, is

therefore incomplete. Moreover, reasonable recipes for choosing between specifications will sometimes choose non-linear ones too, and once we allow for non-linear specifications, we basically are back to square one, and have to decide what the costs and benefits are in the presence of non-zero coefficients for non-linear terms. For this, the counterfactual method seems the most logical choice. But if we use the counterfactual method anyway, there is no reason not to apply it to f directly.

In the remainder of this section we would like to look at a few examples. The first examples are meant to illustrate the difference between regressions in models and regressions in statistics. Then we will reconnect the regression method to the examples from Section 3, and see how the regression method leads to costs and benefits that differ from the costs and benefits of the counterfactual method.

4.5.1 Example 1



$g_1(x) = x$ $\qquad$ $g_1(x) = 0.03 + 0.94x$

Figure 11. In the theory model on the left $\int (f - g_1)^2 \, d\mu = 0$, whereas on the right the randomness in the data causes $\int (f - g_1)^2 \, d\mu > 0$.

Suppose we have a theoretical model with $f(x) = x$, and let μ be the uniform distribution on $[0, 1]$. If we minimize the squared difference with $g_1(x) = a_0 + a_1 x$ we, trivially, find $a_0 = 0$, $a_1 = 1$, and $\int (f - g_1)^2 \, d\mu = 0$ (Fig. 11a).

If instead we generate data with a process in which the number of offspring is normally distributed around x, then the randomness in the observations implies that if we minimize the sum of squares, using g_1, this minimum will typically be larger than 0 (Fig. 11b). Because the data are generated by a linear model, with uncorrelated, homoscedastic errors, the Gauss-Markov theorem implies that the OLS estimator has minimum variance among the class of linear unbiased estimators. In other words, the expected value of the estimates of a_0 and a_1 (i.e., the estimates of slope and intercept) will be equal to their true values, and their variances are minimal. Here we know the true values, because we generated the data ourselves, but typically they will not be known, because if they were, there would be no need to do statistics. Here we also happen to know that the assumptions that the Gauss-Markov

theorem requires hold, but these can, and typically should, also be tested statistically. With the data generated for Fig. 11b, we get estimates $\widehat{a}_0 = 0.03$ and $\widehat{a}_1 = 0.94$, which are both relatively close to their true values, which are 0 and 1.

4.5.2 Example 2

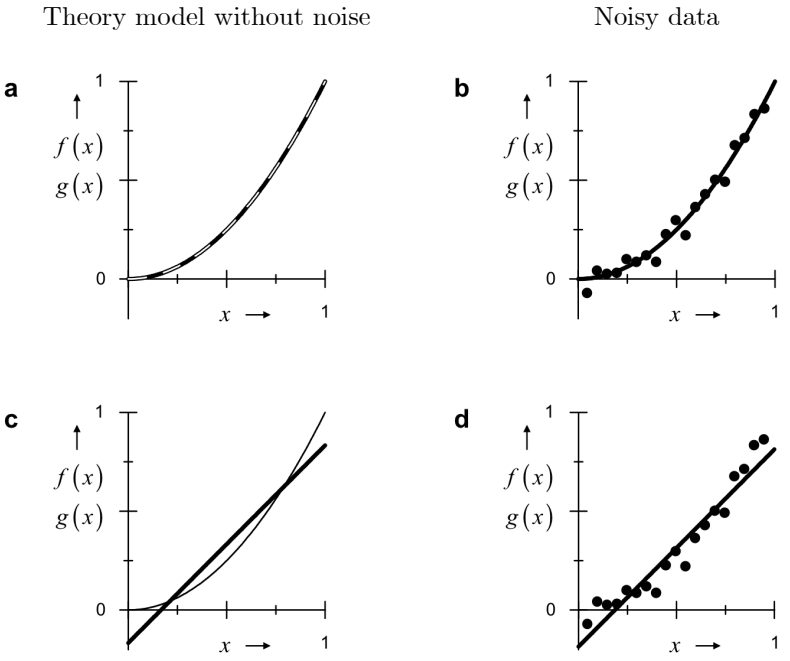


Figure 12. If OLS is applied in a statistical setting, then there are two distinct reasons that can make the sum of least squares be larger than 0; randomness (b and d) and misspecification (d). In a theory context, only misspecification remains (c).

Suppose now that $f(x) = x^2$, and let μ be the uniform distribution on $[0, 1]$ again. If we now minimize the squared difference with $g_1(x) = a_0 + a_1 x$, then we find $a_0 = -\frac{1}{6}$ and $a_1 = 1$ (Fig. 12c). If we minimize the squared difference with polynomials of degree 2 or higher, then we find that only $a_2 = 1$, while all other coefficients are 0 (Fig. 12a). Moreover $\int (f - g_n)^2 \, d\mu = 0$ if $n > 1$, but not if $n = 1$. One could therefore say that if we use a polynomial of degree 1, then the reason that least squares minimization gives us $a_0 = -\frac{1}{6}$ and $a_1 = 1$, instead of both being 0, is that g_1 has no coefficient for x^2 in it.

If we generate data with a process in which the number of offspring is normally distributed around x^2, then the randomness in the observations implies that if we minimize the sum of squares, using g_2, this will typically be minimized at some value larger than 0 (Fig. 12b). If, however, we minimize the sum of squares using g_1, then there are two reasons why the sum of squares is larger than 0. One is again that there is randomness, and the other is that the model is misspecified – we have used g_1 instead of g_2, or, in other words, we have not included a coefficient for x^2 (Fig. 12d). Methods in statistics are all geared

towards avoiding the latter. If we now go back to the theory domain, and use g_1 instead of g_2 there, however, then the only reason that remains for why $\int (f - g_n)^2 \, d\mu > 0$ is that g_1 is not equal to f.

It is therefore important to realize that if the regression method is applied to a deterministic model that is not itself linear, then this is a different exercise than applying it to data that are produced by a linear, but noisy data-generating process. Without randomness, the only reason why $\int (f - g_n)^2 \, d\mu$ is not 0 is misspecification. Properties such as those implied by the Gauss-Markov theorem – which is why OLS is applied in statistics – do not apply here. Equation (4.2) nonetheless remains as true as it was before; the direction of selection in a theory model does not change if we replace f with g_1.

4.5.3 Example 3

A simple example with two variables is $f(x, y) = a_{11}xy$. When the squared difference between f and $g(x, y) = a_{00} + a_{10}x + a_{01}y$ is minimized, this will, for many probability distributions μ, lead to non-zero a_{00}, a_{10} and a_{01}, while all three of them are in fact 0.

4.5.4 More specification issues

Allen et al. (2013) give examples that also indicate that the regression method is blind to specification. They illustrate that if the standard linear specification is used, regardless of its fit, then one easily arrives at a c and b that can certainly not properly be interpreted as describing the costs and benefits of the behavior. One example is "hanger-on" behavior, where individuals seek out high fitness individuals to hang out and interact with, but where this interaction does not confer any benefit to the high fitness individual. The regression method would then nonetheless conclude that $b > 0$, and mistake the hanging on for something with a positive effect on fitness. One could argue that this is due to the fact that the regression method in this case is clearly misspecified, but that only underscores the fact that we apparently apply, and need, criteria for what good and bad specifications are. And as soon as goodness of fit starts to matter for the specification of the model, there is no reason why the goodness of fit might not pick a non-linear model as the winner. We will return to this point in Section 9. Here we continue with applications of the regression method to examples that feature in the previous section.

4.5.5 Back to the replicator dynamics for the prisoners dilemma

Gardner et al. (2011) apply the regression method both to the prisoners dilemma, allowing for unequal gains from switching, and to the rock band game. For the prisoners dilemma, we include the least squares minimization in Appendix A, where it should be noted that we use R, S, T and P for payoffs, as we do throughout this paper, while Gardner et al. (2011) follow Queller (1985) by parametrizing those payoffs with B, C and D.

The values that result from applying the regression method are:

$$c = \frac{1-p+rp}{1+r}(P-S) + \frac{p+r(1-p)}{1+r}(T-R) \tag{4.12}$$
$$b = \frac{1-p+rp}{1+r}(T-P) + \frac{p+r(1-p)}{1+r}(R-S).$$

These are different from the values we find using the counterfactual approach (see Section 3.2.1).

4.5.6 One partner versus many partners

In Section 3 we have assumed that each individual is paired with one partner and with one partner only. Alternatively, one can assume that each individual interacts with a large, effectively infinite, sample of the population, and retains the average payoff from these interactions. In this case every individual cooperator gets the same payoff, equal to the average for cooperators from the "one partner" case ($\overline{\pi}_C$), while every individual defector is assumed to get $\overline{\pi}_D$. This "many partners" setup leads to the same equation for the replicator dynamics as the one partner setup (Eq. 3.1). The counterfactual benefits and costs described in Section 3 would also remain the same, but a change to the many partners setup does have an effect when applying the regression method. The parent population now consist of only two distinct points: $(x, y) = (1, \alpha + (1-\alpha)p)$, with frequency p for cooperators and fitness $f(1, \alpha + (1-\alpha)p) = \overline{\pi}_C$, and $(x, y) = (0, (1-\alpha)p)$, with fitness $f(0, (1-\alpha)p) = \overline{\pi}_D$. Because there are only two distinct points, x and y are linearly related: $y = \alpha x + (1-\alpha)p$. It follows from the discussion in Section 4.2 that the regression method does not produce a unique benefit b and cost c in this case.[8]

4.5.7 Back to the replicator dynamics for the rock band game

For the rock band game, Gardner et al. (2011) find

$$c = 1 - \frac{3p(1-p)}{9p(1-p) - 2(f_1+f_2)} \times \frac{f_3}{p} \times 2$$
$$b = \frac{6-p(1-p)}{9p(1-p) - 2(f_1+f_2)} \times \frac{f_3}{p} \times 2.$$

where we have left out the normalization (this is inconsequential; see footnote 6). Also these are different from the costs and benefits based on counterfactuals (see Section 3.3).

[8] This failure of the regression method was first noticed by Allen and Nowak (2015) in the context of a finite-population model. Rousset (2015) seems to claim that this finding is erroneous, because it does not reproduce the result of Gardner et al (2011). However, Rousset (2015) apparently missed the fact that the "one partner" and "many partners" setups lead to different outcomes in the regression method. Gardner et al (2011) and Rousset (2015) use a "one partner" setup, so it is not surprising that they obtain different results from Allen and Nowak (2015), who used a "many partners" setup.

How we define costs and benefits is therefore consequential for whether or not Hamilton's rule holds; it always does if we choose to define b and c as regression coefficients, but it does not always hold if we define them by comparing fitnesses to their counterfactuals. A related, but different question is if it helps understand the dynamics it describes better, if we rewrite the criterion for cooperators to win with Hamilton's rule, where b and c follow from the regression method. This is at least to some extent a matter of preference. Our preference in this particular case goes to the condition

$$\frac{f_3}{p} > \frac{1}{2}.$$

All that matters in the Rock Band Game is that more cooperators get 2 instead of 1 than get 0 instead of 1 (see Fig. 10 and Section 3). Therefore at least half of the cooperators must find themselves in groups of 3 cooperators. That is exactly what this condition says. For us, replacing this criterion with its equivalent Hamilton's rule alternative, with regression coefficients for costs and benefits, is not a gain in clarity or insight in the condition:

$$\left(\frac{6p\,(1-p)}{9p\,(1-p)-2\,(f_1+f_2)} \times \frac{f_3}{p} \times 2\right) r - \left(1 - \frac{3p\,(1-p)}{9p\,(1-p)-2\,(f_1+f_2)} \times \frac{f_3}{p} \times 2\right) > 0.$$

4.5.8 Different costs, different benefits, different rules

The idea that the disagreement about the generality of Hamilton's rule might be the result of a failure to disambiguate different versions of it was put forth by Birch (2014). He also compares different ways to define costs and benefits in Hamilton's rule. One of the possibilities he considers is the regression method, which, in his terminology, leads to the general version of Hamilton's rule. The other possibility is termed the special version, and it is meant to capture the way b and c are defined in Nowak, Tarnita & Wilson (2010) as well as in Van Veelen (2009) and Van Veelen et al., (2012). Since the latter three papers differ in their treatment of b and c, it is unavoidable that the description of the special version there has features of both, but reflects neither choice perfectly. For discussing this, it will be useful to understand a point made by Grafen (2007b), and since this is discussed in Section 7, we will postpone these more detailed points to the final section.

In the remainder of this paper we will consider Hamilton's rule with costs and benefits based on counterfactuals, and not on the regression method. We will find that even then, there is a sizeable domain within which inclusive fitness works, but also that there is a domain where it does not. With b and c defined according to the regression method, whether or not inclusive fitness always aligns with the direction of selection is no longer a question, as the first order conditions imply that it always does.

5 Comparative statics

Even if Hamilton's rule may not always hold as a *quantitative* prediction, it may still be valid *qualitatively*, in the sense that higher relatedness is conducive to the evolution of cooperation. With appropriate definitions of what it means for relatedness to favour cooperation, this turns out to be true for 2-player games. For games with more players these comparative statics do not apply generally, but may apply in reasonably restricted subsets of all possible population structures.

One of the obvious implications of Hamilton's rule is that relatedness is good for cooperation. The higher relatedness is, the larger the scope for the evolution of cooperation, because with a higher r we can make do with a smaller benefit b to offset the same costs c. In Section 3 we found that Hamilton's rule, with definitions of costs and benefits based on comparisons with counterfactuals, is only accurate for games with (generalized) equal gains from switching. But even if $rb > c$ is not the right criterion to determine whether or not cooperation gets selected, it could still be that an increase in relatedness is typically good news for cooperation. In other words, there may be a whole set of games for which the quantitative prediction does not fit Hamilton's rule, but for which it remains true that an increase in relatedness increases the scope for the evolution of cooperation.

In this section we will explore different ways to formalize what we could mean when we say that an increase in relatedness r is good for cooperation (Matessi & Karlin, 1984, 1986 call this the *qualitative validity of the Hamilton rule*, or the *Hamilton Property*). We will look at "comparative statics" (see also Section 2.4 in Frank, 1998, Milchtaich, 2006, Allen & Nowak, 2015, and Cooney & Veller, 2015) and find out that, indeed, there are many 2-player games for which one can unambiguously say that relatedness fosters cooperation. For games with more than 2 players there are complications, but even there it is possible to use comparative statics for *specific* models within which a similar claim is true.

There is a variety of reasons why this is worth doing. One reason is that in the debate concerning inclusive fitness and the evolution of cooperation, kin selection and Hamilton's rule are sometimes conflated (some examples are Foster et al., 2006a,b, Nowak et al. 2010 and Birch & Okasha, 2015). When the general validity of Hamilton's rule is questioned, it is therefore often assumed that kin selection is under attack. By looking at comparative statics, we show that those should be treated separately; increasing relatedness does favour cooperation in an unambiguous sense in almost all 2-player games, including games for which we already showed that Hamilton's rule does not predict the direction of selection. We therefore would also argue that much of the empirical evidence that is claimed to support Hamilton's rule should really be interpreted as supporting the comparative statics instead.

Another reason for looking into this is that it is just very interesting to see if we can formalize and explore if and how relatedness helps cooperation evolve.

5.1 Definitions, derivatives and isoclines with 2 players

The first question we might ask is how increasing the degree of relatedness affects the speed at which cooperation grows (or shrinks) under the dynamics. From Section 3.2 we know the dynamics – they are given by $\dot{p} = p\,(1-p)\,(\overline{\pi}_C - \overline{\pi}_D)$. From Section 3.2.1 we know that $\overline{\pi}_C = (r + (1-r)\,p)\,R + (1-r)\,(1-p)\,S$ and $\overline{\pi}_D = (r + (1-r)\,(1-p))\,P + (1-r)\,pT$. In order to evaluate the effect of a change in r, holding p fixed, we take the first derivative with respect to r to find that

$$\begin{aligned} \frac{\partial \dot{p}}{\partial r} &= p(1-p)\left(\frac{\partial \overline{\pi}_C}{\partial r} - \frac{\partial \overline{\pi}_D}{\partial r}\right) \\ &= p(1-p)\left[(1-p)(R-S) + p(T-P)\right] \end{aligned} \tag{5.1}$$

When this quantity is positive, an increase in r implies an increase in the rate of increase, or a decrease in the rate of decrease, of cooperators.[9] This is the main comparative static of interest with regard to cooperation, and it allows us to define the first, strongest sense in which increased relatedness might "favour" cooperation:

Definition 1 *Increased relatedness favours cooperation in the first sense if* $\frac{\partial \dot{p}(p,r)}{\partial r} \geq 0$ *for all* (p, r).

It is immediately clear that for all 2-player games with $R \geq S$ and $T \geq P$ – which includes all prisoners' dilemmas – increased relatedness favours cooperation in the strongest sense. It is also clear that if $R < S$, or if $T < P$, there will be frequencies p for which $\frac{\partial \dot{p}(p,r)}{\partial r} < 0$; they are low frequencies p if $R < S$ and high frequencies p if $T < P$.

A second definition would be useful in the case where the dynamics always result in convergence to a pure state comprising only cooperators or only defectors – except when starting on the boundary between the plus- and the minus-region. If an increase in r is to favour cooperation, then the basin of attraction of the cooperative outcome should expand as we increase r. This would follow if the proportions of cooperators at unstable fixed points were decreasing in r. These fixed points typically represent polymorphisms, with both cooperators and defectors present.

Definition 2 *Increased relatedness favours cooperation in the second sense if, writing* $p^*(r)$ *as the locally unstable equilibrium proportion of cooperators for a given* r, $p^*(r)$ *is non-increasing in* r.

[9] It should be noted that the results are not dependent on the linearity of the replicator dynamics; they generalize to any dynamics with the form $\dot{p} = F(p, \overline{\pi}_C - \overline{\pi}) = F(p, (1-p)[\overline{\pi}_C - \overline{\pi}_D])$ where the partial derivative $F_2 > 0$, since then

$$\frac{\partial \dot{p}}{\partial r} = (1-p)F_2(p, (1-p)[\overline{\pi}_C - \overline{\pi}_D])\left(\frac{\partial \overline{\pi}_C}{\partial r} - \frac{\partial \overline{\pi}_D}{\partial r}\right),$$

which is of the same sign as $\frac{\partial \overline{\pi}_C}{\partial r} - \frac{\partial \overline{\pi}_D}{\partial r}$.

Finally, in instances where there is a mixed equilibrium comprising both cooperators and defectors for some r, a third definition is useful. Relatedness favouring cooperation could then mean that the proportion of cooperators in equilibrium is higher for higher r, where the equilibrium typically contains defectors as well as cooperators.

Definition 3 *Increased relatedness favours cooperation in the third sense if, writing $p^*(r)$ as the locally stable equilibrium proportion of cooperators for a given r, $p^*(r)$ is non-decreasing in r.*

It can be shown (see Appendix B) that if increased relatedness favours cooperation under the first definition, then it necessarily does so under the second and third definitions too. In this sense, Definitions 2 and 3 are *weaker* than Definition 1. Appendix B also shows that for every r, there is at most one polymorphic equilibrium, which justifies defining $p^*(r)$ as *the* share of cooperators at that equilibrium.

We will go over the games that feature in the previous section in a systematic way. In all those cases, the line that separates the "plus-region" from the "minus-region" will be important. That line, the isocline where $\overline{\pi}_C = \overline{\pi}_D$, gives those values of (p, r) for which $\dot{p} = 0$, and hence p is a fixed point of the dynamics.

5.1.1 Prisoners' dilemmas

Prisoners' dilemmas are defined by the ordering of payoffs $T > R > P > S$. The first and the third inequality represent that the individual always gains from defecting. The second represents the inefficiency of mutual defection relative to mutual cooperation. As a result, we have $R > S$ and $T > P$, so, as mentioned already in Section 5.1, higher relatedness always favours cooperation under the stronger definition of increasing the growth rate of cooperators relative to defectors. This is a particularly strong result: in precisely those games where cooperation is best defined and most studied, increasing the degree of relatedness promotes cooperation under our strongest definition, with or without equal gains from switching.

Although the first definition is the strongest, and implies the other two, it is still worth confirming that the second and third do indeed hold in the respective cases to which they apply. That is illustrated in Figure 13 (the corresponding calculations are found in Appendix B). In Fig. 13a the intersection of the constant-r arc and the isocline separates the basins of attraction of full defection on the left and full cooperation on the right. As r goes up, and we move to ever lower constant-r arcs, the intersection moves more to the left, which increases the size of the basin of attraction of cooperation. In Fig. 13c, the stable fixed point of the dynamics is a point in the interior of the simplex. As r goes up, we again go to ever lower constant-r arcs, but now the intersection moves more to the right, where the equilibrium proportion of cooperators is higher.

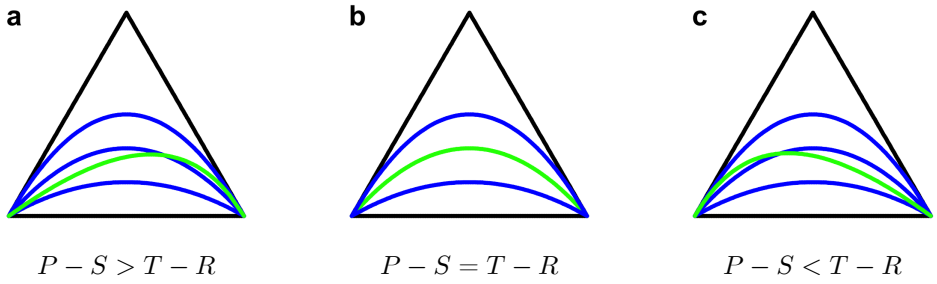


Figure 13. Prisoners' dilemmas. Phase planes with sample constant-r trajectories for the prisoners' dilemma. In each case, the green line represents the isocline $\overline{\pi}_C = \overline{\pi}_D$ on which the proportion of cooperators is stationary. In the region above the green threshold the proportion of cooperators is decreasing; in the region below it the proportion of cooperators is increasing. The blue arcs are constant-r arcs. In the first case the intersection of a constant-r arc and the isocline separates the basins of attraction of full cooperation ($p = 1$) and full defection ($p = 0$) (a). The middle case has equal gains from switching, and no intersections if $r \neq \frac{c}{b}$ (b). In the third case, the intersection of a constant-r arc and the isocline is a stable and attracting fixed point (c). In each case, increasing relatedness favours cooperation under the strongest definition.

5.1.2 Stag hunt games

Stag hunt, or coordination games are defined by the inequalities $R > P$, $P > S$ and $R > T$. For consistency, we again call the strategy which yields R when mutually played the cooperative strategy. The difference with prisoners' dilemmas is that there $T > R$, which, in combination with the other inequalities, implies that playing D rather than C always came with higher payoffs. In stag hunt games the best response to D is D, as it is in the prisoners' dilemma, but the best response to C is C, as $R > T$. Here it is again useful to distinguish three cases: (a) $R > T \geq P > S$, (b) $R > P > T > S$, and (c) $R > P > S \geq T$.

In case (a), it follows immediately that increased relatedness favours cooperation in the strongest sense, because $R > S$ and $T \geq P$. In cases (b) and (c), increased relatedness only increases the growth rate of cooperation if the frequency of cooperators p is below a maximum level $\frac{R-S}{R+P-S-T}$. Cooperation is therefore not favoured by increased relatedness under the strongest definition. However, the fact that selection for cooperation is not everywhere increased by an increase in r here only implies that there is a region where selection for cooperation is slower (right of the rightmost grey lines in Fig. 14b and c). The more important effect of an increase in r is that it increases the basin of attraction of the cooperative equilibrium, which implies that relatedness does favour cooperation by the second definition. Calculations are in Appendix B.

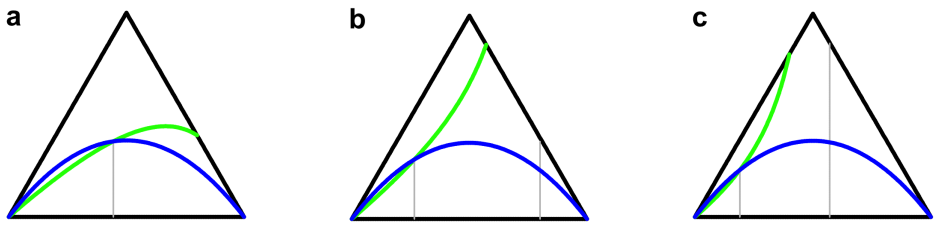


Figure 14. Stag hunt games. Phase planes for stag hunt, or coordination games. In case (a), increasing relatedness favours cooperation under the strongest definition. In cases (b) and (c), increasing relatedness does not favour cooperation under the strongest definition. The rightmost grey line reflects states with $p = \frac{R-S}{R+P-S-T}$. To the right of it, increasing relatedness slows down the growth rate of cooperation. Increasing relatedness however does favour cooperation under the weaker second definition. An increase in r implies that the blue and the green line intersect at a point with a lower fraction of cooperators, and since here the intersection is an unstable mixed equilibrium, this implies that the basin of attraction of full cooperation expands.

5.1.3 (General) hawk dove games

General hawk dove, or snowdrift games are defined by the inequalities $T > R$, $R > P$ and $S > P$. For consistency, we call the strategy which yields R when mutually played the cooperative strategy, although that is not as unambiguous a label as with prisoners' dilemmas. It is useful to distinguish between three cases: (a) $S \geq T > R > P$, (b) $T > S \geq R > P$, and (c) $T > R > S > P$, where the third corresponds to the usual snowdrift game.

In case (c), the usual snowdrift game, it follows immediately that increased relatedness favours cooperation in the strongest sense, because, as in the case of the prisoners' dilemma, $R > S$ and $T > P$. In cases (a) and (b), increased relatedness only increases the growth rate of cooperation if the frequency of cooperators p is above a minimum level $\frac{S-R}{S+T-R-P}$. Cooperation is therefore not favoured by increased relatedness under the strongest definition. However, the fact that selection for cooperation is not everywhere increased by an increase in r only implies that there is a region where selection for cooperation is slower (the region to the left of the left grey lines in Figure 15a and b). The more important effect of an increase in r is that it shifts the mixed equilibrium in favour of cooperators, which implies that relatedness favours cooperation by the third definition. Calculations are in Appendix B.

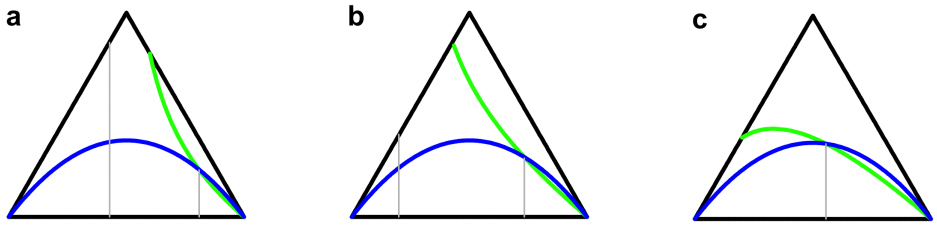


Figure 15. Hawk dove games. Phase planes for the general hawk dove game. In cases (a) and (b), increasing relatedness does not favour cooperation under the strongest definition. The leftmost grey line reflects states with $p = \frac{S-R}{S+T-R-P}$. To the left of it, increasing relatedness slows down the growth rate of cooperation. Increasing relatedness however does favour cooperation under the weaker third definition. An increase in r implies that the blue and the green line intersect at a point with a higher fraction of cooperators, and since here the intersection is a stable mixed equilibrium, this implies that there are more cooperators in equilibrium. In case (c), the usual snowdrift game, increasing relatedness favours cooperation under the strongest definition.

5.2 Comparative statics on efficiency (still with 2 players)

So far, we have simply labeled one strategy as cooperative and the other as defecting, based on the fact that one strategy, when played against itself, yields more than the other does, when played against itself. We included a broad spectrum of games that one could consider to be cooperative dilemmas. Within this broad set, we have shown that, at least under the weaker two definitions, increased relatedness always favours cooperators. For some of the games that are included, however, having more of the strategy that is labeled as cooperative does not always imply higher average payoffs. For some of the general snowdrift games, for instance, cooperators increase the average payoff when rare, but not when abundant. When abundant, playing the strategy we labeled "cooperate" is therefore not necessarily the cooperative thing to do.

A way around this problem is to ask whether relatedness also favours efficiency – where we take efficiency to be the average payoff across the population: $\overline{\pi} = p\overline{\pi}_C + (1-p)\overline{\pi}_D$. Though our discussion will be more brief than that around cooperation, it will again be useful to distinguish between three definitions of 'favouring' efficiency, the first being the strongest, implying the second and third.

Definition 4 *Increased relatedness favours efficiency in the first sense if* $\frac{\partial \overline{\pi}(p,r)}{\partial r} \geq 0$ *for all* (p, r).

Analysis of the system under this strong definition is somewhat intractable. We will therefore turn to the two weaker definitions in what follows.

If we expect convergence to pure states, the basin of convergence (in terms of p) for the efficient outcome should increase with r. Since the fully cooperative pure state is, by definition, more efficient than the fully defecting pure state, this would follow from favouring cooperation under the earlier second definition.

Definition 5 *Increased relatedness favours cooperation (and efficiency) in the second sense if, writing $p^*(r)$ as the locally unstable equilibrium proportion of cooperators for a given r, $p^*(r)$ is non-increasing in r.*

In situations where we expect a stable mixed equilibrium, relatedness favours efficiency if the average payoff at equilibrium increases with r.

Definition 6 *Increased relatedness favours efficiency in the third sense if, writing $p^*(r)$ as the locally stable equilibrium proportion of cooperators for a given r, $\overline{\pi}\left(p^*(r), r\right)$ is nondecreasing in r.*

5.2.1 Prisoners' dilemmas

In the previous section we saw that in case (a) $P - S > T - R$, we have, for some r, an unstable mixed equilibrium, and so the second definition is appropriate. It was shown in Section 5.1.1 that, in this case, increased relatedness favours cooperation under the second definition, so it favours efficiency under the second definition too.

In the other case, (c) $P - S < T - R$, for certain r, there is a stable mixed equilibrium, and so the second definition is applicable. At the stable equilibrium the average payoff of both cooperators and defectors are equal to each other – $\overline{\pi}_C = \overline{\pi}_D$ – and therefore also equal to the overall average payoff:

$$\overline{\pi} = p\overline{\pi}_C + (1-p)\overline{\pi}_D = p\left[\overline{\pi}_C - \overline{\pi}_D\right] + \overline{\pi}_D = \overline{\pi}_D = \overline{\pi}_C$$

The frequency p at the intersection of the isocline and a constant-r arc is found by taking the equation of the isocline – $\overline{\pi}_C = \overline{\pi}_D$ – and isolating p. This way we find $p^*(r) = \frac{S-R+(R-P)/(1-r)}{(T-R)-(P-S)}$. The equilibrium is on the isocline, so if we substitute this for p either in $\overline{\pi}_C = (r + (1-r)\,p)\,R + (1-r)\,(1-p)\,S$ or in $\overline{\pi}_D = (r + (1-r)\,(1-p))\,P + (1-r)\,pT$, we find the average payoff as a function of T, R, P and S, and of r. Either way we find

$$\overline{\pi} = \overline{\pi}_C = \overline{\pi}_D = \frac{r\,(R-S)\,(T-P) - PR + ST}{(T-R) - (P-S)}$$

This is increasing in r, since $R > S$ and $T > P$ in the prisoners' dilemma (the denominator is positive, since $P - S < T - R$). Hence, in this case, increased relatedness favours efficiency under the third definition.

5.2.2 Stag hunt games

In stag hunt, or coordination games, for large enough r, we have an unstable mixed stationary point: starting off the isocline, the population converges either to the fully cooperative or fully defective outcome. Thus, the second definition of favouring is appropriate. We showed in the previous section that, in these games, increased relatedness favours cooperation under the second definition, and so increased relatedness favours efficiency under the second definition as well.

5.2.3 General hawk dove games

In the hawk dove, or snowdrift game, it was shown that, for sufficiently low r, we have a stable mixed equilibrium (for higher r, we always have convergence to the fully cooperative outcome). Thus, the second definition is apposite. Since, in the mixed equilibrium, $\overline{\pi}_D = \overline{\pi}_C$, we may use the same simplification as above in writing $\overline{\pi} = \overline{\pi}_C = \overline{\pi}_D$. On the isocline, again, we have $p^*(r) = \frac{S-R+(R-P)/(1-r)}{(T-R)+(S-P)}$, and $\overline{\pi}_C$ simplifies as in the previous subsection. So, $\frac{\partial \overline{\pi}}{\partial r} = \frac{(R-S)(T-P)}{(T-R)+(S-P)}$. Since $(T-R)+(S-P) > 0$ and $T > P$ in the general snowdrift game, this expression is positive if and only if $R > S$. So, increased relatedness favours efficiency in the third sense if and only if $R > S$ (case (c) in figure 15); for $R < S$, the equilibrium outcome becomes *less* efficient as we increase r. The intuition for this negative result is clear: in these 'unusual snowdrift' games, both off-diagonal payoffs, S and T, are greater than the diagonal payoffs, R and P. Since the effect of increasing r is precisely to increase the instances of diagonal payoffs relative to off-diagonal payoffs, in these games, the effect would be to decrease efficiency.

5.3 Comparative statics with more than 2 players

With more than 2 players things are a bit more complicated. If games are played between n players, a fully general rule that summarizes when cooperators are selected for is going to have to feature $n-1$ parameters for population structure. Using only one parameter for population structure – such as r – opens up the possibility of counterintuitive findings. Even in games where it is unambiguous what the cooperative thing to do is, it can still be that one population structure has a higher relatedness r while the other favours cooperation and efficiency more. So here it is certainly not true that a higher r is always good for cooperation (this is also noted, in a somewhat more complicated setting, in Matessi & Karlin, 1984, 1986).

The possibility of such paradoxical findings exists, because we are looking for a totally general claim, which is to allow for *all* combinations of *all* possible fitness effects on the one hand, and *all* possible population structures on the other hand. We can however impose restrictions on either ingredient. One example, as we have seen in Section 3.3, is to restrict the set of games, or fitness effects, to those that have generalized equal gains from switching.

If we do, then no matter which population structure we choose from the set of all population structures, all that matters is relatedness r. This implies that for a given game that has generalized equal gains from switching, the comparative statics are that an increase in r fosters cooperation.

Another possibility is that we restrict the set of population structures. For a game that does not have equal gains from switching, it is of course still possible to only consider a restricted set of population structures. It could for instance be that a number of species are all the same in almost all respects, and only differ along a single dimension, for instance in the number of matings that a queen has. That implies that within this set of species, not all population structures are possible, and we are only looking at those that are attained by varying the number of matings per queen, for instance, or whatever it is that the one dimension represents. If this particular subset of all possible population structures is combined with a game that does not have equal gains from switching, then it is very well possible that within this subset, population structures with a higher r favour cooperation. This is illustrated in the next two figures.

The game in the figures is the same as in Section 3.3; $\pi_{i,C} = -1$ if $i = 1$ or 2, $\pi_{i,C} = 1$ if $i = 3$, and $\pi_{i,D} = 0$ for all i. There we rewrote the criterion $\overline{\pi}_C > \overline{\pi}_D$ as $\frac{f_3}{p} > \frac{1}{2}$. In Figure 16, the green surface represents that threshold; it separates all population states where $\overline{\pi}_C > \overline{\pi}_D$, and cooperators win, from all population states where $\overline{\pi}_C < \overline{\pi}_D$, and defectors win. The blue surface are all states with relatedness $\frac{1}{4}$. Clearly the dissimilarity in shape implies that one can find points with $r < \frac{1}{4}$ where cooperators nonetheless are selected for, and points with $r > \frac{1}{4}$ where defectors are selected for (see Fig. 16).

Figure 17 shows that within a *restricted* set of population structures, the comparative statics may still hold. As an example we take $f_0 = (1-\alpha)(1-p)^3 + \alpha(1-p)$, $f_1 = (1-\alpha)p(1-p)^2$, $f_2 = (1-\alpha)p^2(1-p)$ and $f_3 = (1-\alpha)p^3 + \alpha p$. For this one-parameter set of population structures, relatedness equals α. The basins of attraction of cooperation and defection meet at frequency $p = \sqrt{\frac{1-2\alpha}{2-2\alpha}}$, which is decreasing in α – with $0 \leq \alpha \leq 1$. The basin of attraction of cooperation therefore increases with relatedness. Many empirical studies may fit such restricted sets of population structures.

