*Ecology*, 94(3), 2013, pp. 560–564


Reports

# The AIC model selection method applied to path analytic models compared using a d-separation test

Bill Shipley[1]

*Département de Biologie, Université de Sherbrooke, Sherbrooke, Quebec J1K 2R1 Canada*

*Abstract*. Classical path analysis is a statistical technique used to test causal hypotheses involving multiple variables without latent variables, assuming linearity, multivariate normality, and a sufficient sample size. The d-separation (d-sep) test is a generalization of path analysis that relaxes these assumptions. Although model selection using Akaike's information criterion (AIC) is well established for classical path analysis, this model selection technique has not yet been developed for d-sep tests. In this paper, I explain how to use the AIC statistic for d-sep tests, give a worked example, and include instructions (supplemental material) to implement the analysis in the R computing language.

*Key words: AIC statistic; d-separation (d-sep) test; path analysis; structural equation modeling (SEM).*

## Introduction

Path analysis is a statistical method of testing multivariate hypotheses concerning the cause–effect links between variables, of subsequently quantifying these links, and of decomposing these causal effects (Bollen 1989, Shipley 2000*a*, Grace 2006). This method is becoming increasingly popular in ecology and evolution because these disciplines study phenomena that are both inherently multivariate and are often concerned with causal relationships. In this paper, I use the term "path analysis" to refer to a general class of acyclic causal models involving only observed (manifest) variables but without any further assumptions about the distributions of such variables or the functional form of the links between them. If we further restrict the term "path analysis" to include the assumptions of multivariate normality and linearity, then a path analysis is simply a structural equations model (SEM) without latent variables. However, many path analytic models in ecology and evolution cannot be analyzed using classical SEM either because some variables are not normally distributed, some links are not linear, the data have a nested structure that creates dependencies between observations, or sample sizes are too small to justify the asymptotic properties of SEM. To overcome these limitations, I developed an alternative statistical test (a d-sep test) of path models that relaxes these assumptions, based on the graph theoretic notion of directed separation (d-separation) of directed acyclic graphs (DAGs). See Pearl (2000), Shipley (2000*a*), and Grace et al. (2012) for extended discussions of graph-theoretic treatments of causal models.

Manuscript received 14 June 2012; revised 8 October 2012; accepted 11 October 2012. Corresponding Editor: B. D. Inouye.
[1] E-mail: Bill.Shipley@USherbrooke.ca

The most common purpose of path analysis is confirmatory: to test for an agreement between specific causal hypotheses and empirical data. After fixing a significance level, a model is judged to be either rejected by the data or else statistically consistent with it. However, for a given sample size, it is possible for more than one competing causal model to be judged consistent with the data, and this can lead to questions of model selection among competing models. One class of such competing models are "equivalent models" (Shipley 2000*a*:264–266); no statistical test can differentiate between them because they place equivalent statistical constraints on the data. Another class of such competing models, and the subject of this paper, are models that are judged consistent with the data because of the limitations of statistical power. One of the most popular methods of model selection for this second class of competing models is Akaike's information criterion (AIC) statistic and its variants (Akaike 1973, Burnham and Anderson 2010). The application of the AIC statistic to classical SEM is well known (Bollen 1989). Classical SEM consists of maximizing the likelihood of the model-predicted covariance matrix, which is itself a function of three types of parameters: variances of exogenous variables, covariances between exogenous variables, and path coefficients. It is not clear if the AIC statistic can be adapted to d-sep tests. For instance Cardon et al. (2011) used AIC statistics to compare competing d-sep models based only on the justification that "... the information criteria is a weighted sum of the measure of badness of fit (i.e., the *C*-value [of the d-sep test]) and of a measure of complexity...". However, this is incorrect. The AIC statistic requires that the measure of "badness of fit" be based on maximum-likelihood estimates and so it is necessary to show that the *C* statistic is a maximum-likelihood estimate. This paper, therefore, has two purposes. First, I show

