Journal of Theoretical Biology 299 (2012) 64–80

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Group selection and inclusive fitness are *not* equivalent; the Price equation vs. models and statistics

Matthijs van Veelen [a,]*, Julián García [b], Maurice W. Sabelis [c], Martijn Egas [c]

[a] CREED, Universiteit van Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
[b] Max-Planck-Institut für Evolutionsbiologie, Germany
[c] Instituut voor Biodiversiteit en Ecosysteem Dynamica, Universiteit van Amsterdam, The Netherlands

ARTICLE INFO

Available online 7 August 2011

Keywords:
Group selection
Inclusive fitness
The Price equation
Price's theorem
Dynamical (in)sufficiency

ABSTRACT

It is often suggested that any group selection model can be recast in terms of inclusive fitness. A standard reference to support that claim is "'Quantitative genetics, inclusive fitness, and group selection" by Queller (1992) in the American Naturalist 139 (3), 540-558. In that paper the Price equation is used for the derivation of this claim. Instead of a general derivation, we try out a simple model. For this simple example, we find that the result does not hold. The non-equivalence of group selection and kin selection is therefore not only an important finding in itself, but also a case where the use of the Price equation leads to a claim that is not correct.

If results that are arrived at with the Price equation are not correct, they can typically be repaired by adding extra assumptions, or explicitly stating implicit ones. We give examples with relatively mild and with less mild extra assumptions. We also discuss why the Price equation is often referred to as dynamically insufficient, and we try to find out what Price's theorem could be.

> **The Dude**: This is a very complicated case, Maude. You know, a lotta ins, lotta outs, lotta what-have-yous. A lotta strands to keep in my head, man. Lotta strands in old Duder's head.
>
> *The Big Lebowski*

1. Introduction

George R. Price produced two of the most influential papers about the evolution of cooperation in the last 50 years. One of them, written together with Maynard Smith (Maynard Smith and Price, 1973) is about why conflicts between animals do typically not escalate. In order to be able to predict which strategies for conflict will evolve, it introduces the notion of an evolutionarily stable strategy (ESS). This has become the central concept in evolutionary game theory, together with the replicator dynamics that was introduced by Taylor and Jonker (1978). There is no doubt that evolutionary game theory in general and the idea of an ESS in particular has been essential for understanding the evolution of cooperation. In models with mutation and selection, the ESS is the most natural refinement of a Nash equilibrium, and to formulate a model and look for evolutionarily stable strategies has become a standard approach.

The other paper—this one single authored—introduces what is now known as the Price equation (Price, 1970). This paper has also been very influential, and the equation is regularly described as giving a simple, but very deep insight into the fundamentals of population genetics (see for instance Frank, 1995; Grafen, 2002; Gardner, 2008). Countless papers have been written using the Price equation, and its fame as the equation that describes the evolution of altruism has given $\overline{w}\Delta z = cov(w,z)$ in biology something of the appeal that $E = mc^2$ has in physics. This appeal is enhanced by Price's remarkable life story, and his equation has therefore become the nucleus of the biography by Harman (2010), where scientific thinking about the evolution of selflessness in general, all the way from Fisher, Haldane and Wright to Maynard Smith and Hamilton, culminates in the discovery of Price's equation.

There is a difference, though. While the ESS is undisputed as a tool for modelling, the Price equation is not, and nor are the results that are arrived at with it. Especially in the debate about the value of inclusive fitness (Nowak et al., 2010; Gardner et al., 2011) and the relation between group selection and inclusive fitness (Queller, 1992; Sober and Wilson, 1998; Wilson and Wilson, 2007; Traulsen and Nowak, 2006; Lehmann et al., 2007; Killingback et al., 2006; Grafen, 2007a; Van Veelen, 2009, 2011a,b; Wild et al., 2009; Wade et al., 2010; Marshall, 2011a,b) results that are derived with the Price equation are contested. In Van Veelen et al. (2010) we claim that the disagreement about these results is partly caused by the use of the Price equation. If we ignore the abuse of the word covariance, then the Price equation is an identity, and can therefore not be wrong. Its typical use

* Corresponding author. Tel.: +31 20 5255293; fax: +31 20 5255283.
E-mail address: C.M.vanVeelen@uva.nl (M. van Veelen).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.07.025

however confuses probability theory and statistics, as well as identity and causality.

If the Price equation indeed is not a proper tool for doing statistics, nor for making models or deriving predictions, as claimed in Van Veelen (2005), then there are a lot of questions that arise concerning the large literature in which the Price equation is used. Has using the Price equation ever lead to incorrect claims? If the Price equation is bad statistics, then what would good statistics be? Does that imply that these results are all wrong? Is there such a thing as Price's theorem? And why is it called dynamically insufficient? In this paper we will try to address these issues. The different sections in this paper are therefore somewhat loosely connected, as they answer different questions concerning the Price equation and the literature using it. The central part however concerns the question whether or not using the Price equation has ever lead to incorrect results.

Queller (1992) is regularly referred to as support for the widely held belief that models of group selection and inclusive fitness are equivalent (see for instance Okasha, 2010). The paper uses the Price equation to show that both group selection models and inclusive fitness work for the same reasons if they do, and fail for the same reasons if they do not. In Section 3 we will go through all steps of the argument, not with the Price equation, but with an extremely simple example. It turns out that none of the steps of the argument is correct already for a very simple set of models. If the claim is not correct for one example, then it surely cannot be correct in general. This particular result, arrived at with the Price equation, therefore, turns out to be wrong.

Section 4 describes the relation between the Price equation and the statistics literature.

Section 5 looks at the issue of dynamic sufficiency. We argue that the fact that the Price equation is regularly described as limited by dynamic insufficiency is really a symptom of the real problem with the Price equation. An identity itself cannot be dynamically sufficient or insufficient. Models can. We claim that the lack of rigour concerning what the numbers are that go into the Price equation can induce people using it to make implicit assumptions that amount to dynamically insufficient models.

The literature mentions not only the Price equation, but also Price's theorem and Price's rule. While the Price equation can be traced back to Price's work (1970, 1972), this is not true for Price's theorem or Price's rule. Section 6 discusses what Price's theorem could be.

Of course not all results arrived at with the Price equation are wrong—even if the Price equation does not provide a proper proof. In Section 7 we therefore look at the scope for repair of results "derived" with the Price equation. For some results one can simply write down a proper proof without the Price equation. For other results it turns out that we need to make some extra assumptions to repair the result. This indicates that using the Price equation induces assumptions being swept under the carpet. Rederiving results without the Price equation then forces one to get them back from under there.

2. The Price equation

What can go wrong when the Price equation is used for the derivation of a theoretical result can best be explained with the words of a famous Dutch football player. When he was once asked what you should do in order to win a game, Johan Cruijff replied that you should score [at least] one more goal that your opponent. This of course is a funny reply (although it is not sure if it was actually meant as such) because it is both indisputably correct as well as completely useless. It is quite possible that it was Cruijff's way of saying that the question was rather unspecific and broad, but then again, it is equally possible that Cruijff himself actually thought that he had stumbled upon a deep truth. It is also possible that what he really meant to say is "don't play too defensive; you can concede a goal and still win the game". What is important for the analogy with the Price equation is that it is certainly not an answer to the question as the journalist meant it; he or she expected an answer like "play 4-3-3" or "train less" or "don't play too defensive", preferably with an explanation of why that would be the key to winning a game. Johan Cruijff's answer just rephrased what it is to win, and did not suggest how to do it. Still it was correct as any answer can be. But it is not an answer that is of too much use.

Price formulated his equation well before Cruijff formulated this particular footballogism. But even though he cannot possibly be inspired by Cruijff, Price's famous equation and Cruijff's (locally) famous answer share the same basic logic, although with the Price equation this is much harder to see. The Price equation does not concern what happens in a football game, but it is about what happens between two subsequent generations. The numbers that it uses are the genetic compositions of the two generations. Van Veelen (2005) goes into more detail here, but what is most important is that we realize that the numerical input of the Price equation is a list of numbers. It is a list that concerns two generations, and which tracks who is whose offspring. But whatever it reflects, it is crucial to realize that the point of departure is nothing but a list of numbers. This list of numbers is used twice. First we use it to compute the frequencies of the gene under consideration in generations 1 and 2, respectively, and subtract the latter from the former. This amounts to the change in gene frequency. Then we use the same list to compute a few other, slightly more complex quantities. The essence of the Price equation is that these quantities also add up to the change in gene frequency. One way of computing the change in frequency therefore can be rewritten as the other and vice versa. What they are, therefore, is nothing but two equivalent ways to compute the change in gene frequency, given a list of numbers concerning genes in two subsequent generations (see Fig. 1).

What is important to realize, is that this equivalence is tautological. Therefore it is true whatever the numbers are that are on the list. Whether this particular second generation is likely to follow the first or not, the two ways of computing the change in frequency return the same number. Had the list of numbers been

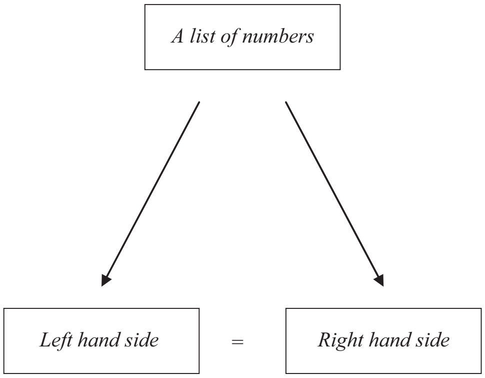

Fig. 1. In its most simple form, the list contains (1) per individual in the parent population the dose of a gene, (2) the same for individuals in the offspring generation, and (3) who in the offspring population got which gene from which parent. The simple way to compute the change in gene frequency (left hand side) is just to calculate the gene frequency in the parent population, calculate the gene frequency in the offspring population, and subtract the latter from the former. The right hand side is much more complex, but nonetheless it is the same change in gene frequency that follows from the list of numbers being what it is; see Box 1, Van Veelen (2005) and www.evolutionandgames.com/price for details. Less simple forms of the Price equation exist, but they are not fundamentally different.

different, then everything, on both sides of the equation, would have been different. While the Price equation tautologically holds for any thinkable transition from one generation to the other, real models should be informative about whether or not such a transition is likely to occur, or, more precisely, whether one transition is relatively likely compared to other transitions.

In football, we are interested in whether or not we are likely to win a game. In biology, we want to know what the chances are that the gene frequency goes up. Cruijff's statement basically amounts to "a game is won if a game is won". That is true, but it is not particularly useful; we want to know what determines the odds of winning. Price's statement has the same form; "the gene frequency goes up if the gene frequency goes up". Or, more precisely: "the change in gene frequency is the change in gene frequency". This is also as true as can be, but again not very useful. The reason that this is harder to see, is that the right hand side of the Price equation has a shape that suggests that there is more to this equality than there really is. It has a covariance-like term in it and a few other terms that look like—but really are not—terms that are used in statistics. This has lead many into temptation to think that the one side of the equation (the one that looks simple and obviously is the change in gene frequency) is explained by the other (the one that contains a covariance-like term, but that, although that is harder to see, really is nothing but the very same change in gene frequency). Something along those lines could have been correct if there would be a real covariance

Box 1

If we restrict ourselves to population states with one and the same population size, and assume a haploid species which reproduces asexually, we get an extremely simple version of the Price equation. For all transitions, we have the following identity, in which N is the population size, q_i is the genotypic value of individual i, and z_i is the number of offspring in the second generation of individual i from the first generation. This makes $\sum_i q_i$ the sum of genotypic values in the parent generation, $\sum_i z_i q_i$ the sum of genotypic values in the offspring generation, and with the observation that $\sum_i z_i = N$ because of the constant population size, the simple version of the Price equation follows (see also www.evolutionandgames.com/pricefor an interactive tutorial with this and other, less simple versions of the Price equation).

$$\Delta Q = \left[\frac{\sum_i z_i q_i}{N} - \left(\frac{\sum_i z_i}{N}\right)\left(\frac{\sum_i q_i}{N}\right)\right] \quad \text{(i)}$$

The right hand side looks like a covariance, but it is very important to realize that it is not. A covariance is a property of a joint distribution of two random variables. The right hand side here is not that; it is a function of numbers, or variables. If the numbers q_i and z_i are random variables, then the right hand side is also a random variable - and not a covariance. The properties of this random variable can be derived if we know the properties of q_i and z_i. Alternatively, if the numbers are data (realizations of random variables), then the right hand side is the *sample* covariance. This can certainly play a role in statistics as an estimate of the true covariance, but the Price equation does not add anything to what the statistics literature offers here; $\Delta Q =$ "sample covariance" does not add anything to our knowledge of the bias of this estimate, nor does it help us see how it should be used for statistical tests.

Maybe the most unfortunate thing about the Price equation is that the term on the right hand side is denoted as a covariance, even though it is not. The equation thereby turns into something that can easily set us off in the wrong direction, because it now resembles equations as they feature in other sciences, where probabilistic models are used that do use actual covariances. A correct treatment with an actual covariance in it would then be as follows. Suppose we are in a certain population state. For that population state we can *assume* a probability distribution over all possible transitions, and any probability distribution implies a covariance between genotype and number of offspring. This covariance is a number, it reflects the properties of the chance experiment in which a new generation is produced from the current state, and we can leave this number in there as a variable; $Cov(z,q)$. In this simple setting we can show that whatever the value of the true covariance is,

$$\mathbb{E}[\Delta Q] = Cov(z,q) \quad \text{(ii)}$$

where $\mathbb{E}[\Delta Q]$ is the expected change in gene frequency, and $Cov(z,q)$ is the true, assumed covariance. When we do statistics, and the numbers are data, then the idea is that we do not know the true value of $Cov(z,q)$, in which case the *sample* covariance can be a good estimate, depending on the number of observations. Note, however, that this is not what the Price equation does.

If we interpret z_i and q_i as random variables, then (ii) follows from (i) if we take expectations on both sides of the equation, since $\mathbb{E}[\text{"sample covariance"}] = Cov(z,q)$.[1] One could then also say that by replacing the right hand side of (i) by $Cov(z,q)$, but not taking an expectation on the left hand side, Eq. (i) is turned into a meaningless equation, with a random variable on the one, and a number on the other side.