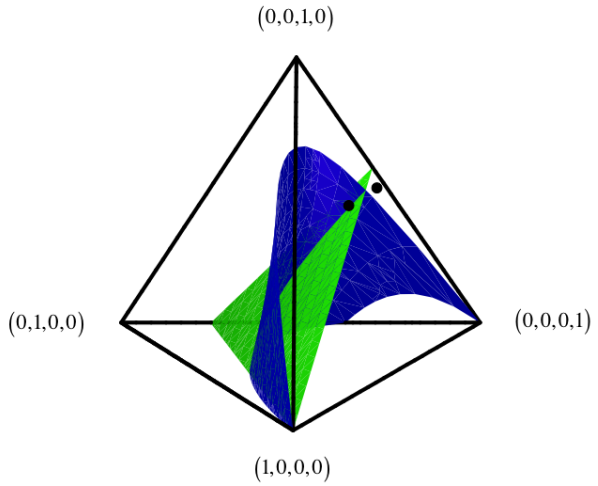


Figure 16. Comparative statics in general do not apply. The vertex closest to us is $f_0 = 1$, the rightmost vertex is $f_3 = 1$, the leftmost vertex is $f_1 = 1$ and the top vertex is $f_2 = 1$. The surface that separates the plus-region from the minus region (green) does not have the same shape as a fixed-r surface (blue; $r = \frac{1}{4}$). This implies that we can find two population states, one with $r < \frac{1}{4}$ where cooperation is nonetheless selected, and one with $r > \frac{1}{4}$ where defection is nonetheless selected.

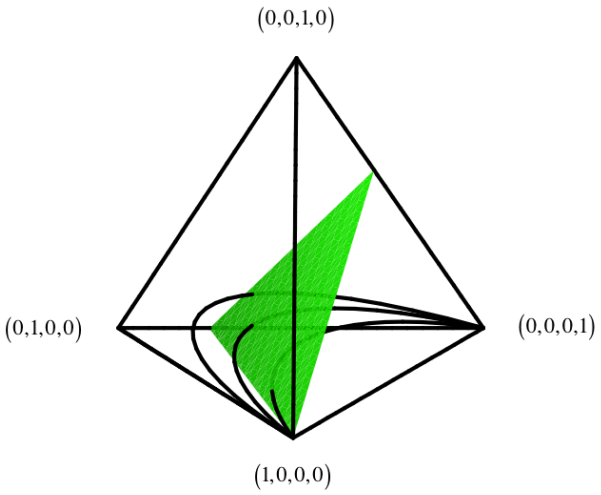


Figure 17. Comparative statics in a restricted set of population structures. The black lines reflect three specific population structures; one with $r = 0$ (the most outward), one with $r = 0.25$ and one with $r = 0.5$ (the most inward). The green surface separates the basins of attraction of full defection at $f_0 = 1$ and full cooperation at $f_3 = 1$, for those three population structures. The figure shows that the basin of attraction of cooperation is larger for higher r.

6 Adaptive dynamics

In this section we consider adaptive dynamics in structured populations. For fitness functions that exhibit "equal gains from switching" globally, Hamilton's rule matches the direction of selection at any point in time along the trajectory. Fitness functions that are differentiable will exhibit equal gains from switching locally, and for those that do not have bifurcations, the same will be true. With bifurcations, Hamilton's rule matches the direction of selection up to the bifurcation. We furthermore generalize the canonical equation from Allen (2013) to non-differentiable payoff functions. For those, Hamilton's rule does not apply. For some fitness functions, moreover, the assumption of monomorphic populations that adaptive dynamics makes is hard to justify.

Rather than assuming that there are two types to begin with – cooperators and defectors – one could assume instead that there is a whole continuum of possible levels of cooperation. Moreover, one could assume that at any point in time the population is close to being monomorphous. We get a stylized version of being close to monomorphous if we assume that selection is much faster than mutation. Mutations then either go to fixation or go extinct before the next mutation arises, so at any point in time there are at most two strategies present: an incumbent, and a (more recent) mutant. One can furthermore assume that mutations are typically local; any mutation is most likely taking only a very small step on this continuum.

Although the profusion of possible strategies implies an enormous scope for deviations from equal gains from switching at the global level, the assumption of local mutations can bring us back to a setting where *locally* equal gains from switching is restored. If this is the case, inclusive fitness will describe the success or failure of a succession of mutants, and one can easily imagine a dynamics that keeps moving up to the point where inclusive fitness is maximized. There are however also exceptions, as we will see below.

A continuous trait space and local mutations are assumptions that feature in many inclusive fitness papers (some examples are Taylor, 1989, Taylor & Frank, 1996, Rousset & Billiard, 2000, Roze & Rousset, 2004, and Lehmann, 2012). They are also the basic assumptions in adaptive dynamics (Metz et al., 1996, Dieckmann & Law, 1996, Geritz et al., 1998, Champagnat et al., 2001, 2006, 2007, Dercole and Rinaldi, 2008). Besides sharing a basic setup, there are also differences between this part of the inclusive fitness literature and adaptive dynamics. The adaptive dynamics literature typically assumes a well-mixed population, and focusses on non-social traits. The inclusive fitness papers are about traits that do have fitness effects on others, and typically do not assume a well-mixed population. More recently, some authors have introduced population structure in adaptive dynamics (Champagnat & Méléard, 2007, Allen et al., 2013). This is a nice cross-over, and it turns out that this approach is also very instructive in describing how inclusive fitness works, and what its limitations are.

The actual results pertaining to adaptive dynamics with population structure are mostly

in Appendix C. Here in the main text we will apply them to a few instructive examples. The first two show that inclusive fitness can describe the evolution of a trait value in a variety of cases. In the third example inclusive fitness no longer works, but one version of adaptive dynamics remains a relatively accurate description of dynamics with reasonable parameter choices. The fourth example shows that adaptive dynamics can also cease to be a good description of evolution altogether. The examples in Section 6.6 illustrate limitations of inclusive fitness as pointed out by Doebeli & Hauert (2006), based on Doebeli, Hauert & Killingback (2004).

6.1 Four games with continuous trait space

Even for discussing some instructive examples, it is unavoidable to introduce some notation. Traits will be values in $\mathbb{R}$. The first examples reflect interactions between 2 players that both exhibit a trait in $\mathbb{R}$. We assume, as always, that the game is symmetric. In a 2-player game this means that, if $\pi_j(x, y)$ was to denote the payoff to player j if player 1 plays x and player 2 plays y, then $\pi_2(x, y) = \pi_1(y, x)$. This implies that it is in fact redundant to have vector valued payoff function; it is sufficient to have a simple payoff function $\pi : \mathbb{R}^2 \to \mathbb{R}$ which describes how much player 1 gets as a function of what player 1 and 2 do. How much player 2 gets immediately follows from interchanging the variables.

Later we will also consider interactions between n players that all have traits in $\mathbb{R}$. Again, a payoff function $\pi : \mathbb{R}^n \to \mathbb{R}$ is sufficient to describe a symmetric game; the payoff of player j in the action profile $(x_1, ..., x_n)$ is the value of payoff function π evaluated in the action profile where x_1 and x_j have swapped places. Symmetry does, however, impose a restriction on the payoff function π now, and that is that $\pi(x_1, x_2, ..., x_n) = \pi(x_1, p(x_2, ..., x_n))$ for all permutations p.

The first payoff function generates a prisoners' dilemma with equal gains from switching for every combination of two different strategies in $\left[0, \frac{a}{2}\right]$, with $a > 0$.

$$\pi(x, y) = ay - x^2 \tag{6.1}$$

This function is called *additively separable*, which means that there are functions $b(y)$ and $c(x)$ such that $\pi(x, y) = b(y) - c(x)$. Additive separability guarantees that all possible restrictions to 2 by 2 matrix games exhibit equal gains from switching. If a player changes from playing x to playing $x + \delta$, then the opponent gains $b(x + \delta) - b(x)$, while that change implies a loss of $c(x + \delta) - c(x)$ to the player itself, *regardless of the action of its opponent.* This independence of the action of the opponent is what defines equal gains from switching. Fig. 18 illustrates that property. We get this picture by choosing different values for x and y and plotting the payoffs that they result in. The dotted lines keep the action of one player fixed and vary the actions of the other continuously, while the solid lines depict a succession of bimatrix games.

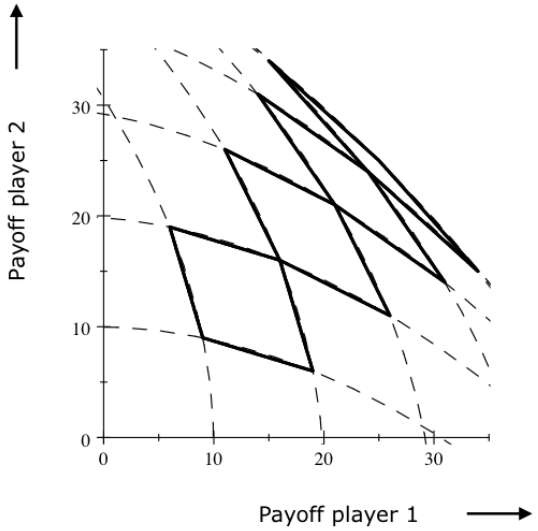


Figure 18 The dashed lines give combinations of payoffs that are attained by keeping one player's trait value fixed (at $1, 2, 3, 4$ and 5, respectively) and continuously varying the trait value of the other, all with payoff function $\pi(x, y) = 10y - x^2$. With this payoff function, any two given trait values constitute a bimatrix game with equal gains from switching. The solid lines represent four such games; with trait values $\{1, 2\}, \{2, 3\}, \{3, 4\}$ and $\{4, 5\}$, respectively. The figure, as well as the formula, shows that an increase in cooperation gets ever more expensive as the trait value increases.

The second example is a slight variation; for every combination of two strategies in $\left[0, \frac{a}{2}\right)$ we still get a prisoners' dilemma.

$$\pi(x, y) = ay - xy \tag{6.2}$$

Again we assume that $a > 0$. This function is not additively separable and it generates a sequence of prisoners' dilemmas that does not have equal gains from switching. However, for ever smaller mutations we get ever closer to a game that does have equal gains from switching. In other words, in the limit of weak selection in phenotype space (or δ-weak selection; see Wild & Traulsen, 2007), we do arrive at a game with equal gains from switching. This is visible if we take the limit for $\delta \to 0$ of the appropriately rescaled payoff matrix that comes with resident t and mutant $t + \delta$:

$$\lim_{\delta \to 0} \frac{1}{\delta} \left(\begin{bmatrix} at - t^2 & a(t+\delta) - t(t+\delta) \\ at - (t+\delta)t & a(t+\delta) - (t+\delta)^2 \end{bmatrix} - \left(at - t^2\right) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right) =$$

$$\lim_{\delta \to 0} \frac{1}{\delta} \begin{bmatrix} 0 & a\delta - t\delta \\ -\delta t & a\delta - \left(2t\delta + \delta^2\right) \end{bmatrix} = \lim_{\delta \to 0} \begin{bmatrix} 0 & a - t \\ -t & a - 2t - \delta \end{bmatrix} = \begin{bmatrix} 0 & a - t \\ -t & a - 2t \end{bmatrix}$$

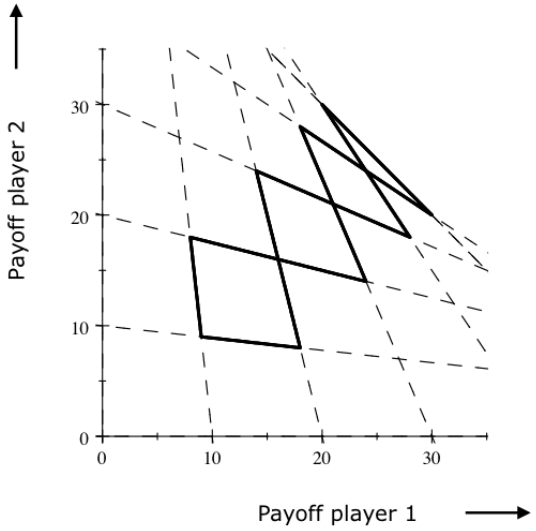


Figure 19. With payoff function $\pi(x, y) = 10y - xy$ the bimatrix games with trait values $\{1, 2\}, \{2, 3\}, \{3, 4\}$ and $\{4, 5\}$ are games with unequal gains from switching. In the limit of δ-weak selection, however, one can say that it does have equal gains from switching.

The third example is a 2-player version of the minimum effort game (Van Huyck, Battalio & Beil, 1990). This is a well known game in economics, and the only difference with the standard minimum effort game is that here we have quadratic costs.

$$\pi(x, y) = a \min(x, y) - x^2 \tag{6.3}$$

Again we assume that $a > 0$. For every combination of two different strategies in $\left[0, \frac{a}{2}\right]$ this gives us a stag hunt (or coordination) game[10], which by definition does not have equal gains from switching. This remains true, even in the limit of δ-weak selection (weak selection in phenotype space).

This function is obviously not additively separable. The quadratic costs again ensure that whatever the population structure, there will always be a point where increases in costs inhibit the evolution of ever higher values of the trait.

[10] The matrix is: $\begin{bmatrix} ax - x^2 & ax - x^2 \\ ax - (x+\delta)^2 & a(x+\delta) - (x+\delta)^2 \end{bmatrix}$. It is a stag hunt game for any $\{x, x+\delta\}$ for which $ax - x^2 < a(x+\delta) - (x+\delta)^2$ or $2x\delta + \delta^2 < a\delta$ or $2x + \delta > a$. This is certainly true if both x and $x + \delta$ are smaller than $\frac{a}{2}$.

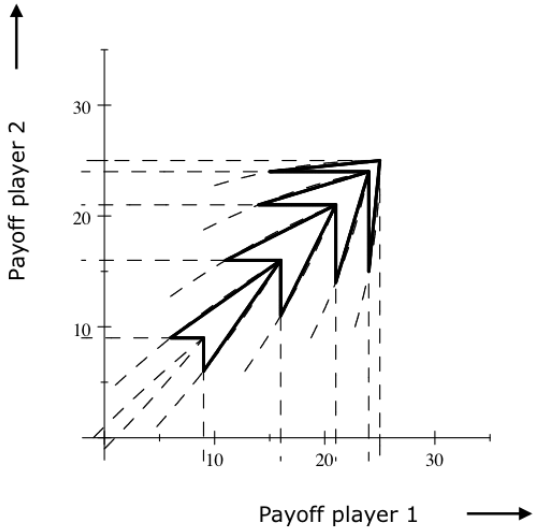


Figure 20. With payoff function $\pi(x,y) = 10\min(x,y) - x^2$, the bimatrix games with trait values $\{1,2\},\{2,3\},\{3,4\}$ and $\{4,5\}$ do not have equal gains from switching.

The fourth example could be dubbed a maximum effort game.

$$\pi(x,y) = a\max(x,y) - x^2 \tag{6.4}$$

In this game it is sufficient for the production of the benefit if only one individual contributes. This implies that if the other would not contribute at all, one would be better off choosing a positive trait value rather than 0, but between the two players, both would obviously prefer the other to be the one that does the contributing.

Again we assume that $a > 0$, and for every combination of two different strategies in $\left[0, \frac{a}{2}\right]$ this gives us a hawk-dove game[11], which by definition does not have equal gains from switching. This remains true, even in the limit of weak selection in phenotype space (δ-weak selection).

This function is obviously not separable either. The quadratic costs again ensure that whatever the population structure is, there will always be a point where increases in costs inhibit the evolution of ever higher values of the trait.

Straightforward n-player versions of these four games are: the n-player linear public goods game, with $\pi(x_1,...,x_n) = a\sum_{i=2}^{n} x_i - (n-1)(x_1)^2$, an n-player non-linear public goods game,

[11] The matrix is: $\begin{bmatrix} ax - x^2 & a(x+\delta) - x^2 \\ a(x+\delta) - (x+\delta)^2 & a(x+\delta) - (x+\delta)^2 \end{bmatrix}$. Because with $\delta > 0$ it is always the case that $a(x+\delta) - x^2 > a(x+\delta) - (x+\delta)^2$, this game is a hawk-dove game for any $\{x, x+\delta\}$ for which $ax - x^2 < a(x+\delta) - (x+\delta)^2$ or $2x\delta + \delta^2 < a\delta$ or $2x + \delta < a$. This is certainly true if both x and $x+\delta$ are smaller than $\frac{a}{2}$.

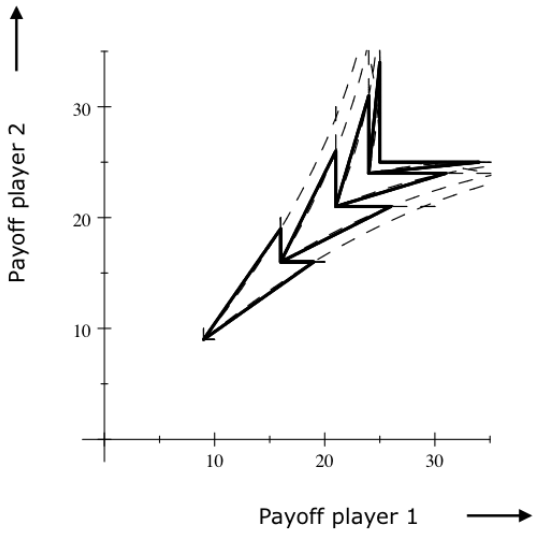


Figure 21. With payoff function $\pi(x,y) = 10\max(x,y) - x^2$ the bimatrix games with strategies $\{1,2\}, \{2,3\}, \{3,4\}$ and $\{4,5\}$ do not have equal gains from switching.

with $\pi(x_1,...,x_n) = a\sum_{i=2}^{n} x_i - \prod_{i=1}^{n} x_i$, the n-player minimum effort game, with $\pi(x_1,...,x_n) = a\min\{x_1,...,x_n\} - (x_1)^2$, and the n-player maximum effort game, with payoff function $\pi(x_1,...,x_n) = a\max\{x_1,...,x_n\} - (x_1)^2$.

6.1.1 Hamilton's rule with a continuous trait space

If we want to depict Hamilton's rule in a similar figure, we can first go back to the original setup, where the problem is described as an individual choice problem. For any given status quo, and from the viewpoint of player 1, Hamilton's rule defines a straight line with slope $-\frac{1}{r}$ through the fitnesses that belong to the status quo. If π_1 represents the fitness, or payoff, of the agent, and π_2 the fitness of it's interaction partner, then the line is given by the equation $\pi_1 + r\pi_2 = K + r$. This can be rewritten as $\pi_2 = \frac{K+r-\pi_1}{r}$, which makes π_2 is a function of $\pi_1$with slope $-\frac{1}{r}$. Inclusive fitness $rb - c$ remains constant on that line, and it separates mutants that are selected for (right/up from the line, with positive inclusive fitness) from mutants that are not (left/down from it, with negative inclusive fitness). This is depicted in Fig. 22a, for a status quo with fitness 1 and relatedness $r = 0.5$. If a mutant fixes, then we have a new status quo, because every individual is now both making the transfer and receiving it. This new status quo is given in Fig. 22b, where the mirror image of the original situation is also drawn, because that is what the original situation looks like from the perspective of player 2. Through the new status quo, there is of course a new line that separates further mutants that would, and mutants that would not be selected for. In adaptive dynamics, the status quo changes all the time, so we can fill $\mathbb{R}^2$ with those separator lines (Fig. 22c). Fig 22d includes those from the viewpoint of player 2.

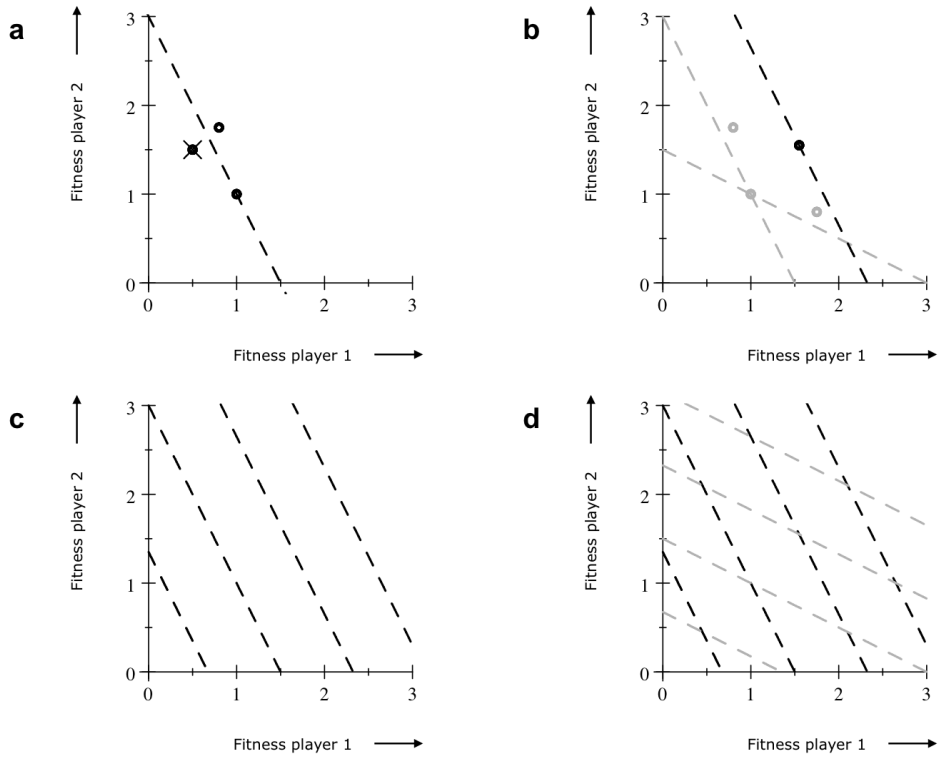


Figure 22. Hamilton's rule for different choices. A line with equal inclusive fitness through the fitnesses in the status quo separates mutations with an advantage from mutations with a disadvantage (a). The mutation with a disadvantage is the one with a cross in the figure. After an advantageous mutation has gone to fixation, the status quo has changed, and we have a new line separating advantageous from disadvantageous mutants. In grey the previous lines, the previous status quo, and mutant fitnesses, both from the perspectives of player 1 and 2 (b). Lines of equal inclusive fitness from the viewpoint of player 1 (c) and from the viewpoint of both players (d).

Of course not all combinations of payoffs or fitnesses are feasible; Hamilton's rule only tells us which mutants, if they were to appear, would be selected for, and which not. What is and what is not feasible, is given by the fitness function. In order to illustrate what happens with a given relatedness r and a given payoff function π, we will therefore superimpose the "Hamilton's rule picture" for that value of r over the game-figure for that π. The combined figure will then illustrate up to what point we should expect the trait value to increase (see for instance Fig. 23a, 24a and 27).

6.2 Game 1: global equal gains

With the first payoff function, Hamilton's rule should work perfectly well in describing the dynamics. This payoff function is additively separable, and therefore *any* two types play a game with equal gains from switching. That implies that we would not even have to assume that the typical step size of mutations is small for the dynamics to converge to the point where $rb = c$. If we do assume small mutations, we expect the dynamics to slowly approach this point from below, or from above, depending on which side of the rest point it happens to have started.

The game payoffs for the first example were given by $\pi(x, y) = ay - x^2$. If a player changes from playing x to playing $x + \delta$, then the opponent gains $a\delta$, and the player itself loses $(x + \delta)^2 - x^2 = 2\delta x + \delta^2$, which is approximately $2\delta x$ for small δ. Inclusive fitness is therefore positive if $x < \frac{ra}{2}$ and negative if $x > \frac{ra}{2}$. At $x = \frac{ra}{2}$ the dynamics should not be expected to move any further, and the (individual) payoffs of all individuals in the population are $\frac{ra^2}{2} - \left(\frac{ra}{2}\right)^2$. In Fig. 23, $a = 10$ and $r = 0.5$, so the equilibrium value of x is 2.5, and the payoff there is $\frac{75}{4}$. The outcome of a simulation with only somewhat rare, local mutations (Fig. 23b) indeed matches what we would expect (Fig 23a; lines drawn from player 1's perspective only).

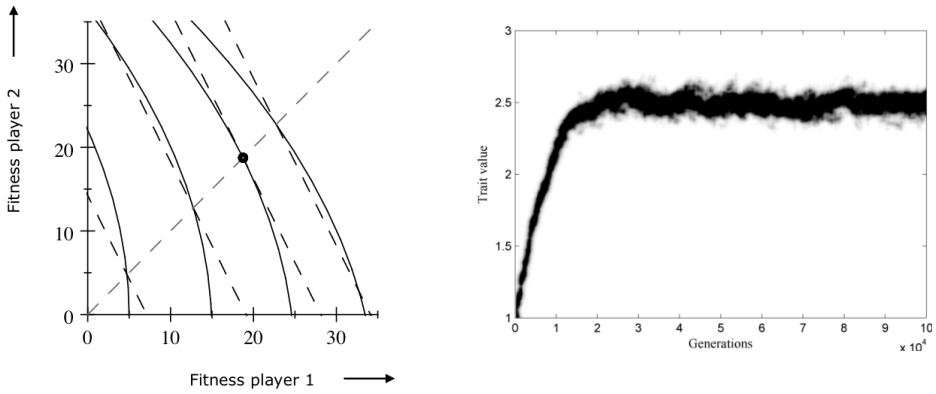


Figure 23. With $\pi(x, y) = 10y - x^2$ and $r = 0.5$ inclusive fitness is maximized at trait value $x = \frac{5}{2}$. The corresponding payoffs are $\pi\left(\frac{5}{2}, \frac{5}{2}\right) = 18\frac{3}{4}$. The solid lines in the left picture depict payoffs for player 1 and 2 for a fixed trait value of player 2 and a varying trait value of player 1, as in Fig 18–21. Here the fixed trait values for player 2 are $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$ and $\frac{7}{2}$, respectively. The broken lines reflect Hamilton's rule, from the perspective of player 1, as explained in Fig. 22. Simulations on the right indeed show an increase in trait value to $x = 2.5$. The simulations use a Wright-Fisher process with relatedness r, as described in Appendix C1.

Adaptive dynamics are typically described with a differential equation that is referred to as the canonical equation. For a setting with population structure, which is what we have

here, Allen et al. (2013) arrived at the following canonical equation.

$$\dot{x} = N_e \frac{N-1}{N} \frac{u\epsilon^2}{\pi(x,y)} \left(\left. \frac{\partial \pi(x',x)}{\partial x'} \right|_{x'=x} + r \left. \frac{\partial \pi(x,x')}{\partial x'} \right|_{x'=x} \right) \tag{6.5}$$

This is equation (5) from Allen et al. (2013), with a few variables relabeled (see Appendix C; see also Champagnat & Méléard, 2007, Champagnat & Lambert, 2007, and Lehmann, 2012). The population size is N, N_e is the 'effective population size', u is the individual mutation probability when producing an offspring, and ϵ is the standard deviation of the distribution from which the step size of the mutation is drawn. What the effective population size is depends on the reproduction process. In the standard Wright-Fisher process, a new generation is created by choosing N new individuals independently, and at every draw every individual in the parent generation has a probability proportional to their payoffs of being drawn as a parent. For this process the effective population size N_e equals N. If one single individual is chosen to produce the entire next generation, and every individual has a probability proportional to their payoffs of being that one individual, then the effective population size N_e equals 1. If the payoffs just do not matter at all for the probabilities with which individuals reproduce, then the effective population size is 0. More details about effective population size can be found in Allen et al. (2013) and in Appendix C. For the simulations, we use a version of the Wright-Fisher process that allows for positive relatedness. This process is also used in García & Van Veelen (2016) and Van Veelen et al. (2012), and is described in Appendix C1. For this process the effective population size N_e equals $\frac{N}{1+r}$.

Arriving at the canonical equation (6.5) involves three steps. The first is that we imagine the following hypothetical process. Suppose that mutations arise at a rate Nu. The step size of a mutation is drawn from some distribution with expected value 0 and standard deviation ϵ. If a mutant does arise, then instantly it is determined whether it goes to fixation or goes extinct. The probability with which it fixes is taken to be the fixation probability that we would get for the actual reproduction process, given the size of the mutation that is drawn. One difference with more detailed and less stylized processes such as the one that we use for the simulations is that there time is discrete, so that mutations can only arise at times $1, 2, ...$, when new generations are formed, while here they can arrive at any moment $t > 0$. Also, the uncertainty concerning whether or not the mutation fixes is not resolved immediately in the simulations, but in however long it takes the mutant to fix or go extinct.

The second step is that the expected change in trait value in one generation is computed in the limit of $\Delta t \downarrow 0$ and $\epsilon \downarrow 0$. This expected change is proportional to u times ϵ^2, so with small u and small ϵ this is going to be a very small number.

The third step is that we consider a deterministic approximation, where the time derivative of the trait value is set equal to its expected change.

Whether or not the canonical equation describes actual processes of evolution well depends on how innocuous the assumption of instantaneous resolution of the uncertainty is, and how much is lost in step three, where we go to a deterministic approximation. For both steps the payoff function matters, and also how small u and ϵ really are can make a difference. The process that treats the dynamics as if the uncertainty about the fate of a mutant is resolved immediately is only a good representation if we can be relatively sure that the fate of one mutant is decided before the next one arises. For this to be the case, we need to have a sufficiently low mutation probability u, so that it is very unlikely that a next mutant arises before the previous one has either gone to fixation or gone extinct. What sufficiently low is also depends on the payoff function π.

In this first example, we see that the population is certainly not always monomorphous, implying that mutants typically do not fix or go extinct before the next one arises. The speed with which the average trait value in the population moves is therefore somewhat different from the speed that the canonical equation would give (the simulations move a bit slower). The direction of selection, however, matches the adaptive dynamics very well. We can therefore conclude that the adaptive dynamics describe the stochastic dynamics relatively well, even with not so small mutation rates, and that inclusive fitness does determine the direction of selection and the rest point of the dynamics.

6.3 Game 2: local equal gains

The game payoffs for the second example were given by $\pi(x,y) = ay - xy$. This game is not additively separable, but if we assume small mutations, the game between resident and mutant is very close to having equal gains from switching. If a player changes from playing x to playing $x+\delta$, then the opponent gains $(a-y)\,\delta$, and the player itself loses $y\delta$. This is evaluated at $y = x$, so inclusive fitness is positive if $x < r(a-x)$ and negative if $x > r(a-x)$. At $x = \frac{ra}{1+r}$ the dynamics do not move any further, and the (individual) payoffs of both are $\frac{ra^2}{1+r} - \left(\frac{ra}{1+r}\right)^2$. The outcome of a simulation with rare, local mutations (Fig. 24b) again matches this prediction (Fig 24a; lines from player 1's perspective only).

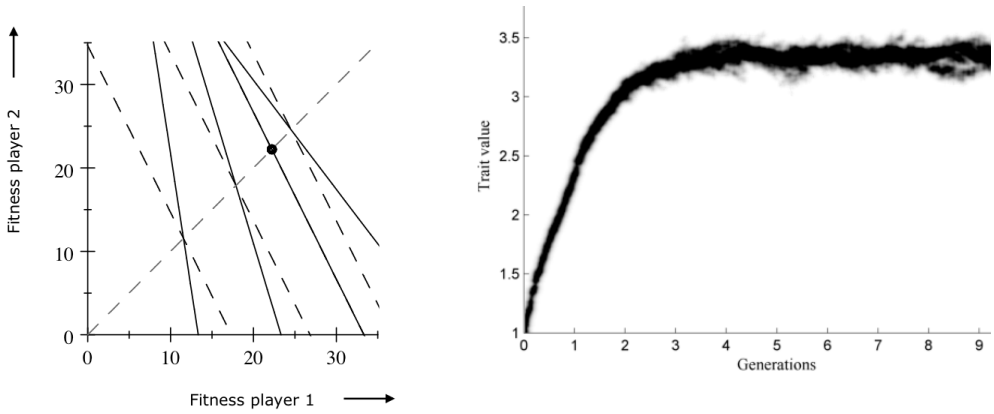


Figure 24. With $\pi(x,y) = 10y - xy$ and $r = 0.5$ inclusive fitness is maximized at trait value $x = \frac{10}{3}$. The corresponding payoffs are $\pi\left(\frac{10}{3}, \frac{10}{3}\right) = 22\frac{2}{9}$. The fixed trait values for player 2 in the left figure are $\frac{4}{3}$, $\frac{7}{3}$, $\frac{10}{3}$ and $\frac{13}{3}$, respectively. Simulations on the right indeed show an increase in trait value to $x = \frac{10}{3}$.

It is not only for this particular game that we recover equal gains from switching locally. This is the case for all differentiable payoff functions. If we take the limit for $\delta \to 0$ of the appropriately rescaled payoff matrix that comes with resident t and mutant $t+\delta$ for general differentiable payoff functions – as we did for this particular game above, where game 2 was introduced – then we find

$$\lim_{\delta\to 0} \frac{1}{\delta} \begin{bmatrix} \pi(t,t) - \pi(t,t) & \pi(t,t+\delta) - \pi(t,t) \\ \pi(t+\delta,t) - \pi(t,t) & \pi(t+\delta,t+\delta) - \pi(t,t) \end{bmatrix} = \begin{bmatrix} 0 & \frac{d\pi}{dy} \\ \frac{d\pi}{dx} & \frac{d\pi}{dx} + \frac{d\pi}{dy} \end{bmatrix}$$

In other words, locally the benefits-to-costs ratio is well defined; we can think of $\frac{d\pi}{dy}$ as the benefit to the other of a small increase in my trait value, and of $\frac{d\pi}{dx}$ as its costs to me. Not just for this example, but with differentiable payoff functions in general, we expect the trait value to go up if $rb - c > 0$, with $b = \frac{d\pi}{dy}$ and $c = \frac{d\pi}{dx}$, as is also reflected in the canonical equation. The link between differentiability and (local) additivity was pointed out early

on in the literature; see for instance Taylor, 1988, p. 151–152, Taylor, 1989, p. 140, and Rousset, 2004, p. 95.

That does not clear us from all problems, even with differentiable payoff functions, as we will see below in Game 5. But it does imply that there is a much larger set of payoff functions, on top of the additively separable ones, for which inclusive fitness also works, if we assume local and small mutations. Additively separable functions have equal gains from switching built in there right from the beginning, while differentiable ones get equal gains in the limit of very small mutations.

6.4 Game 3: minimum effort

The third payoff function is not differentiable; for $\pi(x,y) = a\min(x,y) - x^2$ the derivative does not exist at the point where it is most needed, which is at $x = y$. But differentiable or not, there are of course still dynamics to be studied. Determining the dynamics here is complicated by the fact that if an increase in trait value is favoured, a decrease is no longer automatically disfavoured, and vice versa. Also the lack of differentiability implies that however small we choose the mutation size δ, the game between resident and mutant never has equal gains from switching. As we will see, this implies that the intuition from Hamilton (1964) that worked in Game 1 and Game 2 no longer works here. The key is in the fact that the loss of equal gains from switching implies that it is no longer possible to assume that any effect I have on others is mirrored by their effect on me.

The game between resident x and mutant $x+\delta$ is given by the payoff matrix below. We assume that $\delta > 0$ in order to make the mutant a proper increase in trait value.

$$\begin{array}{ccc} & x & x+\delta \\ x & ax - x^2 & ax - x^2 \\ x+\delta & ax - (x+\delta)^2 & a(x+\delta) - (x+\delta)^2 \end{array}$$

With relatedness r and a frequency of the mutant that is approaching 0, the average payoff to the resident x is simply $\pi(x,x) = ax - x^2$. The average payoff to the mutant when rare is $r\pi(x+\delta, x+\delta) + (1-r)\pi(x+\delta, x) = ax + ra\delta - (x+\delta)^2$. Therefore the mutant $x+\delta$ can invade if $ra\delta - 2x\delta - \delta^2 > 0$. Assuming that δ is sufficiently small, this boils down to $x < \frac{ra}{2}$.

The game between resident x and mutant $x-\delta$ is given by the next payoff matrix – where the mutant now represents a proper decrease in trait value.

$$\begin{array}{ccc} & x & x-\delta \\ x & ax - x^2 & a(x-\delta) - x^2 \\ x-\delta & a(x-\delta) - (x-\delta)^2 & a(x-\delta) - (x-\delta)^2 \end{array}$$

With relatedness r and a frequency of the mutant that is approaching 0, the average payoff to the resident x is still $\pi(x,x) = ax - x^2$. The average payoff to the mutant when rare is $r\pi(x-\delta, x-\delta) + (1-r)\pi(x-\delta, x) = a(x-\delta) - (x-\delta)^2$. The mutant $x-\delta$ therefore can invade if $-a\delta + 2x\delta - \delta^2 > 0$, that is, if $x > \frac{a}{2}$, assuming that δ is sufficiently small. (We would get the same answer if we considered the first matrix, and check how x does when $x+\delta$ is the resident. The version with $x-\delta$ for $x+\delta$ we thought may be a bit more intuitive though).