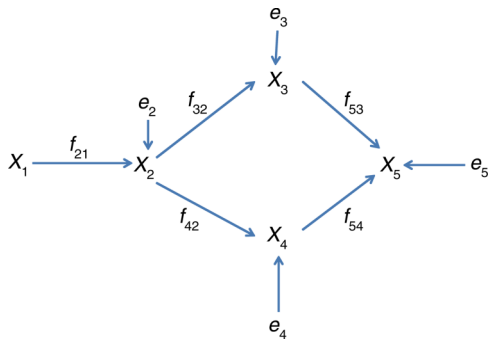


FIG. 1. A directed acyclic graph showing the hypothesized causal links between five measured random variables ($X_i$); the errors ($e_i$) represent unknown and mutually independent causes of each measured variable, and the probability distributions of these errors (and of $X_1$) are unspecified. The unspecified functions linking each effect–cause pair are indicated as $f_{ji}$, where $i$ is the the number of the effect and $j$ is the number of the cause.

(Appendix A) that the $C$ statistic is indeed a maximum-likelihood estimate if the null probabilities upon which it is based are, themselves, maximum-likelihood probabilities. Second, I describe how to apply the AIC statistic to models tested using d-sep tests. As will be seen, the analysis by Cardon et al. (2011) is indeed correct, although their stated justification is not.

### *D-sep tests*

The mathematical justification for a d-sep test is given in Shipley (2000*b*) and described in detail in Shipley (2003, 2004, 2009). The six steps are as follows.

1) Specify the causal hypothesis in the form of a directed acyclic graph, which is a diagram showing the variables involved in the hypothesis and the direct causal links between them as arrows; an example is shown in Fig. 1 in which I include the non-standard notation "$f_{ij}$" to indicate the functional form of the equation linking variables $i$ and $j$.
2) Find each unique pair of variables ($X$, $Y$) in the graph that do not have a single arrow going from one to the other; i.e., $X \rightarrow Y$ or $X \leftarrow Y$.
3) Find the direct hypothesized causes of each variable in the pair (the "causal parents"). The set of causal parents for each pair is called the conditioning set. For instance, variables $X_1$ and $X_3$ in Fig. 1 are not joined by an arrow and the set of causal parents of $X_1$ (i.e., none) and $X_3$ (i.e., $X_2$) are $\{\varphi, X_2\} = \{X_2\}$ where $\varphi$ is a null set. Therefore $\{X_2\}$ is the conditioning set for the pair ($X_1$, $X_3$).
4) Convert each unique pair plus their conditioning set into an independence claim. Specifically, given a pair ($X_i$, $X_j$) and the set $\mathbf{Q}$ of causal parents of this pair that define the conditioning set, the independence claim is "$X_i \perp\!\!\!\perp X_j \,|\, \mathbf{Q}$": i.e., "variables $X_i$ and $X_j$ are probabilistically independent conditional of the set of variables in $\mathbf{Q}$." The entire set of $c$ d-separation claims constructed in this way is called a basis set. Table 1 lists the $c = 5$ d-separation claims in the basis set for the DAG in Fig. 1. This basis set has five important properties (Shipley 2000*b*). First, all dependence/independence relationships between the variables in the DAG can be derived from those in the basis set. Second, no independence relationship in the basis set can be derived from some combination of other d-sep claims in the basis set (i.e., the basis set is the minimal set needed to specify the DAG). Third, if data are generated following the causal structure specified by the DAG, then each d-separation claim in the DAG implies a statistical independence in the data; furthermore, there can exist no independence relationships in the data that are not implied by the d-separation claims in the basis set. Stated differently, the basis set defines the constraints placed on the data by the hypothesized causal model. Fourth, the independence claims of the elements in this basis set are mutually independent. Finally, these independence claims hold for any sampling distribution of the variables in the data and for any functional form of the causal relationships.
5) Calculate the null probability ($p_i$) associated with each of the predicted independence claims in the basis set using statistical tests whose assumptions are appropriate for variables and relationships in question.

