Ecological Archives E094-047-A1

Bill Shipley. 2013. The AIC model selection method applied to path analytic models compared using a d-separation test. *Ecology* 94:560-564. http://dx.doi.org/10.1890/12-0976.1

APPENDIX A. Proof that the C statistic of a d-sep test is a maximum likelihood estimate.

This appendix proves the following claim made in the main paper: if we calculate a series of null probabilities, $x = \{x_1, x_2, ..., x_k\}$, based on mutually independent tests, and if the underlying observations in fact obey the probability densities or distributions assumed by the null hypotheses, then the C statistic of the d-sep test is twice the negative of the log-likelihood of $y = \{y_1, y_2, ..., y_k\}$, where $y_i = \ln(x_i)$. Since y is a one-to-one mapping of x, then C is also is twice the negative of the log-likelihood of $x = \{x_1, x_2, ..., x_k\}$. If the null probabilities (x_i) are based on maximum likelihood estimates of the parameters in the underlying probability models used in calculating the null probabilities, then C is also twice the negative of the maximum of the log-likelihood.

Step 1

The uniform probability density is defined as:

$$p(x) = \frac{1}{\beta - \alpha} \quad \text{for} \quad \alpha < x < \beta; \quad p(x) = 0 \quad \text{elsewhere and} \quad \int_{-\infty}^{\infty} p(x)dx = 1 \tag{A.1}$$

The standard uniform probability density is therefore $p(x) = 1$ for $0 < x < 1$ and $p(x) = 0$ elsewhere. Equivalently, the standard uniform probability density is defined as the probability, $p(x)$, of observing a random value from this distribution falling in the infinitesimal interval between x and $x + dx$:

> $p(x)dx = dx$ for $0 < x < 1$ $p(x) = 0$ elsewhere. If the interval is $dx = 1$ (i.e. x is anywhere between 0 and 1) then $p(x) = 1$. If the interval is $dx = 0.5$ (for example, if x is anywhere between 0 and 0.5) then $p(x) = 0.5$.

Step 2

Now, let $y = \ln(x)$ be a transformation of the standard uniform variate x. The fundamental transformation law of probabilities(for example Freund 1962) states that if the probability density of x is given by $p(x)$ and the transformation function, given by $y = h(x)$, is differentiable and either increasing or decreasing for all values within the range of x, then the probability density of y is given by:

$$p(y) = p(x)\left|\frac{dx}{dy}\right|, \quad \frac{dy}{dx} \neq 0. \quad Now, \frac{dy}{dx} = \frac{1}{x}; \quad \left|\frac{dx}{dy}\right| = x = e^{\ln(x)} = e^{y} \tag{A.2}$$

Therefore, $p(y) = e^y$. This is the well-known fact that if x is a random variable from a standard uniform distribution then $\ln(x)$ follows an exponential distribution. Furthermore, if $y = \{y_1, y_2, ..., y_k\}$ is a vector of mutually independent random observations drawn from $p(y)$ then the likelihood of y, given $p(y)$ is:

$$\mathcal{L}(\mathbf{y}) = \prod_{i=1}^{k} p(y_i) = \prod_{i=1}^{k} e^{y_i} = e^{\sum_{i=1}^{k} y_i} = e^{\sum_{i=1}^{k} \ln(x)} \tag{A.3a}$$

$$-2\ln\left(\mathcal{L}(\mathbf{y})\right) = -2\sum_{i=1}^{k} \ln(x_i) = C \tag{A.3b}$$

Therefore, if $x = \{x_1, x_2, ..., x_k\}$ is a series of mutually independent values drawn from a standard uniform distribution,

and $y_i = \ln(x_i)$, then C is twice the negative of the log-likelihood of $y = \{y_1, y_2, ..., y_k\}$.

Step 3

We now combine this with the fact that, if we calculate the probability of observing at least as extreme a value as x_i given a probability model when x in fact follows this probability model (i.e. when the null hypothesis is true), and repeat this in a series of mutually independent tests, then this series of null probabilities, $x = \{x, x_2, ..., x_k\}$, follows a uniform distribution (Fisher 1925). This gives our fundamental result.

LITERATURE CITED

Fisher, R. A. 1925. Statistical methods for research workers. Oliver & Boyd, Edinburgh.

Freund, J. E. 1962. Mathematical statistics. Prentice-Hall, Englewood Cliffs, N.J.

[Back to E094-047]

Ecological Archives E094-047-A2

Bill Shipley. 2013. The AIC model selection method applied to path analytic models compared using a d-separation test. *Ecology* 94:560-564. http://dx.doi.org/10.1890/12-0976.1

APPENDIX B. Calculating the AIC statistic for d-sep tests.