Taking those two thresholds together, we find that for values of x between $\frac{ra}{2}$ and $\frac{a}{2}$, both an increase and a decrease in trait value is disfavoured. When we draw a pairwise invasion plot – which is a traditional way to visualize this in the adaptive dynamics literature; see

for instance Brännström, Johansson, & von Festenberg, (2013) – then any x between those bounds is suggested to be stable (see Fig. 25a).

A possible alternative approach might take into consideration that some disadvantageous mutations are more disadvantageous than others. Even though for any $x \in \left(\frac{ra}{2}, \frac{a}{2}\right)$ both increases and decreases are disfavoured, it is still possible that mutants with an increased trait value are disfavoured more (or disfavoured less) than mutants with an equally large change, but in the opposite direction. With sufficient time, it could therefore be that however sticky a trait value, it might still be more likely to be replaced by a mutant with a higher trait value than it is to be replaced by a mutant with a lower trait value – or vice versa.

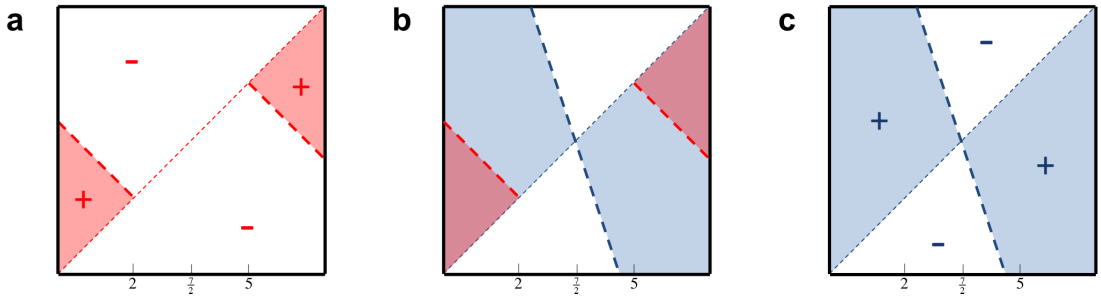


Figure 25. Pairwise invasion plot for $\pi(x, y) = 10 \min(x, y) - x^2$ with $r = 2/5$. It describes, given a trait value of the incumbent (the variable on the horizontal axis), whether a trait value of a mutant (on the vertical axis) would give that mutant an advantage or a disadvantage. For x between 2 and 5 both increases and decreases are disadvantageous (a). One can also use the σ-result from Tarnita et al. (2009) to determine which mutant is favoured. There, the two balance at trait value $x = 3.5$ (c). Both criteria are combined in the middle panel (b).

In order to determine which direction is *more* likely, or less unlikely, it seems natural to use the σ-result from Tarnita et al. (2009). This result does not (just) look at whether or not a mutant has an advantage or a disadvantage when rare. The σ-result gives a criterion that indicates, for two strategies, which one has the larger fixation probability when appearing as a single mutant in a population where the other is the incumbent. It therefore by definition does have the property that if one has the smaller fixation probability, the other has the larger one, and vice versa. The σ-result assumes weak selection in payoff contribution (which is sometimes called w-weak selection) but also works with small mutation size δ (a.k.a. δ-weak selection; see Wild & Traulsen, 2007, for the difference). We do have to assume that π is continuous, which it is here, but not that it is differentiable.

With strategies A and B and our simple population structure with assortment parameter r, the fixation probability of a single A mutant is larger than the fixation probability of a single B mutant if and only if $\frac{1+r}{1-r}\pi(A, A) + \pi(A, B) > \pi(B, A) + \frac{1+r}{1-r}\pi(B, B)$ – see

Appendix C6. If we now take trait values x and $x+\delta$, then the second is favoured when $x < \frac{(1+r)a}{4}$ and δ is sufficiently small.[12] With $a = 10$ and $r = 2/5$ this threshold value is 3.5 (see Fig. 25c).

The two different possibilities for what to expect from the dynamics, as depicted in Fig. 25, can also be described somewhat more formally. The first approach would depend on the fact that increasing the population size exaggerates even the smallest disadvantages that a mutant may have. If we keep the distribution of mutants, and therefore the ϵ constant, as well as mutation probability u, then increasing the population size would make mutants with even a very small disadvantage have a fixation probability that is ever closer to 0 compared to $1/N$ (or, in other words, it would make N times the fixation probability approach 0). Choosing a very large population would then create a marked separation between the middle part of the trait space, where $x \in [2,5]$, and the two other parts, with $x < 2$ or $x > 5$. In the middle part, the speed with which the trait moves would be much lower than in the other two, and this difference in relative speed could be made ever more pronounced by increasing the population size (see Appendix C9).

The second option generalizes the canonical equation from Allen (2013) to non-differentiable payoff functions. This approach keeps the population size constant, and implies choosing a small u and taking a limit of $\Delta t \downarrow 0$ and $\epsilon \downarrow 0$. This is what we do in Appendix C, where we arrive at the same canonical equation, but now with the symmetric derivative replacing the normal derivative.

$$\dot{x} = N_e \frac{N-1}{N} \frac{u\epsilon^2}{\pi(x,y)} \left(\frac{\partial_s \pi(x',x)}{\partial_s x'}\bigg|_{x'=x} + r \, \frac{\partial_s \pi(x,x')}{\partial_s x'}\bigg|_{x'=x} \right) \tag{6.6}$$

The symmetric derivative is the average of the left- and right derivative, both of which exist in our case, and is defined as:

$$\frac{\partial_s f(x)}{\partial_s x} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x-\epsilon)}{2\epsilon}.$$

Simulations show that the generic case is relatively well described by the latter approach. For mutation probabilities that are not actually very small, and a population size that is not very large, the trait evolves to $x = 3.5$, without much noticeable change in speed at $x = 2$. The third panel from Fig. 25, with the pairwise invasion plot based on the σ-result,

[12]

$$\begin{aligned}
\frac{1+r}{1-r}\left(a(x+\delta) - (x+\delta)^2\right) + ax - (x+\delta)^2 > ax - x^2 + \frac{1+r}{1-r}\left(ax - x^2\right) \Leftrightarrow \\
\frac{1+r}{1-r}\left(a\delta - 2\delta x - \delta^2\right) - 2\delta x - \delta^2 > 0 \Leftrightarrow \\
(1+r)(a - 2x - \delta) - (1-r)(2x+\delta) > 0 \Leftrightarrow \\
x < \frac{(1+r)a}{4} - (1-r)\delta
\end{aligned}$$

therefore describes the direction of selection best, even if the mutation probability is not actually very small. In order to observe a difference between the speed for $x < 2$ and $x > 2$, we have to choose a very large population size.

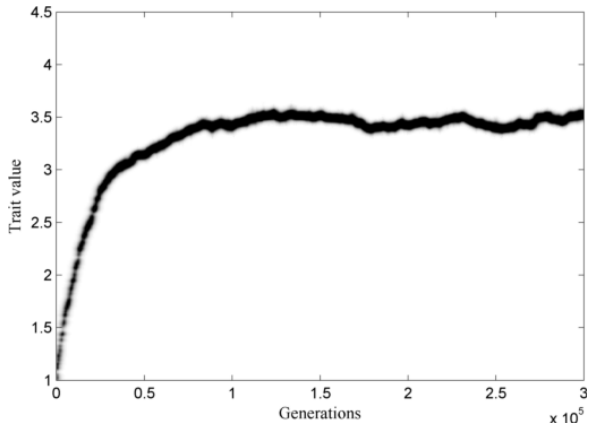


Figure 26. With $\pi(x, y) = 10 \min(x, y) - x^2$ and $r = 2/5$ the simulations with take the population to a trait value of around 3.5. This is the trait value where Tarnita's σ-result applied to this game suggests that neither increases nor decreases in trait value have a selective advantage. The corresponding payoffs are $\pi\left(\frac{7}{2}, \frac{7}{2}\right) = 22\frac{3}{4}$.

The simple inclusive fitness intuition that worked for Games 1 and 2 turns out not to work for Game 3. If we draw a figure similar to Fig. 23 (for Game 1) and Fig. 24 (for Game 2), with the effects of the changes in trait value on self and on the other, then we get Fig. 27. The idea "*I weigh the effect I have on myself with 1 and the effect on the other with relatedness r*" is represented by the dotted lines, which did serve us before to separate the mutants with an advantage from those with a disadvantage. This picture now suggests that at trait values $1, 2, 3$ and 4 both increases and decreases in trait value reduce inclusive fitness, as all the solid lines are below their respective broken counterparts. The catch is that at the diagonal, all games between mutant and incumbent, however small δ is, are coordination games. This figure keeps the strategy of the other player constant. With equal gains that is not a problem, because the effects then do not depend on what the other is. Without equal gains, however, keeping the strategy of the other player constant is not inconsequential. For those games we have seen in Section 3 that inclusive fitness does not work. For the mutant, playing against a copy of itself really is quite different from playing against the incumbent – and this difference would not have been there if the payoff function π had been differentiable.

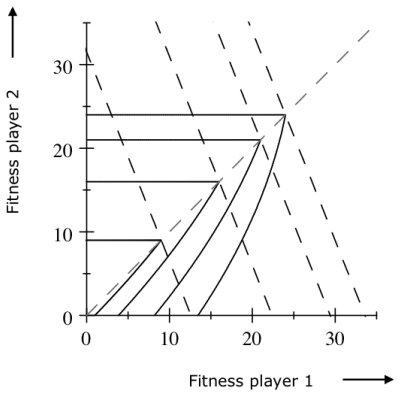


Figure 27. The solid lines are payoff combinations for a fixed trait value of player 2, and a varying trait value of player 1, for $\pi(x, y) = 10 \min(x, y) - x^2$. When the trait value of player 1 is below the trait value of player 2, increases in trait value by player 1 increase the minimum. This always increases the payoff of player 2, and when the derivative of x^2 is not too high, the payoff of player 1 increases as well, but less so. This is reflected in the upward sloping part. When the trait value of player 1 is above the trait value of player 2, then increases in trait value by player 1 do not increase the minimum. Hence, the payoff to player 2 remains the same, while the payoff to player 1, which includes the x^2 term, decreases. This is reflected in the horizontal part. The dotted lines reflect points with equal inclusive fitness for $r = 2/5$. If we follow the logic of Fig. 23 and 24, then for trait values of $1, 2, 3$ and 4 both increases and decreases in trait value would be at a disadvantage. The analysis of the dynamics shows that this is not correct; for $y = 1$ an increase is actually favoured, and for $y = 2, 3$ and 4 the dynamics also go up.

6.5 Game 4: maximum effort

The fourth payoff function is also not differentiable; for $\pi(x,y) = a\max(x,y) - x^2$ the derivative does not exist, again, at $x = y$. Again there are complications for the dynamics, although somewhat different ones.

The game between resident x and mutant $x+\delta$ is given by the payoff matrix below, where we still assume $\delta > 0$.

	x	$x+\delta$
x	$ax - x^2$	$a(x+\delta) - x^2$
$x+\delta$	$a(x+\delta) - (x+\delta)^2$	$a(x+\delta) - (x+\delta)^2$

With relatedness r and a frequency of the mutant that is approaching 0, the average payoff to the resident x is simply $\pi(x,x) = ax - x^2$. The average payoff to the mutant when rare is $\pi(x+\delta, x+\delta) = a(x+\delta) - (x+\delta)^2$. The mutant $x+\delta$ therefore can invade if $a\delta - 2x\delta - \delta^2 > 0$, that is, if $x < \frac{a}{2}$, assuming that δ is sufficiently small.

The game between resident x and mutant $x-\delta$ is given by the next payoff matrix – where the mutant now represents a proper decrease in trait value.

	x	$x-\delta$
x	$ax - x^2$	$ax - x^2$
$x-\delta$	$ax - (x-\delta)^2$	$a(x-\delta) - (x-\delta)^2$

With relatedness r and a frequency of the mutant that is approaching 0, the average payoff to the resident x has not changed: it is still $\pi(x,x) = ax - x^2$. The average payoff to the mutant when rare is $r\pi(x-\delta, x-\delta) + (1-r)\pi(x-\delta, x) = ax - ra\delta - (x-\delta)^2$. The mutant $x-\delta$ therefore can invade if $-ra\delta + 2x\delta - \delta^2 > 0$, that is, if $x > \frac{ra}{2}$, assuming that δ is sufficiently small.

Taking those two thresholds together, we find that for values of x between $\frac{ra}{2}$ and $\frac{a}{2}$, both an increase and a decrease in trait value are favoured.

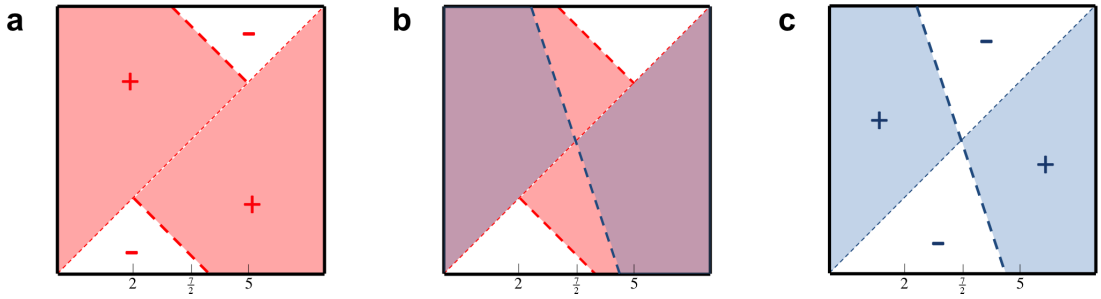


Figure 28. Pairwise invasion plot for $\pi(x,y) = 10\max(x,y) - x^2$ with $r = 2/5$. It describes, given a trait value of the incumbent (the variable on the horizontal axis), whether a trait value of a mutant (on the vertical axis) would give that mutant an advantage or a disadvantage. For x between 2 and 5 both increases and decreases are advantageous (a). One can also use the σ-result from Tarnita et al. (2009) to determine which mutant is favoured. There, the two balance at trait value $x = 3.5$ (c). Both criteria are combined in the middle panel (b).

We can obviously hope to get around this in a way that is similar to the way we did for Game 3. With both increases and decreases in trait value being advantageous, we can again use the σ-result in order to determine which direction is *more* likely. With trait values x and $x+\delta$, increases in trait value are favoured when $x < \frac{(1+r)a}{4}$ and δ sufficiently small.[13] With $a = 10$ and $r = 2/5$ this threshold value is 3.5 (see Fig. 28c).

The trouble with this, though, is that the adaptive dynamics limit must assume mutation rates that are sufficiently small such that previous mutants typically have either gone extinct or gone to fixation before the next mutant arises. With games between mutants and incumbents that are anti-coordination games, this requires mutation rates that are spectacularly low. Moreover, if we are not that close to the limit, the dynamic behaviour is very different, as we can see in the simulations. The starting point of adaptive dynamics – monomorphic populations on the move – therefore turns out to be inappropriate for such a game, where also mixtures are prone to invasions beyond the extremities of the mixture. The canonical equation therefore is not a good approximation of the dynamics..

In Section 6.6 we will see more examples, including differentiable ones, where heterogeneity is to be expected, and where the adaptive dynamics framework requires such extremely

[13]

$$\frac{1+r}{1-r}\left(a(x+\delta)-(x+\delta)^2\right)+a(x+\delta)-(x+\delta)^2 > a(x+\delta)-x^2+\frac{1+r}{1-r}\left(ax-x^2\right) \Leftrightarrow$$

$$\frac{1+r}{1-r}\left(a\delta-2\delta x-\delta^2\right)-2\delta x-\delta^2 > 0 \Leftrightarrow$$

$$(1+r)(a-2x-\delta)-(1-r)(2x+\delta) > 0 \Leftrightarrow$$

$$x < \frac{(1+r)a}{4}-(1-r)\delta$$

low mutation rates that the limit results that the adaptive dynamics framework offers are no match with dynamic behaviour with reasonable mutation rates. These and similar issues are also pointed out by Barton & Polechevá (2005).

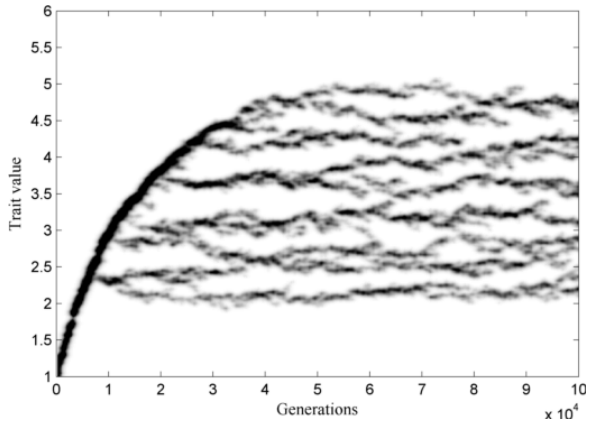


Figure 29. With $\pi(x,y) = 10\max(x,y) - x^2$ and $r = 2/5$ the simulations take the population to distributions of the trait value stretching from 2 to 5. Heterogeneity is the rule for this game; one needs extremely low mutation rates to maintain the assumption of monomorphic populations, even more so than with Game 3. Moreover, when mutation rates are not actually sufficiently low, the dynamics nonetheless still followed what the σ-result in Tarnita et al. (2009) implied with Game 3 (see Figure 28), while here that is not the case.

6.5.1 Life without adaptive dynamics

Game 4 suggests that there are cases in which it is not reasonable to assume that populations are almost always relatively close to being monomorphous. With the point of departure of adaptive dynamics out the door, what else can we do to describe where we should expect evolutionary dynamics will take us? Monomorphous populations are one extreme; the other extreme is a distribution of strategies on a continuum. The alternative approach is therefore to look for stable distributions (see Van Veelen & Spreij, 2009).

A necessary condition for a distribution to be stable is that every strategy in it should earn the same payoff (if that would not be the case, some strategies would be selected for, and hence the distribution could not have been stable in the first place). This requires that the derivative, taken with respect to the trait, must be 0 for every trait value that is in the distribution. Let the probability distribution be given by the function $f : \mathbb{R} \to \mathbb{R}$. The payoff function is defined for interactions between trait values x and y (in this case it is $\pi(x,y) = a\max(x,y) - x^2$), but we would like to extend the definition to include the expected payoff of trait value z against the entire distribution f, and include relatedness r. For this, we write $\pi_r(z,f)$.

Against $x \leq z$, the payoff of z is $az - z^{\beta}$, and against $x > z$ the payoff of z is $ax - z^{\beta}$. The payoff therefore is

$$\begin{aligned}\pi_r(z,f) &= r\pi(z,z) + (1-r)\pi_0(z,f) \\ &= r\left(az - z^2\right) + (1-r)\left[a\int_{-\infty}^{z} zf(x)\,dx + a\int_{z}^{\infty} xf(x)\,dx - z^2\right] \\ &= a\left[rz + (1-r)\left(z\int_{-\infty}^{z} f(x)\,dx + \int_{z}^{\infty} xf(x)\,dx\right)\right] - z^2\end{aligned}$$

This implies that we are looking for a distribution that satisfies

$$\frac{d\pi_r(z,f)}{dz} = a\left(r + (1-r)F(z)\right) - 2z = 0$$

where $F(z) = \int_{-\infty}^{z} f(x)\,dx$.

Because $F(z)$ must lie between 0 and 1, the uniform distribution on the interval $\left[\frac{ar}{2}, \frac{a}{2}\right]$ must be invariant:

$$F(z) = \begin{cases} 1 & \text{if } z \geq \frac{a}{2} \\ \frac{2z-ar}{a(1-r)} & \text{if } \frac{ar}{2} \leq z < \frac{a}{2} \\ 0 & \text{if } z < \frac{ar}{2} \end{cases} \qquad f(z) = \begin{cases} \frac{2}{a(1-r)} & \text{if } \frac{ar}{2} \leq z < \frac{a}{2} \\ 0 & \text{elsewhere} \end{cases}$$

Although the population is of course at no point an actual uniform distribution, this does describe the simulation results much better than the adaptive dynamics does (see Fig. 29).[14]

[14] An analytical stability check could be done by showing global superiority of the uniform distribution on the interval $\left[\frac{ar}{2}, \frac{a}{2}\right]$. Then local superiority is obviously implied in any metric, which is sufficient to imply asymptotic stability in any metric; see Spreij & van Veelen (2009). Here we just rely on the simulations, which suggest that the uniform distribution is not only invariant, but also stable.

6.6 Game 5: bifurcations

There are two good reasons why it is worth looking at another example. The first is that both differentiability and non-differentiability are of course stylized characteristics of a model with a continuous trait space, where we allow for mutations of any size. Mutations are typically not infinitesimally small though, and discussing whether differentiability is a reasonable assumption, or a particularly unreasonable one, is therefore a little artificial from the get go. We can try to circumvent that issue altogether by looking at different (continuous, differentiable) payoff functions that differ gradually as for how strongly the payoff function is bent at $x = y$, and see if the things we see in Game 4 really depend on non-differentiability per se. Within the following family of payoff functions, the "curvature" at the diagonal varies with a parameter β. In the limit of $\beta \to \infty$, the function ceases to be differentiable, as it converges to the payoff function of Game 4. Therefore we can approach Game 4 with a sequence of payoff functions that themselves are all differentiable.

An even better reason is that with this family of payoff functions, we can illustrate how, even if we make all the assumptions that are required for adaptive dynamics to work in examples like Game 1 and Game 2 – differentiable payoff function, local, small and infrequent mutations – we can still end up in a situation where inclusive fitness does not determine entirely what happens in an evolving population. The problem is caused by the fact that Hamilton's rule may help find a "singular point" – a point where $rb = c$, and where a monomorphous population ceases to move – but that once at such a singular point, there are still different possible scenarios, depending on the properties of the payoff function – as pointed out by Ajar (2003), Doebeli & Hauert (2006) and Wakano & Lehmann (2014); see also Doebeli, Hauert & Killingback (2004). One is that the population remains where it is forever. Another is that we see a bifurcation. After a bifurcation, we get two divergent subsets of the population. In that case, despite the fact that we have local and small mutations, and a differentiable payoff function, the population ceases to be monomorphic, and will consist of individuals with increasingly different trait values. In this case we therefore get substantial deviations from equal gains from switching.

The family of payoff functions we consider is $a\left(x^{\beta}+y^{\beta}\right)^{\frac{1}{\beta}}-x^{2}$. For $\beta = 1$, the payoff function simplifies to $a\left(x+y\right)-x^{2}$. For $\beta \to \infty$ the payoff function simplifies to $a\max\left(x,y\right)-x^{2}$. For ease of comparison, we choose $a = 10\left(2^{-\frac{1}{\beta}}\right)$ in all examples below. If we do, then the derivatives are $\frac{d\pi}{dx}\big|_{y=x} = 5-2x$ and $\frac{d\pi}{dy}\Big|_{y=x} = 5$, which implies that the adaptive dynamics would see an increasing trait value for $x < \frac{5}{2}\left(1+r\right)$, regardless of β.

Fig. 30 shows that we do indeed get ever more bifurcations as β increases, and the payoff functions get ever more curved around the diagonal. After the bifurcations, equal gains from switching no longer applies, and both the dynamics and the equilibrium distribution can no longer be described with inclusive fitness.

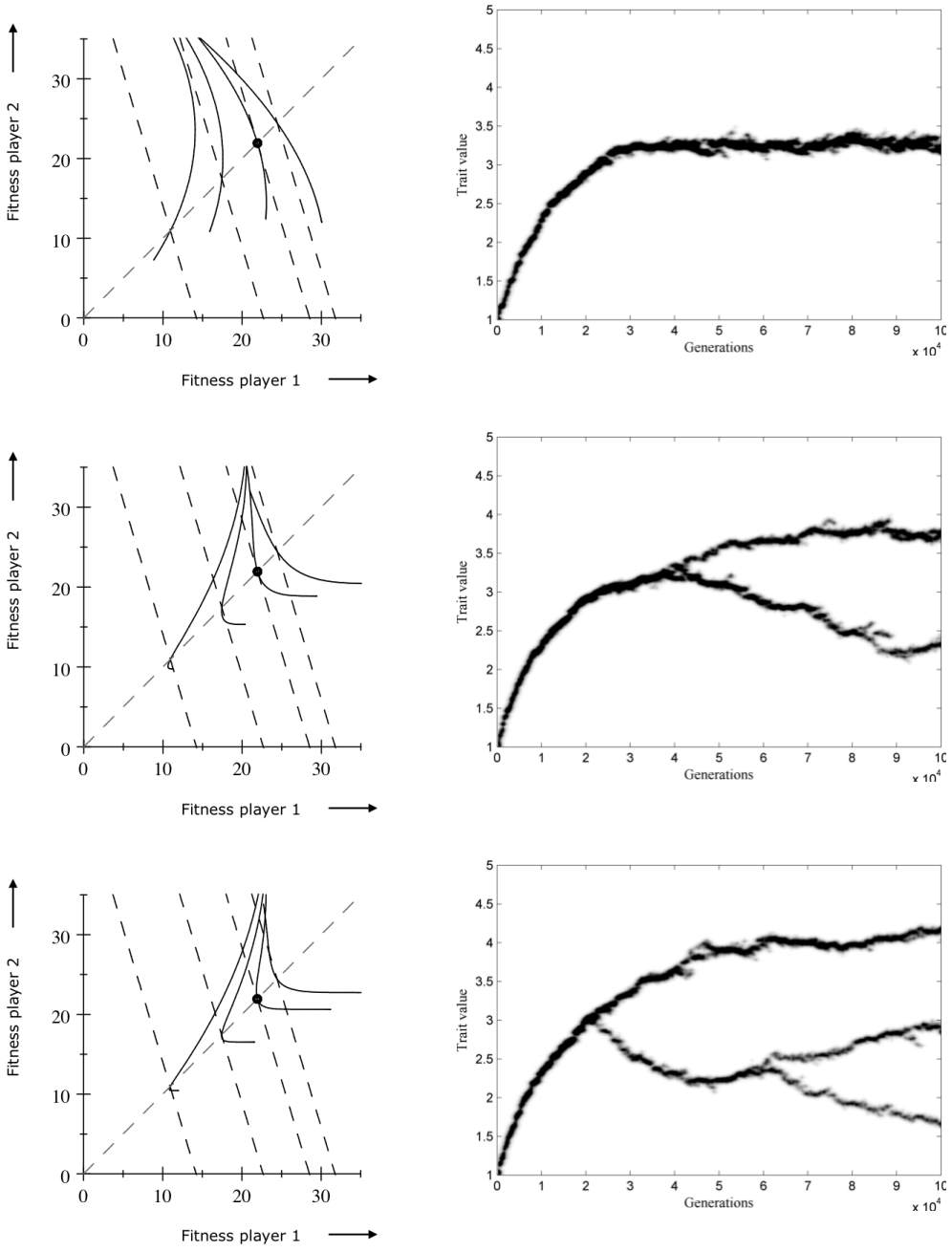


Figure 30. With $\pi(x,y) = 10 \cdot 2^{-\frac{1}{\beta}} \left(x^{\beta} + y^{\beta}\right)^{\frac{1}{\beta}} - x^2$ inclusive fitness is maximized at trait value $x = \frac{5}{2}(1+r)$. The corresponding payoffs are $\pi\left(\frac{5}{2}(1+r), \frac{5}{2}(1+r)\right) = 25(1+r) - \frac{25}{4}(1+r)^2$. These figures all have have $r = 0.3$, and go from $\beta = 2$ (top row), to $\beta = 7$ (middle row), to $\beta = 17$ (bottom row).

6.7 An example of how equal gains is sometimes implicitly assumed

One of the classic references in the inclusive fitness literature is Rousset & Billiard (2000). Given a population structure of n demes, located on a circle, Rousset & Billiard (2000) allow for a structure of effects on the fecundity of others that is very general in one sense. The average trait value in a deme j is allowed to have an effect on fecundity of individuals in all other demes, with the only restriction that these effects have to be symmetric.

> "Migration rates between demes depend only on their relative position, so that the relative contribution of an individual in deme j to gametes competing in deme l may be written as g_{l-j} (where $g_{l-j} = g_{j-l}$)" (page 817).

Our setting, where individuals only have an effect on the individuals they are interacting with, is a special case of Rousset & Billiard (2000) in the sense that effects on fecundity of individuals on demes other than the one an individual is on itself are assumed not to exist. This implies a simplification to the island model, which is also treated by Rousset & Billiard (2000) as a special case (page 820–821).

There is, however, one aspect in which their setting is less general. Both in the general formulation (equations (2) and (3) on page 817) and in the island model (page 820) the arguments of the functions g_r and w are *average* trait values (g_r is a function that returns the relative contribution of an individual to gametes competing in a deme r steps apart on the circle, and w is the fitness function). If we take their island model, where $w \equiv w(z_\bullet, z_0, z_1)$ is the fitness function, and try to match it with Game 2, we see that we cannot, because it is not enough only to know averages. In their function $w \equiv w(z_\bullet, z_0, z_1)$ the variables $z_\bullet, z_0$ and z_1 are the focal individual's own phenotype, the average phenotype on the focal individual's own deme (excluding the focal individual) and the average phenotype of all individuals in other demes. In our Game 2, where the payoff function is $\pi(x, y) = ay - xy$, we choose again $a = 10$. The cases below all have three groups of size 2.

Case 1:	Trait values	$1, 1$	$1, 3$	$1, 3$
	Payoffs	$9, 9$	$27, 7$	$27, 7$

If we take as a focal individual the first individual in group 1, then its fitness is $\frac{9}{86}$. Furthermore $z_\bullet = 1, z_0 = 1$ and $z_1 = 2$. Below, the $z_\bullet, z_0$ and z_1 are the same, but the fitnesses are different; for the focal individual it is $\frac{9}{78}$.

Case 2:	Trait values	$1, 1$	$1, 1$	$3, 3$
	Payoffs	$9, 9$	$9, 9$	$21, 21$

The reason why they are different is that games with this payoff function do not satisfy equal gains from switching. The effect on total payoff of the two individuals with trait value

3 is not the same if they both are matched with an individual with trait value 1 compared to if they are matched with each other.

This particular example can easily be encompassed by reformulating their islands model as a special case of their general model, by replacing the variable z_1, which now is the average over *all* demes other than the one the focal individual is on, by a vector which gives per deme averages, as is possible in their more general setup. This, however, would still not allow a fitness function based on averages to capture everything. If we compare Case 1 with Case 3 below, we find that the deme averages are still the same ($1, 2$ and 2, respectively) but the fitness of the focal individual is still different; now it is $\frac{9}{82}$.

Case 3:	Trait values	$1, 1$	$2, 2$	$2, 2$
	Payoffs	$9, 9$	$16, 16$	$16, 16$

Note that if we want to make a payoff function for which we can in fact derive a fitness function for the island model that depends only on the trait value of the individual and on the average trait values per deme, we are restricted to a set of payoff functions that not only are additively separable (which implies that all games satisfy equal gains from switching) but also has constant costs and benefits of (more) cooperation; only $\pi(x, y) = c(y) - b(x) = \mathbf{c}y - \mathbf{b}x$ works here, where the bold $\mathbf{b}$ and $\mathbf{c}$ are constants. This is not the most interesting case, because the value added of a model with a continuum of trait values lies exactly in the possibility that marginal costs and benefits of cooperation are not constant. If the benefit/cost ratio of an increase in cooperation is the same everywhere, we expect that an increase in cooperation is selected for (or against) everywhere, whereas we are most interested in finding a point up to which increases in the level of cooperation are selected for, but beyond which further increases are not.

Again, differentiability of payoff function π offers an escape here. In the limit of small mutations, and with the fitness function $\pi(x, y)$ differentiable at all points where $x = y$, the games between mutant and incumbent do exhibit equal gains from switching, and payoffs are even locally linear in x and y. One way of encompassing that is to acknowledge that with a differentiable payoff function, the true fitness of a focal individual can be approximated properly with a fitness function w that uses averages, if, of course, trait values are sufficiently close to each other (which we get if we assume few mutants, and small mutations). It is worth observing, though, that equal gains from switching is built in in this model from the get go.

7 Local interaction vs. local competition

In this section we discuss cancellation effects. Wilson, Pollock and Dugatkin (1992) and Taylor (1992a,b) found that cooperation may not evolve in populations with local reproduction, even though that makes relatedness between interacting agents positive. The reason for this is that in such models not only the opportunities for cooperation are local, but competition is local too. We illustrate this by combining the cycle – which is a very simple stylized population structure – with different update rules. This is the most important deepening of our understanding of kin selection since Hamilton (1964). What matters is not so much that interacting agents are related; what matters is that there is a *discrepancy* between how assorted the opportunities for cooperation are, and how assorted competition is. A classic way to generate such a discrepancy is kin recognition.

One of the ways in which interactants could end up being related is by reproduction not being a global affair, but a local one – which it typically is. Assuming that individuals also find their opportunities to cooperate locally, this implies that if two individuals interact, they are close by, and if they are close by, they are more likely to share common ancestry than with individuals that are further apart. Limited dispersal therefore may seem to be a good way to get cooperation to evolve. In Hamilton (1964a, page 10) "population viscosity" is therefore suggested to foster cooperation:

> With many natural populations it must happen that an individual forms the centre of an actual local concentration of his relatives which is due to a general inability or disinclination of the organisms to move far from their places of birth. In such a population, which we may provisionally term "viscous", the present form of selection may apply fairly accurately to genes which affect vagrancy. It follows from the statements of the last paragraph but one that over a range of different species we would expect to find giving-traits commonest and most highly developed in the species with the most viscous populations whereas uninhibited competition should characterize species with the most freely mixing populations.

Whether or not viscosity would have that effect was investigated in a paper by Wilson, Pollock and Dugatkin (1992). With a specific choice of how to generate viscosity, they found that there was no such effect. This was confirmed in a more formal and general way in Taylor (1992a,b), and the reason why the suggested effect was not found, was that limited dispersal also implies that competition is local, and the effect of that was left out of the equation.[15] What is needed to have altruism evolve, is therefore not just that opportunities

[15] Considerations that do include local competition are also found earlier; Hamilton (1971), Boyd (1982), Grafen (1983) and other references in Wilson, Pollock and Dugatkin (1992) and Taylor (1992a,b).

for cooperation are local, and with related individuals, but that the symmetry between local competition and local opportunities for cooperation is broken.

In this section we will illustrate this with a simple example of a local interaction structure: the cycle. The cycle is sufficiently stylized to make for a good, insightful illustration, and also illustrates how different approaches have advantages and limitations of their own. The setup was introduced by Ellingsen (1993), Eshel et al. (1998), Lieberman et al. (2005) and Ohtsuki et al. (2006a,b), and the analysis was repeated, but now in inclusive fitness terms, in Grafen (2007b) and Taylor et al. (2007).

7.1 Different intuitions

There are different ways to form an intuition about how kin selection works. One is the core intuition from Hamilton (1964a,b), which looks at the effects of the behaviour on the actor itself, and at the effects on other individuals. It then considers whether or not those combined effects result in a net plus or a net minus for the gene. Alternatively, since many settings are symmetric in some relevant sense, all behaviours can also be mirrored. That is, instead of considering the effect that I have on, say, my sibling, and include his or her change in fitness in my books, one could also consider the effect that my sibling has on me instead. With equal gains from switching, we can be sure that this is an unambiguous swap, and that the numbers in the overall accounting system do not change. This different way of accounting does however foster a different intuition – where different is neither superior nor inferior – and the second way of balancing the books is typically referred to as "neighbour modulated fitness".

The "neighbour modulated fitness" intuition says that population structure can make cooperation evolve, because it gives those that cooperate with others an increased probability of also receiving cooperation. With population structure, what you are will be informative about what you are likely to face, and if you are a cooperator yourself, you are extra likely to also face a cooperator and get the benefits of being on the receiving end too.

The intuition for the counterbalancing effect of local competition can now be described as a complication within this neighbour modulated fitness idea. With local dispersion, you find not only your possible cooperators close by, but also your competition. And one can easily imagine that the competition in a cluster of cooperators is also extra intense. Those that you are interacting with have an increased probability of being cooperators too – which is good for you, because you will benefit from their cooperation. But those that you are competing with also have an increased probability of being cooperators too, and, more importantly, they have an increased chance of being surrounded by cooperators too, just like you do. Which implies that the competition is also enjoying increased fitness benefits because of the proximity of other cooperators – and that is bad for you.