The final step (6) is to combine these null hypotheses using Fisher's $C$ statistic (Eq. 1). If the data are generated following the causal hypothesis specified in the DAG, this $C$ statistic will follow a chi-squared distribution whose degrees of freedom are equal to $2c$ (Fisher 1925):

$$C = -2\sum_{i=1}^{c} \ln(p_i). \qquad (1)$$

### *AIC and the d-sep test*

Given a statistical model ($M$), which assumes a parametric probability distribution and which has been fitted to empirical data ($D$) by choosing parameter values ($\hat{\Theta}$) that maximize the likelihood of the model given the data, the AIC statistic is defined as

TABLE 1. The d-separation claims in the basis set implied by the directed acyclic graph (DAG) shown in Fig. 1.

| Claim number | d-separation claim |
|---|---|
| 1 | $X_1 \perp\!\!\!\perp X_3 \,\|\, \{X_2\}$ |
| 2 | $X_1 \perp\!\!\!\perp X_4 \,\|\, \{X_2\}$ |
| 3 | $X_1 \perp\!\!\!\perp X_5 \,\|\, \{X_3, X_4\}$ |
| 4 | $X_2 \perp\!\!\!\perp X_5 \,\|\, \{X_1, X_3, X_4\}$ |
| 5 | $X_3 \perp\!\!\!\perp X_4 \,\|\, \{X_2\}$ |

*Note:* The notation "$X_i \perp\!\!\!\perp X_j \,|\, \{\mathbf{Q}\}$," where $\mathbf{Q}$ is a set of variables not including $X_i$ or $X_j$, means that variable $X_i$ must be independent of $X_j$ conditional of the set of conditioning variables in $\mathbf{Q}$.

TABLE 2. Translation of the DAG in Fig. 1 into structural equations.

| Nonparametric equations | Parametric equations | Number of free parameters |
|---|---|---|
| $X_{1j} = \varepsilon_{1j}$ | $\varepsilon_{1j} \sim \mathcal{N}(\mu = 0, \boldsymbol{\sigma}_1)$ | 1 |
| $X_{2j} = f(X_{1j}) + \varepsilon_{2j}$ | $X_{2j} = \mathbf{a}_{21}X_{1j}$; $\varepsilon_{2j} \sim \mathcal{N}(\mu = 0, \boldsymbol{\sigma}_2)$ | 2 |
| $X_{3j} = f(X_{2j}) + \varepsilon_{3j}$ | $X_{3j} = \mathbf{a}_{32}X_{2j}$; $\varepsilon_{3j} \sim \mathcal{N}(\mu = 0, \boldsymbol{\sigma}_3)$ | 2 |
| $X_{4j} = f(X_{2j}) + \varepsilon_{4j}$ | $X_{4j} = \mathbf{a}_{42}X_{2j}$; $\varepsilon_{4j} \sim \mathcal{N}(\mu = 0, \boldsymbol{\sigma}_4)$ | 2 |
| $X_{5j} = f(X_{3j}) + f(X_{4j}) + \varepsilon_{5j}$ | $X_{5j} = \mathbf{a}_{53}X_{3j} + \mathbf{a}_{54}X_{4j}$; $\varepsilon_{5j} \sim \mathcal{N}(\mu = 0, \boldsymbol{\sigma}_5)$ | 3 |

*Notes:* The nonparametric equations link the variables but do not specify the functional forms of the links or the distributions of the random components. The parametric form specifies the functional forms of the links and the distributions of random components in terms of free parameters (**a** and **σ**). In this example, we assume linearity for the functional forms and multivariate normality for the random components ($\varepsilon_{.ij}$) and so $\mathcal{N}(\mu,\boldsymbol{\sigma}_i)$ is the univariate normal distribution of each, whose standard deviations ($\boldsymbol{\sigma}_i$) must be estimated from the data (mean = μ). The number of free parameters included in each parametric equation is counted in the last column. There are a total of 10 free parameters that must be estimated.

$$\mathrm{AIC} = -2\ln\left[L\left(M(\hat{\Theta});D\right)\right] + 2K \tag{2a}$$

$$\mathrm{AIC_c} = -2\ln\left[L\left(M(\hat{\Theta});D\right)\right] + 2K\left(\frac{n}{n-K-1}\right). \tag{2b}$$

Eq. 2b involves a correction for small sample sizes ($n$). In this equation, $L(M(\hat{\Theta})$; D) is the likelihood of the model, given the data, when the estimated parameters of the probability distribution maximize this likelihood. $K$ is the number of maximum-likelihood parameters that are estimated using the empirical data. Given competing models, the one with the smallest AIC value is preferred and the relative support of the different models is based on the differences in the AIC values relative to the preferred model. See Burnham and Anderson (2010) for details of model selection via the AIC methodology.

