Processing and Analysis of Biological Data

Path analysis and causal inference

Øystein H. Opedal

2 Dec 2023

Cause and correlation in biology

Correlation does not imply causation. This statement is central to scientific thinking, and underscores the importance of interpreting results from observational studies carefully, and ideally confirming any inferred relationship experimentally. Experiments are indeed a powerful way of separating the effects of multiple correlated variables. In this chapter, we will discuss an alternative approach to inferring causality tracing back to the work of Sewall Wright a hundred years ago (Wright 1921 and later). Broadly speaking, the method can be used to infer causality by combining knowledge about the natural history/mechanics of the study system with estimated statistical parameters such as correlation coefficients and regression slopes.

For further reading I strongly recommend Bill Shipley´s book "Cause and Correlation in Biology".

As an example, we will work with the alpine plants dataset.

```
plants = read.csv(file="datasets/alpineplants.csv")
```

Wrightian Path analysis

In its simplest form, a path analysis consists of a series of correlations combined with linear regressions fitted to standardized variables (zero mean, unit variance), thus obtaining *path coefficients*. Before going into technical aspects, a critical point is that before estimating any parameters, causal inference through path analysis or related methods requires formulating a graphical model in the form of a *directed graph* showing the assumed causal (and non-causal) relationships between a set of variables.

As an example, we will consider two different models for how snow depth, minimum winter temperature and soil moisture affect the distribution and abundance of *Carex bigelowii*. In the first model, we will assume independent effects of each predictor, thus building a path model on the form

$$snow \rightarrow Carex.bigelowii$$

$$min.T.winter \rightarrow Carex.bigelowii$$

$$soil.moist \rightarrow Carex.bigelowii$$

An alternative model is that snow cover affects winter temperature and soil moisture, which is turn affects the plant.

$$snow \rightarrow soil.moist$$

$$snow \rightarrow min.T.winter$$

$$min.T.winter \rightarrow Carex.bigelowii$$

$$snow \rightarrow Carex.bigelowii$$

In path analysis, we call the response variables (with arrows coming into them) *endogeneous* variables, and the predictors (with arrows only going out of them) *exogeneous* variables.

The first model can be fitted as a standard multiple regression, while the second model will involve fitting three different component models. Before fitting the models, we remove some `NA`s and z-transform all variables (including the response variables).

```
plants = na.omit(plants)
plants = as.data.frame(scale(plants))

round(colMeans(plants), 2)
```

```
##    Carex.bigelowii Thalictrum.alpinum      mean_T_winter       max_T_winter
##                  0                  0                  0                  0
##       min_T_winter      mean_T_summer       max_T_summer       min_T_summer
##                  0                  0                  0                  0
##              light               snow         soil_moist           altitude
##                  0                  0                  0                  0
```

```
round(apply(plants, 2, sd), 2)
```

```
##    Carex.bigelowii Thalictrum.alpinum      mean_T_winter       max_T_winter
##                  1                  1                  1                  1
##       min_T_winter      mean_T_summer       max_T_summer       min_T_summer
##                  1                  1                  1                  1
##              light               snow         soil_moist           altitude
##                  1                  1                  1                  1
```

```
m1 = lm(Carex.bigelowii ~ snow + min_T_winter + soil_moist, data=plants)

m2a = lm(min_T_winter ~ snow, data=plants)
m2b = lm(soil_moist ~ snow, data=plants)
m2c = lm(Carex.bigelowii ~ min_T_winter + soil_moist, data=plants)
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = Carex.bigelowii ~ snow + min_T_winter + soil_moist,
##     data = plants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3948 -0.4935 -0.2902  0.2450  3.7531
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  4.996e-16  9.775e-02   0.000    1.000
## snow         1.772e-01  1.629e-01   1.088    0.280
## min_T_winter 2.065e-01  1.647e-01   1.254    0.213
## soil_moist   2.254e-02  1.120e-01   0.201    0.841
##
## Residual standard error: 0.9427 on 89 degrees of freedom
## Multiple R-squared:  0.1404, Adjusted R-squared:  0.1114
## F-statistic: 4.844 on 3 and 89 DF,  p-value: 0.003612
```

The first model suggests positive but weakly supported effects of both snow cover and minimum winter temperature on the abundance of *Carex bigelowii.* Keep in mind though that snow cover and minimum winter temperature are strongly positively correlated, so that we may have some issues with multicollinearity in this model.

EXERCISE: Draw (on paper) the path diagram corresponding to this model, and add the estimated path coefficients, including the correlations between the exogeneous (predictor) variables. We can calculate the unexplained variance ("U") in the response as $\sqrt{(1-r^2)}$ (which places it on the standardized [correlation] scale like the path coefficients). Interpret the results.