7.2 Cooperation on the cycle

In order to illustrate this, we will look at three examples. The first example is the Birth Death process on the cycle. Here local competition completely washes out the effect of local opportunities for cooperation. The second example is Death Birth on the cycle, and there the cancelling out is only partial. The last example is a mixture of the two, with a twist that allows for a complete breaking of the symmetry, and no cancelling at all.

Individuals are organized on a circle. They play a game with their two neighbours. Here we will consider a simple prisoners' dilemma with equal gains from switching:

$$\begin{array}{cc} & \begin{array}{cc} D & C \end{array} \\ \begin{array}{c} D \\ C \end{array} & \begin{array}{cc} 0 & b \\ -c & b-c \end{array} \end{array}$$

These payoffs are then translated into scaled payoffs to allow for a parameter that reflects the intensity of selection. In the earlier papers, the scaled payoffs were: $1 - w + w{\cdot}payoffs$, where w is the intensity of selection (Ohtsuki et al., 2006a, Ohtsuki & Nowak, 2006b). In more recent papers $e^{w \cdot payoffs}$ is also used (Gokhale & Traulsen, 2010, Van Veelen & Nowak, 2012). These scaled payoffs determine probabilities in the update step.

In the Birth Death process, the probability with which an individual is chosen to reproduce is proportional to its scaled payoff. In other words, the probability with which any specific individual begets an offspring is its scaled payoff divided by the sum of all scaled payoffs. Then one of the neighbours of the individual that reproduces dies – where both neighbours are chosen with equal probability – and is replaced by the new individual.

In the Death Birth process, first an individual is chosen to die, where all individuals have equal probability to be chosen. The two neighbours of the just vacated spot then compete to put an offspring there, and their chances are again proportional to their scaled payoffs.

In the Shift process (Allen & Nowak, 2012), the probability with which an individual is chosen to reproduce is again proportional to its scaled payoffs. Then either the offspring is placed between the parent and its left neighbour, or between the parent and its right neighbour, both with probability one half. One individual is also chosen to die, where all have equal probability of being chosen. The individuals in between the new offspring and the vacated spot move one position, so that we again have a full circle with no empty spots. The offspring thereby pushes the neighbours away towards the vacated place. If the parent is chosen to die, then the offspring just takes the parent's place.

The first two processes are analysed in Ohtsuki et al. (2006a) and Ohtsuki & Nowak (2006b) using the Moran process. The central measure of expected evolutionary success is the fixation probability of a mutant. There are two criteria that can be used to classify mutants as advantageous or disadvantageous. One can compare the fixation probability of a mutant cooperator in a world of defectors to the fixation probability of a neutral mutant,

which is one over the population size N. Another possibility is to compare the fixation probability of a mutant cooperator in a world of defectors to the fixation probability of a mutant defector in a world of cooperators. For the Death Birth and the Birth Death process on the cycle, Ohtsuki et al. (2006a) find that these criteria give the same results. These results imply that, in the limit of weak selection, cooperation is never favoured in the Birth Death process, and cooperation is favoured in the Death Birth process if $\frac{c}{b} > \frac{N-4}{2N-4}$, which, for large N, comes down to $\frac{1}{2}b > c$.

It is tempting to see Hamilton's rule already in there. Ohtsuki et al. (2006a) did notice the similarity, but it is important to realize, as Grafen (2007b) pointed out, that b and c are only payoff parameters, which in this case cannot be equated to fitness effects. Also relatedness is not $\frac{1}{2}$; one can easily imagine that in a process where the cooperators are always grouped together in one string, and defectors in another, with only two boundaries, relatedness should be larger than $\frac{1}{2}$. For large population size N it should actually get close to 1. The fitness effect of being a cooperator instead of a defector are found by looking at how these payoffs affect reproduction. They are given, along with relatedness for both the Birth Death and the Death Birth process, in Table 1 of Grafen (2007b). Below, this table is reproduced, and an extra row is added to also give the fitness effects in the third process. Relatedness for this process may be different from the relatedness for the Birth Death and the Death Birth process.

Individual	$j-3$	$j-2$	$j-1$	j	$j+1$	$j+2$	$j+3$
Relatedness to j	$\frac{N^2-18N+53}{N^2-1}$	$\frac{N^2-12N+23}{N^2-1}$	$\frac{N^2-6N+5}{N^2-1}$	1	$\frac{N^2-6N+5}{N^2-1}$	$\frac{N^2-12N+23}{N^2-1}$	$\frac{N^2-18N+53}{N^2-1}$
$\approx$	$1-\frac{18}{N}$	$1-\frac{12}{N}$	$1-\frac{6}{N}$	1	$1-\frac{6}{N}$	$1-\frac{12}{N}$	$1-\frac{28}{N}$
Effect on payoff	0	0	$+b$	$-2c$	$+b$	0	0
Fitness effect (BD)	0	$-b/2$	$+b+c$	$-b-2c$	$+b+c$	$-b/2$	0
Fitness effect (DB)	$-b/4$	$+c/2$	$+b/4$	$-c$	$+b/4$	$+c/2$	$-b/4$
Fitness effect (Shift)	$-\frac{2(b-c)}{N}$	$-\frac{2(b-c)}{N}$	$+b-\frac{2(b-c)}{N}$	$-2c-\frac{2(b-c)}{N}$	$+b-\frac{2(b-c)}{N}$	$-\frac{2(b-c)}{N}$	$-\frac{2(b-c)}{N}$

Table 1. These examples illustrate how cancellation effects work. Relatedness is computed in the limit of low mutation; $r_k = \lim_{u \downarrow 0} \frac{q_k - \overline{q}}{1-\overline{q}}$, where q_k is the stationary IBD probability of neighbors at distance k – which depends on mutation probability u – while $\overline{q}$ is the average IBD probability among all pairs (see also Grafen, 2007b, and Allen & Nowak, 2012).

Birth Death on the cycle. In the first case the opportunities for cooperation that one has are with the exact same individuals that one is competing with. In this case cooperation

never evolves – that is, not as long as $c > 0$. Somewhat more formally: if $c > 0$, there is always a population size N such that the fixation probability of a cooperator is larger than the fixation probability of a defector (this actually holds for any intensity of selection; see Proposition 5 in Van Veelen & Nowak, 2012). Inclusive fitness suggests the same; going over all affected neighbours, and weighing the effects on them with relatednesses, Grafen (2007b) finds that the inclusive fitness effect of cooperation is

$$-\frac{N^2-12N+23}{N^2-1}b+\frac{N^2-6N+5}{N^2-1}(2b+2c)-b-2c=\frac{12\left(b-c\left(N-1\right)\right)}{N^2-1}$$

For every $c > 0$ this will be negative from some N onwards. In this case local competition completely cancels out local opportunities for cooperation.

One can think of the effects in this formula as effects a player has on itself and its neighbours. The effect I have on myself is $-b-2c$; I twice lower my chance of being picked for reproduction by c, and I twice lower it by another $b/2$ through adding benefit b to my neighbours. The effect I have on my left and right neighbour is $+b+c$ on both sides; me being a cooperator increases their fitness with b plus half of the decrease in my payoff, which is $2c$). The effect I have on my left and right neighbours twice removed is $-b/2$ on both sides; by cooperating, I lower their chances by $b/2$, as I did to myself, since I am the other neighbour of my neighbour.

These effects can of course also be mirrored. My neighbour being a cooperator would increase my fitness by $b+c$, for the exact same reason why me being a cooperator would increase his or hers. Similarly, the neighbour one further removed being a cooperator hurts me $-b/2$. The latter perfectly captures the cancellation effect. With this type of local interaction, if I am a cooperator, it is quite likely that my neighbour is too. But my neighbour also has an increased opportunity to border with another cooperator on the other side as well, which makes me face increased competition. In this case these two effects cancel out exactly.

Death Birth on the cycle. In the second case the opportunities for cooperation and competition do not overlap anymore; one competes with neighbours twice removed, and interacts for cooperation with direct neighbours. This discrepancy helps avoid full cancellation, although there is still some. In this case cooperation evolves if $c/b > 1/2$. Somewhat more formally: if $c/b > 1/2$, there is always a population size N such that the fixation probability of a cooperator is larger than the fixation probability of a defector (see Proposition 7 in Van Veelen & Nowak, 2012). Inclusive fitness suggests the same:

$$-\frac{N^2-18N+53}{N^2-1}\frac{b}{2}+\frac{N^2-12N+23}{N^2-1}c+\frac{N^2-6N+5}{N^2-1}\frac{b}{2}-c=\frac{6\left(\left(N-4\right)b-\left(2N-4\right)c\right)}{N^2-1}$$

If $c/b > 1/2$ this will be positive from some N onwards.

In this case local competition does not completely cancel out local opportunities for cooperation. I might benefit from my direct neighbours on both sides in competing with my two neighbours twice removed, who both may also benefit from their two direct neighbours, one of which is my neighbour too, and one of which is my neighbour three times removed.

Shift on the cycle. The third example is extreme, in that there is no cancelling at all. The opportunities for cooperation are local – with the direct neighbours – but competition is global, because who reproduces and who dies is not linked, and any increase in aggregate payoffs hurts every individual on the cycle equally. It will come as no surprise that here cooperation can evolve as soon as it implies an efficiency gain, that is, when $b > c$. A standard computation shows that, indeed, if $b > c$ then there is a population size N such that cooperation is favoured from that N onwards[16]. The process is analyzed in much more detail in Allen & Nowak (2012).

The relatednesses for the shift process may not be the same as for the Birth Death and the Death Birth process. A nice property of relatednesses on the cycle, however, is that they by definition add up to 0.[17] This implies that if we take all $-\frac{2(b-c)}{N}$ terms and weigh them by relatedness to the actor, they also add up to 0. Inclusive fitness is therefore positive if

$$r_2 b > c$$

where r_n is the relatedness to the individual $n-1$ spots removed. With $\lim_{N\to\infty} r_2 = 1$, and with $b > c$, this will also be true from some N onwards.

[16] The core ingredient of the computation is the ratio of two probabilities; T_i^+ is the probability of going up one state, from i to $i+1$, and T_i^- is the probability of going down one state, from i to $i-1$, where i is the number of cooperators and $N-i$ the number of defectors (see Lieberman, Hauert & Nowak, 2005, Ohtsuki et al, 2006a, Ohtsuki & Nowak, 2006b, Nowak, 2006, Van Veelen & Nowak, 2012).

$$\frac{T_i^+}{T_i^-} = \frac{1-w+w\left(\left(2i-2\right)b-2ic\right)}{1-w+w2b} = \frac{1-w+2w\left(i\left(b-c\right)-b\right)}{1-w+2wb}$$

The fixation probability of a mutant cooperator is larger than $1/N$ if $\prod_{i=1}^{N-1}\frac{T_i^+}{T_i^-} > 1$. In the limit of weak selection,

$$\prod_{i=1}^{N-1}\frac{T_i^+}{T_i^-} = 1 + w\left(\sum_{i=1}^{N-1}\left(2i\left(b-c\right)-2b\right) - \sum_{i=1}^{N-1}2b\right) =$$
$$= 1 + w\left(N\left(N-1\right)\left(b-c\right) - 4b\left(N-1\right)\right)$$

hence

$$\prod_{i=1}^{N-1}\frac{T_i^+}{T_i^-} > 1 \text{ if } N > \frac{4b}{b-c}$$

[17]

$$\sum_{k=1}^{N} r_k = \lim_{u\downarrow 0}\sum_{k=1}^{N}\frac{q_k - \overline{q}}{1-\overline{q}} = 0$$

See also the caption of Table 1.

8 Preference evolution

In this section we consider evolution of preferences, first in the absence of population structure, and later with possibly positive relatedness. If agents can infer each other's preferences, then these preferences can serve as a commitment device. Depending on the fitness function this can lead to the evolution of altruistic preferences, or spiteful ones, even without population structure. When combined with relatedness, Hamilton's rule still describes the equilibrium outcome.

The first known reference to what later became known as Hamilton's rule was made by J.B.S. Haldane. When asked if he would jump into a river to save a brother from drowning, he is said to have answered that he would to save two brothers, but not one, or to save eight cousins, but not seven. With the hindsight of the cancellation effect, the example of siblings and cousins turns out to be a particularly good one. Identifying siblings, and singling them out for altruism, is a perfect way to break the symmetry between local opportunities for cooperation and local competition. Once past a certain age, siblings don't compete with each other any more intensely than they do with other individuals their own age. Yet, if they can identify each other, they can single each other out for altruism and cooperation, and if they do, the benefits accrue to related individuals, without coming back at the agents through increased competition. Breaking the symmetry using kin recognition therefore works (see also papers on kin recognition in humans by Lieberman, Tooby & Cosmides, 2003, 2007).

There are also reasons why that may not be the end of the story for the model species "humans". We do have a theory of mind, and engage in all kinds of strategic behaviours, given what we know, or think we know, about the world and about each other. That may open the door for complications. In this section we will discuss a few of those complications – which are not necessarily limited to humans, but which are certainly easiest to think of in our own species.

Let's go back to the basic model as proposed by Hamilton. That model begins with a behaviour that comes with a given cost c to the actor and a given benefit b to the recipient. Every individual in every generation is thought of as facing the same choice: to give or not to give, always at the same cost c and always for the same benefit b. For any possible combination of c and b, the model gives a prediction; either the altruistic act is selected for, or it is not.

In reality, individuals may face a variety of choices. Sometimes the altruistic act confers a large benefit on the other at a low cost to the actor, sometimes only a small benefit at high costs, and everything in between. If the individual has no way of identifying the differences between those cases, it will have to make a generic choice, and evolution will select what is best on average. But if an individual can tell different costs and benefits apart, then it can also differentiate, and decide to be cooperative/altruistic in some, and not cooperative/selfish in other cases. If the individual can differentiate perfectly, then the

prediction is that the decision rule that the individual will use will in fact coincide with Hamilton's rule; they will act altruistically only in those cases where $rb > c$ (Theorem 1, Van Veelen, 2006).

The mathematical toolbox in economics provides just the right framework to describe such decision rules. A decision rule, assuming that it satisfies some basic consistency conditions, can be summarized by a *utility function*. With two individuals, a utility function has two arguments: my fitness and your fitness. We will denote those with π_1 and π_2, taking the perspective of player 1. When faced with a bunch of possible combinations of fitness for self and other, the decision rule that goes with a specific utility function is to choose the one that comes with the highest value of the utility function. The utility function that goes with Hamilton's rule is: $u_1(\pi_1, \pi_2) = \pi_1 + r\pi_2$.

Such a decision rule is typically depicted with "indifference curves", or, in other words, lines of equal utility. For the utility function $u_1(\pi_1, \pi_2) = \pi_1 + \frac{1}{2}\pi_2$, some such lines are drawn in Fig. 31.

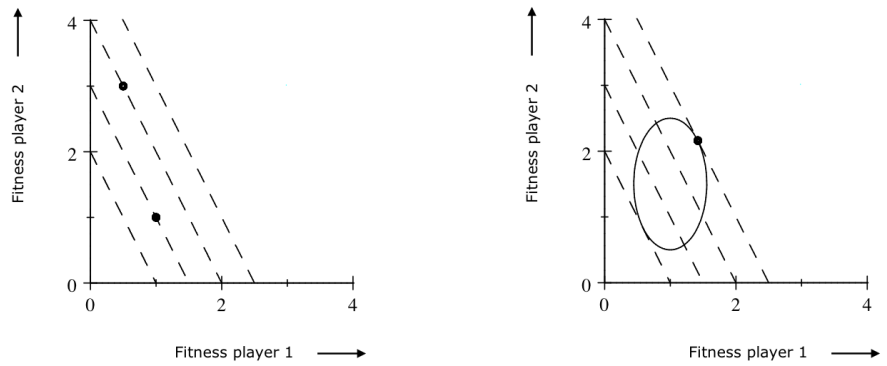


Figure 31. (a) Faced with a choice between $(1, 1)$ and $(0.5, 3)$, a player 1 with utility function $u_1(\pi_1, \pi_2) = \pi_1 + \frac{1}{2}\pi_2$ prefers the latter option, as $1 + \frac{1}{2} \cdot 1 = 1\frac{1}{2} < 2 = 0.5 + \frac{1}{2} \cdot 3$. This is reflected in the figure, where $(0.5, 3)$ lies on a higher iso-utility line, or indifference curve. If $(1, 1)$ is the status quo, then the other option implies a cost of 0.5 to self and a benefit of 2 for the other. (b) Faced with a choice between a whole set of options, the point on the highest indifference curve is chosen.

The payoffs π_1 and π_2 are not chosen directly, but indirectly. Players 1 and 2 choose actions x and y, the combination of which leads to payoffs $\pi_1 = \pi(x, y)$ and $\pi_2 = \pi(y, x)$. (As in Section 3, we assume that the game is symmetric). Players 1 and 2 are assumed to understand those consequences, and choose x and y so as to maximize their respective utilities u_1 and u_2, given what the other player is doing. In other words, they are playing a Nash equilibrium, where each evaluates the outcomes according to their utility function.

While individuals choose their actions x and y on the basis of their own utility functions, and the utility function of the individual they are paired with, evolution chooses utility

functions on the basis of the payoffs that they imply for the player that has them. This is called preference evolution (examples include Robson, 2001, Samuelson, 2001, Weibull & Salomonsson, 2006, Van Veelen, 2006, Dekel, Ely, Yilankaya, 2007, Akçay et al., 2009, Robson & Samuelson, 2011, Akçay & Van Cleve, 2012, Alger & Weibull, 2012) and sometimes it is also referred to as the indirect evolutionary approach (Güth & Yaari, 1992, Güth, 1995).

Now suppose that the phenotypes that evolution can choose from are all linear utility functions of the type depicted in Fig. 31. In other words, what evolves is the altruism parameter α in the utility function $u(\pi_1, \pi_2) = \pi_1 + \alpha\pi_2$, where α is the weight attached to the payoff of the interaction partner. In a world with very simple games, where all that happens is that individuals at times have the opportunity to make transfers of different kinds, as in Hamilton (1964), selection would always favour $\alpha = r$ over all other values of α (see Theorem 1, Van Veelen, 2006). But now let's look at a world where individuals play slightly more complicated games, and where both players are aware of the utility function of the other, and where both individuals make inferences as to what that implies for the action that the other player is going to choose. Below, we will write α for the altruism parameter of player 1 and β for the altruism parameter of player 2.

Again, there are games where nothing special happens. If we take a prisoners' dilemma with equal gains from switching, and two players that have a utility function of this form, then for both players the preferred action is independent of the action of the other, whatever their respective altruism parameters are. But there are also games where what I would prefer to do depends on what the other does. In such cases interesting things can happen, as was pointed out by Akçay et al. (2009) and Alger & Weibull (2012).

Let's consider the payoff function $\pi(x, y) = 4(xy)^{\frac{1}{2}} - x^2$. With this payoff function, and with two selfish individuals – $\alpha = 0$ and $\beta = 0$ – what one would prefer to do, depends on what the other does. For both players, the utility-maximizing choice is to play the cube root of the choice of the other.[18] Since $\alpha = \beta = 0$, this is also the payoff-maximizing choice for both. Therefore in Nash equilibrium, we would have $x = \sqrt[3]{y}$ and $y = \sqrt[3]{x}$, and hence both players will choose $x = y = 1$. Both players get a payoff of $\pi(1, 1) = 3$.

But what happens if a mutant with a positive level of altruism enters? If player 1 is a mutant with altruism parameter $\alpha = \frac{1}{3}$, then its best response[19] is to play $x = \sqrt[3]{\frac{16}{9}y}$. The best response of player 2, with altruism parameter $\beta = 0$ is still $y = \sqrt[3]{x}$. The Nash equilibrium between those two players is $x = \left(\frac{16}{9}\right)^{\frac{3}{8}} > 1$ and $y = \left(\frac{16}{9}\right)^{\frac{1}{8}} > 1$. What has happened now is that player 1 increased its choice of x, because it cares about player 2. Player 2 subsequently increased its choice of y too, not because player 2 cares about player 1 too, but only because increasing y is the self-interested best thing to do in response to an increase in x of player 1.

[18] With $\alpha = 0$ the first derivative of player 1's utility fnction is $\frac{du}{dx} = 2\left(\frac{y}{x}\right)^{\frac{1}{2}} - 2x$. This is 0 if $x = \sqrt[3]{y}$.

[19] With $\alpha = \frac{1}{3}$ the first derivative of player 1's utility function is $\frac{du}{dx} = 2\left(\frac{y}{x}\right)^{\frac{1}{2}} - 2x + \frac{1}{3}\left(2\left(\frac{y}{x}\right)^{\frac{1}{2}}\right)$. This is 0 if $x = \sqrt[3]{\frac{16}{9}y}$.

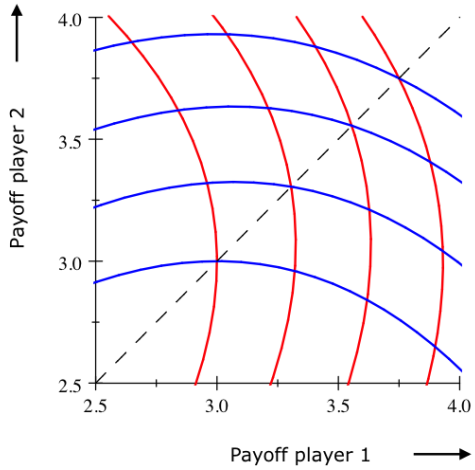


Figure 32. The payoff function $\pi(x,y) = 4(xy)^{\frac{1}{2}} - x^2$. A red line represents combinations of payoffs to player 1 and player 2 that player 1 can effectively choose from by choosing x, given a fixed choice of $\widehat{y}$ of player 2. It is described by $(\pi(x,\widehat{y}), \pi(\widehat{y},x))$, where $\widehat{y}$ is chosen to be $1, 1\frac{1}{6}, 1\frac{2}{6}$ and $1\frac{3}{6}$, respectively, and x varies continuously. A blue line does the same from the perspective of player 2. The graph thereby takes the action of one player constant, and pictures the effect of the other's possible actions on the payoffs of both players.

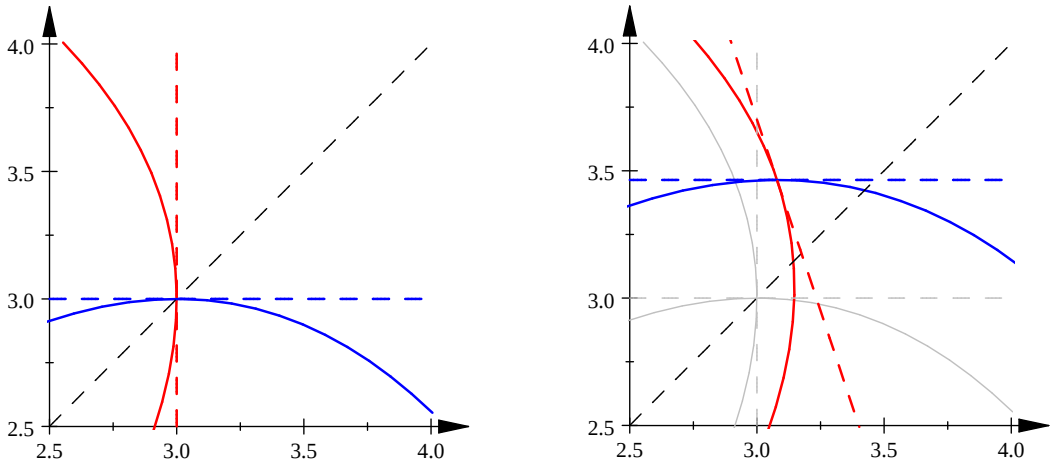


Figure 33. (a) The Nash equilibrium, given that both are selfish ($\alpha = \beta = 0$). (b) If player 1 changes to an altruistic preference (for instance $\alpha = \frac{1}{3}$, as in the picture) it induces the other player to increase the level of cooperation, even though player 2 still has $\beta = 0$. The possibility to credibly commit to playing a more cooperative action pays off with strategic complements; player 1 ends up with a higher payoff at $\alpha = \frac{1}{3}$ than at $\alpha = 0$. Therefore

$\alpha = \beta = 0$ is not an equilibrium in preference evolution (for the record: $\alpha = \beta = \frac{1}{3}$ is).

It turns out that this is actually good for the material payoff of player 1. The payoff of player 1 is now $\pi\left(\left(\frac{16}{9}\right)^{\frac{3}{8}}, \left(\frac{16}{9}\right)^{\frac{1}{8}}\right) = 4\left(\frac{16}{9}\right)^{\frac{1}{4}} - \left(\frac{16}{9}\right)^{\frac{3}{4}} > 3$, which therefore is more than a player with $\alpha = 0$ would earn against a player with $\beta = 0$. Caring for the other individual now has worked as a way to credibly commit to increasing one's action x, which induced the other to increase her action y, which in this case turned out to pay off at the ultimate level.

With utility functions as phenotypes, evolution acts on the altruism parameter α. It increases until α reaches $\frac{1}{3}$, where it increases no further. Note that this still is all within a well-mixed setting, with relatedness 0. In spite of this, altruism has evolved, not because of positive relatedness, but because altruism serves as a commitment device in games with strategic complements. The next example shows that the opposite can happen with strategic substitutes.

For this we consider the payoff function $\pi(x, y) = 8(x + y)^{\frac{1}{2}} - \sqrt{2}x^2$. With this payoff function, and with two selfish individuals – that is, $\alpha = 0$ and $\beta = 0$ – what one would prefer to do also depends on what the other does. Here also the Nash equilibrium is that both players choose $x = y = 1$.[20] Since $\alpha = \beta = 0$, this is also the payoff-maximizing choice for both. The payoff of both players is $\pi(1, 1) = 7\sqrt{2}$.

But what happens if a mutant with a negative level of altruism enters? If player 1 is a mutant with altruism parameter $\alpha = -\frac{1}{5}$, then the Nash equilibrium between those two players is[21] $x = \frac{19}{20}\left(\frac{40}{39}\right)^{\frac{1}{3}} < 1$, and $y = \left(\frac{40}{39}\right)^{\frac{1}{3}} > 1$. What has happened now is that player 1 decreased its choice of x, because it dislikes payoff going to player 2. Player 2 subsequently increased its choice of y, because increasing y is the self-interested best thing to do in response to an increase in x by player 1.

It turns out that this is actually good for the material payoff of player 1. The payoff of player 1 increases, compared to what a player with $\alpha = 0$ would earn against a player with $\beta = 0$. Being jealous of the other individual, or spiteful, now has worked as a way to credibly commit to decrease one's action x, which induced the other to increase her action y, which in this case turned out to pay off at the ultimate level.

With utility functions as phenotypes, evolution acts on the altruism parameter α. For this example it decreases until α reaches $-\frac{1}{5}$, where it will decrease no further. Again this still is all within a well-mixed setting. In spite of this, spite has evolved, not because of negative relatedness, but because spite serves as a commitment device in games with strategic substitutes.

---

[20] With $\alpha = 0$ the first derivative of player 1's utility fnction is $\frac{du}{dx} = 4(x + y)^{-\frac{1}{2}} - 2\sqrt{2}x$. This is 0 if $x^2(x + y) = 2$. In symmetric equilibrium $x = y = 1$.

[21] With $\alpha = -\frac{1}{5}$ the first derivative of player 1's utility function is $\frac{du}{dx} = 4(x + y)^{-\frac{1}{2}} - 2\sqrt{2}x - \frac{1}{5}\left(4(x + y)^{-\frac{1}{2}}\right)$. This is 0 if $x\sqrt{x + y} = \frac{19}{20}\sqrt{2}$.

With $\beta = 0$, the first derivative of player 2's utility function is 0 if $y\sqrt{x + y} = \sqrt{2}$. In equilibrium therefore $x = \frac{19}{20}y$, hence $x = \frac{19}{20}\left(\frac{40}{39}\right)^{\frac{1}{3}}$, $y = \left(\frac{40}{39}\right)^{\frac{1}{3}}$.

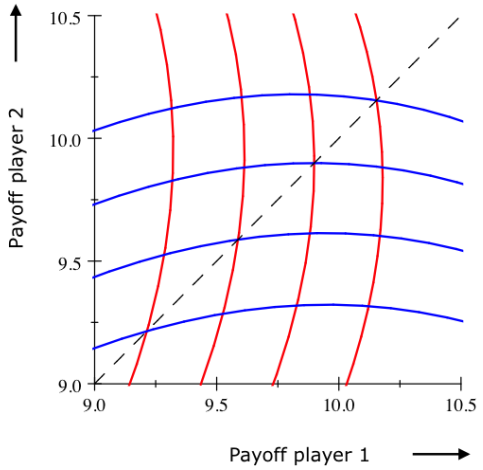


Figure 34. The payoff function $\pi(x,y) = 8(x+y)^{\frac{1}{2}} - \sqrt{2}x^2$. The red lines represent the payoff combinations $(\pi(x,\widehat{y}), \pi(\widehat{y},x))$ where $\widehat{y}$ is a fixed value (here chosen to be $0.8, 0.9, 1.0$ and 1.1) and x varies continuously. The blue lines represent the payoff combinations $(\pi(\widehat{x},y), \pi(y,\widehat{x}))$ where $\widehat{x}$ is a fixed value (again $0.8, 0.9, 1.0$ and 1.1) and y varies continuously.

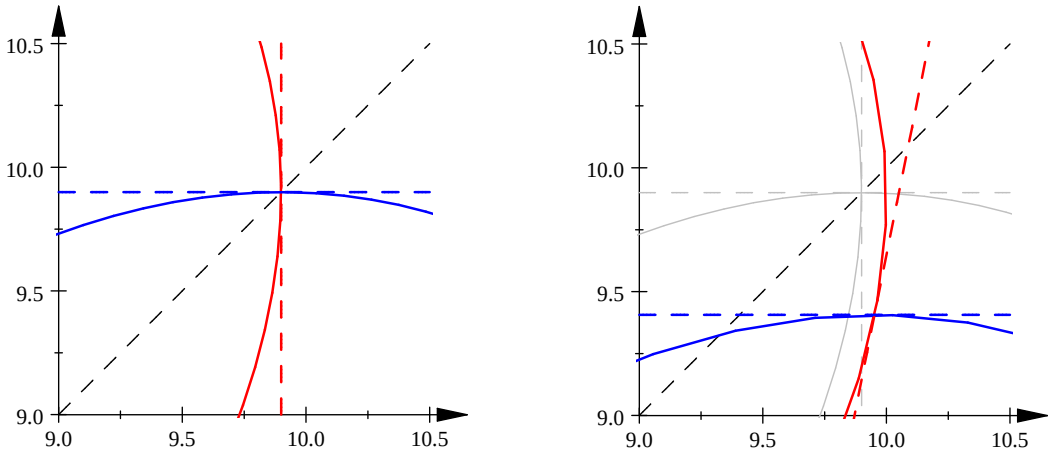


Figure 35. (a) The Nash equilibrium, given that both are selfish ($\alpha = \beta = 0$). (b) If player 1 changes to a spiteful preference (for instance $\alpha = -\frac{1}{5}$, as in the picture) it induces the other player to increase the level of cooperation, even though player 2 still has $\beta = 0$. The possibility to credibly commit to playing a spiteful action pays off with strategic substitutes; player 1 ends up with a higher payoff at $\alpha = -\frac{1}{5}$ than at $\alpha = 0$. Therefore $\alpha = \beta = 0$ is not an equilibrium in preference evolution (for the record: $\alpha = \beta = -\frac{1}{5}$ is).

What this shows is that altruism – or spite – may also evolve for reasons other than interactants being related and without being matched assortatively otherwise. Of course, on top of room for commitment, there could be relatedness too. In their paper, Alger & Weibull (2012) look at how the combination of both ingredients leads to levels of altruism or spite that depend on the type of game as well as relatedness. Here we would also like to do that, and see how that fits into Hamilton's rule. We will do that in a somewhat more general way than in the previous examples, and comprise the entire Nash-equilibrium finding in a single function, assuming that there is a unique Nash equilibrium. The Nash equilibrium is a combination (x^*, y^*) of actions of players 1 and 2, that obviously depends on their respective altruism parameters α and β, and on the payoff function π. This function is denoted by $N_\pi(\alpha, \beta)$. For brevity, we also write x_1^*, x_2^*, y_1^* and y_2^* for derivatives $\frac{dx^*(\alpha,\beta)}{d\alpha}$, $\frac{dx^*(\alpha,\beta)}{d\beta}$, $\frac{dy^*(\alpha,\beta)}{d\alpha}$ and $\frac{dy^*(\alpha,\beta)}{d\beta}$. The resulting payoffs for player 1 are obviously $\pi(N_\pi(\alpha, \beta)) = \pi(x^*, y^*)$.

One way to write down the indirect evolution method in formulas in this setting, with relatedness r, is to say that it leads to an altruism parameter α^* for which at least the following must hold

$$(1-r)\left.\frac{d\pi\left(N_\pi(\alpha,\beta)\right)}{d\alpha}\right|_{\beta=\alpha=\alpha^*} + r\left.\frac{d\pi\left(N_\pi(\alpha,\alpha)\right)}{d\alpha}\right|_{\alpha=\alpha^*} = 0 \tag{8.1}$$

With x^* and y^* both as functions of both the altruism parameter of player 1 and of player 2, we can rewrite the stability condition as

$$(1-r)\left[\frac{d\pi}{dx}x_1^* + \frac{d\pi}{dy}y_1^*\right]_{\beta=\alpha=\alpha^*} + r\left[\frac{d\pi}{dx}(x_1^* + x_2^*) + \frac{d\pi}{dy}(y_1^* + y_2^*)\right]_{\beta=\alpha=\alpha^*} = 0 \tag{8.2}$$

This can be reorganised as

$$\left.\frac{d\pi}{dx}x_1^*\right|_{\beta=\alpha=\alpha^*} + \left.\frac{d\pi}{dy}y_1^*\right|_{\beta=\alpha=\alpha^*} + r\left[\frac{d\pi}{dx}x_2^* + \frac{d\pi}{dy}y_2^*\right]_{\beta=\alpha=\alpha^*} = 0 \tag{8.3}$$

These three terms now allow for a nice summary of effects. The first term describes how (further) changes in my altruism parameter α would affect my own payoff through its effect on my own choice of a variable x in Nash equilibrium. This term is typically negative; if I start caring more for the other, I am more inclined to trade my own payoff for the payoff of the other. Therefore, if a positive α evolves, this has to be offset by a positive effect through other channels. The second term describes how (further) changes in my altruism parameter α would affect my payoff through its effect on my opponents, or co-player's, choice of her variable y in Nash equilibrium. This captures the commitment effect. It is positive in the first example, where me becoming *more* altruistic makes the other choose a more cooperative y. It is negative in the second example, where me becoming *less* altruistic

makes the other choose a more cooperative y. The last term concerns the effects on my payoff of a corresponding change in the altruism parameter by my opponent, or co-player, which I am facing with probability r. These are, again, subdivided; first there is the effect through my own equilibrium choice x, and then the effect through the choice y of my co-player.

This is not yet inclusive fitness according to the intuition from Hamilton (1964), but rather the neighbour-modulated fitness version. In order for this to be inclusive fitness, we have to realize that the symmetry of the game implies that the effect of a change in my opponent's β on my equilibrium choice x^* equals the effect of a change in my own α on my opponent's equilibrium choice y^*, provided that $\alpha = \beta$. In other words, $y_2^*|_{\alpha=\beta} = x_1^*|_{\alpha=\beta}$. Similarly, $x_2^*|_{\alpha=\beta} = y_1^*|_{\alpha=\beta}$. Therefore we can rewrite the condition as a proper inclusive fitness maximization condition.

$$\left.\frac{d\pi}{dx}x_1^*\right|_{\beta=\alpha=\alpha^*} + \left.\frac{d\pi}{dy}y_1^*\right|_{\beta=\alpha=\alpha^*} + r\left[\frac{d\pi}{dx}y_1^* + \frac{d\pi}{dy}x_1^*\right]_{\beta=\alpha=\alpha^*} = 0 \tag{8.4}$$

The properties of the examples all show up in this formula. In the two previous examples in this section, $r = 0$, so all that happens does so through the first two terms. Strategic complementarities make the two terms balance at a positive α. As long as the sum of the two terms is still positive, then increasing α hurts me less through the effect on myself – reflected in the first term – than it gains me trough the effect through the other – reflected in the second term, which is positive with complementarities. The second example is the mirror image. With strategic substitutes the two terms balance at a negative α. When the sum of the two terms is still negative, then decreasing α hurts me less through the effect on myself – reflected in the first term; I am now actually willing to pay to reduce the other's payoff – than it gains me through the effect on the other player – reflected in the second term, which is negative in case of complementarities.