This document explains how to calculate the AIC value for a path model with a hierarchical data structure. Text in `courier font` represents command lines or output in R (R-Development-Core-Team 2008). Further information concerning the fitting of such models can be found in Shipley (2009). The text file in the suppliment (simulated.data.txt) contains 250 observations involving 10 "individuals" in each of 25 "species". Five variables have been measured on each individual. Figure S1 shows two alternative path models, of which model 1 is the correct data-generating structure. After importing this file using the read.table function, the names in the data frame are:

```
> names(simulated.data)
[1] "species" "x1" "x3" "x4" "x5"
```

The first step is to attach the library ggm of R:

```
> library(ggm)
```

It is possible that you will need to install two other libraries before you can install ggm. If so, then execute the following commands:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("graph")
> biocLite("RBGL")
```

The second step is to enter the two path models using the DAG function in the ggm library:

```
> model1<-DAG(x2~x1,x3~x2,x4~x2,x5~x3+x4)
> model2<-DAG(x2~x1,x3~x2,x4~x2,x5~x2+x3+x4)
```

The third step is to obtain the basis sets of each model that are required to perform the d-sep test. These are given by the basiSet(amat) function in the ggm library, where "amat" is the object created via the DAG function. In our case, we would specify:

```
> basiSet(model1) > basiSet(model2)
```

For instance, the output of basiSet(model1) is:

```
[[1]]
[1] "x1" "x4" "x2"

[[2]]
[1] "x1" "x3" "x2"

[[3]]
[1] "x1" "x5" "x4" "x3"

[[4]]
[1] "x2" "x5" "x1" "x4" "x3"

[[5]]
[1] "x4" "x3" "x2"
```

This gives the five d-separation claims that form the basis set for the first model. Each element first lists the two variables that are d-separated, followed by the set of the causal parents of this pair. For instance, the first d-separation claim ([1] "x1" "x4" "x2") states that variables x1 and x4 are d-separated given variable x2 in model 1.

The fourth step involves obtaining the probabilities of the pair of variables in each d-separation claim being independent conditional on the causal parents, given the data. Obtaining such null probabilities will differ depending on the distributional properties of the variables. In our case, we know that the hierarchical nature of our data creates dependencies between the observations within each species. We will therefore test the independence claims using mixed model linear regression with normally distributed residuals, via the lmer function of the lme4 library in R. For each d-separation claim in the basis set we therefore regress on of the two variables in the pair on the set of conditioning variables plus the second variable in the pair. If the two variables are independent conditional on these causal parents then the slope associated with the second variable in the pair will not be significantly different from zero. The null probability of this slope being zero is given by a t-test. Here is the command to test the first d-separation claim from model 1 (assuming that you have already loaded the lme4 library):

```
> summary(lmer(x4~x2+x1+(x2+x1|species),data=simulated.data))
```

I have not reproduced the entire output of this command here. We need the null probability that x4 is independent of x1, given x2. This requires the t-value associated with the null hypothesis that the slope of variable x1 is zero; here t= -1.173. The degrees of freedom are 250 - 25 = 225 (i.e. the total number of observations minus the number of groups) giving a null probability:

```
> 2*(1-pt(abs(-1.173),225)) [1] 0.2420361
```

Note that the correct number of degrees of freedom in mixed-model regression is currently a point of contention but I am using the degrees of freedom that are output, for instance in the lme function for linear mixed models of the nlme library. We must repeat this for each of the d-separation claims in the basis set for this first model, giving five null probabilities. These five null probabilities are combined using Fisher's C-statistic and the resulting value will be used to calculate the AIC statistic for this first model.

We then repeat these same steps for the second model, giving a second C-statistic. Table B1 lists the basis sets for the two models, the null probabilities associated with each d-separation claim, and the resulting C-statistics. Note that model 2 is nested within model 1 in our example, and so the elements of its basis set are a subset of those in the basis set of model 1, but this will not true in general.

In order to obtain the AIC statistics for the two models we must also count the total number of parameters needs to fit each model. This requires that we again fit a series of mixed models but now we follow the links of each model rather than using the d-separation claims of the basis sets. For instance, the first link in model 1 is x1 ? x2 and so we must regress x2 on x1. For each such regression we count the total number of parameters that are estimated from the data. Since we must use a mixed model regression this includes the variances and covariances of the random part of the model plus the parameters (intercepts and slopes) of the fixed part of the model. Table B2 lists the regressions required to estimate each model (indicated by the notation "y ~ x") and the number of parameters estimated in each regression. The total number of parameters that are estimated in the full path model is the sum of the number of parameters estimated for each separate regression. The results of these regressions give the parameter estimates of the links in each path model. For instance, the x1 ? x2 link of model 1gives the following estimates:

$x2_{ij} = ?_i + ?_i x1_{ij} + ?_{ij}$
$?_i = -0.46 + N(0, 1.09)$
$?_i = 0.97 + N(0, 0.27)$
$\text{correlation}(?_i, ?_i) = -0.36$
$?_{ij} = N(0, 4.67)$

Finally, given the C-statistic and the total number of parameters (K) associated with each path model, we can calculate the AIC statistic, or its bias-corrected version, for each model (Equations B1a, b); *n* is the total number of observations (i.e. $n = 250$).

$$AIC = C + 2K \qquad \text{(B.1a)}$$

$$AIC_c = C + 2K\left(\frac{n}{n-K-1}\right) \quad \text{(B.1b)}$$

These values are:

Model 1: AIC = 8.501 + 2(28) = 64.501, AIC_c = 8.501 + 2(28)(250/221) = 71.849
Model 2: AIC = 8.436 + 2(33) = 74.436, AIC_c = 8.436 + 2(33)(250/221) = 83.097

TABLE B1. The elements of the basis set for each model are listed in the first column; the notation "x4_||_x1|x2" means that the pair (x1, x4) is d-separated conditional on x2. The following columns list the *t*-value associated with the null hypothesis that the slope of the second variable in the pair is zero in the associated mixed model regression and its null probability.

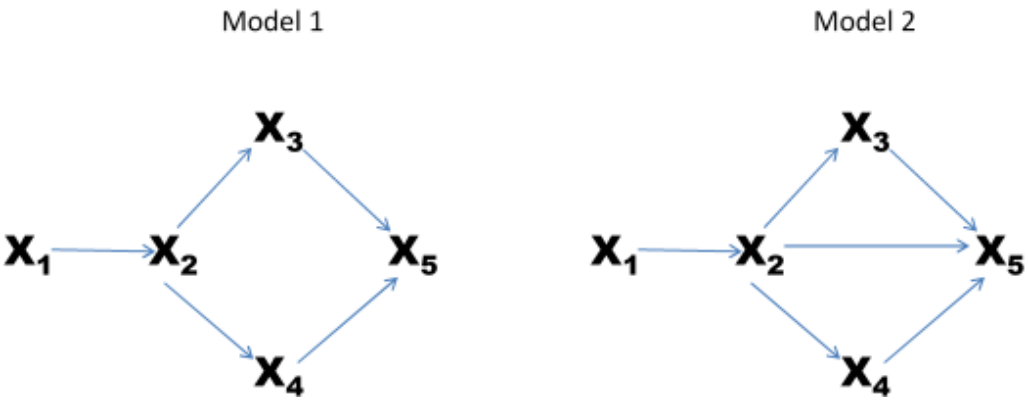
d-sep claim	*t*-value	Null probability
Model 1		
x4_||_x1|{x2}	-1.173	0.242
x3_||_x1|{x2}	1.600	0.111
x5_||_x1|{x3,x4}	-0.330	0.742
x4_||_x3|{x2}	0.334	0.739
x5_||_x2|{x1,x3,x4}	-0.040	0.968
	C-statistic:	8.501, 10 d.f.
Model 2		
x4_||_x1|{x2}	-1.173	0.242
x3_||_x1|{x2}	1.600	0.111
x5_||_x1|{x3,x4}	-0.330	0.742
x4_||_x3|{x2}	0.334	0.739
	C-statistic	8.436, 8 d.f.

TABLE B2. The series of separate mixed model regressions that must be fit to estimate the parameters of each of two competing path models; note that intercepts have been included. The notation "x2 ~ x1" means that x2 is regressed on x1 with both slopes and intercepts being modeled as random. K_i gives the number of parameters that must be estimated in each separate mixed-model regression and K is the total number of parameters that must be estimated for the model in question.

Regression	K_i
Model 1	
x2 ~ x1	6
x3 ~ x2	6
x4 ~ x2	6
x5 ~ x3 + x4	10
K=?K_i	28

Model 2	
x2 ~ x1	6
x3 ~ x2	6
x4 ~ x2	6
x5 ~ x2 + x3 + x4	15
$K = ?K_i$	33

FIG. B1. The two alternative path models whose AIC statistics are sought.



LITERATURE CITED

R-Development-Core-Team 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. Ecology 90:363-368.

[Back to E094-047]