In this model we can calculate the total (net) effect of snow cover on the abundance of *Carex bigelowii* by summing the direct effect and the effects arising through correlations with other variables.

```
summary(m1)$coef[2,1] +
summary(m1)$coef[3,1]*cor(plants$snow, plants$min_T_winter, "pairwise") +
summary(m1)$coef[4,1]*cor(plants$snow, plants$soil_moist, "pairwise")
```

```
## [1] 0.3508336
```

```
cor(plants$snow, plants$Carex.bigelowii, "pairwise")
```

```
## [1] 0.3508336
```

In the second model, there is (as expected) a strong positive effect of snow cover on minimum winter temperature, and in turn a positive effect of winter temperature on *Carex bigelowii.* Thus, under this model, we have strong support for the hypothesized causal links from snow cover to *Carex* abundance.

EXERCISE: Draw the path diagram and interpret the direct and indirect effects of snow cover on *Carex* abundance.

```
summary(m2a)$coef
```

```
##                  Estimate Std. Error       t value     Pr(>|t|)
## (Intercept) -1.106648e-15 0.06354811 -1.741434e-14 1.000000e+00
## snow         7.927891e-01 0.06389254  1.240816e+01 2.829427e-21
```

```
summary(m2b)$coef
```

```
##                 Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 6.074519e-17 0.09345465 6.499965e-16 1.000000e+00
## snow        4.433825e-01 0.09396118 4.718784e+00 8.546112e-06
```

```r
summary(m2c)$coef
```

```
##                 Estimate Std. Error      t value    Pr(>|t|)
## (Intercept)  6.733193e-16 0.09784898 6.881209e-15 1.000000000
## min_T_winter 3.389064e-01 0.11095020 3.054582e+00 0.002964704
## soil_moist   3.985433e-02 0.11095020 3.592092e-01 0.720279994
```

Structural equation modelling

Structural equation modelling is a further development of path analysis that offers greater flexibility compared to traditional path analysis. Below we fit our second candidate model using the `piecewiseSEM` package.

```r
library(piecewiseSEM)

m2 = psem(lm(soil_moist~snow, data=plants),
          lm(min_T_winter~snow, data=plants),
          lm(Carex.bigelowii~min_T_winter+soil_moist, data=plants),
          data=plants)
```

```r
summary(m2)
```

```
##   |                                                                                |

##
## Structural Equation Model of m2
##
## Call:
##   soil_moist ~ snow
##   min_T_winter ~ snow
##   Carex.bigelowii ~ min_T_winter + soil_moist
##
##     AIC      BIC
##  28.445   53.771
##
## ---
## Tests of directed separation:
##
##                      Independ.Claim Test.Type DF Crit.Value P.Value
##      Carex.bigelowii ~ snow + ...      coef 89     1.0875  0.2797
##   min_T_winter ~ soil_moist + ...      coef 90     1.9656  0.0524
##
## Global goodness-of-fit:
##
##   Fisher's C = 8.445 with P-value = 0.077 and on 4 degrees of freedom
##
## ---
## Coefficients:
##
##          Response    Predictor Estimate Std.Error DF Crit.Value P.Value
##        soil_moist         snow   0.4434    0.0940 91     4.7188  0.0000
##      min_T_winter         snow   0.7928    0.0639 91    12.4082  0.0000
```

```
##   Carex.bigelowii min_T_winter   0.3389    0.1110 90     3.0546  0.0030
##   Carex.bigelowii   soil_moist   0.0399    0.1110 90     0.3592  0.7203
##   Std.Estimate
##         0.4434 ***
##         0.7928 ***
##         0.3389  **
##         0.0399
##
##   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
##
## ---
## Individual R-squared:
##
##          Response method R.squared
##        soil_moist   none      0.20
##      min_T_winter   none      0.63
##   Carex.bigelowii   none      0.13
```

```
plot(m2)
```

The `piecewiseSEM` package implements tests of so-called *directed separation* ("d-separation"), which is a test for conditional non-independence of variables. Here, the key test is if *Carex* abundance is really conditionally independent of snow cover, i.e. after we have accounted for winter temperature and soil moisture. The interpretation of these tests is different from what we are used to for normal models. In this case the null hypothesis is that the model represents the data well, and a low p-value therefore indicates *bad* model fit, while a higher p-value indicate decent fit to the data. However, as always, we need to interpret any result in light of the parameter estimates.

Note that for a SEM containing multiple components, the `piecewiseSEM` package also provides a hypothesis test for the entire model, again with a high p-value indicating a decent fit to the data.

