Social Learning

An Introduction to Mechanisms, Methods, and Models

William Hoppitt and Kevin N. Laland

PRINCETON UNIVERSITY PRESS

Princeton and Oxford

Chapter 5

Statistical Methods for Diffusion Data

This is the first of three chapters focusing on statistical techniques for inferring and quantifying social transmission in groups of animals in the wild, or in "captive" groups of animals in naturalistic social environments. Here we focus on techniques for analyzing time-structured data on the occurrence of a particular behavior pattern, or *behavioral trait*, in one or more groups. For the most part we focus on cases where a novel trait spreads through one or more groups. Following standard terminology in the field of social learning, we refer to the spread of a trait through a group as a *diffusion*, and the resulting data as *diffusion data* (regardless of whether there is evidence for social transmission). Such data may arise if the spread of a naturally occurring trait is recorded, or for diffusions that are initiated by a researcher, by presenting some kind of task (for nonhumans, usually a foraging task) that members of a group must learn to solve (see sections 3.2.2 and 7.2.1). The study of diffusion data is likely to be of crucial importance if researchers are to understand how and when novel innovations spread and give rise to traditions.

The level of detail of diffusion data varies. At one extreme, a researcher might possess a complete history of each individual's performance of the trait, along with a history of its observations of others' trait performances. Such data potentially allow rich inferences to be made about the social learning strategies (box 8.2) and mechanisms (J. R. Kendal et al. 2007; Hoppitt et al. 2012; see box 4.2) being utilized. However, this is perhaps only likely for captive groups, or for the diffusion of the solution to a task in which every manipulation of the task can be monitored closely. More commonly, a researcher might only have an estimate of when each individual first performed the trait, with an associated indirect assay of who is likely to have observed whom. In such cases, one can view individuals as moving from a *naïve* state in which they do not perform the trait, to an *informed* state, through *acquisition* of the trait by means of either *asocial learning* or *social transmission* of the trait from an informed individual (or both). A researcher can then use a model in which the *rate of trait acquisition* is modeled (section 5.2). In other cases, a researcher may not be able to identify individuals in the population at all, and only have an estimate of how many individuals have performed the trait so far (section 5.1). Finally, we look at how social transmission can be inferred from the spread of a behavioral trait through space.

5.1 Diffusion Curve Analysis

A diffusion curve is a plot of the number of individuals observed to have performed a behavioral trait against time.[1] For many years, in both the human and nonhuman social learning literature, it was believed that the shape of the diffusion curve could be used to infer whether social transmission was involved in the spread of a behavioral trait. The idea is that asocial learning proceeds at an approximately constant rate, resulting in an r-shaped diffusion curve. In contrast, if the trait is acquired by social transmission, the per capita rate of acquisition will increase as the number of demonstrators increases, giving an acceleratory curve (see fig. 5.1). If a complete diffusion is documented throughout the entire group or population in question, the diffusion curve is expected to be S-shaped, as it levels out with all individuals having acquired the trait. The reason that social transmission is widely thought to generate an s-shaped curve is that it requires both demonstrators and observers, and when either are rare, as occurs early or late in the diffusion, the rate of spread is constrained; however, when both are common, as in the middle of the diffusion, the spread is at its most rapid. Different functional forms are fitted to the data, and their fit compared (see box 5.3), and social learning is inferred if acceleratory curves fit best.

Diffusion curve analysis has been used extensively to infer social transmission in both humans and nonhuman animals (J. Henrich 2001; Lefebvre 1995; Reader 2004; E. Rogers 1995; Roper 1986). Lefevbre (1995) used this method to analyze 21 diffusions of foraging innovations from the primate literature, including cases from Japanese macaques (e.g., fish eating; Watanabe 1989), vervet monkeys (acacia-pod dipping; Hauser 1988), and chimpanzees (mango and lemon eating; Takahata et al. 1986; Takasaki 1983). Lefevbre found an overall trend for accelerating learning rates, seemingly consistent with models of social transmission.

Unfortunately, recent work suggests that that the shape of the diffusion curve is not a reliable signature of social transmission (Laland and J. R. Kendal 2003; Reader 2004; Franz and Nunn 2009; Hoppitt, Kandler, et al. 2010). First, a number of researchers have pointed out that social learning will not necessarily result

[1] Sometimes a proxy for the number of informed individuals is used, such as the number of times a trait is seen being performed during a particular time period.