We can construct another example by taking the first payoff function from Section 6, and consider $\pi_1(x, y) = ay - x^2$. This payoff function has equal gains from switching for every two choices of trait values, which implies that there are no strategic substitutes or complements, or, in other words, that the second term in the formula is always 0. The reason is that with linear altruistic preferences and games with no strategic substitutes or complements, my choice of x has no effect on the optimal choice y of my opponent, or co-player. If I have utility function $u_1(\pi_1, \pi_2) = \pi_1 + \alpha\pi_2$, then with this game my utility is maximized at $x = \alpha \cdot \frac{a}{2}$. This is independent of the other player's choice y, and hence this immediately is x^*. The other player's utility is maximized at $y = \beta \cdot \frac{a}{2}$, which is independent of my choice x. Because both players' optimal choices do not depend on what the other does, also the altruism parameters α and β only affect the behaviour of the one that has them, and not the behaviour of the opponent in Nash equilibrium. In other words, both $y_1^* = 0$ and $x_2^* = 0$. The resulting formula for this payoff function with equal gains from switching therefore is

$$\left.\frac{d\pi}{dx}x_1^*\right|_{\beta=\alpha=\alpha^*} + r\left.\frac{d\pi}{dy}x_1^*\right|_{\beta=\alpha=\alpha^*} = 0 \tag{8.5}$$

In our example, $\frac{d\pi}{dx} = -2x$, which has to be evaluated at $x = x^* = \alpha \cdot \frac{a}{2}$. Furthermore we have $x_1^* = \frac{a}{2}$ and $\frac{d\pi}{dy} = a$. Solving this equation gets us $\alpha^* = r$.

In this last example, altruism evolves because of relatedness. In the other two, even though equation (8.4) is still Hamilton's rule, we get $\alpha^* = \frac{1}{3} \neq 0 = r$ and $\alpha^* = -\frac{1}{5} \neq 0 = r$, and hence in both those cases it does not capture kin selection. Altruism and spite there evolve because it helps individuals commit.

8.1 Commitment issues, secret handshakes and green beards

The mechanism at work in these examples is that altruism and spite are used as commitment devices. The idea that solving commitment issues is what many human emotions are for, is, in a much broader sense, the central thesis in the book *Passions within Reason* by Frank (1988). One relevant question there, and in the above examples, is of course if players can indeed infer each other's level of altruism or spite. Truthful revealing of preferences is complicated by the fact that in both situations players have an incentive to lie. In the first example players would like to be perceived as more altruistic than they really are; in the second they would like to be perceived as more spiteful. One would have to assume that preferences are observable, and players cannot deceive each other, for this commitment argument to work. The extent to which humans can credibly commit has been the subject of some debate; Binmore (1994) for instance disagrees with Frank (1987, 1988) on our capacity to read each other. This is not the place to try to answer that question, but it is clear that at least in principle it is possible that altruism or spite evolves for reasons other than shared genes.

An alternative possibility that observable preferences open up, is that they can serve as a "secret handshake" (Robson, 1990, Samuelson, 2001, Dekel, Ely & Yilankaya, 2007), or, in biological terms, a green beard. A mutant individual could have a preference for being cooperative when matched with someone with the same preference, and not when matched with others. This mutant could invade a selfish population. This possibility comes with similar complications concerning truthfulness. A subsequent mutant that has the right handshake, or beard colour, and defects nonetheless, could in turn invade a population with a sufficiently high share of the first mutant. The green beard effect, although not yet by that name, was suggested as a possibility already by Hamilton (1964b) himself. On page 25 of part II he suggests, with some skepticism, that there could be "*a supergene affecting (a) some perceptible feature of the organism, (b) the perception of that feature, and (c) the social response upon what was perceived.*" The term "green beard" was coined by Dawkins (1976). A larger literature about this possibility developed later; some examples include Keller &

Ross (1998), Riolo, Cohen & Axelrod (2001), Axelrod, Hammond & Grafen (2004), Jansen & Van Baalen (2006), and García, Van Veelen & Traulsen (2014).

8.2 Siblings vs. friends

Haldane's famous example involves relatives saving relatives from drowning. Yet friends might also rescue each other. A core ingredient of friendship seems to be that friends care for the well-being of each other too, and display a willingness to bear certain costs to the benefit of the other. One approach to explain that could be to model that as a repeated game (for example Axelrod & Hamilton, 1981). A somewhat different, and interesting alternative is proposed by Eshel & Motro (1981). Their main observation is that if I know that you are altruistic to me, and that you would save me, or help me, if that is not too risky or costly to you, then your being alive becomes something that is valuable to me. If you are drowning, then my saving you has as an extra benefit that I will have you around to save me, if tables turn and I am in need of saving. That could reinforce the altruism between for instance siblings that is generated by kin selection. Note that this does not have to imply a violation of Hamilton's rule; the saving of the other now just implies an additional benefit to the actor.

Their observation also implies that there could be helping behaviour even between unrelated individuals. The other's possible future help can make it a worthwhile investment to help him or her today. Whether the label "altruism" applies here is a matter of taste. In a strict sense, this is not altruism, because it just serves the individual's best interests to save the other. In a less strict sense, one could also defend calling it altruism, as the proximate mechanisms and the behaviour between a person saving a sibling from drowning and saving a friend from drowning might be relatively close. In the final section we will avoid confusion by referring to all those behaviours as "helping behaviour", and save altruism for use in a stricter sense.

9 Empirical testing

In this section we look at how we can test empirically whether or not Hamilton's rule holds. First we return to the replicator dynamics. With costs and benefits according to the regression method, Hamilton's rule always holds, so there is no need for empirical testing. With costs and benefits according to the counterfactual method, we can get violations, but if we want to observe violations *in equilibrium*, we need to look for equilibria where cooperators and defectors coexist. With no scope for finding violations in equilibria where either one has gone to fixation, the "false negatives" in an overview by Bourke (2014) need explaining too. We furthermore discuss empirical complications if we assume a continuous trait space with adaptive dynamics, and we look at a hypothetical statistical exercise that tries to distinguish between different update rules on the cycle.

There is, obviously, an enormous empirical literature that is inspired by or based on Hamilton's rule. Describing it would be a massive task, well beyond the scope of this paper, and we will not attempt to do that. What we will do in this section, is discuss explicit empirical tests of Hamilton's rule. For that we first need to determine what a violation of Hamilton's rule would look like. This is best done by going back to the setting with the replicator dynamics from Section 3. We will start with that in Section 9.1. Here we can, again, choose to define costs and benefits using the counterfactual method, or the regression method.

With the regression method, no true model would ever violate Hamilton's rule. With the counterfactual method, true models can violate Hamilton's rule, in the sense that Hamilton's rule can disagree with the direction of selection. In many empirical studies, however, it is moreover assumed, tacitly or explicitly, that the system we observe is in equilibrium. In this section we will show that, if this is a monomorphic equilibrium, in which either cooperators or defectors have gone to fixation, then, in equilibrium, Hamilton's rule will also not be violated when we use costs and benefits according to the counterfactual method. It is important to realize that this does not mean that Hamilton's rule generally holds after all. It can still point in the wrong direction concerning out-of-equilibrium dynamics, and it can still give the wrong answer to the question if the behaviour would still be stable if relatedness is increased or decreased by a certain amount.

If the equilibrium is a mixture of cooperators and defectors, and we use the counterfactual method to determine costs and benefits, then we can observe a violation of Hamilton's rule in equilibrium. This does moreover require that we allow our statistical model to be non-linear – provided, of course, that this is what the data tells us.

Bourke (2014) reviews twelve explicit tests of Hamilton's rule. None of these papers look at polymorphisms, and none of these papers use non-linear statistical models. This implies that we should expect no violations of Hamilton's rule. Yet there are quite a few, and therefore we will also discuss what could have caused these "false negatives".

In Section 9.2 we will switch to adaptive dynamics. We will discuss some empirical difficulties that arise when one would want to try to establish empirically whether or not a population finds itself at the equilibrium trait value, for which $rb = c$, as the results in Section 6 predict for a considerable set of fitness functions.

In Section 9.3 we will revisit the cycle. This setup was used in Section 7 to illustrate cancellation effects. Models like the cycle are meant to illustrate a principle, and not as a model that matches the local interaction structure of a specific organism particularly well. It can nonetheless be instructive to imagine an empirical exercise, where we know that individuals are organized on a cycle, but not which update rule is used. Since the different update rules imply different fitness effects, trying to reconstruct the update rule becomes an empirical exercise in measuring fitness effects and model specification.

9.1 Replicator dynamics and actual tests

When considering empirical tests of Hamilton's rule, one relevant question is: which Hamilton's rule are we testing? In Sections 3 and 4 we have seen that there are different definitions of costs and benefits, depending on whether we use the counterfactual method (see Karlin & Matessi, 1983, Matessi & Karlin, 1984, 1986, and Section 3.2.1) or the regression method (see Gardner et al. 2011, and Section 4). With the latter definition of costs and benefits, Hamilton's rule *always* holds.

The fact that with this definition, Hamilton's rule leaves no scope for testing its validity empirically is of course not a bad thing. The idea that natural selection works because fitter genotypes are more likely to survive than less fit genotypes also escapes empirical testing, if the fitness of a genotype is measured by counting how many survive and procreate and how many do not. In this sense, natural selection is also tautologically occurring, and, needless to say, this in no way diminishes the relevance and importance of the idea of evolution by natural selection. What it does imply, however, is that when there are papers that set out to test the validity of Hamilton's rule empirically, then it is to be expected that they will *not* be using the regression method to compute b and c, because that would render the actual data-collection a waste of energy, as we already know that whatever the data, Hamilton's rule will always be confirmed.

The studies in the review by Bourke (2014) cover a range of behaviours; egg dumping in lace bugs (Loeb, 2003); guarding and worker behaviour in a variety of bees (Hogendoorn & Leys, 1993, Stark, 1992, Bourke, 1997, Richards, French & Paxton, 2005); female joining behaviour in a variety of wasps (Queller & Strassmann, 1988, Nonacs & Reeve, 1995, Noonan, 1981, Metcalf & Whitt, 1977, Gadagkar, 2001); kin discrimination in cannibalizing behaviour in larvae of tiger salamanders (Pfennig, Collins & Ziemba, 1999); cooperative lekking in wild turkeys (Krakauer, 2005); and helping at the nest in the white-fronted bee-eater (Emlen & Wrege, 1989). Seven of those studies find that the behaviour has positive inclusive fitness ($rb > c$), one of them finds a negative inclusive fitness ($rb < c$), three

studies have mixed results (some cases or years with $rb > c$, some with $rb < c$) and one finds inclusive fitness equal to 0 ($rb = c$).

These studies are summarized in the review as follows:

> Overall, the studies considered in this review strongly confirm the predictions of Hamilton's rule regarding the conditions and likely causes that underpin social evolution at ecological and evolutionary timescales.

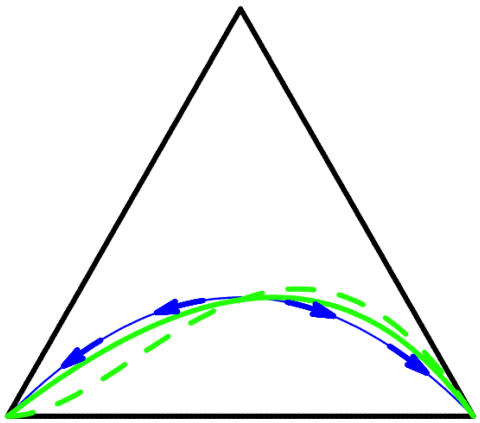
That is a remarkably positive aggregation of these results, which contain quite a few violations of Hamilton's rule. Observed worker behaviour in halictid bees was found to have negative inclusive fitness in Richards, French & Paxton (2005). Guarding behaviour in allodapine bees was found to have positive inclusive fitness for observations from 1987, and negative for observations from 1988 (Stark, 1992, Bourke, 1997). Female joining behaviour was found to have negative inclusive fitness in *Polistes annularis* for 5 out of 5 group sizes observed in 1977, and for 3 out of 5 group sizes observed in 1978 by Queller & Strassmann (1988). Female joining behaviour was found to have negative inclusive fitness in *Polistes fuscatus* for 3 out of 4 group sizes (Noonan, 1981), and to be not statistically different from 0 in *Polistes dominulus* in Nonacs & Reeve (1995). It is true that Hamilton's rule is confirmed quite a few times (7 out of 12 studies, where one study is taken to be a set of two papers on the same behaviour and species). But together these 12 studies certainly do not imply that Hamilton's always holds – assuming that we have confidence in the statistical power of the individual studies. After all, one behaviour in one species for which we are confident that the data imply that inclusive fitness really is negative is enough to reject the claim that Hamilton's rule *always* applies.

There are some statistical concerns though, that imply that the violations of Hamilton's rule that are found do not have to be the final answer in these specific cases. Also it is worth trying to answer the question what could have generated the violations, and whether the authors were using the regression method, the counterfactual method, or neither, to compute the costs and benefits. For that it will be useful to return to Section 3, in which individuals also face a binary choice. That is what we will do below.

9.1.1 Violations of Hamilton's rule in equilibrium

Hamilton's rule always holds if we use the regression method to define costs and benefits. If we use the counterfactual method, and the game has equal gains from switching, then Hamilton's rule also always holds. One would therefore only expect possible violations in an empirical study if the counterfactual method is used, and if the game moreover does not satisfy equal gains from switching. There are two cases to be considered: the case where bistability is possible ($P - S > T - R$) and the case that allows for coexistence ($P - S < T - R$).

$P - S > T - R$, the defecting equilibrium If we expect to find a population in equilibrium, then $P - S > T - R$ implies that the population will either be in the corner of the simplex where the frequency of cooperators is 0, or in the corner where that frequency is 1. At the first corner, the inclusive fitness of the cooperative behaviour, if b and c are defined according to the regression method, is negative, because with this definition, inclusive fitness always matches the direction of selection. With equation (4.12), and filling in $p = 0$, we find that $b_{regr} = \frac{1}{1+r}(T - P) + \frac{r}{1+r}(R - S)$, and $c_{regr} = \frac{1}{1+r}(P - S) + \frac{r}{1+r}(T - R)$. For benefits and costs according to the counterfactual method, we use equation (3.3), and filling in $p = 0$, we find that $b_{count} = T - P$ and $c_{count} = P - S$. Since we assumed that $P - S > T - R$, the costs according to the counterfactual method in this corner are higher than the costs according to the regression method, while the benefits are lower. Therefore, given that in this corner $rb_{regr} < c_{regr}$, certainly $rb_{count} < c_{count}$.



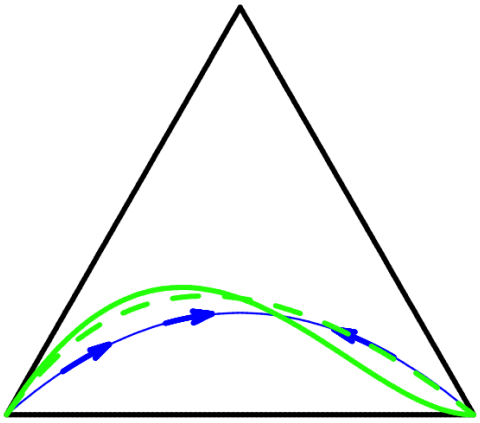
$P - S > T - R$

Figure 36. The solid green line separates the region with rising shares of cooperators (below it) from the region with declining shares of cooperators (above it). The dashed green lines separate the regions where inclusive fitness is positive (below) and negative (above), when benefits and costs are computed according to the counterfactual method. In the left corner, the dashed line is below the solid green line, implying that if defection is stable, then certainly inclusive fitness with the counterfactual method is negative too. In the right corner, the dashed line is above the solid green line, implying that if cooperation is stable, then certainly inclusive fitness with the counterfactual method is positive too.

$P - S > T - R$, the cooperative equilibrium The situation in the other corner is the mirror image with the same conclusion; in equilibrium, there is no scope for observing violations. With equation (4.12), this time filling in $p = 1$, we find that, according to the regression method, $b_{regr} = \frac{r}{1+r}(T - P) + \frac{1}{1+r}(R - S)$, while $c_{regr} = \frac{r}{1+r}(P - S) +$

$\frac{1}{1+r}(T-R)$. For benefits and cost according to the counterfactual method, we use equation (3.3) again, and filling in $p=1$, we find that $b_{count}=R-S$ and $c_{count}=T-R$. Since we assumed that $P-S>T-R$, that implies that in this corner, the costs according to the counterfactual method are lower than the costs according to the regression method, while the benefits are higher. Therefore, given that in this corner $rb_{regr}>c_{regr}$, certainly $rb_{count}>c_{count}$. In equilibrium, such a violation would therefore never be observed. For violations, we would need to see selection in action in the region between the solid and broken green lines in Fig. 36, but in equilibrium, this will not be observed, also not with the counterfactual method.

$P-S<T-R$ The case with coexistence does allow for violations of Hamilton's rule to be observed in equilibrium. The equilibrium fraction of cooperators is $p(r)=\frac{1}{1-r}\frac{(1-r)S+rR-P}{S-R+T-P}$ (see Appendix B). If we fill in this frequency in (3.3), we get $b_{count}=\frac{1}{1-r}(T-S+r(P-R+S-T))$ and $c_{count}=\frac{r}{1-r}(R-P)$, which implies that $rb_{count}-c_{count}=\frac{r}{1-r}(T-R+P-S+r(P-T+S-R))$. With $P-S<T-R$, this is non-zero, except where the solid green line intersects the dashed one in Fig. 37. If we would like to find violations of Hamilton's rule in equilibrium, therefore, we should focus on cases that allow for coexistence.



$P-S<T-R$

Figure 37. The equilibrium is found at the intersection of the solid green line and the blue constant-r arc, and this intersection is below the green dotted line, which means that inclusive fitness with costs and benefits according to the counterfactual method would incorrectly suggest that selection would lead to a further increase of the frequency of cooperators.

9.1.2 What causes the violations in the empirical studies?

Given that there are so many ways not to find violations of Hamilton's rule, it is interesting to find out what causes the relatively large number of studies that find that the prevalent

behaviour has negative inclusive fitness. It will also be interesting to find out how costs and benefits are computed in such empirical studies. In order to answer those questions, we first look at the right/down corner of Fig. 36, where the share of cooperators is 1. In that corner, there are no defectors around. By lack of observations of defectors, both the regression method and the counterfactual method would be at a loss. (The reason why the counterfactual method would also be at a loss there, is that it also depends on estimating a statistical model that describes how fitnesses depend on behaviours. A difference is that the counterfactual model does not restrict the statistical model to being linear, but both need observations of cooperators *and* defectors).

The absence of defectors does not have to imply that it is impossible to recover what the fitnesses would be without giving or receiving help. One can gather observations on individuals that do not give or receive help for reasons other than being a defector, or their partner being one. One such reason could be that they have no interaction partner to receive help from, or to give help to, to begin with. Such observations offer a perfect way to get around a possible lack of defectors.

The studies surveyed in Bourke (2014) do indeed typically use clever workarounds to get at the payoff of defectors. They do, however, use linear specifications. Linearity is a feature that is shared with the regression method, while a discrepancy between Hamilton's rule using the counterfactual method on the one hand, and the direction of selection on the other, would hinge on the true relation between fitnesses and the type of oneself and one's interaction partner not being linear. The question therefore remains what generates the violations. We will discuss three possible reasons.

The observations are only a sample, and the number of offspring is a random variable Suppose the population is not actually in the corner with cooperators only, but we have a population in which there are still some defectors around. Suppose we now draw a number of pairs from that population. Fitness is only the expected number of offspring, and clearly the number of offspring has to be a random variable. In the sample, it might be that the individuals in DD-pairs happen to all have many offspring. This then could lead to inclusive fitness within our sample being negative, even though in the population as a whole inclusive fitness is positive, and the fraction of cooperators in the population as a whole is still rising. Notice that with the regression method, it is still tautologically true that inclusive fitness is positive if and only if the frequency of cooperators is going up. Here that only implies that if randomness gives us a sample in which inclusive fitness is negative, then within this sample, indeed the share of cooperators went down – or vice versa. An ever larger sample size would reduce the probability of this happening ever more.

Relatedness is estimated separately Relatedness in equation (4.9) is computed based on the distribution of cooperators and defectors in that sample; it is defined as $\frac{Cov(X,Y)}{Var(X)}$, where X and Y pertain to with the distribution of cooperators over pairs in the sample.

Note that this is only an estimate of the true relatedness, but one that would make Hamilton's rule work. That, however, is not how relatedness is computed in empirical studies. Some find r using the pedigrees of the interacting individuals (Metcalf and Whitt, 1977; Emlen and Wrege, 1989; Stark, 1992; Hogendoorn and Leys, 1993; Richards et al., 2005; Gorrell et al., 2010). Others use genetic marker testing when gathering genetic information of the organisms is feasible (e.g. in Loeb, 2003; Krakauer, 2005; Hatchwell et al., 2014). Replacing $\frac{Cov(X,Y)}{Var(X)}$ with another measure for r implies that equation (4.9) no longer is a tautology, and it becomes possible that inclusive fitness – with benefits and costs according to the regression method, but $\frac{Cov(X,Y)}{Var(X)}$ replaced with a different estimate of r – is positive, while in the sample the frequency of cooperators goes down. Again, more data will reduce the likelihood of this problem occurring.

The workaround might get the frequencies of pair types wrong For the third reason, we take a closer look at the computation of the costs and benefits according to the regression method, when applied to the replicator dynamics for the prisoners' dilemma. This is done in Appendix A3 (see also Gardner et al., 2011). The important thing to observe, is that the solution to the minimization depends on the shares f_0, f_1 and f_2 of, respectively, DD-pairs, CD-pairs, and CC-pairs, unless we have equal gains from switching. With equal gains from switching, fitnesses of different types in different pairs can be made to match the linear model exactly, which makes the minimization independent of f_0, f_1 and f_2. Without equal gains from switching, this is no longer the case. If we have obtained our observations not from matches with defectors, but from not being matched at all, for instance, then there is no reason to assume that the numbers of the different types of observations happen to be the same as the f_0, f_1 and f_2 that would go with the relatedness r at hand, in combination with a small p. This problem cannot be reduced by just gathering more data.

Because one would not expect to observe a violation of Hamilton's rule in a model with a binary choice and without coexistence, we would think that the violations of Hamilton's rule are there because of these three reasons. Notice that this implies that, while chance events may make for underestimations of the inclusive fitness in some cases, they will make for overestimations in others.

9.2 Adaptive dynamics

For the setup of Section 6, where a continuous trait is assumed to evolve, one could also ask what the empirical implications of the observations there would be. Suppose therefore that we do indeed observe a trait that can naturally be seen as continuous. This would be in line with a fair share of the inclusive fitness literature, such as, for example Taylor & Frank (1996) and Rousset & Billiard (2000), as well as with adaptive dynamics; see for example Metz et al. (1996), Dieckmann & Law (1996), and Champagnat et al. (2001, 2006, 2007).

A first question is, again, if we expect to see evolution in action, and watch a mutant for which rb is larger than c take over the population, or if we expect to observe a population in equilibrium. The first option would imply that we start in a disequilibrium state. In experiments one could get a disequilibrium state by manipulating either the phenotype of the species, for instance by knocking out some genes, or activating others, or one could manipulate the environment to change the selective pressure. Fig. 38a illustrates this; we would like to see a mutant with $rb > c$ appear and succeed. In this figure, the intersection of what is feasible (below the curved solid line) and which mutants would have increased inclusive fitness (above the straight dashed line) is non-empty.

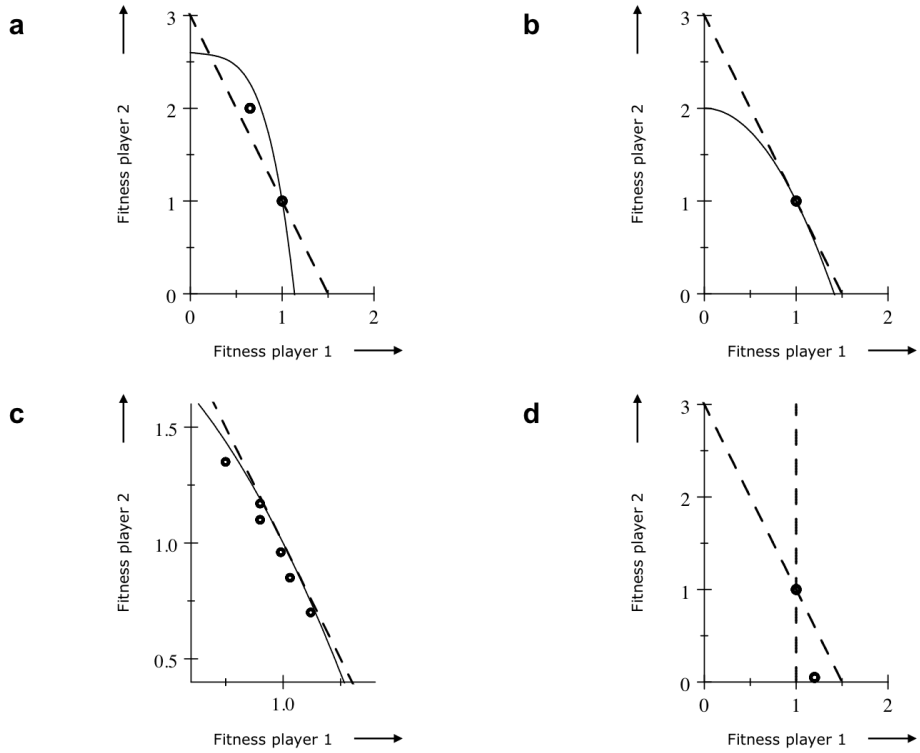


Figure 38 The dotted lines are lines with equal inclusive fitness through the fitnesses in the status quo (which are assumed to be 1 for both players). They separate mutations with an advantage from mutations with a disadvantage. The curved, solid lines separate feasible fitness effects and non-feasible fitness effects. (a) A mutant that is both feasible and that implies an increase in inclusive fitness. (b) An equilibrium state, where no increases in inclusive fitness are feasible. (c) A cloud of observations near equilibrium. (d) A binary choice. Eating a fellow brood member would have positive inclusive fitness when it concerns an unrelated individual (vertical line), but negative inclusive fitness when it is a sibling (downward sloping line).

Alternatively, one might want to assume that the trait is at its equilibrium value. In the cases of games 1 and 2 from Section 6, that would imply that the intersection of what is feasible and what would increase inclusive fitness is empty. This is illustrated by Fig. 38b. A possible empirical test of Hamilton's rule, with costs and benefits according to the counterfactual method, would then need an estimate of the slope of the curved line that separates the feasible from the infeasible phenotypes at the status quo. This slope should equal $-\frac{1}{r}$.[22] Of course, estimating this slope is a problem. One option could be that with a little diversity, one could hope to get a cloud of observations that might serve as a proxy for the trade off. But even if that would lead to an estimate of the slope that is significantly different from $-\frac{1}{r}$, this would not necessarily imply a violation of Hamilton's rule. Suppose that the estimated slope is found to be smaller than $-\frac{1}{r}$. That would suggest that there are traits that are more cooperative than the current ones that have a higher inclusive fitness. It could very well be, however, that this is really caused by the fact that it is easier to trace fitness effects between interactants than it is to pick up cancellation effects. Counterfactuals that pertain to the immediate effects between interactants are easier to establish, while counterfactuals that pertain to cancellation effects (which might consist of many small effects) are harder to pinpoint. If those cancellation effects are not picked up by the statistics, then one would overestimate the inclusive fitness of an increase in trait value.

If the observations are suggestive of a continuously differentiable fitness function, as they are in Fig. 38c, and especially if they suggest a convex set of feasible points, one would also have no reason to expect that Hamilton's rule would actually fail. Only if even locally, cooperative behaviours are strategic complements, or if diversity is too large to maintain the assumption that the population is close to monomorphic, would one expect Hamilton's rule, with costs and benefits according to the counterfactual method, to fail.

None of the empirical tests of Hamilton's rule that are reviewed by Bourke (2014) treat their cooperative trait as continuous. Many choices are also binary by definition. Pfennig et al., (1994, 1999) study cannibalizing larvae, and eating or not eating a fellow brood member is an all or nothing choice, because eating half a fellow brood member is not an interesting option. Such a case could provide a test of Hamilton's rule. One might first of all check qualitatively if they developed kin recognition, and whether kin are less likely to eat kin. One could however also check quantitatively whether the threshold matches what one would expect from Hamilton's rule (see Fig. 38d). Other traits may not be binary by definition; helping at the nest could also be a continuous trait, as the amount of help may

[22] The dotted lines in Fig 38 represent points with inclusive fitnesses equal to the point where both agent and interaction partner have fitness 1. Therefore, if π_1 represents the fitness, or payoff, of the agent, and π_2 the fitness of it's interaction partner, then the line is given by the equation $\pi_1 + r\pi_2 = 1 + r$. This can be rwritten as $\pi_2 = 1 + \frac{1-\pi_1}{r}$, which makes π_2 is a function of $\pi_1$with slope $-\frac{1}{r}$.

vary. One could, however, make a case for treating it as a binary choice if the amount of help provided has one peak at 0 and one at a different amount of help.

9.3 The cycle and the regression method

Suppose we observe a population of individuals that are organized on a cycle of length N. Each individual interacts with both neighbours. We moreover observe that there are two types only, and that there is one string of i consecutive individuals of type C and one string of $N-i$ consecutive individuals, all of type D. This puts us firmly within the setup of Ohtsuki et al. (2006a), Ohtsuki & Nowak (2006b), Grafen (2007b), and Section 7. We do, however, not know what the update process is; it could be Birth-Death, Death-Birth, or Shift. We would like to figure out, by looking at how the population changes from one moment in time to the next, which of these processes we are dealing with, and what the benefits and costs according to the regression method are. Because we assume that the game in payoff terms has equal gains from switching, we should expect that the game in fitness effect terms does too, and hence also that the counterfactual method will result in the same costs and benefits as the regression method.

Every reproduction event gives us a population at two moments in time, which we can feed into the regression method as discussed in Section 4. Let's say that variable x_1 now reflects the type of any individual itself; $x_1 = 1$ for a C-player, while being a D-player implies that $x_1 = 0$. Variable x_2 reflects the average type of the two direct neighbours – which means that it could be 0, $\frac{1}{2}$, or 1 – and variable x_m reflects the average type of the two neighbours $m-1$ steps away. The fitness f of an individual is 0 if it dies and does not reproduce, 2 if it reproduces and does not die, and 1 otherwise.

Looking at one reproduction event will not be very informative; one can of course apply the regression method, and it would give a b and a c, but which b and c that is, will depend on which reproduction event that happens to be. With one transition, that will be a very noisy signal, from which one cannot draw any conclusions. What one would need is very many observations, starting from the same population state, which would then have to be aggregated, such that we get a less noisy measure of how fitnesses of individuals depend on their own type and on the types of their neighbours. The most interesting positions on the cycle to look at will be the ones at or close to the boundaries, where C-players and D-players meet. Fig. 39 below shows the expected values of those fitnesses, if the underlying process is Birth-Death, Death-Birth, and Shift, respectively. These values are found simply by applying the fitness effects from Grafen (2007b); see also Table 1 in Section 7.

The easiest process is the Shift process. If we have a sample that is sufficiently large, we can be confident that the fitnesses are relatively close to the expected values from Fig. 39c. If we were to run the simple two-variable regression on these expected values, we would find $a_{00} = 1 - \frac{2i(b-c)}{N}$, $a_{10} = -2c$ and $a_{01} = 2b$. With a sufficiently large sample generated by the Shift process, the regression run on the data will be likely to return values close

to these. If we were to rerun the regression with more variables – for instance x_1, x_2 and x_3 instead of just x_1 and x_2 – then the fixed term as well as the coefficients of x_1 and x_2 would still be close to the same values, while the newly introduced coefficients would be close to 0. Notice that the expected values from Fig. 39c exactly match the description according to the regression model: $1-\frac{i}{N}2\left(b-c\right)=a_{00}$, $1-\frac{i}{N}2\left(b-c\right)+b=a_{00}+\frac{1}{2}a_{01}$ (where $\frac{1}{2}a_{01}$ represents the benefits of having one out of two neighbours cooperating), $1-\frac{N-i}{N}2\left(b-c\right)-b=a_{00}+a_{10}+\frac{1}{2}a_{01}$, and $1-\frac{N-i}{N}2\left(b-c\right)=a_{00}+a_{10}+a_{01}$.

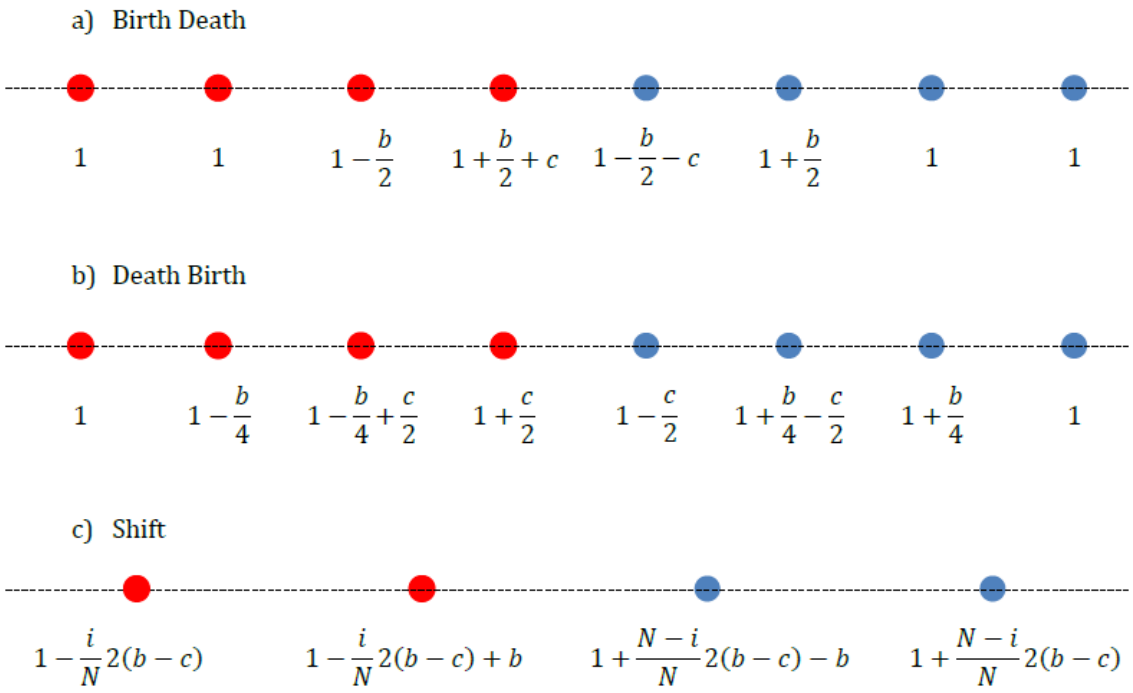


Figure 39. Fitnesses with different update rules. Red individuals are D-players, blue ones play C. Notice that the b and c's here are just parameter values, as in Ohtsuki et al. (2006a), Ohtsuki & Nowak (2006b), Grafen (2007b) and Section 7. Independent of the process, and according to the regression method, $r_m=\frac{Cov(X_1,X_m)}{Var(X_1)}=\frac{N(i+1-m)-i^2}{Ni-i^2}$ for $m<\min\left\{i,N-i\right\}$.

Running the two-variable regression on expected values of the Birth-Death process gives $a_{00}=1+b\frac{i}{N}\left(1-\frac{Ni-i^2}{Ni-i^2-N}\right)$, $a_{10}=-b-2c$ and $a_{01}=b\frac{Ni-i^2}{Ni-i^2-N}+2c$. In this case, however, if we rerun the regression method with variables x_1 to x_3 instead of just x_1 and x_2, we do find differences. While the coefficients of x_1 remains $-b-2c$, the coefficient of x_2 will change to $2b+2c$, and the coefficient of x_3 will become $-b$. After this, adding more variables does not induce more such changes, and the additional coefficients will be 0. Again, with a sufficiently large sample generated by the Birth-Death process, the regressions run on these data, instead of the expected values, will be likely to return values close to these.

One important thing to notice here is that we have now found *two* Hamilton's rules for the Birth-Death process. Both are derived with the regression method; one by using variables x_1 and x_2, and one by using x_1, x_2, and x_3. If we now return to the derivation of the result that Hamilton's rule *always* holds, in Section 4, we see that the same logic applies to both rules equally, even though they return different fitness effects; the effect on a neighbour once removed is $\frac{b}{2}\frac{Ni-i^2}{Ni-i^2-N}+c$ in the first, and $b+c$ in the second, while the effect on a neighbour twice removed is set to 0 in the first, and equal to $-\frac{b}{2}$ in the second.