[1] To be perfectly precise, the sample covariance is defined as

$$\frac{N}{N-1}\left[\frac{\sum_i z_i q_i}{N} - \left(\frac{\sum_i z_i}{N}\right)\left(\frac{\sum_i q_i}{N}\right)\right].$$

For N independent draws of (z_i, q_i)-pairs from a joint distribution with a given covariance $Cov(z,q)$, the expected value of the sample covariance would be equal to $Cov(z,q)$. Here however we do not have N independent draws of (z_i, q_i)-pairs, because (1) the q_i's are not really random, as they together simply make up a given parent population, and (2) the fixed population requires that $\sum_i z_i = N$. What we do have is therefore N independent draws of *individuals* for the next generation, given a parent population $\mathbf{q} = (q_1, \ldots, q_N)$, which makes $\mathbf{z} = (z_1, \ldots, z_N)$ a vector of N random variables that are not independent. In such a setting, one can add a hypothetical chance experiment: draw a parent, with each parent equally likely to be drawn (here, that is: the probability of parent j being drawn is $1/N$). Now let z and q be the random variables that are defined as z_j and q_j if parent j is drawn. These random variables have a properly defined covariance—$Cov(z,q)$—and one can easily show that the expected value, for a given $\mathbf{q}$, of

$$\frac{\sum_i z_i q_i}{N} - \left(\frac{\sum_i z_i}{N}\right)\left(\frac{\sum_i q_i}{N}\right)$$

not multiplied by $N/(N-1)$, equals $Cov(z,q)$. Summarized, that is

$$\mathbb{E}[\Delta Q] = \mathbb{E}\left[\frac{\sum_i z_i q_i}{N} - \left(\frac{\sum_i z_i}{N}\right)\left(\frac{\sum_i q_i}{N}\right)\right] = Cov(z,q)$$

which is Eq. (ii). See also www.evolutionandgames.com/price. With large N the difference between the two disappears, as $N/(N-1)$ goes to 1.

instead of a covariance-like term. A real covariance would be a *given* part of an actual model, which determines the odds for a generation 2 to follow a given generation 1. Then the change in gene frequency would be a random variable, the properties of which are determined by a *given* covariance. But the covariance-term in the right hand side of the Price equation is not given. It is a number that varies with the original list of numbers, just as much as the left hand side does. So neither side of the Price equation follows from the other, nor is either side explained by the other (see Box 1).

What the proper thing to do is depends on what the numbers are, and on what type of question we want to answer. If we want to do modelling, then we can assume a probabilistic model, and see what different assumptions at a disaggregate level will imply for expected (changes in) population measures such as gene frequency. Note that this is not what the Price equation does; it just gives two equivalent ways of computing the change in frequency. If, on the other hand, we want to do empirical research—in which case the list of numbers must be actual data—then we should use the data in the list to estimate parameters of an actual model and use them to perform statistical tests. Again, this is not what the Price equation does. (Why the Price equation does not help answering either type of questions is explained in detail in Van Veelen, 2005. This can also be explored on www.evolutionandgames.com/price. This low-threshold interactive guide draws transitions (lists of numbers) from different distributions on demand, and indicates how probability theory and how statistics would deal with those. It also indicates how the Price equation literature deals with them. Also Section 4 compares the Price equation literature to the standard statistics literature).

The fact that the Price equation is not useful as a formal tool for deriving results does not mean that it is useless in general. Writing changes in gene frequencies in the two ways in which the Price equation writes them can be perfectly useful, but more in the sense that it can inspire us. The right hand side of the equation—which can differ in shape, depending on the setting we think of—can help us think what reasonable assumptions at a disaggregate (individual) level could be. These assumptions can then be formulated in mathematical terms, and from them we can, preferably in a theorem-proof form, derive perhaps at first not obvious implications at the aggregate (population) level. But in order for such a result to actually be shown to follow from those assumptions, we need a proper theorem with a proper proof. It is nothing but basic logic that only results that are, or that can be, stated in a theory–proof form should be considered to hold. Quite a few results that are claimed to follow from a Price equation approach may actually survive this check, in the sense that they can be formulated and properly proven without reference to the Price equation. But some do not, which underscores the importance of actual proofs, rather than "derivations with" or "expansions of" the Price equation.

3. Group selection and kin selection

Queller (1992) compares inclusive fitness models and group selection models using the Price equation. This paper is regularly referred to in order to support the claim that group selection models and inclusive fitness are equivalent, and recently it is also used to interpret experimental results (see for instance Chuang et al., 2010). The claim of the paper is that both group selection models and inclusive fitness work for the same reasons if they do, and fail for the same reasons if they do not. The results can be summarized very shortly as follows (see Fig. 2). If there is non-additivity in the fitness effects—reflecting for instance synergies—then that makes the separation condition fail. This separation condition is a condition that allows for the separation that inclusive fitness makes as well as for the separation that group selection makes. Therefore, if the one separation works, then the other works too, and vice versa. This result is regularly described as proof that group selection and kin selection are equivalent (see for instance Okasha, 2010) or, more precisely, that any group selection model can be recast in terms of inclusive fitness (while it is never invoked to claim that also every inclusive fitness model can be recast as a group selection model).

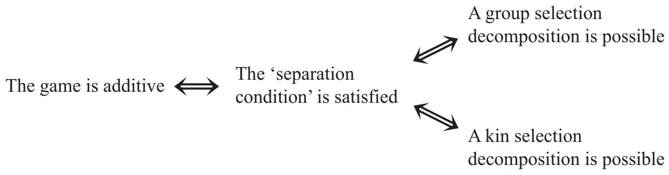
The game is additive
The 'separation condition' is satisfied
A group selection decomposition is possible
A kin selection decomposition is possible

Fig. 2. This summarizes the argument in Queller (1992). Together, these equivalences would imply that a group selection decomposition is possible if, and only if, a kin selection decomposition is possible. However, the example shows that none of the three implications above hold, already for a very simple set of models. Whether or not the game is additive depends on parameter d. Whether or not the separation condition is satisfied depends on parameters r, b and $\mathbf{P}$. A group selection decomposition is always possible, a kin selection decomposition only if $d=0$. Also the overall statement is therefore not correct; additivity of the game is required for the kin selection decomposition to be possible, but not for the group selection decomposition.

One reason to re-examine this result is that Van Veelen (2009, 2011a,b) looks at a simple set of group selection models, and shows that inclusive fitness gives the correct prediction only for a well-defined strict subset of this set of models. That seems at odds with the result from Queller (1992) result, which suggests that inclusive fitness would give the correct prediction for *all* group selection models. The results in Van Veelen (2009) were therefore called into question in Marshall (2011a)—see also Van Veelen (2011a) for a response—and the same argument was repeated in Marshall (2011b) and Gardner et al. (2011). Similar debates are Lehmann et al. (2007) reacting to Traulsen and Nowak (2006) and Grafen (2007b) reacting to Killingback et al. (2006), and also Wild et al. (2009) and Nowak et al. (2010) express opposing views on the reach of inclusive fitness. It therefore seems worthwhile to see what causes these discrepancies and try to find out if it could be that it is the Price equation that sends us barking up the wrong tree.

The core of Queller (1992) is a general claim, arrived at with the Price equation. In order to explore the validity of the general claim, it can be worthwhile to look at a very simple model, and try to see whether or not the claims in Queller actually are true for the simple example. After all, if the claim holds in general, then it should also hold for that simple example. This turns out not to be the case. Fig. 2 depicts Queller's argument, which has a few equivalences in it. All chains in the argument turn out not to be correct. First, for the simple example, whether or not fitness effects are additive turns out not to bear on whether or not his separation condition is satisfied. It is possible that Queller's (1992) separation condition holds, while fitnesses are not additive, and that it does not hold, even if fitnesses are additive. In turn, whether or not Queller's (1992) separation condition is satisfied has no implications for whether or not the decompositions made by inclusive fitness and group selection are possible. An inclusive fitness separation may be possible, while Queller's separation condition is not satisfied, and not possible, even if the condition is satisfied, while a group selection separation is always possible. And not only are the different chains in the argument incorrect, also the overall statement is not true for our example. The separation made by inclusive fitness turns out to be possible if fitnesses are additive, while the group selection separation is always possible. The simple example therefore also shows that the two do not, as claimed, work or fail for the same reasons.

Because the group selection/inclusive fitness debate is controversial, we should emphasize that this should not be taken as an argument in favour of, or against, either group selection theory or inclusive fitness. It is only a check on the claim whether or not separations are *possible*, and makes no claim whatsoever about how *helpful* either separation is.

Also, it should be noted that there are models that are not group selection models, but for which inclusive fitness can give the correct prediction. One example is Grafen (2007a), who reanalysed results on the cycle from Ohtsuki and Nowak (2006), and concluded that there an inclusive fitness approach also leads to the correct prediction. Lion et al. (2011) also argue that there are models that are not group selection models, where an inclusive fitness approach gives the correct prediction.

3.1. The simple example

3.1.1. The dynamics

For the simple example, it is relatively easy to use the replicator dynamics, as defined in Van Veelen (2011b) for n-player games and population structure. This is a perfectly regular, ordinary model (or set of models). It is also relatively easy to work with, because it is deterministic. The shortest way to write this generalized replicator dynamics is as follows:

$$\dot{p} = p(1-p)[\overline{\pi}_C - \overline{\pi}_D] \tag{1}$$

where $\overline{\pi}_C$ and $\overline{\pi}_D$ are the average payoffs of C-players and D-players, respectively, incorporating population structure. In continuous time, $\dot{p}$ is the equivalent of $\Delta\overline{G}$.

The replicator dynamics has a great attraction. Because it assumes an infinite population setting in which the population shares evolve deterministically, there is no actual uncertainty in the dynamics of the population shares of the strategies. This implies that many subtleties and possible sources for confusion concerning covariances and sample covariances disappear (see also Section 7.5).

3.1.2. The population structure

The population structure is defined by frequencies of different types of groups. Groups can be composed of 0 cooperators and n defectors, 1 cooperator and $n-1$ defectors, and so on, and a population state will be characterized by the frequencies of those different types of groups. They are denoted by $f_i, i=0,\ldots,n$, where f_i is the frequency of groups with i cooperators and $n-i$ defectors in it. In order for $f=(f_0,\ldots,f_n)$ to be a consistent population state, these frequencies have to satisfy the following conditions; $0 \le f_i \le 1$ for all i and $\sum_{i=0}^{n} f_i = 1$. Also, obviously, the frequency of cooperators in the population as a whole is given by $p=(1/n)\sum_{i=0}^{n} if_i$.

3.1.3. The game

In Van Veelen (2011a), the payoffs are denoted by $\pi_{C,i}, i=1,\ldots,n$ and $\pi_{D,i}, i=0,\ldots,n-1$, which are the payoffs to a cooperator, resp. defector, if there are in total i cooperators in a group. For the simple example, we will use a payoff matrix from Queller's (1985).[2]

$$\begin{array}{ccc} & C & D \\ C & b-c+d & -c \\ D & b & 0 \end{array}$$

In other words, we take $\pi_{C,1}=-c$, $\pi_{C,2}=b-c+d$, $\pi_{D,0}=0$ and $\pi_{D,1}=b$.

The replicator dynamics can also deal with entries in the payoff matrix that are frequency dependent (see Taylor and Jonker, 1978, in which the replicator dynamics are presented in a much more general form than they are normally used. Van Veelen (2009) also allows for c and b to be frequency dependent, and see also Van Veelen, 2011a,b, and Appendix C). Here we make a counterexample, which implies that if an example with payoffs that are not frequency dependent contradicts the result, examples with frequency dependence will not change that.

3.1.4. Genotype and phenotype

We will, for simplicity, assume throughout that phenotype and genotype are binary; $P,G \in \{0,1\}$. In Sections 3.3.1–3.3.3 we will assume that phenotype and genotype are the same. In Section 3.4 we will assume that having the cooperative genotype only implies a certain probability of expressing it; $\mathbb{P}(P=1)=\mathbf{P}\cdot G$. That is, if your genotype is 0, then your phenotype will be zero too, but if your genotype is 1, then your phenotype is 1 with probability $\mathbf{P}$ and 0 with probability $1-\mathbf{P}$, with $0 \le \mathbf{P} \le 1$. This makes it a very simple, relatively tractable model for which we can explore whether or not Queller's general statements, derived with the Price equation, hold.

3.2. Claims and results

For the simple example it is shown that the separation condition is satisfied if $r=0$, $b=0$, $\mathbf{P}=0$ or $\mathbf{P}=1$, while the separation condition is not satisfied at frequencies $p \in (0,1)$ if $r>0$, $b>0$, and $0<\mathbf{P}<1$. On the other hand, if we look at the possibility of decomposing fitness, then it turns out that the relevant condition for the inclusive fitness decomposition is that $d=0$, while a group selection decomposition, as described in Queller (1992), is always possible. Again, how useful such a decomposition is, is a matter of debate. The important point however is that it is always possible, whether or not $d=0$. This clearly contradicts the claim in Queller (1992), as all of the parameters that determine whether the separation condition holds turn out to be irrelevant for the actual decompositions. Also the overall claim is contradicted; sometimes a group selection decomposition is, and a kin selection decomposition is not possible.

In order to ease our way into the model, we start with the case where phenotype and genotype are the same, or, in other words, where $\mathbf{P}=1$. This implies that the separation condition from Queller (1992) is satisfied. For this case we will compare the possibilities of the two ways of decomposing fitness. Then we will move on to the more general case with $0 \le \mathbf{P} \le 1$, summarize the separation condition for this model, and look again at the two ways of decomposing fitness. This is followed by a discussion of the separation condition and what that implies for this model. This discussion is perhaps the most difficult to follow, but it is also the most important, because it relates the good intuition that is behind it to standard, but rigorous statistics.

A lot of the work is actually rather dull algebra, so the sections below will all refer to appendices where all kinds of variances, covariances and conditions are computed.