In Appendix B, I show that the maximum likelihood of a path model that is tested using the d-sep test, when the individual null probabilities of the d-sep claims in the basis set are based on probability densities evaluated using the maximum-likelihood parameter estimates of these probability densities, is a function of the $C$ statistic. Specifically,

$$-2\ln\left[L\left(M(\hat{\Theta});D\right)\right] = -2\sum_{i=1}^{n}\ln(p_i) = C. \tag{3}$$

In order to use the AIC methodology, we must also know the number of maximum-likelihood estimates (i.e., $K$) needed to evaluate these null probabilities. This will depend on the probability model used to evaluate each d-sep claim.

A DAG can always be translated into a series of functional relationships (structural equations) linking the variables and the distributional form of the random components of the probabilistic model. The empirical fitting of data to the DAG requires the estimation of certain parameters of the probabilistic model. The value of $K$ in Eq. 1 is the total number of free parameters that must be estimated in order to obtain the null probabilities needed to apply the d-sep test. The easiest way to calculate $K$ in the context of path models is to do the following:

1) Translate the DAG into the series of cause–effect relationships implied by the DAG (nonparametric structural equations).
2) Specify the functional form and the distributional assumptions of the random components of these relationships by explicitly specifying the free parameters that must be estimated from the data via maximum-likelihood estimation (parametric structural equations).
3) Sum the number of free parameters over these relationships. This gives the total number of free parameters and thus the value of $K$ in Eq. 2.

Consider the case in which we assume a multivariate normal distribution with $V$ variables and linear relationships (in any) between variables. I make these assumptions in order to compare the result with those of a classic structural equations model without latent variables. The number of structural equations implied by a DAG is equal to the number of variables in the DAG. The DAG in Fig. 1 requires five structural equations. These equations must be linear and the random components must each follow a normal distribution. I fix the means of the variables to zero (i.e., I center each variable about its sample mean) because the DAG only describes the relationships between the variables. I must therefore estimate five slopes (path coefficients) and five variances for a total of $K = 10$ free parameters (Bollen 1989, Shipley 2000*a*). If I was interested in the intercepts of the variables then I would have to estimate five more free parameters. Thus, AIC statistic for this path model, when tested using the d-sep test, is $\mathrm{AIC} = C + 10$.

Note that $K = 10$ is the same result as would be obtained from a classical structural equations model. A classical structural equations model involves maximizing the likelihood of the model with respect to the population covariance matrix (**Σ**). However, the elements of **Σ** can be expressed as a function of the path coefficients and the variances; there are 10 of these free parameters (shown in bold in Table 2) given the DAG in Fig. 1. The five d-separation claims in the basis set (Table 1) are constraints placed on the elements of **Σ**.

TABLE 3. Measured covariances (lower triangle), variances (diagonal, shown in bold italic typeface), and Pearson correlation coefficients (upper triangle) in a simulated data set consisting of 100 observations.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | ***0.917*** | 0.514 | 0.219 | 0.234 | 0.175 |
| $X_2$ | 0.534 | ***1.176*** | 0.536 | 0.496 | 0.290 |
| $X_3$ | 0.202 | 0.560 | ***0.930*** | 0.254 | 0.272 |
| $X_4$ | 0.230 | 0.550 | 0.250 | ***1.046*** | 0.231 |
| $X_5$ | 0.172 | 0.321 | 0.268 | 0.241 | ***1.044*** |

The number of free parameters represent the number of model degrees of freedom ($df_M = 10$) and the constraints represent the residual degrees of freedom ($df_R = 5$) while the total number of degrees of freedom are the number of unique elements of $\mathbf{\Sigma}$, of which there are $df_T = V(V + 1)/2 = 15$.

An important advantage of the d-sep test over classical structural equations models is that one can use different functional forms for the links between the variables and for the distributional assumptions of the random components. The null probabilities associated with the d-separation claims can be obtained using statistical tests appropriate to the nature of the data, including permutation methods if necessary although these null probabilities must be associated with fits between the data and a parametric function that maximize the likelihood in order to apply the AIC statistic. To obtain the number of free parameters, one simply converts the nonparametric structural equations into the appropriate parametric form; these parameter estimates can be obtained using mixed models (including the parameter estimates of the hierarchical random components) or generalized linear methods that assume non-normal probability distributions. The act of fitting these models to data by maximizing the likelihood automatically provides the number of free parameters associated with each structural equation. In fact, one can even use generalized additive models based on regression smoothers, which maximize the expected log likelihood, and use the estimated degrees of freedom of the regression smoothers, which are estimates of the equivalent number of free parameters needed for each smoothed term (Hastie and Tibshirani 1986).