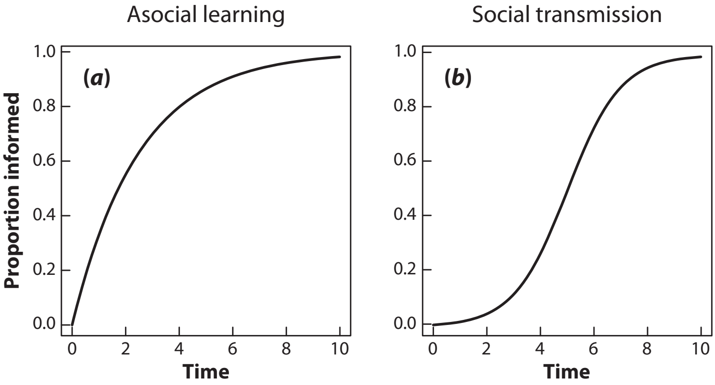

Figure 5.1. Typical diffusion curves traditionally assumed to be characteristic of (*a*) asocial learning (r-shaped) and (*b*) social transmission (s-shaped). Recent theoretical work has cast doubts on these assumptions (see text).

in an S-shaped diffusion curve if the population is structured into subgroups. Laland and J. R. Kendal (2003) and Reader (2004) suggest that directed social learning can result in a step-shaped function, with acceleratory component parts, if the trait spreads more rapidly through closely connected subgroups, such as family units (e.g., Fritz et al. 2000). Furthermore, differences in the rate of acquisition between different subsections of the population might act to obscure any underlying pattern; for example, a strong sex difference might result in a bimodal distribution of latencies to acquire the trait (Reader 2004).

Perhaps even more of a concern, there are good reasons to expect that asocial learning alone can result in an S-shaped diffusion curve, even if populations are homogeneously structured. In general, any process that results in an increase with time in the rate at which individuals acquire the trait will result in an acceleratory diffusion curve. For example, if the trait of interest is the solution to a novel foraging task that is presented to a group of animals, they will often display some neophobia to the task. If the effects of neophobia decrease over time, as we might expect, the rate at which individuals solve the task might also increase with time (Hoppitt, Kandler, et al. 2010). Acceleratory curves can also occur if an individual must move through a number of stages in order to acquire a trait. For example, there may be a number of different steps required to solve a foraging task, such as defenses that need to be removed to access a fruit (Whiten 1998). If the time to complete each step of the task is exponentially distributed, then we would expect the overall time to solve the task to follow an approximate gamma distribution, causing the diffusion curve to become more and more S-shaped as the number of task steps increases (Hoppitt, Kandler, et al. 2010). In conclusion, recent theoretical analyses suggest that researchers cannot reliably infer social learning from the shape of the diffusion curve (Reader 2004 gives further reasons for caution).

5.2 Network-Based Diffusion Analysis (*NBDA*)

Network-based diffusion analysis (*NBDA*), pioneered by Franz and Nunn (2009), infers social transmission if the pattern of spread of a behavioral trait follows the connections of a social network. A social network is a social structure made up of individuals, sometimes called nodes, as well as the connections among them that represent forms of relationship or interdependency, such as patterns of association, interaction, friendship, or kinship (Newman 2010). The assumption here is that a trait will spread sooner between individuals who spend more time together than between less connected individuals. As such, *NBDA* inherently addresses the concern that the pattern of spread of a trait will be influenced by population structure. With social network analysis becoming increasingly popular in both the social (Wasserman and Faust 1994; Newman 2010) and biological (Croft et al. 2008) sciences, the appropriate data for applying *NBDA* to both human and nonhuman populations is often likely to be available. *NBDA*, therefore, offers a viable alternative for inferring social transmission from diffusion data. Though *NBDA* was developed recently in the field of animal social learning, similar methods had previously been developed in the social sciences. Here we start by describing *NBDA* in detail, before discussing how previous approaches relate to it.

Franz and Nunn's original version (2009) of *NBDA* took as data the *times* at which individuals acquire a behavioral trait, which can be considered the time at which an individual was first observed performing the trait in question. Hoppitt, Boogert, and Laland (2010) introduced an alternative version of *NBDA*, which applies to the *order* in which individuals acquire the trait, but not the exact time. These alternative versions of *NBDA* have become known as time of acquisition diffusion analysis (*TADA*) and order of acquisition diffusion analysis (*OADA*), respectively. In both cases, the researcher fits a model including a social transmission component, in which the rate of transmission between an informed and a naïve individual is proportional to the connection between them. If this model is better than a model without social transmission (see section 5.2.2), then this supports the hypothesis that the trait is transmitted through the social network. In box 5.1, we provide the mathematical and technical details underlying *NBDA*. Code to run *NBDA* in the *R* statistical environment (*R* Core Development Team 2011) is available at http://lalandlab.st-andrews.ac.uk/freeware.html. Here we aim to provide a general guide to using *NBDA* for nonmathematical readers.