3.3. Phenotype=genotype

3.3.1. Inclusive fitness

In order to achieve the inclusive fitness decomposition, we follow the section "Inclusive Fitness" from Queller (1992), and apply it to the replicator dynamics example. We will use a natural correspondence for $n=2$ between relatedness and frequency on the one hand and frequencies of different types of groups on the

[2] Section 7.3 and Appendix C discuss the distinction between payoffs and fitness effects. Here the matrix entries are payoffs.

other (see Bergstrom, 2003, or Section 3.3 of Van Veelen, 2011b); if we take $f_0(p)=(1-r)(1-p)^2+r(1-p)$, $f_1(p)=(1-r)2p\,(1-p)$ and $f_2(p)=(1-r)p^2+rp$, then r can naturally be interpreted as relatedness.

If we combine the payoff matrix from Queller (1985)[3] with this population structure in the generalized replicator dynamics, we get the following dynamics (see Van Veelen, 2011a,b)

$$\dot{p}=p(1-p)[rb-c+(r+(1-r)p)d] \tag{2}$$

With the translation provided in Van Veelen (2011a), this is equivalent to Eq. (2) in Queller (1985), assuming that genotype and phenotype are the same, and for $d=0$ to Eq. (16) in Queller (1992), again assuming that genotype and phenotype are the same.

As Marshall (2011a) and Queller (1985) already pointed out, this is *not* Hamilton's rule if $d\neq 0$, and therefore it is better to refer to it as Queller's rule. If indeed $d\neq 0$, then this prevents the separation of fitness effects and population structure. If $d=0$, then we have equal gains from switching—or additive fitness effects—and the separation is possible.

In order to visualize the separation, which is possible if $d=0$, we can write what the formula implies if we assume that d is indeed 0;

$$\dot{p}>0 \Leftrightarrow r>\frac{c}{b} \tag{3}$$

Here we have population structure on the one hand of the inequality sign (r) and fitness effects on the other (b/c). Van Veelen (2011a) shows that inclusive fitness is only a meaningful concept if it means that it allows for such a separation of population structure and fitness effects. Van Veelen (2011b) shows that such a separation is only possible for $d=0$.

3.3.2. Group selection

In order to achieve the most natural separation of between-group selection and within-group selection, we will follow the section "Group Selection" from Queller (1992) and apply his recipe to the same example.

At first there seems to be a problem with the application of the Price equation to our particular replicator dynamics. The equation, as formulated by Price (1970), assumes that we know who in the next generation is whose offspring. This is not specified in the (generalized) replicator dynamics. What we can do, however, is start by simply computing the first term in the separation suggested in Queller (1992), which is the (so-called) covariance between the average phenotype in the group and average fitness (note that phenotype and genotype still are the same here). Once we have the first term, this implies that the other term must be the remainder. It turns out that this gives a perfectly reasonable separation of the effects of between-group and within-group selection.

The more detailed derivation, as well as a suggestion of the more general version, are in Appendix A. The separation we arrive at there is

$$\dot{p}=p(1-p)\overbrace{\left[\frac{b-c}{2}(1+r)+[r+(1-r)p]d\right]}^{\text{between group selection}}-p(1-p)\overbrace{\left[\frac{b+c}{2}(1-r)\right]}^{\text{within group selection}} \tag{4}$$

The first term on the right hand side, including the $p(1-p)$, is the $Cov(G_g,W_g)$ from Queller (1992) (see Queller, 1992, p. 548, and the derivation in Appendix A). It is also clear that (2) and (4) are equivalent.

This is what we get when we apply the group selection separation from Queller (1992) to this simple example. It also makes perfect sense as a separation of the two effects, because (1) $b-c$ is the (baseline) efficiency gain to a group of cooperating, and hence should feature in the between group term, (2) d is the bonus in uniform cooperative groups, and hence should *only* feature in the between-group selection term, (3) $b+c$ is how much the cooperators in mixed groups put themselves at a disadvantage relative to the defectors, hence should feature in the within-group selection term, and (4) a high r increases the between-group selection term, and decreases the within-group selection term.

Again we can visualize this particular separation:

$$\dot{p}>0 \Leftrightarrow \overbrace{\left[\frac{b-c}{2}(1+r)+[r+(1-r)p]d\right]}^{\text{between group selection}} > \overbrace{\left[\frac{b+c}{2}(1-r)\right]}^{\text{within group selection}} \tag{5}$$

This separation is also perfectly consistent with the simple summary of how group selection works in Wilson and Wilson (2007, p. 345).

3.3.3. Do these separations work or fail for the same reasons?

If we look at the first separation—the inclusive fitness one as reflected in (3)—then we see that it is possible for $d=0$ (or, more general, under generalized equal gains from switching; see Van Veelen, 2009, 2011b) and not possible for $d\neq 0$. So it is possible if fitness effects are additive, and not possible if they are not.

If we look at the second separation—the group selection one as reflected in (4)—then it is first of all worth observing that *both* terms actually depend on population structure as well as fitness effects. In that sense they both are compound terms. This implies that perhaps this decomposition may be of limited use. But the more important observation here is that the value of d has no effect whatsoever on the *possibility* of this separation. Whether d equals 0 or not, it is always possible to separate the total effect in a between-group term and a within-group term. This also generalizes to groups larger than 2. Hence, contrary to the claim in Queller (1992, p. 555), non-additive fitness effects do *not* prevent us to separate the total effect into a between- and a within-group effect.

3.4. What if phenotype and genotype are not necessarily the same?

Now suppose that having a gene for cooperation only implies that, independent of the group composition, it is expressed (that is, it leads to cooperative behaviour) with probability $\mathbf{P}$. The kin selection decomposition then becomes (see Appendix B.1)

$$\dot{p}=\mathbf{P}p(1-p)(rb-c+(r+(1-r)p)\mathbf{P}d) \tag{6}$$

Again, this is *not* Hamilton's rule if $d\neq 0$, and therefore it is better to refer to it as a version of Queller's rule. If indeed $d\neq 0$, then this prevents the separation of fitness effects and population structure. If $d=0$, then we have equal gains from switching, and the separation is possible. Note that if $d\neq 0$, then $\mathbf{P}$ not only matters for the speed of selection, but also for what the fixed point of this dynamics is.

In order to visualize the separation, which is possible if $d=0$, we can assume that d is indeed 0 and write an implication of the formula as:

$$\dot{p}>0 \Leftrightarrow r>\frac{c}{b} \tag{7}$$

Of course it is only possible to separate population structure (reflected by the r) from the payoffs or fitness effects (reflected by $\frac{b}{c}$) if $d\neq 0$.

[3] Section 7.3 and Appendix C discuss the distinction between payoffs and fitness effects. Here the matrix entries are payoffs.

The group selection decomposition becomes (see Appendix B.3):

$$\dot{p} = \mathbf{P}p(1-p)\overbrace{\left[\frac{b-c}{2}(1+r)+[r+(1-r)p]\mathbf{P}d\right]}^{\text{between group selection}}-\mathbf{P}p(1-p)\overbrace{\left[\frac{b+c}{2}(1-r)\right]}^{\text{within group selection}} \tag{8}$$

Again, we can visualize this particular separation:

$$\dot{p}>0 \Leftrightarrow \overbrace{\left[\frac{b-c}{2}(1+r)+[r+(1-r)p]\mathbf{P}d\right]}^{\text{between group selection}} > \overbrace{\left[\frac{b+c}{2}(1-r)\right]}^{\text{within group selection}} \tag{9}$$

Again, the formulas show that **P** has an effect on the speed as well as the direction of selection.

These two ways to decompose the change in frequency still imply that whatever was said about separation for the case where phenotype and genotype are equal, still applies here; one cannot say that these separations work or fail for the same reasons.

Because the link between the two decompositions was supposed to run through Queller's separation condition, it is also interesting to look at how this condition relates to these separations. The separation condition (see Section 3.5) is satisfied for our example if $r=0$, or if $b=0$, or if $\mathbf{P}=0$ or $\mathbf{P}=1$, while the separation condition is not satisfied if $r>0$, $b>0$ and $0<\mathbf{P}<1$. The actual separations for our example above however show that all possibilities and impossibilities of separation do not depend on the values of r,b and **P**.

3.5. *The separation condition*

Queller (1992) uses the Price equation, and a familiar problem with the use of the Price equation is that it is not always clear whether this equation is meant to answer a question concerning a (probabilistic) model, or a statistical question. The mix-up between probability theory and statistics is a returning problem in the Price equation literature (Van Veelen, 2005; Van Veelen et al., 2010, see also www.evolutionandgames.com/price, and this paper is no exception. On page 543 we find the following equation (Eq. (3) in Queller, 1992):

$$G=\alpha_G+\beta_{GP}P+\varepsilon_G \tag{10}$$

where P is the phenotypic value and where α_G and β_{GP} are the intercept and slope of the best-fit regression equation. The ε_G's are residuals, which may differ for each individual and which describe the difference between breeding value predicted by the regression and the actual breeding value.

The right hand side of this equation is then used to replace G in Eq. (1) in Queller (1992)—$\Delta\overline{G}=Cov(G,W)$—after which we arrive at Eq. (4) in Queller (1992)

$$\Delta\overline{G}=Cov(\alpha_G,W)+\beta_{GP}Cov(P,W)+Cov(\varepsilon_G,W) \tag{11}$$

The separation condition is then that $Cov(\varepsilon_G,W)=0$.

The first thing to notice is that the presence of a best-fit regression equation and residuals, inevitably implies that *data* have been used to arrive at those α_G, β_{GP} and ε_G.[4] The first two then are estimates of some true values, assuming that the relation between genotype and phenotype is indeed a linear one as suggested by Eq. (10). In a modelling context, on the other hand, best-fit regressions and residuals are out of place; if we are only modelling, we *know* all parameters, and we do not need any best-fit regression equation, nor are there residuals (note that residuals and disturbances are not the same!) This implies that the only setting in which Eq. (10) and its description makes sense is one in which we have been using data in order to do statistics.

Yet everything else in the paper is about modelling; the papers central message concerns properties of different models, which are compared to each other. As such, it is natural that the paper is not at all about standard statistical concerns, such as biases of estimates or hypothesis testing (which are the standard things to look at in statistics). Given these two mutually exclusive ingredients, it is not clear what Eq. (11) is supposed to be. Is it an equation that does statistics, and concerns data? Or is it an equation that describes properties of a model? It is in any case not fully consistent with either option.

Also the separation condition itself does not have a consistent interpretation. On the one hand, ε_G is the result of a combination of an estimation procedure (OLS, which is short for ordinary least squares) and data. The W on the other hand belongs to a model, the properties of which we want to describe, but that have nothing to do with data. As such, it is not a well-defined condition, because it is not clear whether it is a condition that a *model* has to satisfy, or if it is a restriction that the *data* have to satisfy.

That however does not mean that we cannot distill what the underlying intuition could be and see if we can describe a consistent version of the condition that $Cov(\varepsilon_G,W)=0$, or at least find out what it would imply for our simple example. There are actually two routes along which we can find an answer to the question what it would imply for our simple example. Fortunately both routes lead to the same answer.

3.5.1. *Route 1*

Suppose that indeed, as in our example, phenotype and genotype are binary—$P,G\in\{0,1\}$. Suppose furthermore that indeed $\mathbb{P}(P=1)=\mathbf{P}\cdot G$, that is, if the genotype is 0, then the phenotype will be 0 too, and if the genotype is 1, then the phenotype is 1 with probability **P** and 0 with probability $1-\mathbf{P}$, $0\le\mathbf{P}\le 1$. Suppose furthermore that we do not know the actual value of **P**, and that we would like to estimate it statistically—which implies that the numbers are data. A very simple and accurate procedure would be to count all instances where $P=1$ and $G=1$ (under the assumption of the model, $G=1$ is redundant) and divide it by all instances where $G=1$. This is equivalent to the solution of the least square regression of P on G. Note that here we will follow the statistical convention and let α_P and β_{PG} denote the true values and $\widehat{\alpha}_P$ and $\widehat{\beta}_{PG}$ their estimates.

$$P=\alpha_P+\beta_{PG}G+\varepsilon_P \tag{12}$$

With this OLS regression we get

$$\widehat{\beta}_{PG}=\frac{\text{Sample covariance }(G,P)}{\text{Sample variance }(G)}=\frac{\overline{P}-\overline{P}\,\overline{G}}{\overline{G}-\overline{G}\,\overline{G}}=\frac{\overline{P}}{\overline{G}} \tag{13}$$

which is exactly what the straightforward method gave us. The expected value of this estimator is $\mathbb{E}[\widehat{\beta}_{PG}]=\mathbb{E}[\overline{P}/\overline{G}]=\mathbf{P}$. This is also what we find when we use

$$\mathbb{E}[\widehat{\beta}_{PG}]=\frac{Cov(G,P)}{Var(G)}=\frac{p(1-p)\mathbf{P}}{p(1-p)}=\mathbf{P}$$

Furthermore $\widehat{\alpha}_P=\overline{P}-\widehat{\beta}_{PG}\overline{G}$ and

$$\mathbb{E}[\widehat{\alpha}_P]=p\mathbf{P}-\frac{p(1-p)\mathbf{P}}{p(1-p)}p=0$$

The regression in Queller goes in the other direction (this is Eq. (10) above).

$$G=\alpha_G+\beta_{GP}P+\varepsilon_G$$

[4] In statistics it is common practice to distinguish between the true value and the estimate, for instance by denoting the true values by α_G and β_{GP} and the estimates by $\widehat{\alpha}_G$ and $\widehat{\beta}_{GP}$ or a_G and b_{GP}. Here the α_G and β_{GP} are described as the estimates, which is a bit unfortunate given standard statistical notation.

With this OLS regression we get

$$\widehat{\beta}_{GP} = \frac{\text{Sample covariance}(G,P)}{\text{Sample variance}(P)} \tag{14}$$

for which

$$\mathbb{E}[\widehat{\beta}_{GP}] = \frac{Cov(G,P)}{Var(P)} = \frac{p(1-p)\mathbf{P}}{p\mathbf{P}(1-p\mathbf{P})} = \frac{1-p}{1-p\mathbf{P}}$$

Furthermore

$$\widehat{\alpha}_G = \overline{G} - \widehat{\beta}_{GP}\overline{P} \quad \text{and} \quad \mathbb{E}[\widehat{\alpha}_P] = p - \frac{1-p}{1-p\mathbf{P}}p\mathbf{P} = \frac{p-p\mathbf{P}}{1-p\mathbf{P}}$$

The expected value of both estimators is now frequency dependent.