Here, I use a numerical example in which I generated 100 independent "observations" following the causal model in Fig. 1, each being a standard normal deviate. The path coefficients and population Pearson correlation coefficients for the first three links (A → B, B → C, B → D) were 0.5. However, the path coefficients and population Pearson correlation coefficients for the final two links (C → E, D → E) were only 0.2. This simulates a case in which some of the causal links, although present, are very weak. Table 3 lists the sample covariances and correlations for this simulated data set. Imagine that one has five other competing path models (Fig. 2) whose differences are located in those parts of the hypothesized causal structure that have the weakest signal (i.e., the links between $X_5$ and the other variables). Model 1 is the one shown in Fig. 1. With respect to Fig. 1, the remaining models are modified as follows: model 2 does not have the $X_3 \rightarrow X_5$ link; model 3 does not have the $X_4 \rightarrow X_5$ link; model 4 does not have either the $X_3 \rightarrow X_5$ or the $X_4 \rightarrow X_5$ links; model 5 does not have either the $X_3 \rightarrow X_5$ or the $X_4 \rightarrow X_5$ links but does have an $X_2 \rightarrow X_5$ link; model 6 has an $X_2 \rightarrow X_5$ link. Table 4 lists the relevant statistics. None of the six competing models can be rejected at the 0.05 significance level, based on the d-sep test with the statistical power afforded by 100 independent observations. Of course, all but the true model would be rejected if we were able to increase the sample size sufficiently but this is rarely an option in practice. We must conclude that all six models are consistent with the data that we have, given our chosen significance level.

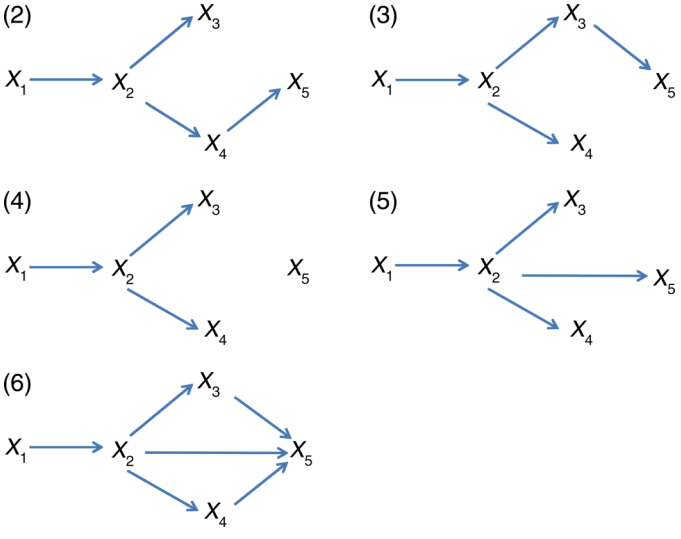


FIG. 2. Five alternative acyclic graphs, relative to the one shown in Fig. 1, are numbered 2–6. For simplicity, the error variables and the unspecified functional links are not shown. The nomenclature is as in Fig. 1.

TABLE 4. Model fit of six competing path models using the data in Table 3; model 1 is the true generating model.