Each version of *NBDA* has its advantages and disadvantages. While *TADA* has more statistical power than *OADA* (Hoppitt, Boogert, and Laland 2010), especially when networks are relatively homogeneous, this comes at the cost of stronger assumptions. In its original form (Franz and Nunn 2009), *TADA* assumes that the *baseline rate of acquisition* (the rate of acquisition in the absence of social transmission) remains constant over time. This can result in false positives in the same circumstances as diffusion curve analysis (i.e., if the asocial acquisition rate increases over time, for example, as a result of a reduction in neophobia to a novel task). *OADA* is not vulnerable to such effects, making the weaker assumption that the baseline rate function is the same for all individuals being analyzed. However,

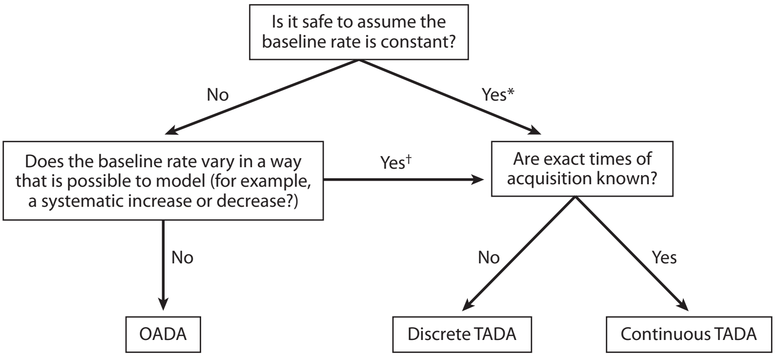
Is it safe to assume the baseline rate is constant?
No
Yes*
Does the baseline rate vary in a way that is possible to model (for example, a systematic increase or decrease?)
Yes†
Are exact times of acquisition known?
No
No
Yes
OADA
Discrete TADA
Continuous TADA

Figure 5.2. Flowchart for selecting the appropriate *NBDA* model. *Researchers should be cautious in assuming the baseline rate of acquisition is constant, because a number of factors can cause increases in the rate (see Hoppitt, Kandler, et al. 2010). †In principle, any function can be used to model the baseline rate. However, the software provided on our website only allows for a systematic increase or decrease. In cases where environmental variables are thought to unpredictably influence the rate of acquisition, but for all individuals in the same way, *TADA* becomes intractable, whereas *OADA* remains appropriate (see Hoppitt, Boogert, and Laland 2010). Based on figure 2 in Hoppitt and Laland (2011).

Hoppitt, Kandler, et al. (2010) extended *TADA* to accommodate a nonconstant baseline rate of acquisition. They suggested use of a baseline rate function corresponding to a gamma distribution of latencies under asocial conditions, which allows for a systematic increase or decrease in the asocial rate of acquisition over time. Nonetheless, *OADA* remains an attractive option if the baseline rate function is thought to follow a form that is difficult to model, or if the researcher does not wish to make any assumptions regarding the shape of the function.

If the researcher chooses to use *TADA*, they also have the choice of treating time as a continuous variable (Hoppitt, Boogert, and Laland 2010), or splitting the diffusion period into a number of discrete units and specifying which individuals acquired the trait in each unit (Franz and Nunn 2009). In practice, the method will fit equivalent models if the number of time units used is large. Typically, computation speeds are faster for the continuous *TADA*, and we recommend that this method be used when exact times of acquisition are known. However, when this is not the case, the discrete *TADA* may be preferable. For example, data might be collected in a series of scans, where, at discrete points in time, the researcher ascertains which individuals are informed. In this case the researcher can only infer a time period during which each individual acquired the behavior, and so the discrete *TADA* is the natural choice.

Even if data is collected individually, in the field it seems likely that there will be observation errors in the recorded time of acquisition. Franz and Nunn (2010) find that this can result in type 1 errors in a discrete *TADA* when the time units are small (and so too, presumably, for the continuous *TADA*), but that if the length of time unit is long enough the problem is alleviated. Franz and Nunn provide a rule of thumb that there should be at least a 50% probability that an individual who has acquired the trait will be observed performing it within any given time unit. A researcher

Box 5.1

Network-based diffusion analysis (*NBDA*)

All forms of *NBDA* can be generalized in the following form:

$$\lambda_i(t) = \lambda_0(t)(1 - z_i(t))\left(s\sum_{j=1}^{N} a_{i,j} z_j(t) + A\right), \tag{5.1.1}$$

where $\lambda_i(t)$ is the rate at which individual i acquires the trait at time t; $\lambda_0(t)$ is a baseline acquisition function determining the distribution of latencies to acquisition in the absence of social transmission; $z_i(t)$ gives the status (1 = informed, 0 = naïve) of individual i at time t; $a_{i,j}$ gives the network connection, or association, between i and j; s is a fitted parameter determining the relative strength of social transmission; and the predetermined form of A determines whether asocial learning of the trait is assumed to occur.