This regression, although the wrong way round if we want to find out from the data how phenotype depends on genotype (which is how the actual causality runs) can nonetheless be given a meaningful interpretation. If we take the expected values of $\widehat{\alpha}_G$ and $\widehat{\beta}_{GP}$ and fill them in the equation—and if we have many observations, we are relatively confident that the estimators will be close to those values—then $\alpha_G + \beta_{GP}P$ gives us the expected value of G given P. If $P=1$, then

$$\alpha_G + \beta_{GP}P = \frac{p-p\mathbf{P}}{1-p\mathbf{P}} + \frac{1-p}{1-p\mathbf{P}} \cdot 1 = 1$$

which is obviously the expected genotype if $P=1$, because $P=1, G=0$ never occurs. If $P=0$, then

$$\alpha_G + \beta_{GP}P = \frac{p-p\mathbf{P}}{1-p\mathbf{P}}$$

while Bayes rule gives us the same answer

$$\mathbb{P}(G=1|P=0) = \frac{\mathbb{P}(G=1, P=0)}{\mathbb{P}(P=0)} = \frac{p(1-\mathbf{P})}{1-p+p(1-\mathbf{P})} = \frac{p-p\mathbf{P}}{1-p\mathbf{P}}$$

This is the expected genotype conditional on $P=0$, because G and P are binary variables.

Therefore, if we are curious what the expected value of the genotype is, conditional on the phenotype, then this regression can be seen as informative. If new data are generated from the same data generating process, but now the genotype is not observed, then this gives the right conditional probabilities.

One might however be able to do a better job at estimating the chances that a data point is $G=0$ or $G=1$ if these new data also include fitnesses. In the extreme case with $r=1$ (and $n=2$), having the gene but not expressing it leads to a payoff of b with probability $\mathbf{P}$ (the probability with which the other, who has it too because $r=1$, expresses it) or of 0 with probability $1-\mathbf{P}$. Not having the gene always leads to a payoff of 0, so in this extreme setting a payoff of b implies that $G=1$ with certainty, and not with probability $\widehat{\alpha}_G$. In other words, in this case there is extra information to be extracted from the fitnesses.

More generally, we can look at the following covariance. Suppose that we know p and $\mathbf{P}$, and that we take the expected values of $\widehat{\alpha}_G$ and $\widehat{\beta}_{GP}$ again, that is, we take $\alpha_G = (p-p\mathbf{P})/(1-p\mathbf{P})$ and $\beta_{GP} = (1-p)/(1-p\mathbf{P})$. Then we can look at the disturbances that result from filling in the data in the equation $G = \alpha_G + \beta_{GP}P + \varepsilon_G$. In other words, define the disturbances as $\varepsilon_G = G - \alpha_G - \beta_{GP}P$ with $\alpha_G = (p-p\mathbf{P})/(1-p\mathbf{P})$ and $\beta_{GP} = (1-p)/(1-p\mathbf{P})$.

Now we can compute a proper covariance (this is done in Appendix (B.5))

$$\begin{aligned} Cov(\varepsilon_G, W) &= \mathbb{E}[\varepsilon_G W] - \mathbb{E}[\varepsilon_G]\mathbb{E}[W] \\ &= \frac{rbp(1-p)\mathbf{P}(1-\mathbf{P})}{1-p\mathbf{P}} \end{aligned} \tag{15}$$

It is clear that this $Cov(\varepsilon_G, W) = 0$ for all p if $r=0,\ b=0,\ \mathbf{P}=0$ or $\mathbf{P}=1$, while $Cov(\varepsilon_G, W) > 0$ for $0<p<1$ if $r>0, b>0$ and $0<\mathbf{P}<1$.

This also implies that the appropriately constructed *sample* covariance between ε_G and W—where ε_G now are actual residuals, using the estimates $\widehat{\alpha}_G$ and $\widehat{\beta}_{GP}$ instead of the true values α_G and β_{GP}, and not the true disturbance terms—will converge in probability to a non-zero limit for the number of observations going to infinity. So with a large enough sample, one could detect a non-zero covariance with high confidence.

If this condition is satisfied, then it means that no information about the value of G can be extracted from the value of W, after we have used the value of P. If this condition is not satisfied, there is information left to extract from W.

3.5.2. Route 2

There is also a second way to arrive at what the separation condition in Queller (1992) must imply for our simple example. The section “The separation condition” in Queller (1992) also states that, given that the separation condition is satisfied, Eqs. (5) and (8) in Queller (1992) are equivalent. Eq. (5) is $\Delta\overline{G} = \beta_{GP}Cov(P,W)$ and Eq. (8) is $\Delta\overline{G} = \beta_{WP}Cov(G,P)$. The reason why they should be equivalent is that β_{GP} is defined as $\beta_{GP} = Cov(G,P)/Var(P)$ while β_{WP} is defined as $Cov(P,W)/Var(P)$. This equivalence can therefore be summarized with the statement that

$$\Delta\overline{G} = \frac{Cov(G,P)Cov(P,W)}{Var(P)} \tag{16}$$

When we compute all terms in this equation (as is done at the bottom of (B.2)) we find that this is only true if $r=0,\ \ b=0,\ \mathbf{P}=0$ or $\mathbf{P}=1$, while it is not true if $r>0,\ \ b>0$ and $0<\mathbf{P}<1$, which is exactly what the separation condition implies for our example if we follow Route 1. This confirms our interpretation of the separation condition, because Eqs. (5) and (8) in Queller (1992) are only claimed to be equivalent if the separation condition is satisfied.

3.6. Group selection and inclusive fitness are not always equivalent

Queller (1992, p. 541) summarizes his results as follows.

> ⋯ inclusive fitness and group selection [⋯] achieve their simplicity in the same way: the separation of fitness parameters from genetic parameters. As a result, the models succeed and fail in ways that pertain to quantitative genetics models in general. Moreover, inclusive fitness models and group selection models are extremely similar to each other. Their only fundamental difference is in how they choose to decompose fitness. Other differences are trivial matters of the form of presentation. The two models work for the same reason, and they fail (to be exact) under the same condition.

The condition is specified at the end of the paper (p. 555)

> Both fail when non-additive fitness raise the level of complexity beyond what can accurately be described by the two pairs of parameters. [⋯] the group selection models suffer from the same non-additivity problem that has caused problems for inclusive fitness models.

When we work out a simple example, this turns out not to fit this picture. The group selection separation suggested in Queller, and applied to a simple example, turns out to be possible, whether or not fitnesses are additive. Because the within-group selection term and the between-group selection term are both compound terms, that have fitness parameters in them as well as parameters that represent population structure, one can perhaps argue that this separation may be of limited use. But the point concerns the possibility, and not the usefulness of the separation. And the possibility is completely unaffected by the additivity of fitness effects (here: the value of d).[5]

The inclusive fitness separation however is only possible if $d=0$. One can make an extension of the rule that incorporates d (Queller's rule instead of Hamilton's rule), but that rule does not achieve the separation that defines inclusive fitness (see Van Veelen, 2011a,b). The inclusive fitness separation can perhaps be seen as more useful than the separation that group selection suggests, because it actually separates fitness parameters from population structure. It is however not always possible.

The argument in Queller (1992) why both separations fail together is that non-additive fitness makes the model fail his separation condition. It is shown that, whichever position one takes on the usefulness of the different separations, this simply cannot be a correct argument. For our example, whether or not Queller's separation condition is satisfied depends on the values of relatedness r, benefit b, and the expression probability $\mathbf{P}$ of the cooperative behaviour for carriers of the cooperative gene. Additivity on the other hand is determined by the value of the synergy parameter d. For our example, the separation condition therefore has nothing to do with additivity, nor with the possibilities of separation.

4. How the Price equation relates to the statistical literature

A key ingredient of the confusion that surrounds the Price equation is that it uses the word covariance to denote a term that is not a covariance. The term can best be described as a function, the value of which depends on a list of numbers (see Section 2 and Box 1). If those numbers are data, then this term is the *sample* covariance. If those numbers are "just numbers", then this term is not even that. Certainly it is not a covariance, which is a well descript property of the joint distribution of two random variables.

When he published his equation in 1970, George Price worked at the Galton Lab at UCL. There is a touch of irony to the fact that he did not use the insights of Karl Pearson, who was the first Galton professor from 1904 to 1933. Pearson was the first to rigorously define covariance as a property of the joint distribution of two random variables, and distinguish that from the *sample* covariance, which is a random variable itself, and an unbiased estimator of a true covariance. These definitions, and the distinction between them, are at the basis of modern statistics. It is very unfortunate that George Price did not couple his good intuition with Pearson's rigorous mathematics, that was readily available in 1970.

Also the subsequent literature on the Price equation has developed almost completely separately from the probability theory and statistics literature. A telling sign of this is that in the vast Price equation literature, we have not been able to find a single instance of a statistical test being performed, nor of a proper parameter estimate.[6] Nor have we found more general statements concerning the properties of statistical tests—such as the chances of false positives or false negatives—or properties of estimators—such as unbiasedness, having minimum variance or asymptotic efficiency. For instance, if the numbers in the Price equation are data, and covariance-like term in the Price equation were to be used as an estimation of the true covariance (which is possible, because it is the *sample* covariance if it uses data), it would be normal to remark that, with a slight change in the denominator, it is unbiased. While these matters are at the core of the statistics literature—it is standard practice to check for these properties for every new estimator or statistical test proposed—-none of these issues feature in the Price equation literature. This should worry us seriously, because there is no reason to think that statistics in biology should be any different from statistics in any other field of science. The models may differ, but what we should look at and worry about when we test them statistically should be the same.

Something quite similar holds for the Price equation and probability theory. Stochastic properties of models that we tend to care for concern expectations and variances of random variables, and describe asymptotic properties, and how properties of stochastic systems can be close to or far away from what they are in the limit. This is not what features in the Price equation literature, in spite of the fact that evolutionary predictions can perfectly well be stated in terms of stochastic systems. In order to get an intuition for why some common practices in the Price equation literature are in fact impossible to reconcile with probability theory, and also impossible to reconcile with statistics, we have made an online tutorial (www.evolutionandgames.com/price). The program will draw transitions (lists of numbers) from a set of different distributions on demand, and indicates how the Price equation literature deals with those. It also indicates how probability theory and how statistics would deal with the same lists of numbers. This should indicate the peculiar place the Price equation has in the probability theory and statistics literature.

That, however, does not at all mean that the intuition for the Price equation itself, nor for ideas inspired by it, cannot be correct. Nor are all results arrived at with the Price equation are wrong. It is just that even after an inspirational phase with the Price equation, one is still in need of normal probability theory and standard statistics in order to get actual results in theorem–proof form.

5. Dynamic insufficiency

Dynamic insufficiency is regularly mentioned as a drawback of the Price equation (see for example Frank, 1995; Rice, 2004). We think that this is not an entirely accurate description of the problem. We would like to argue that the perception of dynamic insufficiency is a symptom of the fundamental problem with the Price equation, and not just a drawback of an otherwise fine way to describe evolution.

To begin with, it is important to realize that the Price equation itself, by its very nature, cannot be dynamically sufficient or insufficient. The Price equation is just an identity. If we are given a list of numbers that represent a transition from one generation to the next, then we can fill in those numbers in both the right and the left hand side of the Price equation. The fact that it is an identity guarantees that the numbers that appear on both sides of the equality sign are the same. There is nothing dynamically sufficient or insufficient about that (this point is also made by Gardner et al., 2007, p. 209).

A model, on the other hand, *can* be dynamically sufficient or insufficient. For simplicity, we can assume that we have a deterministic model. If we then start with generation 1, we know exactly what generation 2 will be. If the model is dynamically insufficient, then that implies that if generation 2 is different from generation 1,

[5] It is also true that the compoundness of the two terms implies that it does not separate fitness parameters from genetic parameters. The separation we find however is totally consistent with what Queller (1992) suggests that this separation should be—it is just what we get if we apply the suggested separation there to the simple example—as well as with what for instance Wilson and Wilson's (2007) simple summary of group selection would suggest.

[6] Two papers in which the Price equation features together with a parameter estimate are Bowles et al. (2003) and Bowles (2009). However, the statistics in Bowles et al. (2003), are just normal regressions, loosely inspired by the Price equation, performed on simulation data, and commented on in Van Veelen and Hopfensitz (2007). Also Bowles (2009) contains the Price equation as well as an estimate. What is estimated there is Wright's F_{ST}. This estimation is also done in a normal, traditional way, without actually using anything from the Price equation.

then the model does not specify what happens when we are departing from this new starting point. If on the other hand the model is dynamically sufficient, then that implies that we also know where we go from generation 2, even if it is different from generation 1, and where we go from generation 3, and so on.

If the underlying model is indeed dynamically insufficient, then we have only one step. Therefore we also have only a Price equation for one transition, that writes the change in gene frequency of this one step in two equivalent ways. If the underlying model is dynamically sufficient, then we have a sequence of steps. For each of those steps we can write the change in frequency in two ways with an equality sign in between, as the Price equation does. Notice that for a dynamically sufficient model, we know for all terms in the Price equation how they change as the generations follow each other, and the population evolves according to the model. The Price equation just processes the numbers that reflect the changes as time goes by. This includes the covariance-like term, which therefore typically will change along the path that the model describes. In Section 7.5 we describe the model by Page and Nowak (2002), which is an example of a dynamically sufficient continuous time model. They also formulate the Price equation for this model at every point in time, completely with a changing covariance-term. This shows that dynamic insufficiency is not in any way a result of the Price equation itself.

The Price equation is not dynamically sufficient or insufficient. The reason why it is nonetheless typically accused of being dynamically insufficient is that, unlike in Page and Nowak (2002), it tends to be used in splendid isolation, and not in combination with an explicitly stated model. This is likely to trigger projection onto the Price equation of what implicit, or explicit, but informally stated assumptions are thought to imply. An example of an assumption that is so natural that it is sometimes thought that there is no need to state it in the form of a mathematical property is fair meiosis. Price (1970) projected what fair meiosis intuitively is thought to imply onto the Price equation, rather than deriving an actual result of a similar purport (see Van Veelen, 2005).