Note that the output of the current model includes a test of directed separation for the relationship between winter temperature and soil moisture, which was not our main interest in this case. We can avoid this by specifying an untested correlation between these variables, corresponding to double-headed arrows in our path analyses above, through the somewhat exotic `%~~%` operator.

```
m2b = psem(lm(soil_moist~snow, data=plants),
          lm(min_T_winter~snow, data=plants),
          lm(Carex.bigelowii~min_T_winter+soil_moist, data=plants),
          min_T_winter %~~% soil_moist,
          data=plants)
summary(m2b)
```

```
##   |                                                                       |

##
## Structural Equation Model of m2b
##
## Call:
##   soil_moist ~ snow
##   min_T_winter ~ snow
##   Carex.bigelowii ~ min_T_winter + soil_moist
##   min_T_winter ~~ soil_moist
##
```

```
##     AIC      BIC
##  22.548   47.874
##
## ---
## Tests of directed separation:
##
##                 Independ.Claim Test.Type DF Crit.Value P.Value
##   Carex.bigelowii ~ snow + ...      coef 89     1.0875  0.2797
##
## Global goodness-of-fit:
##
##   Fisher's C = 2.548 with P-value = 0.28 and on 2 degrees of freedom
##
## ---
## Coefficients:
##
##          Response    Predictor Estimate Std.Error DF Crit.Value P.Value
##        soil_moist         snow   0.4434     0.094 91     4.7188  0.0000
##      min_T_winter         snow   0.7928    0.0639 91    12.4082  0.0000
##   Carex.bigelowii min_T_winter   0.3389     0.111 90     3.0546  0.0030
##   Carex.bigelowii   soil_moist   0.0399     0.111 90     0.3592  0.7203
##    ~~min_T_winter ~~soil_moist   0.2029         - 93     1.9656  0.0262
##   Std.Estimate
##         0.4434 ***
##         0.7928 ***
##         0.3389  **
##         0.0399
##         0.2029   *
##
##   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
##
## ---
## Individual R-squared:
##
##          Response method R.squared
##        soil_moist   none      0.20
##      min_T_winter   none      0.63
##   Carex.bigelowii   none      0.13
```

The AIC values reported by the model summary is not simply the total AIC of the model, but is related to the difference in likelihood associated with the d-separation tests. It can be used to compare alternative models as long as these are nested, yet currently does not work for comparing saturated models (like our first candidate model) to a model with some missing paths (like our second candidate model). As an example, we formulate a third alternative model with different paths missing and compare it to our model 2 using AIC. Specifically, we omit the arrow from snow cover to soil moisture.

```
m3 = psem(soil_moist~1,
          lm(min_T_winter~snow, data=plants),
          lm(Carex.bigelowii~min_T_winter, data=plants),
          min_T_winter %~~% soil_moist,
          data=plants)
summary(m3)
```

```
##   |                                                                  |
```

```
##
## Structural Equation Model of m3
##
## Call:
##   soil_moist ~ 1
##   min_T_winter ~ snow
##   Carex.bigelowii ~ min_T_winter
##   min_T_winter ~~ soil_moist
##
##     AIC      BIC
##  15.351   30.547
##
## ---
## Tests of directed separation:
##
##                        Independ.Claim Test.Type DF Crit.Value P.Value
##   Carex.bigelowii ~ soil_moist + ...      coef 90     0.3592  0.7203
##         Carex.bigelowii ~ snow + ...      coef 90     1.1337  0.2600
##
## Global goodness-of-fit:
##
##   Fisher's C = 3.351 with P-value = 0.501 and on 4 degrees of freedom
##
## ---
## Coefficients:
##
##          Response    Predictor Estimate Std.Error DF Crit.Value P.Value
##      min_T_winter         snow   0.7928    0.0639 91    12.4082  0.0000
##   Carex.bigelowii min_T_winter   0.3573    0.0979 91     3.6497  0.0004
##    ~~min_T_winter ~~soil_moist   0.1239         - 96     1.2037  0.1159
##   Std.Estimate
##         0.7928 ***
##         0.3573 ***
##         0.1239
##
##   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
##
## ---
## Individual R-squared:
##
##          Response method R.squared
##      min_T_winter   none      0.63
##   Carex.bigelowii   none      0.13
```

```
AIC(m2b, m3)
```

```
##   df    AIC
## x 10 22.548
## y  6 15.351
```

The `piecewiceSEM` package allows the component models to be for example GLM´s or mixed models, and is thus very flexible.

EXERCISE: Repeat the analyses above for *Thalictrum alpinum* instead of *Carex*.

EXERCISE: Can you think of other potential models that can be tested? Are there other important environmental variables? Start by drawing the competing models as directed graphs (on paper). Fit and compare the models, and interpret the results.