The $(1 - z_i(t))$ and $z_j(t)$ terms ensures that the trait is only transmitted between informed and uninformed individuals. The model assumes that the rate of transmission between such individuals is proportional to the connection between them, and this rate is scaled by the parameter s. Social transmission is inferred if a model including s is better (see box 5.3) than a model where $s = 0$. One strategy for doing this is to compare a model of "pure" social transmission, where all acquisition is through social transmission ($A = 0$) with a model of asocial learning ($s = 0, A = 1$) (Franz and Nunn 2009). However, this only works if the analysis starts with informed individuals in the group (e.g., trained demonstrators). An alternative strategy is to assume there is always the chance that an individual can acquire the trait asocially. In this case, the most intuitive way to parameterize the model is to set $A = 1$[1]. This means that s gives the rate of social transmission relative to the rate of asocial acquisition. We prefer this approach, and henceforth replace A with 1.

The difference between *TADA* and *OADA* is the way in which the baseline rate function, $\lambda_0(t)$, is treated. In *OADA*, $\lambda_0(t)$ is unspecified, with the assumption that it is the same for all individuals being modeled, whereas in TADA, $\lambda_0(t)$ takes a specified form that is fitted to the data.[2] In both cases, the expression given in equation 5.1.1 leads to a likelihood function (L), which gives the probability of observing the data under the model, given a specific set of parameters, and allowing the model to be fitted by maximum likelihood.[3] The log-likelihood for *OADA* is:

$$\log(L) = \sum_{l=1}^{D}\sum_{i=1}^{N} \log(R_i(t_{l-1}))z_i(t_l)(1 - z_i(t_{l-1})) - \sum_{l=0}^{D} \log\left(\sum_{i=1}^{N} R_i(t_{l-1})\right), \tag{5.1.2}$$

where D is the number of acquisition events observed (where one or more individuals are observed to acquire the trait), $R_i(t)$ is the relative rate of acquisition ($\lambda_i(t)/\lambda_0(t)$), and t_l is the time immediately after the l$^{\text{th}}$ acquisition event (after $z_i(t)$'s are updated). The relative rate of acquisition is used here because the baseline rate function cancels out of the likelihood function. $z_i(t_l)(1 - z_i(t_{l-1}))$ indicates whether i acquired the trait at the l$^{\text{th}}$ acquisition event.

[1] We have previously (Hoppitt, Boogert, and Laland 2010) suggested a bounded paramterization for s, where one sets $A = 1 - s$. This means that s can range between 0 (all asocial acquisition) and 1 (all social tranmission). However, we now suggest that this paramterization is more difficult to interpret when individual-level variables are included (box 5.2).

[2] *NBDA* can be seen as a specialized version of survival analysis (or "event-history analysis"), with *OADA* being equivalent to a modified Cox proportional hazards model, and *TADA* being equivalent to a parametric model (Cox and Oakes 1984). In survival analysis, $\lambda_i(t)$ is referred to as the "hazard function" and $\lambda_0(t)$ as the "baseline hazard function." However, in the context of NBDA, we feel "rate of acquisition function" and "baseline rate function" are more intuitive terms.

[3] An optimization algorithm is run to find the set of parameter values that maximizes the likelihood, or equivalently, minimizes the negative log-likelihood.

In the original forms of *TADA* (Franz and Nunn 2009; Hoppitt, Boogert, and Laland 2010), it is assumed that the baseline rate function is constant ($\lambda_0(t) = \lambda_0$). However, Hoppitt, Kandler, et al. (2010) found that *TADA* is susceptible to false positives in the same circumstances as diffusion curve analysis, and suggested fitting a baseline rate function that allowed for systematic changes over time. Hoppitt, Kandler, et al. (2010) suggest a function corresponding to a gamma distribution of times, under asocial conditions, though here we note a Weibull distribution is commonly used in survival analysis, and might also work well for *NBDA*. In principle, any function can be used so long as the user can provide the "hazard function" $\lambda_0(t)$, and "cumulative hazard function" $\Lambda_0(t)$. For many distributions, these functions are readily available in the R statistical environment (R Core Development Team, 2011).

The user may also choose whether to treat time as a continuous variable (continuous *TADA*) (Hoppitt, Boogert, and Laland 2010), or to divide time into a number of discrete steps of equal (Franz and Nunn 2009) or unequal (Hoppitt, Kandler, et al. 2010) length, specifying which step each individual acquired the behavior (discrete *TADA*).