With or without an explicit model, reasonable researchers will be aware that in a proper model the true covariance might not be constant. This is of course perfectly sensible. It is however very hard to project a covariance that is not constant onto one single Price equation. But it is not the Price equation that is to blame here, it is the fact that we are projecting implicit or informally stated assumptions onto it. The standard idea concerning dynamic insufficiency is that the Price equation may help understanding or explaining what happened in the step from generation 1 to generation 2, but not the subsequent step from generation 2 to generation 3, because the true covariance might change once arrived in generation 2. But that misses the point. If we would have a model that tells us what generation 3 will be, departing from generation 2, then the numbers that come with this new transition can be filled in the Price equation just as well as those reflecting the transition from generation 1 to generation 2 could. The mistake is that the Price equation did not explain or predict for the first transition either. It does not do anything, other than write the change in frequency in two different ways. The intuition that the true covariance might change during a process of selection is very sensible, but it is more important to notice that if we have an actual model that tells us what the dynamics will be, then the Price equation does not provide an answer to any question pertaining to *any* step in the selection process. The conclusion that the Price equation is dynamically insufficient is therefore not correct; it is an identity, and as such it cannot be dynamically sufficient or insufficient. The fact that this conclusion is regularly drawn nonetheless, is informative though; it is a sign that it stimulates us to project our intuition onto it. The Price equation then only feels as a dynamically insufficient straightjacket for our intuition.

6. Is there such a thing as Price's theorem?

The literature sometimes mentions the existence of a Price's theorem (see for instance Rice, 2004; or Gardner et al., 2007, p. 208). Price himself however does not state a theorem, nor have we been able to find an actual theorem by that name anywhere else in the literature. Whenever Price's theorem is mentioned, however, it tends to come with a reference to the paper in which the Price equation is presented (Price, 1970). Since the Price equation is a tautology, and a theorem is a tautology that is not obvious enough to be seen through without a proof, we can conclude that Price's theorem can only be the following.

Theorem 1 (*Price; biology*). *If the left hand side is computed as suggested in Price (1970), and the right hand side too, then they are equal.*

Proof. See Price (1970) or Van Veelen (2005). □

It is a matter of taste if the proof is considered obvious or not so obvious, and therefore whether the term theorem is justified. What is more important is that this theorem ceases to hold if one of the sides is *not* computed as suggested. Crossing off terms on either side—which tends to be done in the right hand side, starting with Price (1970) himself—therefore causes a violation of the condition for the result to hold, unless these terms happen to be zero (see Van Veelen, 2005).

The criticism on the Price equation is not that it is wrong. It is not wrong—as long as no terms are crossed off, and as long as we abide the abuse of the term covariance. The criticism on the Price equation is that, by lack of assumptions, it cannot be used for deriving results that imply predictions. It is equivalent to Cruijff's footballogism; Theorem 1 just states that the change in gene frequency equals the change in gene frequency. It is not incorrect, but it is not very helpful either, equivalent as it is to Johan Cruijff's theorem.

Theorem 2 (*Cruijff; football*). *If team A scores more goals than team B, then team A wins.*

Proof. Follows directly from the definition of winning.

Please note that both Theorem 1 and Theorem 2 do not produce predictions. If a theorem is to produce a prediction, it must have assumptions in it. Starting from those, the theorem then derives a prediction, which can then be tested. If data are collected that do not match the prediction, then we conclude that at least one of the assumptions is not met. In this field, the shape of a theorem that produces a prediction would therefore be as follows.

Theorem 3 (*Biology*). *If the fitness of an individual depends on its own and the other individuals' behaviour according to Assumption* 1, …, *Assumption N, then the behaviour that emerges is more likely to be behaviour A than it is to be behaviour B.*[7]

It is not the fact that Theorem 1 is a tautology that is problematic. All theorems are tautologies, if correct. What is problematic is that the lack of assumptions excludes that it produces a prediction.

The only theorem that qualifies for the label Price's theorem is Theorem 1. We conclude that it indeed holds, but also that it does not help produce a prediction. All it can do is give inspiration as for what interesting or reasonable assumptions could be, and thereby

[7] A possible theorem for football could have the form

Theorem 4 (*football*). *If teams A and B have equally able players, and interactions occur according to Assumption*1, …, *Assumption N, and A plays* 4-3-3 *and B plays* 4-4-2, *then team A is more likely to win than team B.*

help formulating theorems that come in the shape of Theorem 3. The Price equation should remain on the scrap paper though. Once we have assumptions, they should be stated explicitly, and be the starting point for a theorem in the form of Theorem 3.

Queller (1992) also mentions Price's rule. Again, there is no mention of a rule in Price (1970) or Price (1972). There is only an identity.

7. Repairs

A Price equation approach can—perhaps somewhat roughly, and with exceptions—be described as "write (a version of) the Price equation, and project your intuition onto it". If this intuition happens to be spot on, then the "result" is correct, even though the Price equation is no proof. But the Price equation is in no way a guarantee for a correct result, because our intuition is not necessarily correct. If our intuition is just wrong, then there is nothing that can be done. But if our intuition is only a bit off, then we can typically make extra assumptions to make the result correct. The "results" in the Price equation literature therefore come in a riot of colours and flavours, and are to differing degrees and in different ways in need of repair. This section gives a few examples and starts with one that needs no repair at all.

7.1. *Rousset and Billiard (2000)*

The reference paper for inclusive fitness models by Rousset and Billiard (2000) is an example where no repairs are needed. The Price equation is mentioned only in Appendix A, on page 824, and it is clear that nothing would change if the somewhat flexible formulation "obeys a form of Price's equation" would be replaced by a simple "is".

> The change in frequency of A owing to selection obeys a form of Price's equation (Price, 1970),
>
> $$E_p[\Delta p]_{sel} = E_p[\overline{W}\Delta p] = E_p[X_{:A}(i)W(i)] - E_p[X_{:A}(i)]\overline{W}$$
> $$= E_p[X_{:A}(i)W(i)] - p$$
>
> where $E_p[\cdots]$ denotes conditional expectation given allele frequency p in the population. The last equality is correct only with the definition of the fitness function given above.

This statement is not at all incorrect; all it does is assume that $X_{:A}(i)$ and $W(i)$ are random variables, and that $E_p[X_{:A}(i)W(i)]$ and $E_p[X_{:A}(i)]$ exist for any p. The expression $E_p[X_{:A}(i)W(i)] - E_p[X_{:A}(i)]\overline{W}$ can be seen as a proper covariance, which is defined with the choice of a fitness function. That implies that this covariance is a given thing, that follows from model assumptions.

Rousset and Billiard (2000) thereby basically *assume* a covariance, or they assume random variables that imply one, in exactly the same way as the rest of the probability theory discipline does, since well before 1970, and totally oblivious of the existence of the Price equation after 1970. This statement perfectly fits the format described in Section 4.3 of Van Veelen (2005, p. 418). Taking out the reference to the Price equation would not make the slightest difference for the derivation. One is of course free to call this "using the Price equation" but it is worth noting that its use here is atypical, because this paper just starts out with a proper model, the limitations of which are also properly described, and does not claim to 'derive' totally general things, starting with the Price equation. This paper therefore needs no repair at all, other than that the claim that it uses the Price equation could also be dropped.

7.2. *Taylor (1989)*

Only a minute step away from Rousset and Billiard (2000) is Taylor (1989). This is an example of a paper that can be repaired with the smallest of efforts. As suggested in Section 5.2 of Van Veelen (2005), all that needs to be done is that a covariance has to be *assumed* rather than claiming that it is derived. This is easily done, and totally inconsequential for the elegant results in this paper.

7.3. *Queller (1985)*

Some claims arrived at with the Price equation are 'almost' right, in the sense that they are right if they are restated in a different form, and, most importantly, if we formulate the right assumptions, or explicitly state implicit ones, under which the claim is actually correct. That is important, because with the Price equation, it is sometimes suggested that generally true claims are derived that rely on no assumptions whatsoever.

Queller (1985) assumes fitness effects of interactions that can be represented by the following matrix.

$$\begin{array}{lll} & \text{Altruist} & \text{Not altruist} \\ \text{Altruist} & B-C+D & -C \\ \text{Not altruist} & B & 0 \end{array}$$

However natural it seems to just postulate a matrix with fitness effects, one can nonetheless wonder what that means. Can we be certain that there is a dynamical system with parwise interactions that would produce those fitness effects? In Appendix C it is shown that if we think of the replicator dynamics, it is not possible to construct one so that it reproduces these fitness effects of interactions. If we have a matrix with payoffs, then one can combine these with the replicator dynamics, and derive the fitness effects of interactions from them. Those fitness effects will then be frequency dependent. If we would go in the other direction, and start with fitness effects, we can try to reverse engineer what the parameters should have been in order for them to lead to these fitness effects in the replicator dynamics. This turns out to be a problem; it is in general not possible to reverse engineer the payoffs so that they combine with the replicator dynamics to produce fitness effects described by B, C and D. Also, for the degenerate choices of B, C and D for which one can, the replicator dynamics is not well-defined, in the sense that it has no solution, unless $B=C=D=0$.

Instead of parameters that represent fitness effects, we could however consider parameters that represent payoffs. If we do that, with lower case letters for payoffs instead of upper case letters for fitness effects, we get the matrix we also used in Section 3.[8]

$$\begin{array}{lll} & \text{Altruist} & \text{Not altruist} \\ \text{Altruist} & b-c+d & -c \\ \text{Not altruist} & b & 0 \end{array}$$

Van Veelen (2011a) shows that the following equivalence for games with two players and two strategies holds. The paper also contains a translation from way the condition is expressed here—$rb-c+(r+(1-r)p)d>0$—to the notation in Queller (1985), and back.

$$\text{the dynamics are payoff monotonic}$$
$$\Updownarrow$$
$$\text{cooperation is selected if } rb-c+(r+(1-r)p)d>0$$

That means that, whatever the dynamics, as long as they are payoff monotonic, the direction of selection is given by Queller's condition. One could hope that, while a normal derivation gives us this conditional result, a derivation with the Price equation gives us a

[8] We would like to thank an anonymous reviewer for pointing out the issue of payoffs versus fitness effects.

more general, unconditional result. This example however shows that if the impression of deriving a more general result were to be created, then it is misleading. Payoff monotonicity is not only a sufficient, but also a necessary condition for Queller's rule to apply with two players and two strategies. Any dynamics that is not payoff monotonic therefore serves as a counterexample against the general, unconditional version of the result. Luckily, payoff monotonicity is a relatively mild assumption, which implies that Queller's rule applies for a relatively large set of dynamics. Still it is important to realize that it nonetheless does not apply in general. Also it is important to realize that just postulating fitness effects from pairwise interactions—rather than payoffs—is not necessarily enough to formulate a model.

7.4. Queller (1992)

Although the result from Queller (1992) is a bit further away from being correct than the result from Queller (1985), it can also be restored under extra assumptions. Van Veelen (2009) shows that inclusive fitness gives the right prediction for a specific set of group selection models if we assume that (1) the dynamics are payoff monotonic and (2) the game satisfies generalized equal gains from switching (or, in other words, if we assume additive fitness effects).[9] Note that the assumption of payoff monotonicity, which was enough to restore Queller (1985), is not enough to restore Queller (1992); the counterexamples in Section 3 all use the replicator dynamics, which are payoff monotonic.

Price (1972) is also an application of his equation to group selection. Section 6 in Van Veelen (2005) constructs a counterexample that shows that his claim is just not correct and therefore beyond repair.

7.5. Page and Nowak (2002)

In the replicator dynamics (Taylor and Jonker, 1978, see also Weibull, 1995, and Hofbauer and Sigmund, 1998), the evolution of the population share of strategy i is described by

$$\dot{x}_i = [f_i(\mathbf{x}) - \overline{f}(\mathbf{x})]x_i \tag{17}$$

where x_i is the share of strategy i, $\mathbf{x} = [x_1, \ldots, x_n]$, $f_i(\mathbf{x})$ the expected payoff of strategy i in population state $\mathbf{x}$, and $\overline{f}(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})x_i$. (Here we follow the notation from Page and Nowak, 2002, where f_i is the expected payoff of strategy i, and not the frequency of groups with i cooperators in it, as it is elsewhere in this paper). This is a deterministic dynamics, in which the assumption of an infinite population implies that every change in population shares is exactly equal to the expected value of that change.

Now suppose we are in state $\mathbf{x}$, and we draw a random individual. Let $X(i)$ be the 1 if the strategy of this randomly drawn individual is strategy i and 0 otherwise and let Y be the fitness of that individual. Now we can compute the covariance of those two random variables. It is not too hard to see that

$$\begin{aligned} Cov(X(i),Y) &= \mathbb{E}[X(i)f_i(\mathbf{x})] - \mathbb{E}[X(i)]\mathbb{E}[f_i(\mathbf{x})] \\ &= [f_i(\mathbf{x}) - \overline{f}(\mathbf{x})]x_i \end{aligned} \tag{18}$$

Thereby we see that for the replicator dynamics, $\dot{x}_i = Cov(X(i),Y)$. One can therefore say that the dynamics can be described with n Price equations (or $n-1$, since the nth follows from $\sum_{i=1}^{n} x_i = 1$).

It is clear that those covariances depend on, and change with, $\mathbf{x}$ as the population evolves, and equal the derivatives. This shows that dynamical sufficiency is therefore not a problem of the Price equation per se.