The Death-Birth process combined with the regression method gives *three* Hamilton's rules; one for the regression that uses variables x_1 and x_2, one for the regression that uses x_1, x_2 and x_3, and one for the regression that uses x_1 to x_4. Again, the logic of the derivation of the result that Hamilton's rule *always* holds does not depend on which variables are used, as long as the fixed term and x_1 are included (see Section 4, equation 4.10).

It seems natural, though, to think that benefits and costs should be uniquely determined quantities, which would give us one Hamilton's rule per case only. The most natural choice seems to be to choose the regression with a fixed term and x_1 to x_4 for the Death-Birth process, the regression with a fixed term and x_1 to x_3 for the Birth-Death process, and the regression with a fixed term and x_1 and x_2 for the Shift process, because those are the smallest sets of variables that return the true benefits and costs as computed in Grafen (2007b).

We started out, however, with a situation where we did not know what the process is, and only have the data to infer that from. A natural thing to do here would therefore be to do a specification test. First we run a regression including x_1 to x_4, and do a statistical test on whether or not a_{0001} is zero. If a_{0001} is significantly different from 0 then one would conclude that the underlying process is Death-Birth. If not, then one would conclude that it must be one of the other two (assuming that the test has sufficient power). In that case, one would repeat this exercise with independent variables x_1 to x_3 and test if coefficient a_{0010} is significantly different from 0, which would then help decide between Birth-Death and Shift (again assuming that we have sufficiently many observations to give such a test sufficient power). These tests determine which variables should be included in the final regression, that will estimate the fitness effects, from which we then back out the parameters b and c. Notice that the setup with parameters b and c – not to be confused with actual fitness benefits and costs – implies that neighbours play a game with equal gains from switching in payoff terms. This translates to fitness effects that also satisfy equal gains from switching. This, in turn, implies that if we compute benefits and costs with the regression method, we will not be choosing a different specification than we would in a statistical exercise that is not a priori restricted to a linear model. The counterfactual method, combined with an unrestricted, standard statistical search for an appropriate model would therefore result in the same costs and benefits. There would be a difference between the two only if the game between neighbours does not satisfy equal gains from switching.

None of these considerations are anything out of the ordinary. The reason why it is still

worth discussing them is that they illustrate a point made in Section 4. The point there was that nothing in the derivation of the result – that Hamilton's rule always holds, provided that we interpret the regression coefficients as costs and benefits – depends on the choice of variables which are included in the function g, as long as g is assumed to be linear in all of them (see Section 4). That suggests that model choice might be irrelevant for the validity of Hamilton's rule. The example shows, however, that there is no avoiding model choices, if we want Hamilton's rule to be uniquely defined. The example also suggests that if f and μ represent data, a natural criterion would be a statistical test on coefficients a_{0001} and a_{0010}. This also implies that reducing the squared difference between f and g a whole lot by adding one variable only would be a relevant reason to include that variable (f represents the data here, and g the statistical model; see Section 4). There is, however, no difference between statistical tests for whether or not a_{0001} is different from 0, and statistical tests for whether a_{1100} or a_{2000} is non-zero or not. Those tests are not just conceptually the same; also the actual test is exactly the same for variable x_4 as it is for the variable x_1x_2 or the variable $(x_1)^2$. Therefore, when choosing between specifications, there is no reason to treat the question whether or not to include x_4 as any different from the question whether we should include x_1x_2 or $(x_1)^2$. The derivation of the result that Hamilton's rule always holds, however, crucially depends on x_1x_2 and $(x_1)^2$ not being included.

A few more small remarks are in order here. One is that average fitnesses are not the only type of useful information in the data. If we want to figure out which of the three processes we are looking at, then it might be worthwhile, and certainly more efficient, to also look at the variances in fitnesses. Cooperators and defectors in the Birth-Death as well as in the Death-Birth process all have an expected fitnesses of 1 if they are sufficiently far removed from the boundary. In the Death-Birth process, cooperators and defectors have the same birth rate and they also have the same death rate. In the Birth-Death process they do not; cooperators surrounded by sufficiently many cooperators have both a higher birth rate and a higher death rate than defectors. This means that if we wait for a fixed time interval, even though both have the same expected number of offspring (where still being alive oneself counts for a fitness of 1), the variance in the number of offspring in the Birth-Death process is larger for cooperators surrounded by at least two cooperators on either side than it is for defectors surrounded by at least two defectors on either side. More generally: a richer statistical model, which estimates birth and death rates, depending on variables x_1 to x_m, will give statistical tests with much more power, because only the reproduction events at or close to the boundaries count as informative if we depend on estimating a model with fitnesses, while every reproduction event is informative if we depend on estimating birth and death rates.

10 Discussion

Fifty years after the introduction of Hamilton's rule, its generality is still debated. The spectrum of positions stretches all the way from the claim that "*Hamilton's rule almost never holds*" (Nowak, Tarnita & Wilson, 2010) to inclusive fitness being "*as general as the genetical theory of natural selection itself*" (Abbot et al., 2010). The debate seems to be a disagreement about the validity of a well-defined, agreed upon rule. One key to the disagreement, however, is that there are different ways to define the benefits and costs in the rule (Birch, 2014, see also Birch & Okasha, 2015). In Section 4 we have seen that if the regression method is used to determine b and c, then indeed Hamilton's rule always holds, provided that we have a given, linear specification, and that we do not have an underdetermined system. In Section 3 we have seen that if we determine costs and benefits by comparing current fitnesses to what they would have been under alternative behaviour (the counterfactual method), then Hamilton's rule is only guaranteed to match the direction of selection if we assume "equal gains from switching", in which case both definitions result in the same b and c. Finally, some papers have parameters b and c determine the payoffs in a prisoners dilemma with equal gains from switching, and choose those for benefits and costs in Hamilton's rule (Ohtsuki et al., 2006, and the first version of Hamilton's rule in the SI, Part A.7, of Nowak et al., 2010). In this case, Hamilton's rule only applies if, on top of the assumption of equal gains from switching, payoffs translate linearly to fitnesses. This is not the case with many local interaction models (Grafen, 2007b). This third option – b and c as parameters – is not how we think benefits and costs should be defined. Hamilton's rule is about fitness effects, and not model parameters, and therefore we restrict attention to the first two options: b and c according to the regression method; and b and c according to the counterfactual method.

It might be helpful to realize that the difference in definitions drives the difference in claims concerning the generality of Hamilton's rule. That is not going to be the end of the debate, though, because the obvious next point of disagreement is which definition is preferable. Choosing between those two does leave room for individual preferences, and therefore for persisting disagreement. Some authors view the general validity of Hamilton's rule, with the regression method determining b and c, as a deep, fundamental insight (for instance Gardner et al., 2011, West & Gardner, 2013, Marshall, 2011, and Rousset, 2016). Our view is that it "makes" Hamilton's rule work by allowing for just the right kind of model misspecification. Sections 4 and 9.3 explains why we see it that way.

The regression method minimizes the squared difference between a fitness function (either reflecting a model, or observed numbers of offspring) and a linear function (which we, for lack of a better word, refer to as the statistical model). The central result here is that Hamilton's rule always holds, when costs are defined as the coefficient of the variable that represents the individual's own level of cooperation (which is 0 or 1 in models with binary choice), and when benefits to interaction partners in different roles (e.g., siblings,

nephews and nieces, nearby neighbours, faraway neighbours) are defined as the coefficients of the variables that represent the levels of cooperation of the interaction partners in those roles. One observation we make in Section 4 is that the derivation of that result does not assume anything about the specification of the statistical model – that is, it is silent about which variables are to be included in the regression and which are not. With the regression method, Hamilton's rule therefore is not necessarily uniquely defined, as the costs and benefits of the cooperative behaviour may depend on which variables are included. The benefits of having a cooperative sister as computed by the regression method, for instance, may depend on whether or not the level of cooperation of nieces is included as a variable.

In order to overcome the problem that Hamilton's rule is not uniquely defined, one could add a criterion for model choice to the minimization of the squared difference. A natural criterion would be: all the variables that play a role should be included, and not more. If the fitness function is a theoretical model, or follows from one, then that criterion is relatively straightforward to apply. If it reflects data, that requires statistical testing. In neither of the two cases is there a reason why that criterion should apply when we choose whether or not to include different *linear* terms in the model (such as, for instance, the cooperativity of one's nephews or nieces), but not when we choose whether or not to include non-linear terms (such as, for instance, the interaction term between my cooperativity and my sibling's). Also, the statistical test for those two choices is one and the same. The general validity of Hamilton's rule, however, depends on all non-linear terms not being included. This implies that either model choice does matter, in which case non-linear terms should allowed to be included, and Hamilton's rule does not generally hold, or model choice does not matter, in which case Hamilton's rule is not always uniquely defined.

The counterfactual method is also not without complications. Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) compute the costs and benefits of cooperation by going over all cooperators, and comparing their current fitness to what their fitness would have been, had they defected. That does have an intuitive appeal, given that it is the actual cooperators that, in Hamilton's words, "*add to the gene-pool a handful of genes containing* G [the altruistic gene] *in higher concentration than does the gene-pool itself*". A problem with this definition, is that the inclusive fitness of a cooperator is not necessarily minus the inclusive fitness of a defector. A choice that we prefer, and that solves this inconsistency, is to consider everyone, cooperators and defectors. A cooperator then actually incurred those costs, and provided the benefits, and a defector faced them too, but acted differently.

Karlin & Matessi (1983) and Matessi & Karlin (1984, 1986) moreover imposed that costs and benefits should be fixed, and independent of the current frequency of cooperators. They found that Hamilton's rule only applies if fitnesses are linear in the frequency of cooperators, which translates directly to equal gains from switching. We allow for a local definition, where costs and benefits are allowed to change with the frequency of cooperators, but find that the same restriction still applies; also locally, Hamilton's rule applies only with equal gains from switching, provided that we define b and c with either version of the

counterfactual method.

For games with equal gains from switching – that is, in the absence of synergy or the opposite of synergy – the regression method and both versions of the counterfactual method lead to the same costs and benefits. Hamilton's rule then applies, whichever way we define costs and benefits. Equal gains from switching can be a basic assumption of a model, or it can be implied by other choices, such as local mutations in combination with restrictions on the fitness function. In Section 6 we have seen that in an adaptive dynamics context, where we assume local and infrequent mutations, this results in a relatively large domain where inclusive fitness works, with costs and benefits defined according to the counterfactual method. If we have a differentiable fitness function, and no bifurcations, then locally we regain equal gains from switching, and therefore dynamics according to Hamilton's rule.

Birch (2014) also compares different ways to define the benefits and costs in Hamilton's rule. He distinguishes the general version of Hamilton's rule (HRG) from the special version (HRS). The first version uses the regression method to define b and c. The second version uses payoff parameters, which is to capture how b and c are defined in Nowak, Tarnita & Wilson (2010) as well as in Van Veelen (2009). These two papers however differ in their treatment of b and c. In the Hamilton's rule in Part A.7 of the Supplementary Information of Nowak, Tarnita & Wilson (2010), the b and c are indeed parameters that determine the payoffs in a prisoners dilemma with equal gains from switching, and it is those parameters that are indeed considered to be the benefits and costs in Hamilton's rule. This is also done in Ohtsuki et al., (2006), and Grafen (2007b) pointed out that these *payoffs* do not reflect the *fitness effects* of playing C instead of D in the local interaction model from Ohtsuki et al.,(2006). Since the b and c in Hamilton's rule should represent fitness effects, Grafen (2007b) argued that it is not correct to use those parameters instead.

In Van Veelen (2009), on the other hand, the translation from payoff to fitness is not a problem, as the fitness effects there by definition align with the payoffs (see also Van Veelen, 2011b, and Section 3). Moreover, no specific choice for the b and c in Hamilton's rule is made there. All that the counterexamples (pp. 594–595) do, is show that there exists no b and c that are independent from the current population state, and that combine with r to a Hamilton's rule that matches the direction of selection for any frequency. One can therefore say that Van Veelen (2009), like Karlin & Matessi (1983), was looking for a *global* rule, with fixed, frequency-independent costs and benefits. Without stating what the proper definition of b and c would be, it showed that no choice for b and c would produce such a global rule, unless the game has (generalized) equal gains from switching.

In this paper we did allow for b and c to depend on the current population state. Here we found that even if we allow for Hamilton's rule to be a *local* rule, with frequency-dependent b and c, equal gains from switching is required for it to work, if we choose the b and c according to the counterfactual method. Finally, between the counterfactual and the regression method, neither one of the two is more special or more general than the other. For any given game and population state, both methods will simply produce a b and a c.

10.1 Helping behaviour, kin selection, and inclusive fitness

The discussion about kin selection and inclusive fitness is sometimes also clouded by a lack of distinction between kin selection and inclusive fitness (see for example Foster et al., 2006a,b, Nowak et al., 2010, and Birch & Okasha, 2015). We would like to stress that kin selection and inclusive fitness are not the same. The replicator dynamics and the adaptive dynamics typically have equilibrium outcomes that depend on relatedness. If with positive relatedness the outcome differs from what we get at $r = 0$, then it is safe to say that this is the result of kin selection. But, absent equal gains from switching, inclusive fitness may not point to the equilibrium outcome, if we let b and c be determined by the counterfactual method. That was more or less a recurrent theme in this paper. But also the opposite is possible, for instance if positive assortment is not caused by common descent. Individuals can for instance also self-assort on phenotypic similarity (Fletcher & Doebeli, 2009). When the game has equal gains from switching, Hamilton's rule, with $r = \mathbb{P}(C|C) - \mathbb{P}(C|D)$ as a measure for assortment, can still hold, even though this is not kin selection (see Section 3 for the definition of r).

It is also useful to stress that not all explanations for helping behaviour are kin selection explanations or even rely on other sources of assortment. If someone jumps into the river to save her sister, then that can be explained by a kin selection model. If someone jumps into the river to save a friend, then that can be explained by a model in which the value of saving the other lies in the fact that the other is unable to return the favour when dead, as suggested in Eshel & Motro (1981). In settings that are less all-or-nothing, helping behaviour can also be explained with classical repeated interaction models, where only the willingness to help is affected, and not the capacity. Finally, sexual selection models can also explain helping behaviour. If someone jumps into the river to save an unrelated stranger and ends up producing healthy offspring happily ever after with a top quality partner who happened to have witnessed the act of bravery, or gets to spend time with more partners than otherwise, then sexual selection can explain that (Miller, 2001).

In the latter two examples, Hamilton's rule is not necessarily violated, although it is also not the condition to look to for an answer to the question whether or not helping behaviour will evolve. In a typical model of sexual selection, the only thing that matters is whether or not behaviour promotes the fitness of the one that performs it. The one that gets saved is just lucky, and typically not assumed to be related to its saviour. With relatedness 0 to the helped individual, a behaviour that serves the actor well through better mating chances trivially has positive inclusive fitness. But whether or not helping behaviour can evolve depends crucially on whether or not the fitness cost function satisfies the 'single crossing' condition (Kreps & Sobel, 1994, see also Zahavi, 1997). Including the effect on the helped individual and weighing that with 0 does not add to this.

The situation with repeated interactions is somewhat similar. There typically is a multiplicity of equilibria. With relatedness 0, none of the equilibria violates Hamilton's

rule; at all of them, deviating would be bad for the one who deviates, and that would trivially reduce its inclusive fitness. But Hamilton's rule does not help finding equilibria, nor does it provide assistance in determining if some equilibria are perhaps more stable or more likely to be played than others (some references for repeated games are Friedman, 1971, Axelrod & Hamilton, 1981, Boyd & Lorberbaum, 1987, Fudenberg & Maskin, 1986, 1990, Binmore & Samuelson, 1992, Bendor & Swistak, 1995, Cooper, 1996, Volij, 2002, and Imhof, Fudenberg & Nowak, 2005. For a general analysis with relatedness, see Van Veelen, García, Rand & Nowak, 2012).

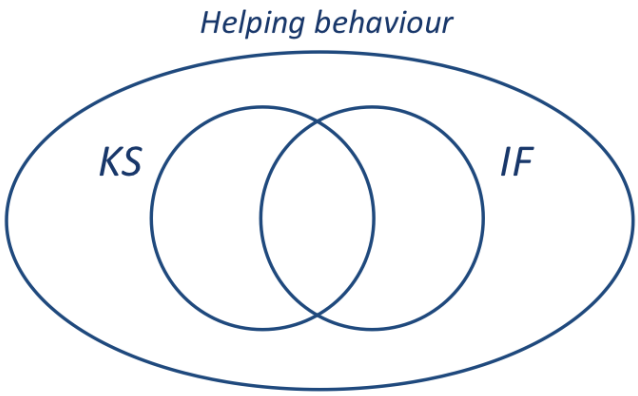


Figure 40. Not all models that explain the evolution of helping behaviour are kin selection models. Sexual selection, or signalling models more generally, may explain helping behaviour. Models of reciprocity with repeated interactions may too. These do not necessarily violate inclusive fitness, but Hamilton's rule does not provide the relevant criterion. Not all models that explain the evolution of helping behaviour have a prediction that follows Hamilton's rule; not even all kin selection ones – unless we use the regression method to determine costs and benefits.

10.2 Understanding cancellation effects is a major step ahead

In Section 7 we have seen that being related is not enough for cooperation to evolve. What matters is that there is a *discrepancy* between with whom there are opportunities for cooperation and with whom there is competition. Cooperation and altruism evolves when these two do not coincide – and of course they have to not coincide the right way; if the discrepancy is in the other direction, spite can evolve (see for instance Boyd, 1982). The setting in Sections 3, 5 and 6 is actually an explicit way to break the symmetry, as interactions in which the games are being played are decoupled from competition, both in the replicator dynamics with population structure, and in the adaptive dynamics with

population structure. Even in Hamilton's original paper, now that we have this new insight, it is true that in order for his setup to be valid, the fitness effects should be interpreted as final, and not imply uneven in- or decreases in competition. Kin recognition therefore is a good fit with those models, and a good tool to break the symmetry. Once past a phase where for instance siblings also compete more intensely for parental attention and resources, they may no longer compete any more intensely with each other than with anyone else. Kin that seek each other out for cooperation therefore will break the symmetry in the exact same way as the stylized replicator dynamics do (see Section 3; see also Lieberman, Tooby &Cos mides, 2003, 2007 for interesting papers on kin recognition in humans). Some of the empirical examples in Section 9 also concern kin recognition. Also life cycles where opportunities for cooperation occur in a phase that is different from the one in which competition happens can help breaking the symmetry. In some settings, one can combine interaction and competition effects in one "effective" payoff matrix (see for instance Lessard, 1997). The insight that local competition can (partially) cancel out local opportunities for cooperation, as described by Wilson, Pollock and Dugatkin (1992) and Taylor (1992a,b) might very well be the most important refinement of our understanding of kin selection since Hamilton's 1964 paper.

10.3 Empirical tests of Hamilton's rule

Hamilton's rule will by definition never be violated if the regression method is used to compute costs and benefits. Therefore, whatever the specification, there is no scope for empirical testing with the regression method. If we use the counterfactual method instead, then there is scope for violations, and hence for empirical testing. It is important, though, to know what a violation would look like. In Section 9 we have seen that in a setting with the replicator dynamics the particulars of the mismatch between Hamilton's rule and the direction of selection imply that we cannot observe violations of Hamilton's rule in a pure equilibrium, whether it consists of cooperators only or of defectors only. Violations can only be observed out of equilibrium, as selection happens, or in equilibrium, if that equilibrium is a polymorphism. In either case the statistics should allow for non-linear terms in order to be able to detect a violation.

References

[1] Abbot, P. et al., 2010. Inclusive fitness theory and eusociality. Nature 471, E1–E4

[2] Ajar, 2003. Analysis of disruptive selection in subdivided populations. BMC Evolutionary Biology 3:22.

[3] Akçay, E., Van Cleve, J., Feldman, M.W., Roughgarden, J., 2009. A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proc. Natl. Acad. Sci. USA, 106 (45), 19061–19066.

[4] Akçay, E., Van Cleve, J., 2012. Behavioral responses in structured populations pave the way to group optimality. Am. Nat. 179, No. 2, 257–269.

[5] Alger, I, Weibull, J.W., 2012. A generalization of Hamilton's rule–Love others how much? J. Theor. Biol. 299, 42–54.

[6] Allen, B. Nowak, M.A., Dieckmann, U., 2013. Adaptive dynamics with interaction structure. Am. Nat. 181(6), E139–E163.

[7] Allen, B. Nowak, M.A., 2012, Evolutionary shift dynamics on a cycle. J. Theor. Biol. 311, 28–39.

[8] Allen, B. Nowak, M.A., 2015, Games among relatives revisited. J. Theor. Biol. 378, 103–116.

[9] Allen, B., Nowak, M.A., Wilson, E.O., 2013. Limitations of inclusive fitness. Proc. Natl. Acad. Sci. USA, 110 (50), 20135–20139.

[10] Archetti, M., Scheuring, I., 2011. Coexistence of cooperation and defection in public goods games. Evolution, 65, 1140–1148.

[11] Archetti, M., Scheuring, I., 2012. Review: Game theory of public goods in one-shot social dilemmas without assortment. J. Theor. Biol. 299, 9–20.

[12] Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.

[13] Barton, N.H., Etheridge, A.H., 2011. The relation between reproductive value and genetic contribution. Genetics 188(4), 953–973.

[14] Barton, N.H., Polechová, J., 2005. The limitations of adaptive dynamics as a model of evolution. J. Evol. Biol. 18, 1186-1190.

[15] Bendor, J., Swistak, P., 1995. Types of evolutionary stability and the problem of cooperation. Proc. Natl. Acad. Sci. USA 92, 3596-360.

[16] Benford, F.A., 1978. Fisher's theory of the sex ratio applied to social hymenoptera. J. Theor. Biol, 72, 701–727.

[17] Binmore, K.G., 1994. *Game Theory and the Social Contract, Volume I: Playing Fair*. Cambridge, MA: MIT Press.

[18] Binmore, K.G., Samuelson, L., 1992. Evolutionary stability in repeated games played by finite automata. J. Econ. Theory 57, 278–305.

[19] Birch, J., 2014. Hamilton's rule and its discontents. Brit. J. Phil. Sci. 65, 381–411.

[20] Birch, J., Okasha, S., 2015. Kin selection and its critics. BioScience 65 (1), 22–32.

[21] Boyd, R., 1982. Density-dependent mortality and the evolution of social interactions. Anim. Beh. 30, 972–982.

[22] Boyd, R., Lorberbaum, J.P., 1987. No pure strategy is stable in the repeated prisoner's dilemma game. Nature 327, 58–59.

[23] Bourke, A.F.G., 1997. Sociality and kin selection in insects. In Behavioural ecology: an evolutionary approach (eds JR Krebs, NB Davies), pp. 203–227, 4th edn. Oxford, UK: Blackwell Science Ltd.

[24] Bourke. A.F.G., 2014. Hamilton's rule and the causes of social evolution. Phil. Trans. R. Soc. B. 369, 20130362

[25] Brännström, Å, Johansson, J, von Festenberg, N., 2013. The hitchhiker's guide to Adaptive Dynamics, Games, 4(3), 304–328.

[26] Castilloux, A.-M., Lessard, S., 1995. The fundamental theorem of natural selection in Ewens' sense (Case of many loci), Theor. Popul. Biol. 48, 306–315.

[27] Champagnat, N., Ferrière, R., Ben Arous, G., 2001. The canonical equation of adaptive dynamics: a mathematical view. Selection 2, 73–83.

[28] Champagnat, N., Ferrière, R., Méléard, S., 2006. Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models. Theor. Popul. Biol. 69, 297–321.

[29] Champagnat, N. Lambert. A., 2007. Evolution of discrete populations and the canonical diffusion of adaptive dynamics. Ann. Appl. Probab. 17(1), 102–155.

[30] Champagnat, N. Méléard, S., 2007. Invasion and adaptive evolution for individual-based spatially structured populations. J. Math. Biol. 55, 147–188.

[31] Cooney, D., Veller, C., 2015. Assortment and the evolution of cooperation in a Moran process with exponential fitness. arXiv:1509.05757 [q-bio.PE]

[32] Cooper, D.J.. 1996. Supergames played by finite automata with finite costs of complexity in an evolutionary setting J. Econ. Theory 68 (1), 266–275.

[33] Charlesworth, B., 1980. *Evolution in Age-Structured Populations.* Cambridge Univ. Press, Cambridge

[34] Darwin, C., 1859. *On the Origin of Species by Means of Natural Selection.* London: John Murray (Reprinted in 1964 by Harvard University Press)

[35] Dawkins, R., 1976. *The selfish gene.* Oxford: Oxford University Press.

[36] Dekel, E., Ely, C.E., Yilankaya, O., 2007. Evolution of preferences. Rev. Econ. Stud. 74, 685–704.

[37] Dieckmann, U., Law, R., 1996. The dynamical theory of coevolution: A derivation from stochastic ecological processes. J. Math. Biol. 34, 579–612.

[38] Doebeli, M., Hauert, C., 2006. Limits to Hamilton's rule. J. Evol. Biol. 19(5), 1386–1388.

[39] Doebeli, M., Hauert, C., Killingback, T., 2004. The evolutionary origin of cooperators and defectors. Science 306, 859–862.

[40] Emlen, S.T., Wrege, P.H., 1989. A test of alternate hypotheses for helping behavior in white-fronted bee-eaters of Kenya. Behav. Ecol. Sociobiol. 25, 303–319.

[41] Eshel, I., Motro, U., 1981. Kin selection and strong evolutionary stability of mutual help. Theor. Pop. Biol. 19, 420–433.

[42] Ewens, W.J., 1989. An interpretation and proof of the fundamental theorem of natural selection. Theor. Pop. Biol. 36 (2), 167–180.

[43] Ewens, W.J., 1992. An optimizing principle of natural selection in evolutionary population genetics. Theor. Pop. Biol. 42 (3), 333–346.

[44] Fisher, R.A., 1930. *The Genetical Theory of Natural Selection* Oxford Univ. Press (Clarendon), London (Reprinted and revised, 1958)

[45] Fletcher, J. A., Doebeli, M. 2009. A simple and general explanation for the evolution of altruism. Proc. R. Soc. B 276, 13–19.

[46] Foster, K.R., Wenseleers, T., Ratnieks, F.L.W., 2006. Kin selection is the key to altruism. Trends Ecol. Evol. 21 (2), 57–60.

[47] Foster, K.R., Wenseleers, T., Ratnieks, F.L.W., Queller, D.C., 2006. There is nothing wrong with inclusive fitness. Trends Ecol. Evol. 21 (11), 599–600.

[48] Frank, S.A., 1986. The genetic value of sons and daughters. Heredity 56, 351–354.

[49] Frank, R.H., 1987. If Homo Economicus could choose his own utility function, would he want one with a conscience? Am. Econ. Rev. 77 (4) 593–604.

[50] Frank, R.H., 1988. *Passions within Reason: The Strategic Role of the Emotions*, W.W. Norton, New York.

[51] Friedman, J., 1971. A noncooperative equilibrium for supergames. Rev. Econ. Stud. 38, 1–12.

[52] Fudenberg, D., Maskin, E., 1986. The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54, 533–554.

[53] Fudenberg, D., Maskin, E., 1990. Evolution and cooperation in noisy repeated games. Am. Econ. Rev. 80 (2), 274–279.

[54] Le Galliard, J.-F., Ferrière, R., Dieckmann, U., 2003. The adaptive dynamics of altruism in spatially heterogeneous populations. Evolution 57,1–17.

[55] Le Galliard, J.-F., Ferrière, R., Dieckmann, U., 2005. Adaptive evolution of social traits: origin, trajectories, and correlations of altruism and mobility. Am. Nat. 165(2), 206–224.

[56] Gadagkar, R., 2001. The social biology of Ropalidia marginata: towards understanding the evolution of eusociality. Cambridge, MA: Harvard University Press.

[57] García, J., van Veelen, M., 2016. In and out of equilibrium I: Evolution of strategies in repeated games with discounting. J. Econ. Theory 161, 161–189.

[58] García, J., van Veelen, M., Traulsen, A., 2014. Evil green beards: Tag recognition can be used to withhold cooperation in structured populations. J. Theor. Biol. 360, 181–186.

[59] Gardner, A., West, S.A., Barton, N.H., 2007. The relation between multilocus population genetics and social evolution theory. Am. Nat. 169, 207–226.

[60] Gardner, A., West, S.A., Wild, G., 2011. The genetical theory of kin selection. J. Evol. Biol. 24, 1020–1043

[61] Gokhale, C.S., Traulsen, A. 2010. Evolutionary games in the multiverse. Proc. Natl. Acad. Sci. USA 107, 5500–5504.

[62] Gokhale, C.S., Traulsen, A. 2011 Mutation-selection equilibrium in evolutionary games with multiple players and multiple strategies. J. Theor. Biol. 283,180–191.

[63] Gorrell, J.C., McAdam, A.G., Coltman, D.W., Humphries, M.M., Boutin, S., 2010. Adopting kin enhances inclusive fitness in asocial red squirrels. Nat. Commun. 1, 22.

[64] Geritz, S.A.H., Kisdi, É., Meszéna, G., Metz, J.A.J., 1998. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. Evol. Ecol. 12, 35–57.

[65] Grafen, A. 1985a. A geometric view of relatedness. Oxford Surveys in Evolutionary Biology 2, 28–90.

[66] Grafen, A. 1985b. Hamilton's rule OK. Nature 318, 310–311.

[67] Grafen, A., 1986. Split sex ratios and the evolutionary origins of eusociality. J. Theor. Biol. 122, 95–121.

[68] Grafen, A., 2004. Willam Donald Hamilton. Biographical memoirs of fellows of the Royal Society, 50, 109–132.

[69] Grafen, A. 2006a. Optimization of inclusive fitness. J. Theor. Biol. 238, 541–563.

[70] Grafen, A. 2006b. A theory of Fisher's reproductive value. J. Math. Biol. 53, 15–60.

[71] Grafen, A. 2007a. Detecting kin selection at work using inclusive fitness. Proc. R. Soc. B 274, 713–719.

[72] Grafen, A. 2007b. An inclusive fitness analysis of altruism on a cyclical network. J. Evol. Biol. 20, 2278–2283.

[73] Grafen, A. 2009. Formalizing Darwinism and inclusive fitness theory. Phil. Trans. R. Soc. B 364, 3135–3141.

[74] Güth, W., 1995. An Evolutionary approach to explaining cooperative behaviour by reciprocal incentives, Int. J. Game Theory, 24, 323–344.

[75] Güth, W., Yaari, M., 1992. Explaining reciprocal behaviour in simple strategic games: an evolutionary approach. In: *Explaining process and change: Approaches to evolutionary economics*. Ed: Witt, U., Michigan University Press, Michigan, 23–34.

[76] Hamilton, W.D., 1963. The evolution of altruistic behavior. Am. Nat. 97 (896), 354–356.

[77] Hamilton, W.D., 1964a. The genetical theory of social behaviour I. J. Theor. Biol. 7, 1–16.

[78] Hamilton, W.D., 1964b. The genetical theory of social behaviour II. J. Theor. Biol. 7, 17–32.

[79] Hamilton, W.D., 1970. Selfish and spiteful behaviour in an evolutionary model. Nature 228, 1218–1220.

[80] Hamilton, W.D., 1972. Altruism and related phenomena, mainly in social insects. Annual Review of Ecological Systems 3, 193-232.

[81] Hamilton, W. D. 1975. Innate social aptitudes of man: Approach from evolutionary genetics. Pages 133–155 *in* Fox, R., ed. *Biosocial Anthropology*, New York: Wiley.

[82] Hatchwell, B.J., Gullett, P.R., Adams, M.J., 2014. Helping in cooperatively breeding long-tailed tits: a test of Hamilton's rule. Phil. Trans. R. Soc. B 369, 20130565.

[83] Hauert, C., De Monte, S., Hofbauer, J., Sigmund, K., 2002. Volunteering as red queen mechanism for cooperation in public goods games. Science 296, 1129–1132.

[84] Hofbauer, J., Sigmund, K., 1990. Adaptive dynamics and evolutionary stability. Appl. Math. Letters 3, 75–79.

[85] Hogendoorn, K., Leys, R., 1993. The superseded female's dilemma: ultimate and proximate factors that influence guarding behaviour of the carpenter bee Xylocopa pubescens. Behav. Ecol. Sociobiol. 33, 371–381.

[86] Hölldobler, B., Wilson, E.O., 2009. *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies*, New York: W. W. Norton.

[87] Van Huyck, J.B., Battalio, R.C., Beil, R.O., 1990. Tacit coordination games, strategic uncertainty, and coordination failure. Am. Econ. Rev. 80(1), 234–248.

[88] Imhof, L.A., Fudenberg, D., Nowak, M.A., 2005. Evolutionary cycles of cooperation and defection. Proc. Natl. Acad. Sci. USA 102 (31),10797–10800.

[89] Iosifescu, M., 1980. *Finite Markov Processes and Their Applications*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1980.

[90] Jansen, V.A.A., Van Baalen, M., 2006. Altruism through beard chromodynamics. Nature 440, 663–666.

[91] Karlin, S., Matessi, C., 1983. The Eleventh R.A. Fisher Memorial Lecture: Kin selection and altruism. Proc. R. Soc. Lond. B 219, 327–353.

[92] Keller, L., Ross, K.G. 1998. Selfish genes: a green beard in the red fire ant. Nature 394, 573–575.

[93] Krakauer, A.H., 2005. Kin selection and cooperative courtship in wild turkeys. Nature 434, 69–72.

[94] Kreps, D.M., Sobel, J., 1994. Handbook of Game Theory with Economic Applications 2, 849–867

[95] Kimura, M., 1964. Diffusion models in population genetics. J. App. Prob. 1, 177–232.

[96] Kingman, J.F.C., 1961a. On an inequality in partial averages. Quart. J. Math. Oxford 12 (1), 78–80.

[97] Kingman, J.F.C., 1961b. A mathematical problem in population genetics. Math. Proc. Camb. Phil. Soc. 57 (3), 574–582

[98] Kurokawa, S., Ihara, Y., 2009. Emergence of cooperation in public goods games. Proc. R. Soc. B, 276, 1379–1384.

[99] Lehmann L., 2012. The stationary distribution of a continuously varying strategy in a class-structured population under mutation-selection-drift balance. J. Evol. Biol. 25(4), 770–787.

[100] Leslie, P.H., 1948. Some further remarks on the use of matrices in population mathematics. Biometrika 35, 213–245.

[101] Lessard, S., 1997. Fisher's Fundamental Theorem of Natural Selection revisited. Theor. Pop. Biol. 52, 119–136.

[102] Lessard, S., Castilloux, A.–M., 1995. The Fundamental Theorem of Natural Selection in Ewens' sense: case of fertility selection. Genetics 141, 3–42.

[103] Lessard, S., 2011. Effective game matrix and inclusive payoff in group-structured populations. Dyn. Games Appl. 1, 301–318.

[104] Lieberman E., Hauert, C., Nowak, M.A., 2005. Evolutionary dynamics on graphs. Nature 433, 312–316.

[105] Lieberman, D., Tooby, J., Cosmides, L., 2003. Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. Proc. R. Soc. B 270 (1517), 819–826.

[106] Lieberman, D., Tooby, J., Cosmides, L., 2007. The architecture of human kin detection. Nature 445, 727–731.

[107] Loeb, M.L.G., 2003. Evolution of egg dumping in a subsocial insect. Am. Nat. 161, 129–142.

[108] Luo, S., 2014. A unifying framework reveals key properties of multilevel selection. J. Theor. Biol. 341, 41–52.

[109] Maciejewski, W., 2014. Reproductive value on evolutionary graphs. J. Theor. Biol. 340, 283–293.

[110] Matessi, C., Karlin, S. 1984. On the evolution of altruism by kin selection. Proc. Natl. Acad. Sci. USA 81, 1754–1758.

[111] Matessi, C., Karlin, S. 1986. Altruistic behavior in sibling groups with unrelated intruders. In: Karlin, S., Nevo, E. (eds.) *Evolutionary Process and Theory*, Orlando, Fla.: Academic Press, 689–724.

[112] Metcalf, R.A., Whitt, G.S., 1977. Relative inclusive fitness in the social wasp Polistes metricus. Behav. Ecol. Sociobiol. 2, 353–360.