The negative log-likelihood for the continuous *TADA* with a constant baseline rate function is:

$$\begin{aligned}\log(L) = &\sum_{l=1}^{D}\sum_{i=1}^{N}(t_{l-1} - t_l)\lambda_0 R_i(t_{l-1})\\ &+\sum_{l=1}^{D}\sum_{i=1}^{N} z_i(t_l)(1 - z_i(t_l))(\log(R_i(t_{l-1}) + z_i(t_{l-1})) + \log(\lambda_0)).\\ &+\sum_{i=1}^{N} R_i(t_D)\lambda_0(t_D - t_{end})\end{aligned} \qquad 5.1.3$$

This can be generalized for a nonconstant baseline rate to give:[4]

$$\begin{aligned}\log(L) = &\sum_{l=1}^{D}\sum_{i=1}^{N} R_i(t_{l-1})(\Lambda_0(t_{l-1}) - \Lambda_0(t_l))\\ &+\sum_{l=1}^{D}\sum_{i=1}^{N} z_i(t_l)(1 - z_i(t_{l-1}))(\log(R_i(t_{l-1}) + z_i(t_{l-1})) + \log(\lambda_0(t_l))),\\ &+\sum_{i=1}^{N} R_i(t_D)(\Lambda_0(t_D) - \Lambda_0(t_{end}))\end{aligned} \qquad 5.1.4$$

where t_{end} is the time at which observation ceased (allowing for individuals who did not acquire the trait during the period of observation), and $\Lambda_0(t)$ is the baseline cumulative hazard, in survival analysis terminology, which is related to the cumulative distribution function of the asocial latency distribution, $F_0(t)$, thus:

$$\Lambda_0(t) = -\log(1 - F_0(t)). \qquad 5.1.5$$

Note that if $\lambda_0(t) = \lambda_0$, $\Lambda_0(t) = \lambda_0 t$, equation 5.1.4 reduces to 5.1.3.

For a discrete *TADA*, the data is provided in *P* discrete time steps, for which the time step in which each individual acquired the behavior is known. The log-likelihood is as follows:[5]

[4] The middle term here is $\sum_{l=1}^{D}\sum_{i=1}^{N} z_i(t_l)(\log(R_i(t_{l-1})) + \log(\lambda_0(t_l)))$ for any individual who is naïve at time t_{l-1}. Here we modify it such that this component is zero for any individual who is informed at time t_{l-1}, avoiding numerical errors arising from $\log(0) = -\infty$, since $R_i(t_{l-1}) = 0$ for such individuals.

[5] The $z_i(t_{start,p})$ term here ensures that the likelihood is zero for individuals who are informed at the start of period *p*, since $R_i(t_{start,p}) = 0$ for such individuals.

(continued)

Box 5.1 *(continued)*

$$\log(L) = \sum_{p=1}^{P}\sum_{i=1}^{N}(1 - z_i(t_{end,p}))R_i(t_{start,p})(\Lambda_0(t_{start,p}) - \Lambda_0(t_{end,p}))$$

$$+\sum_{p=1}^{P}\sum_{i=1}^{N} z_i(t_{end,p})\log(1 - \exp(R_i(t_{start,p})(\Lambda_0(t_{start,p}) - \Lambda_0(t_{end,p}))) + z_i(t_{start,p})), \qquad 5.1.6$$

where $t_{start,p}$ is the start of time period p, and $t_{end,p}$ is the end of time period p. For $\lambda_0(t) = \lambda_0$ this reduces to:

$$\log(L) = \sum_{p=1}^{P}\sum_{i=1}^{N}(1 - z_i(t_{end,p}))R_i(t_{start,p})(t_{start,p} - t_{end,p})\lambda_0$$

$$+\sum_{p=1}^{P}\sum_{i=1}^{N} z_i(t_{end,p})\log(1 - \exp(R_i(t_{start,p})(t_{start,p} - t_{end,p})\lambda_0) + z_i(t_{start,p})). \qquad 5.1.7$$

With time steps of equal length, this is equivalent to the model initially proposed by Franz and Nunn (2009), with their parameter τ, the rate of learning per time step, given by $\tau = \lambda_0 sT$, where T is the length of time step. Note that the discrete *TADA* assumes that individuals who acquire the trait in the same time step do not learn from each other, so this may provide a conservative estimate of the rate of social transmission.

can check this by calculating the proportion of time units in which individuals are observed performing the trait following the time unit when their performance was initially observed. In figure 5.2 we provide a flowchart to aid the choice of *NBDA* method, between *OADA*, continuous *TADA*, and discrete *TADA*. Box 5.1 provides technical details on the different types of *NBDA*, and how each is fitted to the data.