This is the idea in Page and Nowak (2002), although it is stated a bit differently there. They introduce a vector p, which can be chosen to be, but is not restricted to, the unit vectors $e^1, \ldots, e^n$. The average value of p is defined as $\overline{p} = \sum_{i=1}^{n} p_i x_i$, which equals the population shares x_i if $p = e^i$. Then they claim that the Price equation and the replicator dynamics are equivalent. If we read that as the replicator dynamics being equivalent to a set of Price equations for $p = e^1, \ldots, e^n$, then that is a statement similar to what we stated just before, only that it is not entirely correct to say that they are equivalent. The replicator dynamics is a model—it assumes that the population shares evolve according to Eq. (7)—while the Price equation tracks these dynamics. It is only equivalent with the replicator dynamics if it is fed with the replicator dynamics in the first place; if the actual dynamics would be different from (7), that is, if $\dot{x}_i \neq [f_i(\mathbf{x}) - \overline{f}(\mathbf{x})]x_i$, then the Price equation would also turn into a Price inequality; $\dot{x}_i \neq Cov(X(i),Y)$. Therefore the Price equation is only equivalent to the replicator dynamics if you first put the replicator dynamics in it. The Price equation can rewrite the change in gene frequency for a model, but it is not a model itself, unlike the replicator dynamics (see also Price, 1972, pp. A20–A24; Page, 2003; Traulsen, 2010).

8. Conclusion

The life of Price is, to quote Bill Hamilton, "quite a story". He went from applying chemistry to medicine in the Manhattan project to criticizing extra sensory perception research, and from evolutionary biology to biblical exegesis concerning the Passion chronology. Especially the tragic last years, where research concerning the evolution of altruism blurred with a quest for being a good person himself, ending with his disillusioned suicide, appeal to the imagination. A converted atheist goes from helping the homeless to being homeless himself. The equation, which he himself thought of as a revelation from God, against the background of a life that shows how hard it can be, not only to live a life of goodness, but also to live a life at all.

While Maynard Smith and his ESS concept has proven incredibly useful, the Price equation has produced mixed "results". Some results that are claimed to be derived with the Price equation are indeed correct, but others are not, and tend only to be correct under extra assumptions. Derivations with or expansions of the Price equation are therefore typically a reflection of an intuition, rather than a way to prove that this intuition is correct. That of course does not mean that the intuition is wrong, but projecting it on onto the Price equation is just not the same as deriving a result.

Some results that are derived with the Price equation can also be derived without the Price equation. Other results that are derived with the Price equation cannot be derived without the Price equation. The latter results are wrong. The reason why they must be wrong if they cannot be derived without the Price equation, is that the Price equation has no modelling content and makes no assumptions. Without modelling content, nothing can be derived; in the absence of model assumptions, anything could happen. Any claim or prediction must therefore follow from model assumptions. The only definitive way to check if a result arrived at with the Price equation is actually correct, is therefore to simply start with the model assumptions, and derive the result without the Price equation. With this as an ultimate check, it seems that a more efficient way of separating correct results from incorrect ones is to go from model assumptions to predictions without the Price equation in the first place.

One Price equation result that is not correct is the general equivalence of group selection and inclusive fitness models

[9] Van Veelen (2009) is also a little careless here, not stating explicitly that Theorem 1 assumes payoff monotonic dynamics. See also Van Veelen (2011a,b) where payoff monotonicity is stated explicitly as an assumption.

suggested in Queller (1992). For a specific set of group selection models, inclusive fitness gives the correct prediction, if the game has generalized equal gains from switching, that is, if fitness effects are additive (Van Veelen, 2009, 2011b).[10] If the game does not have equal gains from switching, then inclusive fitness does not give the correct prediction. Whether or not games with additive effects capture all, many, or not so many of the relevant evolutionary situations is an empirical question. We do not know the answer to that question. It is however important to realize that this equivalence is not a generally correct result.

Acknowledgments

The comment by James Marshall (2011a) on how the results in Van Veelen (2009) relate to Queller (1985), and subsequently to Queller (1992), is what started this paper; the long discussion was basically the input and starting point. We would therefore like to thank James Marshall as well as Karl Sigmund, who was the editor for that paper, and moderated that discussion with great wisdom. We are also grateful to the participants of the KNAW colloquium on the evolution of conflict and cooperation in Amsterdam, and to the Royal Netherlands Academy of Arts and Sciences for hosting, sponsoring and organizing it. For help on the statistical part—especially with Section 3.5—we would like to thank Maurice Bun, Peter Boswijk and Jan Kiviet. We would also like to thank seminar participants at the Max-Planck Institute for Evolutionary Biology in Plön and the Mathematics and Statistics department at Queen's University in Kingston, as well as Ben Allen, Troy Day, Feng Fu, Andy Gardner, Chaitanya Gokhale, Alan Grafen, Oren Harman, Manfred Milinski, Martin Nowak, Hisashi Ohtsuki, Dave Queller, Jostein Starrfelt, Corina Tarnita, Peter Taylor, Arne Traulsen and Wouter Vahl for comments on this paper and/or discussions about the Price equation.

Appendix A. Algebra I (phenotype=genotype)

A.1. Group selection decomposition I

In Queller, G_g is the average genotypic value of a trait in a group, W_g is the average fitness in a group. Because the replicator dynamics has the convenient property that adding a constant to all payoffs does not change the dynamics, there is an unambiguous choice for what the so-called covariance (as used in the Price equation literature) between the two should be, given that $G=P$:

$$Cov(G_g,W_g)=\sum_{i=1}^{n} f_i \frac{i}{n}\frac{i\pi_{C,i}+(n-i)\pi_{D,i}}{n}-\sum_{i=1}^{n} f_i \frac{i}{n}\sum_{i=1}^{n} f_i \frac{i\pi_{C,i}+(n-i)\pi_{D,i}}{n} \tag{A1}$$

With $n=2$, the payoff matrix from Queller (1992),[11] and $p=\frac{1}{2}f_1+f_2$, that reduces to

$$Cov(G_g,W_g)=\frac{1}{2}f_1\frac{b-c}{2}+f_2\frac{2(b-c+d)}{2}-p\left[f_1\frac{b-c}{2}+f_2\frac{2(b-c+d)}{2}\right]$$

$$=(b-c)\left(p-\frac{1}{4}f_1\right)+f_2 d-p[(b-c)p+f_2 d]$$

$$=(b-c)\left(p-p^2-\frac{1}{4}f_1\right)+(1-p)f_2 d \tag{A.2}$$

We use following natural correspondence for $n=2$ between relatedness and frequency on the one hand and group compositions on the other, $f_0(p)=(1-r)(1-p)^2+r(1-p)$, $f_1(p)=(1-r)2p(1-p)$ and $f_2(p)=(1-r)p^2+rp$. With it we get

$$Cov(G_g,W_g)=(b-c)p(1-p)-\frac{b-c}{2}(1-r)p(1-p)$$

$$+(1-p)((1-r)p^2+rp)d$$

$$=p(1-p)\left[b-c-\frac{b-c}{2}(1-r)+(r+(1-r)p)d\right]$$

$$=p(1-p)\left[\frac{b-c}{2}(1+r)+(r+(1-r)p)d\right] \tag{A.3}$$

In order to arrive at the second term in the decomposition, we can just subtract the first term from the total:

$$\dot{p}-Cov(G_g,W_g)=p(1-p)[rb-c+(r+(1-r)p)d]-p(1-p)$$

$$\left[b-c-\frac{b-c}{2}(1-r)+(r+(1-r)p)d\right]=p(1-p)\left[-(1-r)b+\frac{b-c}{2}(1-r)\right]$$

$$=-p(1-p)\left[\frac{b+c}{2}(1-r)\right] \tag{A.4}$$

Together, this gives us the following decomposition, which is Eq. (4) in the main text:

$$\dot{p}=p(1-p)\overbrace{\left[\frac{b-c}{2}(1+r)+[r+(1-r)p]d\right]}^{\text{between group selection}}-p(1-p)\overbrace{\left[\frac{b+c}{2}(1-r)\right]}^{\text{within group selection}}$$

This approach, where the within-group selection term is computed as the remainder, will also work for $n>2$, because both $\dot{p}$ and $Cov(G_g,W_g)$ are defined for all n and the within group term must be the remainder in order to be a proper decomposition.

Appendix B. Algebra II (genotype and phenotype may differ)

B.1. Inclusive fitness decomposition

We start with payoff monotonicity. Now, the expression with probability $\mathbf{P}$ implies that not everyone with the gene for cooperation does indeed cooperate, which complicates the computation of the average payoffs. Assuming that expression is independent from group composition, and using Eq. (1) from the main text, we get

$$\frac{\dot{p}}{p(1-p)}=\frac{f_2\cdot\{\mathbf{P}^2(b-c+d)\cdot 2+2\mathbf{P}(1-\mathbf{P})(b-c)+(1-\mathbf{P})^2\cdot 0\cdot 2\}+f_1\cdot\{\mathbf{P}(-c)+(1-\mathbf{P})0\}}{2p}$$

$$-\frac{f_1\{\mathbf{P}(b)+(1-\mathbf{P})0\}+f_0\cdot 0\cdot 2}{2(1-p)}$$

$$=\mathbf{P}\left(\frac{f_2\cdot\{\mathbf{P}(b-c+d)\cdot 2+2(1-\mathbf{P})(b-c)\}+f_1\cdot\{-c\}}{2p}-\frac{f_1\{b\}}{2(1-p)}\right)$$

$$=\mathbf{P}\left(\frac{f_2\cdot\{2(b-c)+2\mathbf{P}d\}+f_1\cdot\{-c\}}{2p}-\frac{bf_1}{2(1-p)}\right) \tag{B.1}$$

Singling out the c, the b and the d, this is rewritten as

$$\dot{p}=p(1-p)\mathbf{P}\left(-\frac{2f_2+f_1}{2p}c+\left(\frac{2f_2}{2p}-\frac{f_1}{2(1-p)}\right)b+\frac{2f_2}{2p}\mathbf{P}d\right) \tag{B.2}$$

Now we use $p=(2f_2+f_1)/2$, $r=2f_2/2p-f_1/2(1-p)$ and $f_2/p=\mathbb{P}(T|T)=r+(1-r)p$ (see Van Veelen, 2009, and Appendix A) to

[10] Gardner et al. (2007) and Marshall (2011a,b) have reacted to examples of games that do not have equal gains from switching by adjusting the r between games in order to make Hamilton's rule work. It should be noted that in the model in Van Veelen (2009, 2011b) the interaction structure, that is, who interacts with whom for any given frequency, is unaffected by the change in the game. Changing the r between games therefore prevents Hamilton's rule from being an actual rule, because, as the definition of a rule suggests, a rule is not a rule if it changes from case to case (Van Veelen, 2011a).

[11] Again, Section 7.3 and Appendix C discuss the distinction between payoffs and fitness effects.

arrive at:

Queller's condition plus

$$\dot{p}=p(1-p)\mathbf{P}(-c+rb+(r+(1-r)p)\mathbf{P}d)$$

This is Eq. (6) from the main text.

B.2. Various covariances and route 2 to the separation condition

Because $Cov(P,W)$, $Cov(G,P)$ and $Var(P)$ are relatively easy to compute, we begin with those.

$$\begin{aligned}Cov(P,W)&=\mathbb{E}[PW]-\mathbb{E}[P]\mathbb{E}[W]\\&=\sum_{i=0}^{n}\left(f_i\frac{\sum_{j=0}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}j\pi_{C,j}}{n}\right)\\&\quad-\sum_{i=0}^{n}\left(f_i\sum_{j=0}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}\right)\\&\quad\sum_{i=0}^{n}\left(f_i\sum_{j=0}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j\pi_{C,j}+(n-j)\pi_{D,j}}{n}\right)\end{aligned}\tag{B.3}$$

For our $n=2$ example, that is

$$\begin{aligned}Cov(P,W)&=f_1\left\{\mathbf{P}\frac{1}{2}(-c)\right\}+f_2\left\{2\mathbf{P}(1-\mathbf{P})\frac{1}{2}(-c)+\mathbf{P}^2\frac{2}{2}(b-c+d)\right\}\\&\quad-\sum_{i=0}^{n}\left(f_i\frac{i\mathbf{P}}{n}\right)\left[f_1\left\{\mathbf{P}\frac{b-c}{2}\right\}\right.\\&\quad\left.+f_2\left\{2\mathbf{P}(1-\mathbf{P})\frac{b-c}{2}+\mathbf{P}^2\frac{2(b-c+d)}{2}\right\}\right]\\&=\mathbf{P}\left[f_1\frac{1}{2}(-c)+f_2\{(1-\mathbf{P})(-c)+\mathbf{P}(b-c+d)\}\right]\\&\quad-\mathbf{P}^2p\left[f_1\left\{\frac{b-c}{2}\right\}+f_2\{(1-\mathbf{P})(b-c)+\mathbf{P}(b-c+d)\}\right]\\&=\mathbf{P}\left[f_1\frac{1}{2}(-c)+f_2\{(-c)+\mathbf{P}(b+d)\}\right]\\&\quad-\mathbf{P}^2p\left[f_1\left\{\frac{b-c}{2}\right\}+f_2\{(b-c)+\mathbf{P}d\}\right]\\&=\mathbf{P}[-pc+f_2\{\mathbf{P}(b+d)\}]-\mathbf{P}^2p[p(b-c)+f_2\mathbf{P}d]\\&=\mathbf{P}p\left[-c+\frac{f_2}{p}\{\mathbf{P}(b+d)\}\right]-\mathbf{P}^2p^2\left[(b-c)+\frac{f_2}{p}\mathbf{P}d\right]\\&=\mathbf{P}p[-c+(r+(1-r)p)\{\mathbf{P}(b+d)\}]\\&\quad-\mathbf{P}^2p^2[(b-c)+(r+(1-r)p)\mathbf{P}d]\end{aligned}\tag{B.4}$$

We also compute $Cov(G,P)$.

$$Cov(G,P)=\mathbb{E}[GP]-\mathbb{E}[G]\mathbb{E}[P]=p\mathbf{P}-p^2\mathbf{P}=p(1-p)\mathbf{P}\tag{B.5}$$

We also compute $Var(P)$. Because for every individual, $P^2=P$

$$\begin{aligned}Var(P)&=\mathbb{E}[P^2]-\mathbb{E}^2[P]=\mathbb{E}[P]-\mathbb{E}^2[P]\\&=\sum_{i=0}^{n}\left(f_i\sum_{j=0}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}\right)\\&\quad-\left[\sum_{i=0}^{n}\left(f_i\sum_{j=0}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}\right)\right]^2\\&=\sum_{i=0}^{n}\left(f_i\frac{i\mathbf{P}}{n}\right)-\left[\sum_{i=0}^{n}\left(f_i\frac{i\mathbf{P}}{n}\right)\right]^2\\&=p\mathbf{P}-(p\mathbf{P})^2\end{aligned}\tag{B.6}$$

These variances and covariances can be used to check under what conditions the equivalence in Queller (1992) holds, as summarized by Equation (16) in the main text.

$$\dot{p}=\frac{Cov(G,P)Cov(P,W)}{Var(P)}$$

The right hand side of this equation is

$$\begin{aligned}&\frac{Cov(G,P)Cov(P,W)}{Var(P)}\\&\quad=\frac{p(1-p)\mathbf{P}[\mathbf{P}p[-c+(r+(1-r)p)\{\mathbf{P}(b+d)\}]-\mathbf{P}^2p^2[(b-c)+(r+(1-r)p)\mathbf{P}d]]}{p\mathbf{P}-(p\mathbf{P})^2}\\&\quad=p(1-p)\mathbf{P}\frac{[[-c+(r+(1-r)p)\{\mathbf{P}(b+d)\}]-\mathbf{P}p[(b-c)+(r+(1-r)p)\mathbf{P}d]]}{1-p\mathbf{P}}\\&\quad=p(1-p)\mathbf{P}\left(-c+r\frac{(1-p)\mathbf{P}}{1-p\mathbf{P}}b+(r+(1-r)p)\mathbf{P}d\right)\end{aligned}\tag{B.7}$$

The left hand side is given by equation (6) from the main text:

$$\dot{p}=p(1-p)\mathbf{P}(-c+rb+(r+(1-r)p)\mathbf{P}d)$$

It is clear that these are only equal if $r=0$, $b=0$, $\mathbf{P}=0$ or $\mathbf{P}=1$, while they are unequal if $r>0$, $b>0$ and $0<\mathbf{P}<1$.