| Model | $C$ (df, $P$) | $K$ | $AIC_c$ | $\Delta AIC_c$ | $W$ | A→B | B→C | B→D | C→E | D→E | B→D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 3.239 (8, 0.92) | 11 | 28.239 | 0 | 0.357 | 0.585 | 0.473 | 0.472 | 0.18 | 0 | 0.128 |
| 1 | 6.070 (10, 0.81) | 10 | 28.542 | 0.303 | 0.307 | 0.585 | 0.473 | 0.472 | 0.243 | 0.168 | 0 |
| 5 | 9.092 (12,0.70) | 9 | 29.092 | 0.853 | 0.233 | 0.585 | 0.473 | 0.472 | 0 | 0 | 0 |
| 3 | 11.224 (12,0.51) | 9 | 31.224 | 2.985 | 0.080 | 0.585 | 0.473 | 0.472 | 0.288 | 0 | 0 |
| 2 | 13.983 (12,0.31) | 9 | 33.983 | 5.744 | 0.020 | 0.585 | 0.473 | 0.472 | 0 | 0.226 | 0 |
| 4 | 21.523 (14,0.09) | 8 | 39.105 | 10.866 | 0.002 | 0.585 | 0.473 | 0.472 | 0 | 0 | 0 |
| Model-averaged estimator | | | | | | 0.585 | 0.473 | 0.472 | 0.162 | 0.056 | 0.046 |

*Notes:* $C$ (df, $P$) gives the $C$ statistic and, in parentheses, its degrees of freedom and the null probability. $K$ is the number of parameters needed to fit the model. $AIC_c$ and $\Delta AIC_c$ are the Akaike values and the difference in $AIC_c$ relative to model 6. $W$ gives the model weights, and the last row gives the model-averaged estimate of each path coefficient. The last six columns give the estimated path coefficients associated with each link for each model.

We now apply the procedures of multimodel inference via the $AIC_c$ statistic. Rather than comparing each model to a fixed significance level, we compare them to each other by ranking each relative to the model with the lowest value. The model with the lowest $AIC_c$ is model 6, not model 1 (the true model), although the $\Delta AIC_c$ value for model 1 (0.303) indicates that it is essentially tied with model 6. Models whose $\Delta AIC_c$ values are less than 3 are generally considered to have "substantial" support, models whose $\Delta AIC_c$ are between 3 and 7 are considered to have "considerably less" support, and models whose $\Delta AIC_c$ are >10 have "essentially no" support relative to the best model of the set (Burnham and Anderson 2010). Therefore models 6, 1, 5, and 3 cannot be distinguished while models 2 and 4 would normally be excluded from consideration. Any further choice among the remaining models cannot be based on statistical criteria but might be based on external biological knowledge about the different cause–effect relationships assumed by them.

The supplemental material shows each of these steps for two of the alternative path models that are fit to data having a more complicated nested hierarchical structure using the R statistical package (R Development Core Team 2008).

## ACKNOWLEDGMENTS

This research was financially supported by the Natural Sciences and Engineering Research Council of Canada.

## LITERATURE CITED

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary.

Bollen, K. A. 1989. Structural equations with latent variables. John Wiley and Sons, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2010. Model selection and multimodel inference. A practical information-theoretic approach. Second edition. Springer, New York, New York, USA.

Cardon, M., G. Loot, G. Grenouillet, and S. Blanchet. 2011. Host characteristics and environmental factors differentially drive the burden and pathogenicity of an ectoparasite: a multilevel causal analysis. Journal of Animal Ecology 80:657–667.

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

Grace, J. B. 2006. Structural equation modeling and natural systems. Cambridge University Press, Cambridge, UK.

Grace, J. B., D. R. Schoolmaster, G. R. Gutenspergen, A. M. Little, B. R. Mitchell, K. M. Miller, and E. W. Schweiger. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3:1–44.

Hastie, T., and R. Tibshirani. 1986. Generalized additive models. Statistical Science 1:297–318.

Pearl, J. 2000. Causality. Cambridge University Press, Cambridge, UK.

R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.com

Shipley, B. 2000*a*. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge, UK.

Shipley, B. 2000*b*. A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling 7:206–218.

Shipley, B. 2003. Testing recursive path models with correlated errors using d-separation. Structural Equation Modeling 10:214–221.

Shipley, B. 2004. Analysing the allometry of multiple interacting traits. Perspectives in Plant Ecology, Evolution and Systematics 6:235–241.

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

## SUPPLEMENTAL MATERIAL

### Appendix A

Proof that the C statistic of a d-sep test is a maximum-likelihood estimate (*Ecological Archives* E094-047-A1).

### Appendix B

Calculating the AIC statistic for d-sep tests (*Ecological Archives* E094-047-A2).

### Supplement

Data set used in Appendix B (*Ecological Archives* E094-047-S1).