5.2.1 Inclusion of individual-level variables

A potential problem with *NBDA* is that false positives for social transmission can arise if individuals prefer to associate with others who have a similar asocial rate of acquisition (Hoppitt, Boogert, and Laland 2010). For example, higher-ranking individuals might acquire a trait at a higher rate asocially, and also disproportionately associate with each other, making it appear that the trait is being transmitted among them. As in other statistical models, a researcher can control for the effect of such confounding variables by including them in the model (Hoppitt, Boogert, and Laland 2010; Shipley 1999). There are good reasons for doing this, because even when a variable is not confounded with the social network, statistical power to detect social transmission can be improved by accounting for the variables' effects (Hoppitt, Boogert, and Laland 2010). In addition, it will often be of interest which variables influence the diffusion dynamics (e.g., Boogert et al. 2008).

Hoppitt, Boogert, and Laland (2010) extended *NBDA* to include such "*individual-level variables*" affecting the rate of asocial learning. They recognize two ways in which individual-level variables might be incorporated into the

Box 5.2

Inclusion of individual level variables in *NBDA*

In general, we can expand *NBDA* generally to include *V* continuous individual level variables as follows:

$$\begin{aligned} R_i(t) &= (1 - z_i(t))\left(s\exp(\Gamma_i)\sum_{j=1}^{N} a_{i,j}z_j(t) + \exp(\mathrm{B}_i)\right) \\ \mathrm{B}_i &= \sum_{k=1}^{V}\beta_k x_{k,i} \qquad , \qquad 5.2.1 \\ \Gamma_i &= \sum_{k=1}^{V}\gamma_k x_{k,i} \end{aligned}$$

where $\lambda_i(t) = \lambda_0(t)R_i(t)$, $x_{k,i}$ is the value of the k^{th} variable for individual *i*; β_k is the coefficient giving the effect of variable *k* on asocial learning; and γ_k is the coefficient giving the effect of variable *k* on the rate of social transmission. This general formulation allows the effects of individual level variables on asocial learning and social transmission to differ. In principle, these variables can be fitted in an unconstrained way; alternatively one can fit the additive model defined by Hoppitt, Boogert, and Laland (2010) by constraining $\gamma_k = 0$ for all *k*, or the multiplicative model by constraining $\gamma_k = \beta_k$ for all *k*. Categorical variables, or *factors*, with *F* levels can be fitted by defining $F - 1$ indicator variables determining which category each individual lies in, in the same way as for a standard regression analysis (e.g., see Weissberg 2005). The log-likelihood functions given in box 5.1 remain appropriate.

Interpretation of effects requires some explanation. For individual-level variables, β_k gives the additive effect of an increase of one unit of variable *k* on the log scale, so exp(β_k) gives the multiplicative effect on the rate of acquisition (this is the same as for most standard survival analyses). For example, if we find a coefficient of 1 per cm of body length, this means that, all other things being equal, the model would predict that if an individual A is 1 cm longer than another individual B, and then A would asocially acquire the trait at a rate 2.7 times faster than B.

In contrast, social transmission is modeled as a linear effect, such that it gives the rate of social transmission per unit of connection to informed individuals. In the additive model, this is relative to the baseline level of asocial acquisition (i.e., when $\mathrm{B}_i = 0$. We suggest researchers standardize any continuous variables (subtract the mean, then divide by the standard deviation), meaning *s* can be interpreted as the rate of social transmission relative to the average rate of asocial acquisition. If factors are included in the analysis, *s* is relative to the asocial rate for an individual at the baseline level, for each factor. In the multiplicative model, *s* is invariant to the scale of the individual-level variables (see fig. 5.3).

model. The additive model assumes that the absolute difference in the rate of acquisition between any two individuals remains constant, for any level of social transmission. The multiplicative model instead assumes that the ratio in the rate of acquisition between any two individuals remains constant for any level of social transmission (see fig. 5.3 and box 5.2).[2]

[2] Previously we have suggested that a best fit of the additive model might indicate direct social learning mechanisms, and a best fit of the multiplicative model might indicate indirect mechanisms (see chapter 4; Hoppitt, Boogert, and Laland 2010). However, simulations using algorithms representing either direct or indirect mechanisms (similar to those used in Hoppitt and Laland 2011) did not support this distinction. Researchers should let their data decide whether the additive or multiplicative assumption is most appropriate, or fit a more general model (see box 5.2).