B.3. Group selection decomposition

In Queller (1992), G_g is the average genotypic value of a trait in a group, W_g is the average fitness in a group. The so-called covariance (as used in the Price equation literature) between the two should now incorporate the expressed behaviour.

$$\begin{aligned}Cov(G_g,W_g)&=\mathbb{E}[G_gW_g]-\mathbb{E}[G_g]\mathbb{E}[W_g]\\&=\sum_{i=1}^{n}\left(f_i\frac{i}{n}\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}W_g(j)\right)\\&\quad-\sum_{i=1}^{n}\left(f_i\frac{i}{n}\right)\sum_{i=1}^{n}\left(f_i\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}W_g(j)\right)\\&=\sum_{i=1}^{n}\left(f_i\frac{i}{n}\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}W_g(j)\right)\\&\quad-p\sum_{i=1}^{n}\left(f_i\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}W_g(j)\right)\end{aligned}\tag{B.8}$$

With $n=2$, the payoff matrix from Queller (1992),[12] and $p=\frac{1}{2}f_1+f_2$, that reduces to

$$\begin{aligned}Cov(G_g,W_g)&=\frac{1}{2}f_1\left\{\mathbf{P}\frac{b-c}{2}+(1-\mathbf{P})0\right\}\\&\quad+f_2\left\{\mathbf{P}^2\frac{2(b-c+d)}{2}+2\mathbf{P}(1-\mathbf{P})\frac{b-c}{2}+(1-\mathbf{P})^20\right\}\\&\quad-p\left[f_1\left\{\mathbf{P}\frac{b-c}{2}+(1-\mathbf{P})0\right\}\right.\\&\quad\left.+f_2\left\{\mathbf{P}^2\frac{2(b-c+d)}{2}+2\mathbf{P}(1-\mathbf{P})\frac{b-c}{2}+(1-\mathbf{P})^20\right\}\right]\\&=\mathbf{P}\left[\frac{1}{2}f_1\left\{\frac{b-c}{2}\right\}+f_2\left\{\mathbf{P}\frac{2(b-c+d)}{2}+2(1-\mathbf{P})\frac{b-c}{2}\right\}\right]\\&\quad-\mathbf{P}p\left[f_1\left\{\frac{b-c}{2}\right\}+f_2\left\{\mathbf{P}\frac{2(b-c+d)}{2}+2(1-\mathbf{P})\frac{b-c}{2}\right\}\right]\\&=\mathbf{P}\left[(b-c)\left(p-\frac{1}{4}f_1\right)+f_2\mathbf{P}d\right]-\mathbf{P}p[(b-c)p+f_2\mathbf{P}d]\\&=\mathbf{P}\left[(b-c)\left(p-p^2-\frac{1}{4}f_1\right)+(1-p)f_2\mathbf{P}d\right]\end{aligned}\tag{B.9}$$

With the following natural correspondence for $n=2$ between relatedness and frequency on the one hand and group compositions on the other, $f_0(p)=(1-r)(1-p)^2+r(1-p)$, $f_1(p)=(1-r)2p(1-p)$ and

[12] Again, Section 7.3 and Appendix C discuss the distinction between payoffs and fitness effects.

$f_2(p)=(1-r)p^2+rp$, we get

$$\begin{aligned}Cov(G_g,W_g)&=\mathbf{P}\left[(b-c)p(1-p)-\frac{b-c}{2}(1-r)p(1-p)+(1-p)((1-r)p^2+rp)\mathbf{P}d\right]\\&=\mathbf{P}p(1-p)\left[b-c-\frac{b-c}{2}(1-r)+(r+(1-r)p)\mathbf{P}d\right]\\&=\mathbf{P}p(1-p)\left[\frac{b-c}{2}(1+r)+(r+(1-r)p)\mathbf{P}d\right]\end{aligned}\tag{B.10}$$

In order to arrive at the second term in the decomposition, we can just subtract the first term from the total:

$$\begin{aligned}\dot{p}-Cov(G_g,W_g)&=\mathbf{P}p(1-p)[rb-c+(r+(1-r)p)\mathbf{P}d]\\&\quad-\mathbf{P}p(1-p)\left[b-c-\frac{b-c}{2}(1-r)+(r+(1-r)p)\mathbf{P}d\right]\\&=\mathbf{P}p(1-p)\left[-(1-r)b+\frac{b-c}{2}(1-r)\right]\\&=-\mathbf{P}p(1-p)\left[\frac{b+c}{2}(1-r)\right]\end{aligned}\tag{B.11}$$

Together, this gives us the following decomposition, which is Eq. (8) from the main text:

$$\dot{p}=\mathbf{P}p(1-p)\overbrace{\left[\frac{b-c}{2}(1+r)+[r+(1-r)p]\mathbf{P}d\right]}^{\text{between group selection}}-\mathbf{P}p(1-p)\overbrace{\left[\frac{b+c}{2}(1-r)\right]}^{\text{within group selection}}$$

This approach, where the within-group selection term is computed as the remainder, will also work for $n>2$, because both $\dot{p}$ and $Cov(G_g,W_g)$ are defined for all n and the within group term must be the remainder in order to be a proper decomposition.

B.4. Other so-called covariances we find in Queller (1992)

In Queller (1992) we encounter a few *Cov*-terms. Here we compute them for our example.

$$\begin{aligned}Cov(G_g,P_g)&=\mathbb{E}[G_gP_g]-\mathbb{E}[G_g]\mathbb{E}[P_g]\\&=\sum_{i=1}^{n}\left(f_i\frac{i}{n}\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}\right)\\&\quad-\sum_{i=1}^{n}f_i\frac{i}{n}\sum_{i=1}^{n}\left(f_i\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}\right)\\&=\sum_{i=1}^{n}\left(f_i\frac{i}{n}\frac{i\mathbf{P}}{n}\right)-\sum_{i=1}^{n}f_i\frac{i}{n}\sum_{i=1}^{n}f_i\frac{i\mathbf{P}}{n}\\&=\mathbf{P}(\mathbb{E}[G_g^2]-\mathbb{E}^2[G_g])\\&=\mathbf{P}\cdot Var(G_g)\end{aligned}\tag{B.12}$$

For $n=1$—the individual covariance between G and P—that is

$$Cov(G,P)=\mathbb{E}[GP]-\mathbb{E}[G]\mathbb{E}[P]=p\mathbf{P}-pp\mathbf{P}=p(1-p)\mathbf{P}\tag{B.13}$$

For our $n=2$ example, that is

$$\begin{aligned}Cov(G_g,P_g)&=\mathbf{P}\left(\sum_{i=1}^{n}f_i\frac{i}{n}\frac{i}{n}-\sum_{i=1}^{n}f_i\frac{i}{n}\sum_{i=1}^{n}f_i\frac{i}{n}\right)=\mathbf{P}\left(\sum_{i=1}^{2}f_i\frac{i^2}{4}-p^2\right)\\&=\mathbf{P}(\tfrac{1}{4}f_1+f_2-p^2)=\mathbf{P}(p-\tfrac{1}{4}f_1-p^2)=\mathbf{P}(p(1-p)-\tfrac{1}{4}f_1)\end{aligned}\tag{B.14}$$

With $f_1(p)=(1-r)2p(1-p)$ that is

$$\begin{aligned}Cov(G_g,P_g)&=\mathbf{P}(p(1-p)-\tfrac{1}{2}(1-r)p(1-p))\\&=p(1-p)\mathbf{P}(\tfrac{1}{2}(1+r))\end{aligned}\tag{B.15}$$

We also compute $Cov(P_g,W_g)$.

$$\begin{aligned}Cov(P_g,W_g)&=\mathbb{E}[P_gW_g]-\mathbb{E}[P_g]\mathbb{E}[W_g]\\&=\sum_{i=1}^{n}\left(f_i\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}W_g(j)\right)\\&\quad-\sum_{i=1}^{n}\left(f_i\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}\frac{j}{n}\right)\\&\quad\sum_{i=1}^{n}\left(f_i\sum_{j=1}^{i}\binom{i}{j}\mathbf{P}^j(1-\mathbf{P})^{i-j}W_g(j)\right)\end{aligned}\tag{B.16}$$

For our $n=2$ example, that is

$$\begin{aligned}Cov(P_g,W_g)&=f_1\left\{\mathbf{P}\frac{1}{2}\frac{b-c}{2}\right\}+f_2\left\{2\mathbf{P}(1-\mathbf{P})\frac{1}{2}\frac{b-c}{2}+\mathbf{P}^2\frac{2}{2}\frac{2(b-c+d)}{2}\right\}\\&\quad-\sum_{i=1}^{n}\left(f_i\frac{i\mathbf{P}}{n}\right)\left[f_1\left\{\mathbf{P}\frac{b-c}{2}\right\}\right.\\&\quad\left.+f_2\left\{2\mathbf{P}(1-\mathbf{P})\frac{b-c}{2}+\mathbf{P}^2\frac{2(b-c+d)}{2}\right\}\right]\\&=\mathbf{P}\left[f_1\frac{1}{2}\left\{\frac{b-c}{2}\right\}+f_2\left\{(1-\mathbf{P})\frac{b-c}{2}+\mathbf{P}(b-c+d)\right\}\right]\\&\quad-\mathbf{P}^2p\left[f_1\left\{\frac{b-c}{2}\right\}+f_2\{(1-\mathbf{P})(b-c)+\mathbf{P}(b-c+d)\}\right]\\&=\mathbf{P}\left[f_1\frac{1}{2}\left\{\frac{b-c}{2}\right\}+f_2\left\{\frac{b-c}{2}+\mathbf{P}\left(\frac{b-c}{2}+d\right)\right\}\right]\\&\quad-\mathbf{P}^2p\left[f_1\left\{\frac{b-c}{2}\right\}+f_2\{b-c+\mathbf{P}d\}\right]\\&=\mathbf{P}\left[p\left\{\frac{b-c}{2}\right\}+f_2\left\{\mathbf{P}\left(\frac{b-c}{2}+d\right)\right\}\right]\\&\quad-\mathbf{P}^2p[p(b-c)+f_2\mathbf{P}d]\end{aligned}\tag{B.17}$$

With $f_2(p)=(1-r)p^2+rp$, we get

$$\begin{aligned}Cov(P_g,W_g)&=\mathbf{P}\left[p\left\{\frac{b-c}{2}\right\}+((1-r)p^2+rp)\left\{\mathbf{P}\left(\frac{b-c}{2}+d\right)\right\}\right]\\&\quad-\mathbf{P}^2p[p(b-c)+((1-r)p^2+rp)\mathbf{P}d]\\&=\mathbf{P}p\left[\left\{\frac{b-c}{2}\right\}+((1-r)p+r)\left\{\mathbf{P}\left(\frac{b-c}{2}+d\right)\right\}\right]\\&\quad-\mathbf{P}^2p^2[(b-c)+((1-r)p+r)\mathbf{P}d]\\&=\mathbf{P}p\left[\left\{\frac{b-c}{2}\right\}(1-r\mathbf{P})+((1-r)p+r)\mathbf{P}d\right]\\&\quad-\mathbf{P}p^2\left[(b-c)\mathbf{P}-(1-r)\mathbf{P}\left(\frac{b-c}{2}\right)+((1-r)p+r)\mathbf{P}^2d\right]\\&=\mathbf{P}p\left[\left\{\frac{b-c}{2}\right\}(1+r\mathbf{P})+((1-r)p+r)\mathbf{P}d\right]\\&\quad-\mathbf{P}p^2\left[(\frac{b-c}{2})(1+r)\mathbf{P}+((1-r)p+r)\mathbf{P}^2d\right]\end{aligned}\tag{B.18}$$

Check that for $\mathbf{P}=1$ this is the same as $Cov(G_g,W_g)$ for $\mathbf{P}=1$, as it should.

B.5. Route 1 to the separation condition

Here we compute the following covariance:

$$Cov(\varepsilon_G,W)=\mathbb{E}[\varepsilon_GW]-\mathbb{E}[\varepsilon_G]\mathbb{E}[W]$$

The disturbance term is $\varepsilon_G=G-\alpha_G-\beta_{GP}P$. The fact that

$$\alpha_G+\beta_{GP}=\frac{p-p\mathbf{P}}{1-p\mathbf{P}}+\frac{1-p}{1-p\mathbf{P}}=1$$

implies that $\varepsilon_G=0$ if $G=1$ and $P=1$.