[113] Metz, J.A.J., Geritz, S.A.H., Meszéna, G., Jacobs, F.J.A., van Heerwaarden, J.S., 1996. Adaptive dynamics: A geometrical study of the consequences of nearly faithfull reproduction. In: S.J. van Strien & S.M. Verduyn-Lunel (eds.) *Stochastic and spatial structures of dynamical systems*. North Holland, Elsevier, 183–231.

[114] Milchtaich, I., 2006. Comparative statics of games between relatives. Theor. Pop. Biol. 69, 203–210.

[115] Milinski, M., Sommerfeld, R.D., Krambeck, H.-J., Reed, F. A., J. Marotzke., 2008. The collective-risk social dilemma and the prevention of dangerous climate change. Proc. Natl. Acad. Sci. USA 105, 2291–2294.

[116] Miller, G., 2001. *The mating mind; how sexual choice shaped the evolution of human nature.* New York: Anchor Books.

[117] Mulholland, H.P., Smith, C.A.B., 1959. An inequality arising in genetical theory Am. Math. Monthly 66 (8), 673-683.

[118] Noonan, K.M., 1981. Individual strategies of inclusive fitness-maximizing in Polistes fuscatus foundresses. In: Natural selection and social behavior (eds RD Alexander, DW Tinkle), pp. 18–44. New York, NY: Chiron Press.

[119] Nonacs, P., Reeve, H.K., 1995. The ecology of cooperation in wasps: causes and consequences of alternative reproductive decisions. Ecology 76, 953–967.

[120] Nowak, M.A., 2006. *Evolutionary dynamics: exploring the equations of life*. Harvard University Press, Cambridge, MA.

[121] Nowak, M.A., Sasaki, A., Taylor, C., Fudenberg, D., 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428, 646–650.

[122] Nowak, M.A., Sigmund, K., 1990. The evolution of stochastic strategies in the prisoner's dilemma. Acta Appl. Math. 20, 247–265.

[123] Okasha, S., 2016. On Hamilton's rule and inclusive fitness theory with non-additive payoffs. Phil. Sci, *to appear.*

[124] Okasha S, Martens J. 2016a. The causal meaning of Hamilton's rule. R. Soc. Open Sci. 3, 160037.

[125] Okasha S, Martens J. 2016b. Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. J. Evol. Biol. 29, 473–82.

[126] Ohtsuki, H., 2012. Does synergy rescue the evolution of cooperation? – An analysis for homogeneous populations with non-overlapping generations. J. Theor. Biol. 307, 20–28.

[127] Ohtsuki H., Hauert, C., Lieberman, E., Nowak, M.A., 2006. A simple rule for the evolution of cooperation on graphs and social networks. Nature 441: 502–505.

[128] Ohtsuki H., Nowak, M.A., 2006. Evolutionary games on cycles. Proc. R. Soc. B 273, 2249–2256.

[129] Pacheco, J.M., Santos, F.C., Souza, M.O., Skyrms, B., 2009. Evolutionary dynamics of collective action in N-person stag hunt dilemmas. Proc. R. Soc. B 276 (1655), 315–321.

[130] Pamilo, P., Crozier, R.H., 1982. Measuring genetic relatedness in natural populations: Methodology. Theor. Pop. Biol. 21, 171–193.

[131] Pfennig D.W., Collins, J.P., Ziemba, R.E., 1999. A test of alternative hypotheses for kin recognition in cannibalistic tiger salamanders. Behav. Ecol. 10, 436–443.

[132] Price, G.R., 1970. Selection and covariance. Nature 227 (5257), 520–521.

[133] Price, G.R., 1972a. Extension of covariance selection mathematics. Annals of Human Genetics 35, 485–490.

[134] Price, G.R., 1972b. Fisher's fundamental theorem made clear. Annals of Human Genetics 36, 129–140.

[135] Proulx, S.R., Day, T., 2001. What can invasion analyses tell us about evolution under stochasticity in finite populations? Selection: Molecules, Genes, and Memes, 2,2–15.

[136] Queller, D.C., 1985. Kinship, reciprocity and synergism in the evolution of social behaviour. Nature 318, 366–367.

[137] Queller, D.C., 1992a. A general model for kin selection. Evolution 46 (2), 376–380.

[138] Queller, D.C., 1992b. Quantitative genetics, inclusive fitness, and group selection. Am. Nat. 139 (3), 540–558.

[139] Queller, D.C., 2011.Expanded social fitness and Hamilton's rule for kin, kith, and kind. Proc. Natl. Acad. Sci. USA 108, 10792–10799.

[140] Queller, D.C., Strassmann, J.E., 1988. Reproductive success and group nesting in the paper wasp, Polistes annularis. In Reproductive success (ed. T.H. Clutton-Brock), pp. 76–96. Chicago, IL: University of Chicago Press.

[141] Richards, M.H., French, D., Paxton, R.J., 2005. It's good to be queen: classically eusocial colony structure and low worker fitness in an obligately social sweat bee. Mol. Ecol. 14, 4123–4133.

[142] Richerson, P.J., Boyd, R., 2004. *Not by genes alone; how culture transformed human evolution.* University of Chicago Press, Chicago.

[143] Riolo, R.L., Cohen, M.D., Axelrod, R., 2001. Evolution of cooperation without reciprocity. Nature 414, 441–443.

[144] Robson, A.J., 1990. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. J. Theor. Biol. 144, 379–396.

[145] Robson, A.J., 2001. The biological basis of economic behavior. J. Econ. Lit. 29, 11–33

[146] Robson, A.J., Samuelson, L., 2011. The evolutionary foundations of preferences. Handbook of Social Economics 1, 221–310.

[147] Rousset, F., 2004. *Genetic Structure and Selection in Subdivided Populations.* Princeton University Press, Princeton, NJ

[148] Rousset, F., 2015. Regression, least squares, and the general version of inclusive fitness. Evolution 69, 2963–2970.

[149] Rousset, F., Billiard, S., 2000. A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. J. Evol. Biol. 13, 814–825.

[150] Roze, D., Rousset, F., 2004, The robustness of Hamilton's rule with inbreeding and dominance: kin selection and fixation probabilities under partial sib mating, Am. Nat. 164(2), 214–231.

[151] Santos, F.C., Santos, M.D., Pacheco, J.M., 2008. Social diversity promotes the emergence of cooperation in public goods games. Nature 454, 213–216.

[152] Santos, F.C., Pacheco, J.M., 2011. Risk of collective failure provides an escape from the tragedy of the commons. Proc. Natl. Acad. Sci. USA 108(26), 10421–10425.

[153] Schreuer, P.A.G., Mandel, S.P.H., 1959. An inequality in population genetics. Heredity 13 (4), 519–524.

[154] Simon, B., 2010. A dynamical model of two-level selection. Evol. Ecol. Res. 12, 555–588.

[155] Simon, B., Fletcher, J.A., Doebeli, M., 2013. Towards a general theory of group selection, Evolution 67, 1561–1572.

[156] Sober, E. Wilson, D.S. 1998. *Unto Others; the evolution and psychology of unselfish behavior,* Cambridge, MA: Harvard University Press.

[157] Souza, M.O., Pacheco, J.M., Santos, F.C., 2009. Evolution of cooperation under N-person snowdrift games, J. Theor. Biol. 260, 581–588.

[158] Stark, R.E., 1992. Cooperative nesting in the multivoltine large carpenter bee Xylocopa sulcatipes Maa (Apoidea: Anthophoridae): do helpers gain or lose to solitary females? Ethology 91, 301–310.

[159] Tarnita C.E., Ohtsuki, H., Antal, T., Fu, F., Nowak, M.A., 2009. Strategy selection in structured populations. J. Theor. Biol. 259, 570–581.

[160] Tarnita, C.E., Antal, T, Ohtsuki, H., Nowak, M.A., 2009. Evolutionary dynamics in set structured populations. Proc. Natl. Acad. Sci. USA 106, 8601.

[161] Taylor, P.D., 1988. Inclusive fitness models with two sexes. Theor. Pop. Biol. 34 (2), 145–168.

[162] Taylor P.D., 1989. Evolutionary stability in one-parameter models under weak selection. Theor. Pop. Biol. 36, 125–143.

[163] Taylor P.D., 1992a. Altruism in viscous populations – an inclusive fitness model. Evol. Ecol. 6, 352–356

[164] Taylor P.D., 1992b. Inclusive fitness in a homogeneous environment. Proc R. Soc. Lond. B 249, 299–302.

[165] Taylor, P.D., Day, T., Wild, G., 2007. Evolution of cooperation in a finite homogeneous graph, Nature 447, 469–472

[166] Taylor P.D., Day, T. Wild, G., 2007. From inclusive fitness to fixation probability in homogeneous structured populations J. Theor. Biol. 249, 101–110

[167] Taylor, P.D., Frank, S.A., 1996. How to make a kin selection model. J. Theor. Biol. 180, 27–37.

[168] Taylor, P., Jonker, L., 1978. Evolutionary stable strategies and game dynamics. Math. Biosciences 40, 145-156.

[169] Taylor, P., Maciejewski, W., 2014. Hamilton's inclusive fitness in finite-structured populations. Phil. Trans. R. Soc. B., 369, 20130360.

[170] van Veelen, M., 2005. On the use of the Price equation. J. Theor. Biol. 237, 412-426.

[171] van Veelen, M., 2006. Why kin and group selection models may not be enough to explain human other-regarding behaviour. J. Theor. Biol. 242, 790-797.

[172] van Veelen, M., 2007. Hamilton's missing link. J. Theor. Biol. 246, 551–554.

[173] Van Veelen, M., 2009. Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong. J. Theor. Biol. 259, 589–600.

[174] van Veelen, M., 2011a. A rule is not a rule if it changes from case to case (a reply to Marshall's comment). J. Theor. Biol. 270, 189–195.

[175] van Veelen, M., 2011b. The replicator dynamics with n player games and population structure. J. Theor. Biol. 276, 78–85.

[176] van Veelen, M., García, J., Sabelis, M. W. & Egas, M. 2012. Group selection and inclusive fitness are not equivalent; the Price equation vs. models and statistics. J. Theor. Biol. 299, 64–80.

[177] van Veelen, M., García, J., Rand, D.G., Nowak, M.A., 2012. Direct reciprocity in structured populations. Proc. Natl. Acad. Sci. USA 109, 9929–9934.

[178] van Veelen, M, Luo, S, Simon, B., 2014. A simple model of group selection that cannot be analyzed with inclusive fitness. J. Theor. Biol. 360, 279–289.

[179] van Veelen, M., Nowak, M.A., 2012. Multi-player games on the cycle. J. Theor. Biol. 292 116–128.

[180] van Veelen, M., Spreij, P., 2009. Evolution in games with a continuous action space. Econ. Theory 39, 355–376.

[181] Volij, 0., 2002. In defense of DEFECT. Games Econ. Beh. 39, 309–321.

[182] Wakano, J.Y., Lehmann, L., 2014. Evolutionary branching in deme-structured populations. J. Theor. Biol. 351, 83–95.

[183] Weibull, J.W., Salomonsson, M., 2006. Natural selection and social preferences. J. Theor. Biol. 239 (1), 79–92.

[184] Wang J., Fu, F., Wu, T., Wang, L., 2009. Emergence of social cooperation in threshold public goods game with collective risk. Phys. Rev. E 80, 016101.

[185] West, S.A., Griffin, A.S., Gardner, A. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J. Evol. Biol. 20, 415–432

[186] West, S.A., Griffin, A.S., Gardner, A. 2008. Social semantics: how useful has group selection been? J. Evol. Biol. 21, 374–385.

[187] Wild, G., Traulsen, A., 2007. The different limits of weak selection and the evolutionary dynamics of finite populations. J. Theor. Biol. 247, 382–390.

[188] Wilson, D.S., Wilson, E.O. 2007. Rethinking the theoretical foundations of sociobiology. Q. Rev. Biol. 82, 327–348.

[189] Wilson, D.S., 2008. Social semantics: towards a genuine pluralism in the study of social behaviour. J. Evol. Biol. 21, 368–373

[190] Wilson, D.S., Pollock, G.B., Dugatkin, L.A., 1992. Can altruism evolve in purely viscous populations? Evol. Ecol. 6, 331–341.

[191] Zahavi, A., 1997. *The handicap principle: A missing piece of Darwin's puzzle*. Oxford: Oxford University Press.

[192] Zheng, D.F., Yin, H., Chan, C-H., Hui, P.M., 2007. Cooperative behavior in a model of evolutionary snowdrift games with N-person interactions. Europhys. Let. 80, 18002.

11 Appendix A: The Regression Method

11.1 One independent variable

In the main text, the function f assigned a fixed fitness to every x. Here we relax that, by letting y be the fitness, and by letting probability measure μ be defined over $\left\{(x,y) \in \mathbb{R}^2 \mid y \geq 0\right\}$. We will also need probability measures that are implied by the marginal distribution of x (which we denote by μ_x) and by the conditional distribution of y given x (denoted by $\mu_{y \mid x}$).

Minimizing least squared differences with $g_n(x) = a_0 + a_1 x + ... + a_n x^n$ now implies that

$$\begin{gathered} \frac{d}{da_i}\int (y - g_n)\, d\mu = 0, \quad i = 0, ..., n \\ \Updownarrow \\ \int x^i y d\mu = \int x^i g_n d\mu, \quad i = 0, ..., n \end{gathered} \tag{A.1}$$

Because there is no y in $x^i g_n$, we can replace probability measure μ by μ_x in the integral on the right; $\int x^i g_n d\mu = \int x^i g_n d\mu_x$. The integral on the left can be rewritten as:

$$\begin{aligned} \int x^i y d\mu &= \int_x \left(\int_y x^i y d\mu_{y \mid x} \right) d\mu_x \\ &= \int_x x^i \left(\int_y y d\mu_{y \mid x} \right) d\mu_x \\ &= \int_x x^i \mathbb{E}[y|x]\, d\mu_x \end{aligned} \tag{A.2}$$

This brings us back to the case from the main text, where $f(x) = \mathbb{E}[y|x]$.

11.2 Two or more independent variables

This is basically the same as the previous subsection, but the presence of more than one independent variable implies that we will need to do this using more notation. We have $m \geq 2$ independent variables, $x_1, ..., x_m$, and one dependent variable; x_{m+1}. For the indexing, we need to define $\mathcal{J}$ as a finite subset of $\mathbb{N}_0^m$, and it represents all the terms for which polynomial $g_{\mathcal{J}}$ allows non-zero coefficients; $g_{\mathcal{J}} = \sum_{j \in \mathcal{J}} a_j x_1^{j_1} x_2^{j_2} ... x_m^{j_m}$. We will also need probability measures that are implied by the marginal distribution of $x_1, ..., x_m$ (denoted by $\mu_{x_1,...,x_m}$) and by the conditional distribution of x_{m+1} given $x_1, ..., x_m$ (denoted by $\mu_{x_{m+1} \mid x_1,...,x_m}$).

Minimizing least squared differences with $g_{\mathcal{J}}$ now implies that

$$\frac{d}{da_j}\int (x_{m+1} - g_{\mathcal{J}})\, d\mu = 0 \qquad \forall j \in \mathcal{J} \tag{A.3}$$

$$\Updownarrow$$

$$\int x_1^{j_1} x_2^{j_2} ... x_m^{j_m} x_{m+1} d\mu = \int x_1^{j_1} x_2^{j_2} ... x_m^{j_m} g_{\mathcal{J}} d\mu \qquad \forall j \in \mathcal{J}$$

Because there is no x_{m+1} in $x_1^{j_1} x_2^{j_2} ... x_m^{j_m} g_{\mathcal{J}}$, we can replace probability measure μ by $\mu_{x_1,...,x_m}$ in the integral on the right; $\int x_1^{j_1} x_2^{j_2} ... x_m^{j_m} g_{\mathcal{J}} d\mu = \int x_1^{j_1} x_2^{j_2} ... x_m^{j_m} g_{\mathcal{J}} \mu_{x_1,...,x_m}$. The integral on the left can be rewritten as:

$$\int x_1^{j_1} x_2^{j_2} ... x_m^{j_m} x_{m+1} d\mu = \int_{x_1,...,x_m} \left(\int_{x_{m+1}} x_1^{j_1} x_2^{j_2} ... x_m^{j_m} x_{m+1} d\mu_{x_{m+1} \mid x_1,...,x_m} \right) d\mu_{x_1,...,x_m} \tag{A.4}$$

$$= \int_{x_1,...,x_m} x_1^{j_1} x_2^{j_2} ... x_m^{j_m} \left(\int_{x_{m+1}} x_{m+1} d\mu_{x_{m+1} \mid x_1,...,x_m} \right) d\mu_{x_1,...,x_m}$$

$$= \int_{x_1,...,x_m} x_1^{j_1} x_2^{j_2} ... x_m^{j_m} \mathbb{E}\left[x_{m+1} \mid x_1, ..., x_m\right] d\mu_{x_1,...,x_m}$$

This brings us back to the case from the main text, where $f(x_1, ..., x_m) = \mathbb{E}\left[x_{m+1} \mid x_1, ..., x_m\right]$.

11.3 Application to prisoners dilemmas allowing for unequal gains from switching

The regression method considers two moments in time, and uses numbers that belong to these two discrete moments in time. The model we are considering is a replicator dynamics with relatedness r and game

	D	C
D	P	T
C	S	R

This is a continuous time model, and therefore the regression method should also be used in a marginal version, using derivatives instead of discrete changes. For a simple population with only one type and constant growth rate a, the fitness w typically depends on how long a time interval one would take, as $w = e^{at}$. At any instance, though, one could instead look at the marginal change, by considering $\lim_{\Delta t \to 0} \frac{w(t+\Delta t) - w(t)}{\Delta t}\Big|_{t=0} = \frac{dw}{dt}\Big|_{t=0} = a$, where $t = 0$ is the moment we are considering.

Here we will do the same, and with two types in the population, we now have four growth rates to consider; the growth rate P of D individuals matched with D individuals, and T,

S, and R for their respective combinations. The regression method writes these growth rates as a linear function of whether the individual itself is a cooperator or a defector and whether its interaction is a cooperator or a defector; $g(x,y) = a_{00} + a_{10}x + a_{01}y$, where $x = 0$ if an individual is a defector, $x = 1$ if an individual is a cooperator, $y = 0$ if its interaction partner is a defector, and $y = 1$ if its interaction partner is a cooperator. The population state is given by the frequencies f_0, f_1, f_2 of the different types of pairs. As before, $-a_{10}$ will be interpreted as costs, a_{01} as benefits, and a_{00} as baseline fitness, so we will write this as $g(x,y) = w_0 - cx + by$. A perfect fit with least squares 0 would be achieved if we would fit this to the functional form $g(x,y) = a_{00} + a_{10}x + a_{01}y + a_{11}xy$; with $a_{00} = P$, $a_{10} = S - P$, $a_{01} = T - P$ and $a_{11} = R + P - S - T$ all growth rates are equal to $g(x,y)$. With $g(x,y) = w_0 - cx + by$, on the other hand, we minimize

$$f_0 (P - w_0)^2 + \frac{f_1}{2} (T - (w_0 + a_{10}))^2 + \frac{f_1}{2} (S - (w_0 + a_{01}))^2 + f_2 (R - (w_0 + a_{10} + a_{01}))^2$$

If we denote this sum of squares with E, then the choice of w_0, b and c that minimizes E should satisfy $\frac{dE}{dw_0} = \frac{dE}{db} = \frac{dE}{dc} = 0$. If we take those derivatives, we find the following equations:

$$\begin{aligned}
\frac{dE}{dw_0} = 0 \quad &\Leftrightarrow \quad -2f_0 (P - w_0) - f_1 (T + S - 2w_0 - b + c) - 2f_2 (R - w_0 - b + c) = 0 \\
\frac{dE}{db} = 0 \quad &\Leftrightarrow \quad -f_1 (T - w_0 - b) - 2f_2 (R - w_0 - b + c) = 0 \\
\frac{dE}{dc} = 0 \quad &\Leftrightarrow \quad f_1 (S - w_0 + c) + 2f_2 (R - w_0 - b + c) = 0
\end{aligned}$$

From the last two equations together, we can immediately see that $c + b = T - S$. Using $f_0 + f_1 + f_2 = 1$ and $p = \frac{1}{2} f_1 + f_2$, the conditions can be simplified to

$$\begin{aligned}
w_0 &= f_0 P + \frac{1}{2} f_1 (T + S) + f_2 R + p (c - b) \\
pw_0 &= \frac{1}{2} f_1 (T - b) + f_2 (R - b + c) \\
pw_0 &= \frac{1}{2} f_1 (S + c) + f_2 (R - b + c)
\end{aligned}$$

Finding the solution to this system of equations requires some not very exciting algebra, but if we do it anyway, we find that the solution is:

$$\begin{aligned}
b &= \frac{1 - p + rp}{1 + r} (T - P) + \frac{p + r(1 - p)}{1 + r} (R - S) \\
c &= \frac{1 - p + rp}{1 + r} (P - S) + \frac{p + r(1 - p)}{1 + r} (T - R) \\
w_0 &= \frac{1}{1 + r} \{(1 + p + r - pr)(1 - p + rp) P + (1 - r) p (p + r(1 - p))(S + T - R)\}
\end{aligned}$$

This is Equation (4.12) from the main text. It is equally unexciting, but straightforward, to check that this is a solution indeed; just fill them in in the three equations above, and use $f_0 = (1-r)(1-p)^2 + r(1-p)$, $f_1 = (1-r)2p(1-p)^2$ and $f_2 = (1-r)p^2 + rp$.

11.4 Straightforward construction of b and c according to the regression method

Because we know that Hamilton's rule applies if we define b and c according to the regression method, we can also find those directly. The condition $\overline{\pi}_C > \overline{\pi}_D$ for the prisoners dilemma is rewritten as

$$r\ \left(p(T-P) + (1-p)(R-S)\right) > p(T-R) + (1-p)(P-S)$$

in Section 3. If we multiply by $1+r$ and subtract $r(p(T-R) + (1-p)(P-S))$ left and right we get

$$(1+r)\ \ (r\ \ (p(T-P) + (1-p)(R-S))) > (1+r)\ \ (p(T-R) + (1-p)(P-S))$$

$$(1+r)\ \ (r\ \ (p(T-P) + (1-p)(R-S)))-r(p(T-R) + (1-p)(P-S)) > (p(T-R) + (1-p)(P-S))$$

$$r^2pT + \left(r + r^2(1-p)\right)R - \left(r + r^2p\right)P - r^2(1-p)S > (p(T-R) + (1-p)(P-S)).$$

Then we add $r(1-p)(T-R) + rp(P-S)$ left and right and reorganize to obtain

$$\begin{aligned} r^2pT + \left(r + r^2(1-p)\right)R - \left(r + r^2p\right)P - r^2(1-p)S + r(1-p)(T-R) + rp(P-S) \\ > p(T-R)+(1-p)(P-S)+r(1-p)(T-R)+rp(P-S) \end{aligned}$$

$$r\left((1-p+rp)(T-P) + (p+r(1-p))(R-S)\right) > (p+r(1-p))(T-R)+(1-p+rp)(P-S).$$

Then we divide by $1+r$ again, and find

$$r\cdot\tfrac{1}{1+r}\left((1-p+rp)(T-P) + (p+r(1-p))(R-S)\right) > \tfrac{1}{1+r}\left((p+r(1-p))(T-R) + (1-p+rp)(P-S)\right).$$

This is $rb - c > 0$, with

$$b = \frac{1-p+rp}{1+r}(T-P) + \frac{p+r(1-p)}{1+r}(R-S)$$

and

$$c = \frac{1-p+rp}{1+r}(P-S) + \frac{p+r(1-p)}{1+r}(T-R).$$

12 Appendix B: Comparative statics

12.1 Proofs that Defs. 2 and 3 are implications of Def. 1

Def. 1 ⇒ Def. 3

We prove the stronger claim that, if for all (p, r), $\frac{\partial \dot{p}(p,r)}{\partial r} \geq 0$, then if (p', r') and (p'', r'') are distinct locally stable fixed points, we must have $(p' - p'')(r' - r'') > 0$. We may restrict our attention only to cases of stable fixed points (p', r') with $p' \in (0, 1)$ and $r \in (-1, 1)$.

Claim (1): For each $r \in (-1, 1)$, there is at most one $p' \in (0, 1)$ such that (p', r') is on the isocline. (Note that this justifies the term '*the* locally stable' in the definition.)

Proof: The isocline is defined by $\overline{\pi}_C = \overline{\pi}_D$. Isolating p then gives a unique solution:

$$p(r) = \frac{1}{1-r} \frac{(1-r)S + rR - P}{S - R + T - P}. \tag{B.1}$$

Claim (2a): If (p', r') is locally stable on $r = r'$, then (p', r') is, for $p \in (0, 1)$, globally stable on $r = r'$. That is, if $\exists\ \epsilon > 0$ such that $(p - p')\left[\dot{p}(p, r')\right] < 0$ for all $p \in (p' - \epsilon, p' + \epsilon) \backslash \{p'\}$, then $(p - p')\left[\dot{p}(p, r')\right] < 0$ for all $p \in (0, 1)$.

Proof: Suppose that (p', r') is locally stable, but that there is $p'' > p'$ such that $\dot{p}(p'', r') \geq 0$. If $\dot{p}(p'', r') = 0$, we have a contradiction of Claim (1), so assume $\dot{p}(p'', r') > 0$. Now from local stability of (p', r') and Claim (1), $\exists\ \delta \in (0, p'' - p')$ such that $\dot{p}(p, r') < 0\ \forall\ p \in (p', p' + \delta]$. Now we have $\dot{p}(p' + \delta, r') < 0$ and $\dot{p}(p'', r') > 0$, with $p' + \delta < p''$. Since $\dot{p}(\cdot, r')$ is continuous, the intermediate value theorem requires that $\exists\ p''' \in (p' + \delta, p'')$ such that $\dot{p}(p''', r') = 0$, in contravention of Claim (1). The case where there exists $p'' < p'$ such that $\dot{p}(p'', r') \leq 0$ yields a similar contradiction.

Now suppose that there exist (p', r') and (p'', r'') on the isocline with $r' > r''$ and $p' < p''$ (we need not worry about the case $p' = p''$, since each p defines at most one r on the isocline). Since $\frac{\partial \dot{p}(p,r)}{\partial r} \geq 0$ for all (p, r), we have $\dot{p}(p'', r') \geq \dot{p}(p'', r'') = 0$. If $\dot{p}(p'', r') = \dot{p}(p'', r'') = 0$, we have a contradiction of Claim (1). If $\dot{p}(p'', r') > \dot{p}(p'', r'') = 0$, we have a contradiction of Claim (2a).

Def. 1 ⇒ Def. 2

This follows similarly from Claim (1) and Claim (2b), which is: if (p', r') is locally unstable on $r = r'$, then (p', r') is, for $p \in (0, 1)$, globally unstable on $r = r'$. The proof of Claim (2b) is similar to that of Claim (2a).

12.2 Prisoners' dilemmas

Definition 2 applies to cases where both extreme frequencies $p = 0$ and $p = 1$ are locally stable, which is the case when $P - S > T - R$. Definition 3 requires there to be a stable mixture, which is the case when $P - S < T - R$. The distinction between those two cases shows in the shape of the isocline, which describes an arc in the simplex. This arc hits the corner $f_0 = 1$ (where $p = 0$) at the same slope as the arc $\bar{r} = \frac{P-S}{R-S} \in (0, 1)$, and it hits the corner $f_2 = 1$ (where $p = 1$) at the same slope as $\bar{\bar{r}} = \frac{T-R}{T-P} \in (0, 1)$. Along constant-$r$ arcs that lie above the isocline, the proportion of cooperators is decreasing; below the isocline, the proportion of cooperators is increasing. We can discern three cases: (a) $\bar{r} > \bar{\bar{r}}$, (b) $\bar{r} = \bar{\bar{r}}$, (c), $\bar{r} < \bar{\bar{r}}$ which amount to $P - S > T - R$, $P - S = T - R$, and $P - S < T - R$ respectively. In the case (b), we have equal gains from switching.

In case (a), $\bar{r} > \bar{\bar{r}}$, and so the isocline is a right-skewed arc, with maximum attained for $p > \frac{1}{2}$. Thus, for any given $r \in (\bar{\bar{r}}, \bar{r})$, the constant-$r$ arcs begin above the isocline at $f_0 = 0$, intersect the isocline at some $p = p^*(r)$, and reach $f_2 = 1$ from below the isocline (see Fig. 11a). For these constant-r arcs, cooperation is increasing if $p > p^*(r)$, and decreasing if $p < p^*(r)$. Here, $P - S > T - R$, so now $p^*(r)$ is decreasing in r (with $\bar{\bar{r}} < r < \bar{r}$ ensuring $0 < p^*(r) < 1$). Increasing relatedness therefore favours cooperation under the second definition, that the threshold proportion of cooperation above which full cooperation eventuates decreases as we increase relatedness. Since the first definition of favouring applies and is stronger than the second, this is to be expected.

In case (c), $\bar{r} < \bar{\bar{r}}$, and so the isocline is skewed leftward, with its maximum attained for $p < \frac{1}{2}$ (see Fig. 11c). Since the constant-r arcs are symmetric about $p = \frac{1}{2}$, we have that for all $r \in (\bar{r}, \bar{\bar{r}})$, the constant-$r$ arcs begin below the isocline at $f_0 = 1$, intersect the isocline at some $p = p^*(r) \in (0, 1)$, and reach $f_2 = 1$ from above the isocline (see Fig. 11a). On these constant-r arcs, cooperation is increasing at all $p < p^*(r)$, and decreasing for all $p > p^*(r)$, so that the intersection point $p = p^*(r)$ is the unique stable equilibrium proportion for each r. For $r < \bar{r}$, cooperation decreases for all p, while for $r > \bar{\bar{r}}$, cooperation increases for all p. We find the intersection by solving $\overline{\pi}_C = \overline{\pi}_D$ to p for a given r, and if we do, we get $p^*(r) = \frac{S-R+(R-P)/(1-r)}{S+T-R-P}$. Since $P - S < T - R$, the equilibrium frequency $p^*(r)$ at the intersection is increasing in r (with the further condition $\bar{r} < r < \bar{\bar{r}}$ ensuring that it is between 0 and 1). Increasing relatedness therefore favours cooperation under the third definition, that it never decreases the equilibrium proportion of cooperators. Again, this is to be expected, since we know that the first definition applies in this case, and is stronger than the third.

Finally, in case (b), that of equal gains from switching, the isocline is symmetric about $p = \frac{1}{2}$, so that it coincides with the particular constant-r arc $r = \frac{P-S}{R-S} \in (0, 1)$. For higher r, we have full cooperation in equilibrium; for lower r, we have full defection.

12.3 Stag hunt games

In order to describe the dynamics in the simplex, we will still need the intersection of the isocline and the r-arc, which is still at the point $p^*(r) = \frac{R-S-(R-P)/(1-r)}{R+P-S-T}$ - only slightly rewritten to have both enumerator and denominator positive. This function is decreasing in r, which implies that relatedness favours cooperation in the second sense, as it increases the basin of attraction of cooperators. Moreover, the isocline hits the left corner point, where $f_0 = 1$, and therefore $p = 0$, with a slope equal to the slope of the arc $r = \frac{P-S}{R-S}$. We will furthermore need two more points; the point p^{**} where the isocline hits the edge of the simplex, and the point p^{***} where $\frac{\partial \dot{p}(p,r)}{\partial r}$ changes sign.

In cases (a) and (b), the isocline goes from the corner where $f_0 = 1$, and therefore $p = 0$, to a point on the simplex face where $f_0 = 0$. This intersection of the isocline and the simplex face is at the point $p^{**} = \frac{R-S}{2R-S-T}$, $r^{**} = -\frac{R-T}{R-S}$. In case (c), the isocline intersects the other simplex face, where $f_0 = 0$, and it does so at the point $p^{**} = \frac{P-S}{2P-S-T}$, $r^{**} = -\frac{P-S}{P-T}$.

Left of the isocline (the green lines in Figs. 13a–c), the proportion of cooperators is decreasing. We therefore have two possible outcomes of the dynamics. If $r \geq \frac{P-S}{R-S}$, cooperation is always increasing on constant-r arcs, and the dynamics take the population to the corner where $p = 1$. If $r < \frac{P-S}{R-S}$, then for $p < p^*(r)$ the dynamics take the population to the left down corner, where $p = 0$, and when $p > p^*(r)$ the dynamics take the population to the right down corner, where $p = 1$,.

In all cases, relatedness favours cooperation under the second definition: $p^*(r)$ is decreasing in r, and therefore the basin of attraction of the cooperative equilibrium increases with r. In cases (b) and (c) relatedness does not favour cooperation under the first definition, as for high p, the growth rate of cooperator decreases as r increases. In case (a), increased relatedness does favour cooperation also under the strongest first definition.

12.4 (General) hawk dove games

In order to describe the dynamics in the simplex, we will still need the intersection of the isocline and the r-arc, which is at the point $p^*(r) = \frac{S-R+(R-P)/(1-r)}{S+T-R-P}$. This function is increasing in r, which implies that relatedness favours cooperation in the third sense, as it increases the equilibrium proportion of cooperators. Moreover, the isocline hits the right corner point, where $f_2 = 1$, and therefore $p = 1$, with a slope equal to the slope of the arc $r = \frac{T-R}{T-P}$. We will furthermore need two more points; the point p^{**} where the isocline hits the edge of the simplex, and the point p^{***} where $\frac{\partial \dot{p}(p,r)}{\partial r}$ changes sign.

In case (a), the isocline goes from a point on the simplex face where $f_0 = 0$ to the corner where $f_2 = 1$, and therefore $p = 1$. This intersection of the isocline and the simplex face is at the point $p^{**} = \frac{S-R}{S+T-2R}$, $r^{**} = \frac{T-R}{R-S}$. In cases (b) and (c), the isocline intersects the other simplex face, where $f_2 = 0$, and it does so at the point $p^{**} = \frac{S-P}{S+T-2P}$, $r^{**} = \frac{P-S}{T-P}$.

Right of the isocline (the green lines in Figs. 12a–c), the proportion of cooperators

is decreasing. We therefore have two possible outcomes of the dynamics. If $r \geq \frac{T-R}{T-P}$, cooperation is always increasing on constant-r arcs, and the dynamics take the population to the corner where $p = 1$. If $r < \frac{T-R}{T-P}$, then for given r, there is a stable equilibrium proportion $p^*(r)$ of cooperators. For p below it, the proportion of cooperators is increasing, for p above it, the proportion of cooperators is decreasing.

In all cases relatedness favours cooperation under the third definition: the equilibrium proportion $p^*(r)$ of cooperators is increasing in r. In cases (a) and (b) relatedness does not favour cooperation under the first definition, as for low p, the growth rate of cooperator decreases as r increases. In case (c), increased relatedness does favour cooperation also under the strongest first definition.

13 Appendix C: Adaptive Dynamics with pairwise interactions and fixed relatedness

13.1 A model for games in finite populations with relatedness

The update process is the same as the one used in Van Veelen & García (2010) and Van Veelen et al. (2012). It is a version of the Wright-Fisher that allows for positive assortment.

The parent population consists of individuals $i = 1, ..., N$. We will make a new generation, consisting of $\frac{1}{2}N$ interaction pairs as follows. For the first individual in pair 1, a parent is drawn from the parent population, where every individual from the parent population has a probability of being drawn proportional to their payoff. For the second individual in pair 1, a nested procedure applies. First with probability r, the same parent is chosen. With probability $1 - r$, a parent is drawn from the entire parent generation, where, again, every individual from the parent population has a probability of being drawn proportional to their payoff. This procedure creates one pair with relatedness r. This entire procedure is repeated $\frac{1}{2}N$ times to create an entire new population of interacting pairs.

13.2 Adaptive dynamics with interaction structure for piecewise-differentiable fitness functions

Here we generalize Allen et al. (2013) to games whose payoff functions are piecewise-differentiable rather than differentiable. We consider a class of evolutionary processes for which a trait value x evolves under rare and incremental mutation, with interactions described by the game $\pi(x, y)$. This class is defined by a set of general assumptions specified below. The canonical equation we will arrive at is the following differential equation.

$$\dot{x} = N_{\mathrm{e}} \frac{N-1}{N} \frac{u(x)}{\pi(x;x)} \epsilon^2 \left(\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'} \bigg|_{x'=x} \pi(x', x) + \frac{\sigma - 1}{\sigma + 1} \frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'} \bigg|_{x'=x} \pi(x, x') \right). \tag{1}$$

Above, N is the equilibrium population size, N_{e} is the effective population size, defined in Section 13.3, u is the per capita rate of stochastic mutant appearance (which may depend on the resident trait value x), $\pi(x, y)$ is the payoff to an individual with trait x interacting with an individual with trait y, ϵ^2 is the variance in mutational steps in the trait value, and σ is the structure coefficient of Tarnita et al. (2009), also defined in Section 13.3. The notation $\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} \bullet}$ refers to the symmetric partial derivative, as defined below.