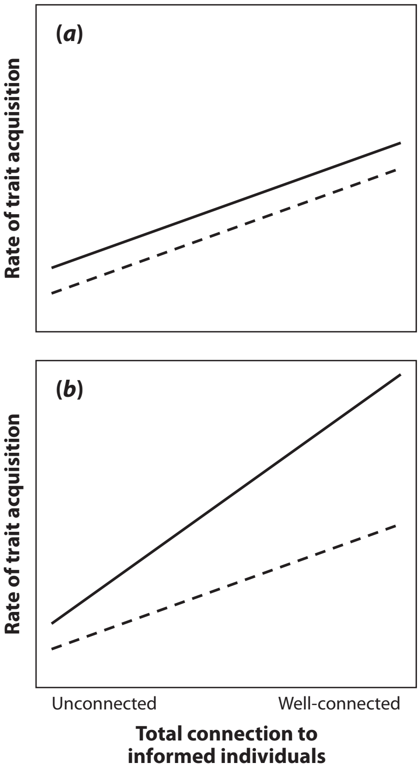

Figure 5.3. A graphical depiction of (*A*) the additive *NBDA* and (*B*) multiplicative NBDA, showing the rate of trait acquisition for two individuals (that differ in their asocial rate of acquisition) as a function of the total connection to informed individuals. At the extreme left of the range, individuals spend no time with any informed individuals, whereas at the extreme right, individuals are extremely well connected to those who have acquired the trait. When unconnected to informed individuals, acquisition is by asocial learning; for all other cases, the rate of acquisition is a combination of social transmission and asocial learning. For both (*A*) and (*B*), the asocial rate of acquisition for individual A (*solid line*) is double that for individual B (*dashed line*). In the additive model, the absolute difference in the rate of acquisition remains constant as the total connection increases, whereas in the multiplicative model, the ratio between the two remains constant. Based on figure 1 in Hoppitt and Laland (2011).

5.2.2 Model selection and inference

To test for social transmission, a researcher must compare a model containing social transmission (henceforth a *social model*) with a model not containing social transmission (an *asocial model*[3]). However, in each case the researcher must decide (*i*) whether to include a constant or nonconstant baseline function (if they wish to employ *TADA*), (*ii*) which individual-level variables to include in each model, and (iii) whether to consider an additive or multiplicative model of social transmission. Model selection for *NBDA* is directly analogous to model selection for a general linear model, and so the same methods can be used. We favor an information theoretic approach (Burnham and Anderson 2002), since in many cases the best asocial model might not be nested in the best model that includes social transmission, meaning that a classical hypothesis test such as a likelihood ratio test (*LRT*) cannot be used. For example, the best asocial model might include a nonconstant baseline (accounting for an acceleratory spread of the trait), whereas the best social model might have a constant baseline function,

[3] We note in passing that it is possible for diffusions that are well described by asocial models to reveal evidence for social learning, but not social transmission (Atton et al., 2012). This *prima facie* surprising observation reflects the breadth of the definition of social learning, which allows for forms of social influence on learning that do not qualify as social transmission.

with social transmission providing an alternative explanation for the acceleratory effect. In such cases a hypothesis test cannot be used, and thus information theoretic approaches come into their own. In addition, information theoretic approaches enable a researcher to use a model averaging approach, which allows taking into account model selection uncertainty (see box 5.3).

We suggest fitting models containing every combination of individual-level variables the researcher wishes to consider, with both a constant and nonconstant baseline rate, for an asocial model, additive social model, and multiplicative social model. In each case, the relative support for each individual model can be judged using Akaike's Information Criterion (AIC), or in practice, AIC_c, which is corrected for sample size (see box 5.3). These criteria assess each model based on how well they fit the data, after penalizing for the number of parameters used, with smaller values indicating that a model has greater predictive power. The models can then be ranked according to AIC_c, and the support for each model, or Akaike weight, is calculated from the difference in AIC_c from the best model (this procedure is implemented automatically in the *R* code provided at http://lalandlab.st-andrews.ac.uk/freeware.html).

The evidence for or against social transmission can be assessed by the total Akaike weights for models including social transmission and asocial models. To make this a fair comparison we suggest that a three-way[4] comparison be made between the asocial models, and the additive and multiplicative models of social transmission. The model with the greatest total Akaike weights is the one best supported by the data. The difference in weight between this and the other models indicates the level of support.[5]

Model averaging methods (box 5.3) enable researchers to estimate the strength of social transmission, and to calculate confidence intervals in a way that allows for model selection uncertainty (Burnham and Anderson 2002). Confidence intervals are especially important in cases where there is little support either way for or against social transmission, because they allow a researcher to set an upper plausible limit for the strength of social transmission. This might enable the researcher to make the stronger conclusion that social transmission is unlikely to be important in the acquisition of a trait.

5.2.3 Modeling multiple diffusions

Sometimes a researcher might have access to data from multiple diffusions, either a single trait spreading through multiple groups, or the spread of multiple traits through one or more groups. It might be preferable to include these in a single

[4] A two-way comparison between asocial models and models containing social transmission is not a fair comparison if twice as many of the latter are considered. For the same reason, more comparisons should be made, if additional models of social transmission are considered (see section 5.2.4).