It is relatively simple to compute $\mathbb{E}[\varepsilon_G]$:

$$\begin{aligned}\mathbb{E}[\varepsilon_G]&=\mathbb{P}[G=0,P=0]\cdot(\varepsilon_G\text{ given }G=0,P=0)\\&\quad+\mathbb{P}[G=0,P=1]\cdot(\varepsilon_G\text{ given }G=0,P=1)\\&\quad+\mathbb{P}[G=1,P=0]\cdot(\varepsilon_G\text{ given }G=1,P=0)\\&\quad+\mathbb{P}[G=1,P=1]\cdot(\varepsilon_G\text{ given }G=1,P=1)\\&=(1-p)\cdot-\alpha_G+0\cdot(-\alpha_G-\beta_{GP})+p(1-\mathbf{P})\cdot(1-\alpha_G)+p\mathbf{P}\cdot 0\\&=-(1-p)\frac{p-p\mathbf{P}}{1-p\mathbf{P}}+p(1-\mathbf{P})\cdot\left(1-\frac{p-p\mathbf{P}}{1-p\mathbf{P}}\right)=0\end{aligned}\tag{B.19}$$

This implies that $Cov(\varepsilon_G,W)=\mathbb{E}[\varepsilon_G W]$. It is a bit more elaborate though to compute $\mathbb{E}[\varepsilon_G W]$. We will denote by (G,G) the genotype of self and other, respectively, and by (P,P) the phenotype of self and other.

$$\begin{aligned}\mathbb{E}[\varepsilon_G W]=\ &\mathbb{P}[(G,G)=(0,0)]\cdot(\varepsilon_G W \text{ given } (G,G)=(0,0))\\ &+\mathbb{P}[(G,G)=(0,1),(P,P)=(0,0)]\\ &\cdot(\varepsilon_G W \text{ given}(G,G)=(0,1),(P,P)=(0,0))\\ &+\mathbb{P}[(G,G)=(0,1),(P,P)=(0,1)]\\ &\cdot(\varepsilon_G W \text{ given } (G,G)=(0,1),(P,P)=(0,1))\\ &+\mathbb{P}[(G,G)=(1,0),(P,P)=(0,0)]\\ &\cdot(\varepsilon_G W \text{ given } (G,G)=(1,0),(P,P)=(0,0))\\ &+\mathbb{P}[(G,G)=(1,0),(P,P)=(1,0)]\\ &\cdot(\varepsilon_G W \text{ given } (G,G)=(1,0),(P,P)=(1,0))\\ &+\mathbb{P}[(G,G)=(1,1),(P,P)=(0,0)]\\ &\cdot(\varepsilon_G W \text{ given}(G,G)=(1,1),(P,P)=(0,0))\\ &+\mathbb{P}[(G,G)=(1,1),(P,P)=(0,1)]\\ &\cdot(\varepsilon_G W \text{ given } (G,G)=(1,1),(P,P)=(0,1))\\ &+\mathbb{P}[(G,G)=(1,1),(P,P)=(1,0)]\\ &\cdot(\varepsilon_G W \text{ given } (G,G)=(1,1),(P,P)=(1,0))\\ &+\mathbb{P}[(G,G)=(1,1),(P,P)=(1,1)]\\ &\cdot(\varepsilon_G W \text{ given } (G,G)=(1,1),(P,P)=(1,1))\end{aligned}$$

$$\begin{aligned}&=(1-p)(r+(1-r)(1-p))\cdot-\alpha_G\cdot 0\\ &\quad+(1-p)(1-r)p(1-\mathbf{P})\cdot-\alpha_G\cdot 0\\ &\quad+(1-p)(1-r)p\mathbf{P}\cdot-\alpha_G\cdot b\\ &\quad+p(1-r)(1-p)(1-\mathbf{P})\cdot(1-\alpha_G)\cdot 0\\ &\quad+p(1-r)(1-p)\mathbf{P}\cdot 0\cdot-c\\ &\quad+p(r+(1-r)p)(1-\mathbf{P})^2\cdot(1-\alpha_G)\cdot 0\\ &\quad+p(r+(1-r)p)(1-\mathbf{P})\mathbf{P}\cdot(1-\alpha_G)\cdot b\\ &\quad+p(r+(1-r)p)\mathbf{P}(1-\mathbf{P})\cdot 0\cdot-c\\ &\quad+p(r+(1-r)p)\mathbf{P}^2\cdot 0\cdot(b-c+d)\end{aligned}\tag{B.20}$$

$$\begin{aligned}&=(1-p)(1-r)p\mathbf{P}\cdot-\frac{p-p\mathbf{P}}{1-p\mathbf{P}}\cdot b\\ &\quad+p(r+(1-r)p)(1-\mathbf{P})\mathbf{P}\cdot\frac{1-p}{1-p\mathbf{P}}\cdot b\\ &=-\frac{p^2(1-p)(1-r)\mathbf{P}(1-\mathbf{P})}{1-p\mathbf{P}}\cdot b\\ &\quad+\frac{p(1-p)(r+(1-r)p)\mathbf{P}(1-\mathbf{P})}{1-p\mathbf{P}}\cdot b\\ &=bp(1-p)\mathbf{P}(1-\mathbf{P})\left(\frac{r+(1-r)p-p(1-r)}{1-p\mathbf{P}}\right)\\ &=bp(1-p)\mathbf{P}(1-\mathbf{P})\frac{r}{1-p\mathbf{P}}\\ &=\frac{rbp(1-p)\mathbf{P}(1-\mathbf{P})}{1-p\mathbf{P}}\end{aligned}$$

Appendix C. Payoffs and fitness effects

In the replicator dynamics (Taylor and Jonker, 1978) the derivative of the frequency in a game between strategies 1 and 2 is then given by

$$\begin{aligned}\dot{p}_1&=p_1(\overline{\pi}_1-\overline{\overline{\pi}})\\ &=p_1(1-p_1)(\overline{\pi}_1-\overline{\pi}_2)\end{aligned}$$

where p_1 is the frequency of strategy 1, $p_2=(1-p_1)$ is the frequency of strategy 2, $\overline{\pi}_1$ is the average payoff for 1, $\overline{\pi}_2$ the average payoff for strategy 2 and $\overline{\overline{\pi}}=p_1\overline{\pi}_1+p_2\overline{\pi}_2$ is the overall average payoff. Suppose for simplicity we look at random matching in combination with the following matrix (see Van Veelen, 2011a,b) for replicator dynamics in setting with population structure)

$$\begin{array}{ccc} & 1 & 2\\ 1 & a & a\\ 2 & 0 & 0\end{array}$$

1. The payoff matrix

In this simple example, strategy 1 players get a payoff of a, whatever the type of their opponent is, and strategy 2 players get a payoff of 0, whatever the type of their opponent. This implies that

$$\overline{\pi}_1=a$$

$$\dot{p}_1=p_1(1-p_1)a$$

Another way to represent that is to say that

$$p_{1,t+\Delta t}\approx p_{1,t}(1+\Delta t(1-p_{1,t})a)$$

This allows us to write this with *fitness effects* (which are frequency dependent) resulting from interactions, rather than payoffs, resulting from interactions. Those fitness effects are $(1-p_1)a$ for any strategy 1 player, regardless of the strategy of the player it interacts with

$$\begin{array}{ccc} & 1 & 2\\ 1 & (1-p_1)a & (1-p_1)a\\ 2 & 0 & 0\end{array}$$

2. The fitness effects matrix

If we however are given a matrix with fitness effects already, a natural question would be if we could reverse engineer what the payoffs should be

$$\begin{array}{ccc} & 1 & 2\\ 1 & A & A\\ 2 & 0 & 0\end{array}$$

3. The fitness effects matrix

Reverse engineering now gives us the following payoff matrix:

$$\begin{array}{ccc} & 1 & 2\\ 1 & \dfrac{A}{1-p_1} & \dfrac{A}{1-p_1}\\ 2 & 0 & 0\end{array}$$

4. The payoff matrix

Here the fitness effects are constant, but the payoffs are frequency dependent. That in itself does not have to be a problem, if only it would lead to a well-defined differential equation. Unfortunately that is not the case; the derivative is now $\dot{p}_1=p_1A$, and that implies that (0,1) is no longer invariant under the dynamics (the increase in p_1 does not slow down or stop near $p_1=1$).

More complications arise when we have a more general matrix

$$\begin{array}{ccc} & 1 & 2\\ 1 & a & b\\ 2 & c & d\end{array}$$

5. The payoff matrix

for which

$$p_{1,t+\Delta t}\approx p_{1,t}(1+\Delta t(1-p_{1,t})(b-d+p(a-b-c+d)))$$

Now it is no longer possible in general to write this as

$$p_{1,t+\Delta t}\approx p_{1,t}(1+\Delta t(pA+(1-p)B))$$

Also reverse engineering is not possible, unless we put restrictions on what the fitness effects are allowed to be.

When we are presented with the following matrix, it is therefore not clear if we can think of a dynamics in which these numbers are indeed fitness effects of interactions between individuals. At least the replicator dynamics do not facilitate that. That does not mean it is not possible to make a dynamics for which the B, C and D are fitness effects. But given that the replicator dynamics are the standard way to model pairwise interactions, and the replicator dynamics do not help here, an alternative is required.

$$\begin{array}{ccc} & 1 & 2 \\ 1 & B-C+D & -C \\ 2 & B & 0 \end{array}$$

6. The fitness effects matrix

References

Bergstrom, T., 2003. The algebra of assortative encounters and the evolution of cooperation. Int. Game Theory Rev. 5 (3), 211–228.

Bowles, S., 2009. Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? Science 324, 1293–1298.

Bowles, S., Choi, J.-K., Hopfensitz, A., 2003. The co-evolution of individual behaviors and social institutions. J. Theor. Biol. 223, 135–147.

Chuang, J.S., Rivoire, O., Leibler, S., 2010. Cooperation and Hamilton's rule in a simple synthetic microbial system. Mol. Syst. Biol. 6, 398.

Frank, S.A., 1995. George Price's contributions to evolutionary genetics. J. Theor. Biol. 175, 373–388.

Gardner, A., 2008. The Price equation. Curr. Biol. 18, R198–R202.

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

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

Grafen, A., 2002. A first formal link between the Price equation and an optimization program. J. Theor. Biol. 217, 75–91.

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

Grafen, A., 2007b. An inclusive fitness analysis of altruism on a cyclical network. Journal of Evolutionary Biology 20, 2278–2283.

Harman, O., 2010. The Price of Altruism: George Price and the Search for the Origins of Kindness. WW Norton, New York.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge, UK.

Killingback, T., Bieri, J., Flatt, T., 2006. Evolution in group-structured populations can resolve the tragedy of the commons. Proc. R. Soc. B 273, 1477–1481.

Lehmann, L., Keller, L., West, S.A., Roze, D., 2007. Group selection and kin selection: two concepts but one process. Proc. Natl. Acad. Sci. 104, 6736–6739.

Lion, S., Jansen, V.A.A., Day, T., 2011. Evolution in structured populations: beyond the kin versus group debate. Trends Ecol. Evol. 26, 193–201.

Marshall, J.A.R., 2011a. Queller's rule ok: comment on van Veelen 'when inclusive fitness is right and when it can be wrong'. J. Theor. Biol. 270, 185–188.

Marshall, J.A.R., 2011b. Group selection and kin selection: formally equivalent approaches. Trends Ecol. Evol. 26, 325–332.

Maynard Smith, J., Price, G.R., 1973. The logic of animal conflict. Nature 246, 15–18.

Nowak, M.A., Tarnita, C.E., Wilson, E.O., 2010. The evolution of eusociality. Nature 466, 1057–1062.

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

Okasha, S., 2010. Altruism researchers must cooperate. Nature 467, 653–655.

Page, K.M., 2003. Unifying evolutionary dynamics and a mathematical definition of selection. In: Mathematical Modelling & Computing in Biology and Medicine, vol. 1. Milan Research Centre for Industrial and Applied Mathematics, Bologna, Esculapio, pp. 303–309.

Page, K.M., Nowak, M.A., 2002. Unifying evolutionary dynamics. J. Theor. Biol. 219, 93–98.

Price, G.R., 1970. Selection and covariance. Nature 227, 520–521.

Price, G.R., 1972. Extension of covariance selection mathematics. Ann. Hum. Genet. 35, 485–489.

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

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

Rice, S.H., 2004. Evolutionary Theory: Mathematical and Conceptual Foundations. Sinauer.

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.

Sober, E., Wilson, D.S., 1998. Unto Others; the Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

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

Taylor, P., Jonker, L., 1978. Evolutionary stable strategies and game dynamics. Math. Biosci. 40, 145–156.

Traulsen, A., 2010. Mathematics of kin- and group-selection: formally equivalent? Evolution 64, 316–323.

Traulsen, A., Nowak, M., 2006. Evolution of cooperation by multilevel selection. Proc. Natl. Acad. Sci. 103, 10952–10955.

Van Veelen, M., 2005. On the use of the Price equation. J. Theor. Biol. 237, 412–426.

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.

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

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

Van Veelen, M., García, J., Sabelis, M.W., Egas, M., 2010. Call for a return to rigour in models. Nature 467, 661 (correspondence).

Van Veelen, M., Hopfensitz, A., 2007. In Love and War; altruism, norm formation, and two different types of group selection. J. Theor. Biol. 249, 667–680.

Wade, M.J., Wilson, D.S., Goodnight, C., Taylor, D., Bar-Yam, Y., de Aguiar, M.A.M., Stacey, B., Werfel, J., Hoelzer, G.A., Brodie III, E.D., Fields, P., Breden, F., Linksvayer, T.A., Fletcher, J.A., Richerson, P.J., Bever, J.D., Van Dyken, J.D., Zee, P., 2010. Multilevel and kin selection in a connected world. Nature 463, E8–E9.

Wild, G., Gardner, A., West, S.A., 2009. Adaptation and the evolution of parasite virulence in a connected world. Nature 459, 983–986.

Weibull, J.W., 1995. Evolutionary Game Theory. MIT Press, Cambridge MA.

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