This equation is arrived at in three steps, as described in the main text. We assume that mutations arise at intensity $Nu(x)$, and that the uncertainty about whether it will fixate or go extinct is resolved instantly, of course according to the true fixation probability. If Δx is the change in the population trait value from time t to time $t + \Delta t$, then with those

assumptions, we show that $E[\Delta x]$ satisfies

$$\frac{\mathbb{E}\left[\Delta x\right]}{\Delta t} = N_{\mathrm{e}} \frac{N-1}{N} \frac{u(x)}{\pi(x;x)} \epsilon^2 \left(\left. \frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'} \right|_{x'=x} \pi(x',x) + \frac{\sigma-1}{\sigma+1} \left. \frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'} \right|_{x'=x} \pi(x,x') \right) + \epsilon^2 Q\left(x, \Delta t, \epsilon\right), \quad (2)$$

where $Q\left(x, \Delta t, \epsilon\right)$ is a function satisfying

$$\lim_{\epsilon \to 0} \lim_{\Delta t \to 0} Q\left(x, \Delta t, \epsilon\right) = 0 \quad \text{for all } x \in \mathbb{R}.$$

Section 13.3 introduces the symmetric derivative and other basic concepts. In Section 13.4 we define the classes of evolutionary models to which our result applies, and prove basic results about these models. The derivation of Eq. (2) appears in Section 13.5.

13.3 General definitions and lemmas

13.3.1 One-sided and symmetric derivatives

Definition 7 *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. We define*

- *The* left derivative *of f at $x \in \mathbb{R}$ as*

$$\frac{d_- f}{d_- x}(x) = \lim_{\epsilon \to 0^-} \frac{f(x+\epsilon) - f(x)}{\epsilon},$$

- *The* right derivative *of f at $x \in \mathbb{R}$ as*

$$\frac{d_+ f}{d_+ x}(x) = \lim_{\epsilon \to 0^+} \frac{f(x+\epsilon) - f(x)}{\epsilon},$$

- *The* symmetric derivative *of a continuous function $f : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ is defined as*

$$\frac{d_s f}{d_s x}(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x-\epsilon)}{2\epsilon},$$

where each derivative above is defined only if the corresponding limits exist. If so, we say f is left-, right-, or symmetric-differentiable at x, respectively.

If f is instead a function of n real arguments $x_1, \ldots, x_n$, then the left, right, and symmetric partial derivatives of f in the kth argument, denoted

$$\frac{\partial_- f}{\partial_- x_k}, \qquad \frac{\partial_+ f}{\partial_+ x_k}, \qquad \text{and} \qquad \frac{\partial_{\mathrm{s}} f}{\partial_{\mathrm{s}} x_k},$$

are defined as the corresponding derivatives of the function

$$x_k \mapsto f(x_1, \ldots, x_k, \ldots, x_n).$$

We state without proof the following elementary lemmas:

Lemma 8 (Multivariate chain rule for one-sided derivatives) *Let $f(x_1, \ldots, x_n)$ be a differentiable function of n real arguments, and let $a_1(t), \ldots, a_n(t)$ be continuous functions which are left- (resp., right-, symmetric-) differentiable at $t = 0$. Then the function g defined by*

$$g(t) = f\big(a_1(t), \ldots, a_n(t)\big)$$

is left- (resp., right-, symmetric-) differentiable at $t = 0$, and

$$\frac{d_- g}{d_- t}(0) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}\big(a_1(0), \ldots, a_n(0)\big) \frac{d_- a}{d_- t}(0)$$

$$\left(\textit{resp.,} \quad \frac{d_+ g}{d_+ t}(0) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}\big(a_1(0), \ldots, a_n(0)\big) \frac{d_+ a}{d_+ t}(0), \right.$$

$$\left. \frac{d_s g}{d_s t}(0) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}\big(a_1(0), \ldots, a_n(0)\big) \frac{d_s a}{d_s t}(0) \right).$$

Lemma 9 *If f is left- and right-differentiable at $x \in \mathbb{R}$, then f is symmetric-differentiable at x and*

$$\frac{d_s f}{d_s x}(x) = \frac{1}{2}\left(\frac{d_- f}{d_- x}(x) + \frac{d_+ f}{d_+ x}(x)\right).$$

Lemma 10 *If f is differentiable at $x \in \mathbb{R}$, then f is left-, right-, and symmetric-differentiable at x and*

$$\frac{d_- f}{d_- x}(x) = \frac{d_+ f}{d_+ x}(x) = \frac{d_s f}{d_s x}(x).$$

13.3.2 Scaling of functions

Definition 11 *Let $U : \mathbb{R} \to \mathbb{R}$ be an integrable function. For $\epsilon > 0$, we define the* ϵ-scaling of U *to be the function $U_\epsilon : \mathbb{R} \to \mathbb{R}$ with*

$$U_\epsilon(x) = \frac{1}{\epsilon} U\left(\frac{x}{\epsilon}\right).$$

Note that U_ϵ is integrable and

$$\int_{-\infty}^{\infty} U_\epsilon(x)\, dx = \int_{-\infty}^{\infty} U(x)\, dx,$$

for all $\epsilon > 0$.

13.4 Models of adaptive dynamics in structured populations

Our results apply to a class of models representing evolution in populations with interaction structure. Here we define this class by specifying the assumptions that each model in this class must satisfy. We separate our assumptions into those describing competition between resident and mutant types (C1-C6), those describing the game (G1), and those describing the process of evolution by trait substitutions (M1-M3).

13.4.1 Resident-mutant competition

We first describe the class of resident-mutant competition models by stating the assumptions that define this class. In the definition, S refers to the set of all states, $F_{\mathrm{M}} \subset S$ is a subset of states that corresponds to mutant fixation, and the probability distribution μ quantifies the likelihood of a being in state at the moment after a mutation first appears (at which point resident-mutant competition is initiated). For a given resident-mutant competition model and payoff matrix G, we define the fixation probability ρ as the probability the Markov chain associated to G hits F_{M}, given that its initial state is sampled according to μ.

C1. There is a finite set S with an associated probability distribution μ, and a distinguished subset $F_{\mathrm{M}} \subset S$ which is assigned zero probability by μ.

C2. For any payoff matrix G of the form

$$G = \begin{pmatrix} a_{\mathrm{MM}} & a_{\mathrm{MR}} \\ a_{\mathrm{RM}} & a_{\mathrm{RR}} \end{pmatrix}, \tag{3}$$

where the entries reflect payoffs from interactions between mutants (M) and residents (R), there is a collection $\{p_{s'|s}\}_{s,s' \in S}$ of transition probabilities, giving S the structure of a Markov chain.

C3. The Markov chain associated to any such payoff matrix G has the following properties:

a) There is zero probability of transitioning from a state in F_{M} to a state not in F_{M}.

b) For any $s \in S$ which is assigned positive probability by μ, and any $s' \in F_{\mathrm{M}}$, there is a positive integer n for which the probability of transitioning from s to s' in n steps is positive.

C4. The transition probabilities $p_{s'|s}$ vary twice differentiably with respect to the entries of G.

C5. If the payoff matrix G is multiplied by a constant $K > 0$, the probability that the Markov chain hits F_{M}, given that its initial state is sampled from μ, is unaffected.

C6. The probability ρ is increasing in a_{MM} and a_{MR}, and decreasing in a_{RM} and a_{RR}, for all values of a_{MM}, a_{MR}, a_{RM}, and a_{RR} sufficiently close to 1.

We define the reverse matrix $\tilde{G}$ to be the payoff matrix in which the roles of resident and mutant are switched,

$$\tilde{G} = \begin{pmatrix} a_{\mathrm{RR}} & a_{\mathrm{RM}} \\ a_{\mathrm{MR}} & a_{\mathrm{MM}} \end{pmatrix}.$$

and the reverse fixation probability $\tilde{\rho}$ is the fixation probability associated to $\tilde{G}$.

Lemma 12 *For any resident-mutant competition model satisfying Assumptions C1–C4, the fixation probability ρ varies twice differentiably with respect to the entries of the game matrix G.*

Proof. Assumption C3 implies that ρ varies smoothly with respect to the transition probabilities $p_{s'|s}$ (see, for example, Theorem 3.3 of Iofescu, 1980). By Assumption C4, ρ varies twice differentiably with respect to the entries of G. ■

For given G and $\delta > 0$, we denote

$$G_\delta = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \delta G.$$

This allows us to analyze the case of weak selection ($\delta \ll 1$). We let ρ_δ and ρ'_δ denote the fixation probabilities associated to G_δ and G'_δ, respectively.

Theorem 13 (Structure coefficient theorem of Tarnita et al., 2009) *For any resident-mutant competition model satisfying Assumptions C1–C6, there is a positive constant σ such that, given any payoff matrix G, $\rho_\delta > \tilde{\rho}_\delta$ for all sufficiently small $\delta > 0$ if*

$$\sigma a_{\mathrm{MM}} + a_{\mathrm{MR}} > a_{\mathrm{RM}} + \sigma a_{\mathrm{RR}}. \tag{4}$$

Conversely, $\rho_\delta < \tilde{\rho}_\delta$ for all sufficiently small $\delta > 0$ if

$$\sigma a_{\mathrm{MM}} + a_{\mathrm{MR}} < a_{\mathrm{RM}} + \sigma a_{\mathrm{RR}}.$$

Definition 14 *For a fixed resident-mutant competition model, we define the* effective population size *as*

$$N_e = \frac{N^2}{N-1} \left. \frac{\partial \rho}{\partial s} \right|_{s=0}, \tag{5}$$

where ρ is the fixation probability of mutants in the game

$$G = \begin{pmatrix} 1+s & 1+s \\ 1 & 1 \end{pmatrix}. \tag{6}$$

In the game G above, mutants and residents have constant payoff $1+s$ and 1, respectively (regardless of interaction partners), so that s can be identified as the mutant's selection coefficient. For models that are amenable to the diffusion approximation, N_{e} is equal to the variance effective population size (Kimura, 1964).

13.4.2 Piecewise-differentiable games

The definitions, assumptions, and results in Section 13.4.1 apply to 2×2 matrix games. We will use these results to study the long-term evolution of real-valued strategies in a continuous game, using the adaptive dynamics approach. We assume that the game payoff function $\pi(x, y)$ satisfies

G1. For each $x \in \mathbb{R}$, $\pi(x, y)$ is positive and left- and right-differentiable in both arguments at $y = x$.

13.4.3 Evolution by trait substitution

We now describe the class of models representing long-term evolution by a stochastic trait substitution sequence. Suppose a resident-mutant competition model has been fixed. For $x, x' \in \mathbb{R}$, let $\rho(x'; x)$ denote the fixation probability ρ for this model, as defined in Assumption C5, with the entries of G given by

$$a_{MM} = \pi(x', x'), \quad a_{MR} = \pi(x', x),$$
$$a_{RM} = \pi(x, x'), \quad a_{RR} = \pi(x, x).$$

The class of long-term evolution models is defined by the following assumptions:

M1. In a monomorphic population with trait value x, mutants appear as a Poisson process with rate $Nu(x)$ per unit time, where u is a positive-valued function.

M2. When a mutant appears, a mutational step $y \in \mathbb{R}$ is sampled from a probability distribution with density function $U_\epsilon(y) = (1/\epsilon)\, U(y/\epsilon)$, where ϵ is a positive constant and U is an integrable, compactly supported function that is symmetric about 0 and has unit variance:

$$\int_{-\infty}^{\infty} y^2 U(y)\, dy = 1. \tag{7}$$

The trait value $x' \in \mathbb{R}$ of the mutant is then assigned to be $x' = x + y$.

M3. If a mutant of trait value x' arises in a monomorphic population with trait value x, then with probability $\rho(x'; x)$ the population becomes monomorphic with trait value x'; otherwise it reverts to being monomorphic with trait value x. (The fixation or disappearance of trait value x' is regarded as instantaneous.)

We note that the distribution U_ϵ of mutational steps that appears in Assumption M2 has variance ϵ^2.

Overall, Assumptions M1–M3 imply a Markov jump process representation of evolution, in which the population jumps stochastically from one monomorphic trait value to another. The rate of transition from a trait value x to x' is given by the density $Nu(x)\, U_\epsilon(x')\, \rho(x; x')\, dx$.

13.5 Results

Lemma 15 *For any model satisfying C1–C4 and G1, the fixation probability $\rho(x'; x)$ is left- and right-differentiable in both arguments at $x' = x$.*

Proof. This follows from Lemmas 8 and 12 and Assumption G1. ■

Theorem 16 *For any model satisfying Assumptions C1–C4, G1, and M1–M3, the expected change in trait value $\mathbb{E}[\Delta x]$ from a given value x in the time window $[t, t + \Delta t)$ satisfies*

$$\frac{\mathbb{E}[\Delta x]}{\Delta t} = Nu(x)\epsilon^2 \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho(x', x) + \epsilon^2 Q(x, \Delta t, \epsilon),$$

where $Q(x, \Delta t, \epsilon)$ is a function satisfying

$$\lim_{\epsilon \to 0} \lim_{\Delta t \to 0} Q(x, \Delta t, \epsilon) = 0 \quad \textit{for all } x.$$

Proof. Assumptions M1–M3 imply that the expected change in trait value, $\mathrm{E}[\Delta x]$, satisfies

$$\frac{\mathbb{E}[\Delta x]}{\Delta t} = Nu(x) \int_{-\infty}^{\infty} y\, \rho(x+y; x)\, U_\epsilon(y)\, dy + S(x, \Delta t, \epsilon),$$

where $S(x, \Delta t, \epsilon)$ is a function satisfying

$$\lim_{\Delta t \to 0} S(x, \Delta t, \epsilon) = 0 \quad \text{for all } x \in \mathbb{R} \text{ and all } \epsilon > 0.$$

We separate the positive and negative values of y:

$$\frac{\mathbb{E}[\Delta x]}{\Delta t} = Nu(x) \int_{-\infty}^{0} y\, \rho(x+y, x)\, U_\epsilon(y)\, dy + Nu(x) \int_{0}^{\infty} y\, \rho(x+y, x)\, U_\epsilon(y)\, dy + S(x, \Delta t, \epsilon). \quad (8)$$

By changing variables and invoking the symmetry of U_ϵ, we can rewrite the first integral on the right-hand side of (8) as follows:

$$\begin{aligned} \int_{-\infty}^{0} y\, \rho(x+y, x)\, U_\epsilon(y)\, dy &= \int_{\infty}^{0} (-y)\, \rho(x-y, x)\, U_\epsilon(-y)\, d(-y) \\ &= -\int_{0}^{\infty} y\, \rho(x-y, x)\, U_\epsilon(y)\, dy \end{aligned}$$

This allows us to recombine the two terms on the right-hand side of (8) to yield

$$\frac{\mathbb{E}[\Delta x]}{\Delta t} = Nu(x) \int_{0}^{\infty} y\, [\rho(x+y, x) - \rho(x-y, x)]\, U_\epsilon(y)\, dy + S(x, \Delta t, \epsilon).$$

The integrand above is symmetric around $y = 0$, allowing us to write

$$\begin{aligned} \frac{\mathbb{E}[\Delta x]}{\Delta t} &= Nu(x) \int_{-\infty}^{\infty} y\, \frac{\rho(x+y, x) - \rho(x-y, x)}{2}\, U_\epsilon(y)\, dy + S(x, \Delta t, \epsilon) \\ &= Nu(x) \int_{-\infty}^{\infty} y^2\, \frac{\rho(x+y, x) - \rho(x-y, x)}{2\, y}\, U_\epsilon(y)\, dy + S(x, \Delta t, \epsilon). \end{aligned}$$

We now substitute $U_\epsilon(y) = (1/\epsilon)\, U(y/\epsilon)$ and change variables from y to $z = y/\epsilon$. This yields

$$\begin{aligned} \frac{\mathbb{E}[\Delta x]}{\Delta t} &= Nu(x) \int_{-\infty}^{\infty} y^2\, \frac{\rho(x+y, x) - \rho(x-y, x)}{2\, y}\, U(y/\epsilon)\, d(y/\epsilon) + S(x, \Delta t, \epsilon) \\ &= Nu(x)\epsilon^2 \int_{-\infty}^{\infty} z^2\, \frac{\rho(x+z\epsilon, x) - \rho(x-z\epsilon, x)}{2\, z\epsilon}\, U(z)\, dz + S(x, \Delta t, \epsilon). \end{aligned} \quad (9)$$

Since $\rho(x', x)$ is symmetric-differentiable in x' at $x' = x$, we can write

$$\frac{\rho(x+z\epsilon, x) - \rho(x-z\epsilon, x)}{2\, z\epsilon} = \left. \frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'} \right|_{x'=x} \rho(x', x) + R(x, z\epsilon),$$

where

$$\lim_{\epsilon \to 0} R\left(x, z\epsilon\right) = 0$$

for each fixed $z \in \mathbb{R}$. Substituting this into (9), we obtain

$$\frac{\mathbb{E}\left[\Delta x\right]}{\Delta t} = Nu(x)\epsilon^2 \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) \int_{-\infty}^{\infty} z^2 U(z)\, dz + Nu(x)\epsilon^2 \int_{-\infty}^{\infty} z^2 R\left(x, z\epsilon\right) U(z)\, dz + S\left(x, \Delta t, \epsilon\right),$$

which simplifies, using (7), to

$$\frac{\mathbb{E}\left[\Delta x\right]}{\Delta t} = Nu(x)\epsilon^2 \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) + Nu(x)\epsilon^2 \int_{-\infty}^{\infty} z^2 R\left(x, z\epsilon\right) U(z)\, dz + S\left(x, \Delta t, \epsilon\right).$$

We now define

$$Q\left(x, \Delta t, \epsilon\right) = Nu(x) \int_{-\infty}^{\infty} z^2 R\left(x, z\epsilon\right) U(z)\, dz + \frac{1}{\epsilon^2} S\left(x, \Delta t, \epsilon\right) \tag{10}$$

so that

$$\frac{\mathbb{E}\left[\Delta x\right]}{\Delta t} = Nu(x)\epsilon^2 \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) + \epsilon^2 Q\left(x, \Delta t, \epsilon\right)$$

as required. It only remains to consider the limit

$$\lim_{\epsilon \to 0} \lim_{\Delta t \to 0} Q\left(x, \Delta t, \epsilon\right) = Nu(x) \lim_{\epsilon \to 0} \int_{-\infty}^{\infty} z^2 R\left(x, z\epsilon\right) U(z)\, dz. \tag{11}$$

Above, we have used the fact that $\lim_{\Delta t \to 0} S\left(x, \Delta t, \epsilon\right) = 0$ for all x and all $\epsilon > 0$; thus

$$\lim_{\epsilon \to 0} \lim_{\Delta t \to 0} \frac{1}{\epsilon^2} S\left(x, \Delta t, \epsilon\right) = 0.$$

Since U is compactly supported, there exists some $\epsilon_0 > 0$ such that the integrand in (11) is bounded, for each $\epsilon < \epsilon_0$, by the integrable function

$$B(x, z) = z^2 U(z) M(x),$$

with

$$M(x) = \sup_{\substack{z \in \mathrm{Supp}\, U \\ \epsilon < \epsilon_0}} R(x; z\epsilon).$$

The supremum above is finite as long as ϵ_0 is sufficiently small, since $R(x, z\epsilon)$ converges to zero for each fixed x as $z\epsilon$ approaches zero. Therefore, by the Lebesgue dominated convergence theorem, the limit and integral in (11) can be interchanged, yielding

$$\lim_{\epsilon \to 0} \lim_{\Delta t \to 0} Q(x, \Delta t, \epsilon) = Nu(x) \int_{-\infty}^{\infty} z^2 \left(\lim_{\epsilon \to 0} R(x; z\epsilon)\right) U(z)\, dz = 0,$$

for each x, completing the proof of the theorem. ■

By Theorem 16, the approximation

$$\frac{\mathbb{E}\left[\Delta x\right]}{\Delta t} \approx Nu(x)\epsilon^2 \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right)$$

becomes increasingly accurate as $\epsilon \to 0$. We therefore consider the following deterministic approximation to the dynamics of x:

$$\dot{x} = Nu(x)\epsilon^2 \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right).$$

Theorem 17 *For any model satisfying Assumptions C1–C6 and G1, the fixation probability $\rho(x'; x)$ satisfies*

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) = N_e \frac{N-1}{N^2 \pi(x;x)} \left(\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x', x\right) + \frac{\sigma - 1}{\sigma + 1} \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x, x'\right) \right),$$

where σ is the structure coefficient from Tarnita et al. (2009) and Theorem 13.

Proof. This proof is organized in three steps.

Step 1: Obtain ρ in terms of partial derivatives of π

Using Lemma 8 we obtain

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) = \sum_{j,k \in \{M,R\}} \left.\frac{\partial \rho}{\partial a_{jk}}\right|_{G=\left(\begin{smallmatrix} \pi(x;x) & \pi(x;x) \\ \pi(x;x) & \pi(x;x) \end{smallmatrix}\right)} \left.\frac{\partial_s a_{jk}}{\partial_s x'}\right|_{x'=x}. \tag{12}$$

By Assumption C5, each of the payoff values a_{jk} can be divided by $a_{RR} = \pi\left(x, x\right)$ without changing the value of ρ. This implies that

$$\left.\frac{\partial \rho}{\partial a_{jk}}\right|_{G=\left(\begin{smallmatrix} \pi(x,x) & \pi(x,x) \\ \pi(x,x) & \pi(x,x) \end{smallmatrix}\right)} = \frac{1}{\pi\left(x, x\right)} \left.\frac{\partial \rho}{\partial a_{jk}}\right|_{G=\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)},$$

so that, combining with (12),

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) = \frac{1}{\pi\left(x, x\right)} \sum_{j,k \in \{M,R\}} \left.\frac{\partial \rho}{\partial a_{jk}}\right|_{G=\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)} \left.\frac{\partial_s a_{jk}}{\partial_s x'}\right|_{x'=x}. \tag{13}$$

By Lemma 8, the partial derivatives of the a_{jk} at $x' = x$ are given by

$$\begin{aligned}
\left.\frac{\partial_s a_{MM}}{\partial_s x'}\right|_{x'=x} &= \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x', x\right) + \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x, x'\right), \\
\left.\frac{\partial_s a_{MR}}{\partial_s x'}\right|_{x'=x} &= \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x', x\right), \\
\left.\frac{\partial_s a_{RM}}{\partial_s x'}\right|_{x'=x} &= \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x, x'\right), \\
\left.\frac{\partial_s a_{RR}}{\partial_s x'}\right|_{x'=x} &= 0.
\end{aligned}$$

We can therefore rewrite (13) in the form

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \rho\left(x', x\right) = \frac{\kappa}{\pi\left(x, x\right)} \left(\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x', x\right) + \kappa' \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi\left(x, x'\right) \right). \tag{14}$$

The quantities κ and κ' are linear combinations of partial derivatives of ρ with respect to the payoff values a_{jk} at $a_{MM} = a_{MR} = a_{RM} = a_{RR} = 1$. In particular, κ and κ' are independent of π, x, and x'.

It now only remains to relate κ and κ' to N_{e} and σ. We can establish these relationships by considering particularly simple payoff functions π, and substituting these payoff functions into (14). Since κ, κ', σ and N_{e} are all independent of π, any relationship derived using a particular π will hold generally.

Step 2: Relate κ and N_{e}

To relate κ and N_{e} we choose a specific payoff function π to substitute into (14). We consider $\pi\left(x, y\right) = 1 + x$, describing frequency-independent selection. Note that $\rho\left(x', x\right)$ is differentiable in both arguments for this game. For $x = 0$ and $x' = s$, the game between residents and mutants is described by the matrix (6), which we used to define the effective population size. Substituting this payoff function and $x = 0$ in (14) yields

$$\left.\frac{\partial}{\partial x'}\right|_{x'=0} \rho\left(x', 0\right) = \kappa. \tag{15}$$

Identifying x' with s and comparing with (5), we obtain

$$\kappa = N_{\mathrm{e}} \frac{N-1}{N^2}. \tag{16}$$

Since the values of κ and N_{e} do not depend on the game being played, (16) holds for all games.

We can also use this game to show that κ and N_{e} must be positive, a fact which we use in relating κ' to σ. For the mutant type x' and resident type $x = 0$ we have $a_{MM} = a_{MR} = 1 + x'$, $a_{RM} = a_{RR} = 1$. By Assumption C6,

$$\left.\frac{\partial}{\partial x'}\right|_{x'=0} \rho\left(x', 0\right) = \left.\left(\frac{\partial \rho}{\partial a_{MM}} + \frac{\partial \rho}{\partial a_{MR}}\right)\right|_{G=\left(\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right)} > 0.$$

Subsequently (15) implies that $\kappa > 0$, and it follows from (16) that $N_{\mathrm{e}} > 0$.

Step 3: Relate κ' and σ

To relate κ' and σ we first observe that for $x = x'$,

$$G = \begin{pmatrix} \pi\left(x', x'\right) & \pi\left(x', x'\right) \\ \pi\left(x', x'\right) & \pi\left(x', x'\right) \end{pmatrix} = \pi\left(x', x'\right) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

By Assumption C5, the fixation probability $\rho(x', x')$ does not depend on the value of $\pi(x', x')$, and is therefore constant with respect to x'. Taking the right derivative at $x' = 0$, and making use of Lemma 8 we obtain

$$0 = \left.\frac{d_+}{d_+x'}\right|_{x'=0} \rho(x', x') = \left.\frac{d_+}{d_+x'}\right|_{x'=0} \rho(x', 0) + \left.\frac{d_+}{d_+x'}\right|_{x'=0} \rho(0, x'),$$

and therefore

$$\left.\frac{\partial_+}{\partial_+x'}\right|_{x'=0} \rho(x'; 0) = -\left.\frac{\partial_+}{\partial_+x'}\right|_{x'=0} \rho(0; x').$$

Since the two above derivatives have opposite signs, there exists an interval $[0, \delta)$ such that if $\rho(x'; 0)$ is increasing in $x' \in [0, \delta)$, then $\rho(0; x')$ is decreasing on this interval, and vice versa. It follows that for x' in this interval,

$$\rho(x'; 0) > \rho(0; x') \quad \Longleftrightarrow \quad \left.\frac{\partial_+}{\partial_+x'}\right|_{x'=0} \rho(x'; 0) > 0. \tag{20}$$

As in Step 2, we now choose a particular payoff function π to substitute into (14). We consider the linear Prisoner's Dilemma with $\pi(x, y) = -cx + by + 1$, $b > c > 0$, with resident and mutant trait values $x = 0$ and $x' > 0$. The payoff matrix G can then be written as

$$G = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + x' \begin{pmatrix} b - c & -c \\ b & 0 \end{pmatrix}.$$

The mutant trait value x' can therefore be interpreted as a selection-strength parameter (Nowak et al., 2004, Tarnita et al., 2009). By the defining condition of the structure coefficient, (4), we have that for sufficiently small $x' > 0$,

$$\rho(x', 0) > \rho(0, x') \quad \Longleftrightarrow \quad \sigma(b - c) - c > b. \tag{21}$$

On the other hand, (14) implies

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}}\ x'}\right|_{x'=0} \rho(x', 0) = \kappa(-c + \kappa' b).$$

Since $\rho(x', x)$ is differentiable in both arguments for this game, we also have

$$\left.\frac{\partial_+}{\partial_+x'}\right|_{x'=0} \rho(x', 0) = \left.\frac{\partial}{\partial x'}\right|_{x'=0} \rho(x', 0) = \kappa(-c + \kappa' b).$$

Applying (20) and the positivity of $\kappa > 0$ (proven in Step 2), we obtain that for sufficiently small $x' > 0$.

$$\rho(x', 0) > \rho(0, x') \quad \Longleftrightarrow \quad -c + \kappa' b > 0. \tag{22}$$

Comparing (21) and (22), we see that

$$\kappa' = \frac{\sigma - 1}{\sigma + 1}. \tag{23}$$

Again, since the values of κ' and σ do not depend on the game being played, this identity holds for all games.

Substituting (23) and (16) into (14) yields the desired result. ■

Combining Theorems 16 and 17 yields:

Corollary 18 *For any model satisfying Assumptions C1–C6, G1, and M1–M3, the expected change in trait value* $\mathbb{E}[\Delta x]$ *from a given value* x *in the time window* $[t, t+\Delta t)$ *satisfies*

$$\frac{\mathbb{E}[\Delta x]}{\Delta t} = N_e \frac{N-1}{N} \frac{u(x)}{\pi(x,x)} \epsilon^2 \left(\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\bigg|_{x'=x} \pi(x', x) + \frac{\sigma - 1}{\sigma + 1} \frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\bigg|_{x'=x} \pi(x, x') \right) + \epsilon^2 Q(x, \Delta t, \epsilon),$$

where $Q(x, \Delta t, \epsilon)$ *is a function satisfying*

$$\lim_{\epsilon \to 0} \lim_{\Delta t \to 0} Q(x, \Delta t, \epsilon) = 0 \quad \textit{for all } x \in \mathbb{R}.$$

13.6 Adaptive dynamics of the r-process

We now apply the general results of Section 13.4 to the evolutionary process defined in Section 13.1.

13.6.1 Structure coefficient

For the r-process with large population size ($N \gg 1$), we have

$$\sigma = \frac{1+r}{1-r}. \tag{24}$$

This is easily verified by using theorem 13, and noting that in the game

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} \end{array},$$

cooperation is favoured (in the sense $\rho_C > \rho_D$ for all sufficiently small $\delta > 0$) if and only if $br > c$.

13.7 Effective population size

We compute the effective population size N_{e} using the approach of Kimura (1964). We consider neutral drift between types A and B under the r-process.

Let the random variable $X(t)$ represent the (relative) frequency of A individuals at time t. (Here, time t refers to time-steps in the r-process, rather than the continuous-time process considered in Section 13.2). To find N_{e} we must compute the variance of $X(t+1)$ conditioned on $X(t) = p$.

We let the random variable $Y_j(t)$ denote the number of A's (0, 1, or 2) among the jth pair, for $j = 1, \ldots, N/2$ at time t. Thus $X(t) = \frac{1}{N} \sum_{j=1}^{N/2} Y_j(t)$. Conditioned on $X(t) = p$, we have, for each $j = 1, \ldots, N/2$,

$$Y_j(t+1) = \begin{cases} 0 & \text{with probability } r(1-p) + (1-r)(1-p)^2, \\ 1 & \text{with probability } (1-r)\,p\,(1-p), \\ 2 & \text{with probability } rp + (1-r)\,p^2. \end{cases}$$

Based on the above probabilities, we have

$$\mathrm{Var}[Y_j\,(t+1)\ \mid X(t)=p] = 2(1+r)p\,(1-p)\,.$$

Now, since the $Y_j\,(t+1)$ are independent when conditioned on $X(t)=p$, we have

$$\begin{aligned}
\mathrm{Var}[X\,(t+1)\ \mid X\,(t)=p] &= \mathrm{Var}\left[\frac{1}{N}\sum_{j=1}^{N/2} Y_j\,(t)\ \mid X\,(t)=p\right] \\
&= \frac{1}{N^2}\,\mathrm{Var}\left[\sum_{j=1}^{N/2} Y_j\,(t)\ \mid X\,(t)=p\right] \\
&= \frac{1}{N^2}N(1+r)p(1-p) \\
&= (1+r)\,\frac{p\,(1-p)}{N}.
\end{aligned}$$

The effective population size is defined by equating the above variance to the corresponding variance in a haploid Wright-Fisher model with population size N_{e}. That is, we set

$$(1+r)\,\frac{p\,(1-p)}{N} = \frac{p\,(1-p)}{N_{\mathrm{e}}}.$$

This yields

$$N_{\mathrm{e}} = \frac{N}{1+r}. \tag{25}$$

13.8 Adaptive dynamics

We turn now to the adaptive dynamics of x under the r-process. We suppose that mutants arrive at a constant rate per unit time: $u(x)=u$. Substituting the values of σ and N_{e} from Eqs. (24) and (25) into Eq. (1), we obtain the following deterministic approximation to the adaptive dynamics of game strategy in the r-process:

$$\dot{x} = \frac{N-1}{1+r}\,\frac{u\epsilon^2}{\pi\,(x,x)}\left(\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}}x'}\right|_{x'=x}\pi\,(x',x) + r\,\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}}x'}\right|_{x'=x}\pi\,(x,x')\right). \tag{26}$$

We now apply this result to the games introduced in the main text.

13.8.1 Game 1

Game 1 has payoff function

$$\pi\,(x,y) = ay - x^2.$$

The partial derivatives are

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}}x'}\right|_{x'=x}\pi\,(x',x) = -2x, \qquad \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}}x'}\right|_{x'=x}\pi\,(x,x') = a.$$

So Eq. (26) becomes

$$\dot{x} = \frac{N-1}{1+r}u\epsilon^2\,\frac{-2x+ar}{ax-x^2}.$$

13.8.2 Game 2

Game 2 has payoff function

$$\pi(x, y) = ay - xy.$$

The partial derivatives are

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi(x', x) = -x, \qquad \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi(x, x') = a - x.$$

So Eq. (26) becomes

$$\dot{x} = \frac{N-1}{1+r} u\epsilon^2 \frac{(-x + r(a-x))}{10x - x^2}.$$

13.8.3 Game 3

Game 3 has payoff function

$$\pi(x, y) = a \min(x, y) - x^2. \tag{27}$$

The partial derivatives are

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi(x', x) = a/2 - 2x, \qquad \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi(x, x') = a/2.$$

So Eq. (26) becomes

$$\dot{x} = \frac{N-1}{1+r} u\epsilon^2 \frac{a/2 - 2x + ar/2}{ax - x^2}. \tag{28}$$

13.8.4 Game 4

Game 3 has payoff function

$$\pi(x, y) = a \max(x, y) - x^2. \tag{29}$$

The partial derivatives are

$$\left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi(x', x) = a/2 - 2x, \qquad \left.\frac{\partial_{\mathrm{s}}}{\partial_{\mathrm{s}} x'}\right|_{x'=x} \pi(x, x') = a/2.$$

So Eq. (26) becomes

$$\dot{x} = \frac{N-1}{1+r} u\epsilon^2 \frac{a/2 - 2x + ar/2}{ax - x^2}. \tag{30}$$

13.9 Adaptive dynamics in the large population limit

The canonical equation is formulated for a fixed population size, where we look at a limit in which Δt and ϵ are going to 0, and where we assume that the mutation rate u is sufficiently small, so that we can treat fixation or extinction of mutants as instantaneous. In the main text we see that this describes the dynamics in the simulations in some games (Games 1, 2 and 3) better than in others (Games 4 and 5). In the latter two cases, this is due to mutation rate u not being small enough for the population to be close to monomorphic

almost all of the time. In Game 4 they would actually have to be outrageously small to keep the population monomorphic, and in Game 5 the same would be necessary to prevent bifurcations.

There are however also other limits one could consider. One of the variables we could include in the limit taking is population size N. Such alternative limits would not always give different results, but in particular for Game 3 there would be a noticeable difference. One example is the order of limits described by Champagnat et al. (2006). This involves two steps:

1. The limits $N \to \infty$ (large population) and $u \to 0$ (rare mutation) are taken so that the inequalities
$$e^{-CN} \ll u \ll \frac{1}{N \log N} \qquad \text{for all } C > 0$$
are maintained. Simultaneously, time is rescaled by the factor $1/(Nu)$, so that the expected time until the appearance of a new mutation remains constant under the above limits.

2. The limit $\epsilon \to 0$ (small mutational steps) is taken. Simultaneously, time is rescaled by the factor $1/\epsilon^2$ so that the expected change in trait value $\mathbb{E}[\Delta x]$ remains constant to first order in Δt.

Under these limits, we expect that in Game 3, the dynamics within the interval $ra/2 < x < a/2$, where both increases and decreases in trait value are disadvantageous, would be an order of magnitude slower than outside this interval, where mutations in one direction are actually advantageous. This is because for most evolutionary models, the fixation probability of a given disadvantageous mutation goes to zero exponentially fast as $N \to \infty$. We expect this result to hold for the r-process as well; thus we expect no fixation of new mutations in the interval $ra/2 < x < a/2$ once the limit in Step 1 is taken.