[5] If the asocial model has an only slightly lower Akaike weight, this means that there is not strong evidence for social transmission. However, it would not make sense for researchers to "reject" social transmission on grounds of parsimony under such circumstances, since the Akaike weights already factor in model complexity when quantifying the level of support.

statistical model, in order to improve the power to detect social transmission and/or allow comparison between groups. For example, researchers may wish to test whether the rate of social transmission is higher in one context than another. This can be done using *NBDA*, although the exact inference can vary depending on which version of *NBDA* is fitted, and how. In addition, for the model to be meaningful, the social networks for each diffusion must be of the same type in each case (see below).

One option is to fit an *OADA*, assuming a separate baseline rate function, for each diffusion (see box 5.4). Here, minimal assumptions are made, and social

Box 5.3

Akaike's information criterion (AIC)

Akaike's Information Criterion (AIC) provides a means to compare the fit of different statistical models that are fitted to the same data. Unlike *p* values, AIC can be used to compare models that are not nested (i.e., when one model is not a constrained version of another). A full description of the theoretical basis for AIC, and a guide to its use are provided by Burnham and Anderson (2002). Here we give a brief summary.

Akaike's Information Criterion is calculated from the log-likelihood for the model (L), where the model parameters have been optimized by maximum likelihood (e.g., see box 5.1):

$$AIC = -2L + 2k$$

where *k* is the number of parameters in the model. A version of AIC is also available that corrects for sample size:

$$AIC_c = -2L + 2k + \frac{2k(k+1)}{n-k-1}.$$

This should be used unless the sample size is large (in which case there will be little difference).

Models with lower AIC explain the model better after appropriately penalizing for the number of parameters in the model. The degree to which additional parameters are penalized is not arbitrary, since the difference in AIC between any two models fitted to the same data estimates the difference in Kullback-Leibler (K-L) information for the two models. K-L information measures the extent to which the predicted distribution for the dependent variable differs from its true distribution (i.e., the information lost when moving from the true distribution to the model).

Note that AIC only gives a measure of the relative fit of candidate models, not a measure of absolute fit, so it is the absolute *difference* in AIC that determines the relative performance of two models. Given a set of *R* candidate models, a researcher can obtain a relative measure of support for each model *i* by calculating the Akaike weights:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta_r)}$$

where Δ_i gives the difference in AIC between model *i* and the best model in the set; w_i can be thought of as the probability the *i* is the model with best K-L information in the set, accounting for sampling error.

learning is inferred purely from whether the order of acquisition tends to follow the network in each diffusion. Researchers cannot compare whether the rate of acquisition varies between diffusions, and so can only test for whether individual-level variables have an effect if they vary with diffusions. It is possible to test for differences in the relative rate of social transmission (per unit of network strength), but statistical power may be lower than using alternative methods.

Conversely, *TADA* is sensitive to the times at which individuals in each group acquire the trait(s), as well as the order they learn. Therefore, if all individuals

One can obtain measures of support for various features of the model, such as the presence of a variable, by summing Akaike weights over those models that include that feature. For example, imagine we run an *NBDA* with size as an individual-level variable, also considering a number of other individual-level variables, as well as additive, multiplicative, and asocial versions of the model. We get a measure of support for an effect of size by summing the Akaike weights of models that include size as a variable. This gives the probability that size is in the model with the best K-L information of those considered, after accounting for sampling variation. The same process could be used to compare different models of social learning with each other, and with models of asocial learning.

AIC also gives us a method of estimating parameter values that is not subject to the same problems as traditional model selection procedures. The traditional method is to select a "best" model, perhaps based on adjusted R-squared, AIC, or stepwise approaches using *p* values and an arbitrary significance level. Inferences are then based on the best model (i.e., they are *conditional on* that model being true). Such approaches do not take into account the uncertainty in the model selection procedure (i.e., which is *really* the best model). An alternative is to use a *model-averaging* procedure, which uses all the models considered to estimate parameter values, but the contribution of each is weighted by its Akaike weight. One obtains a model-averaged estimate $\hat{\hat{\theta}}$ for a parameter θ as follows:

$$\hat{\hat{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

where $\hat{\theta}_i$ is the maximum likelihood estimator for θ for model *i*.[1]

Traditional measures of the precision of a parameter estimate, standard errors, and confidence intervals are also conditional on the final model being correct. There are additional methods for adjusting measures of precision such that they take into account model selection uncertainty, yielding unconditional standard errors and confidence intervals.[2] For details, see Burnham and Anderson (2002, 153–167).

[1] It may sometimes be desirable to conduct model-averaging across only those models in which the parameter is present, in other cases it may make sense to take the value of the parameter to be zero in such models. See Burnham and Anderson (2002, 150–153) for details.

[2] Though they are still conditional on the set of models considered.