Processing and Analysis of Biological Data

Path analysis and causal inference

Øystein H. Opedal

8 Dec 2022

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 comfirming 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). Broaly speaking, the method can be used to infer causality by combining knowledge about the natural history/mechanics of the study species with estimated statistical parameters such as correlation coefficients and regression slopes.

For more in-depth 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 elated methods required 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 -> Carex.bigelowii

min_T_winter -> Carex.bigelowii

soil_moist -> Carex.bigelowii

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

snow -> min_T_winter

snow -> soil_moist

min_T_winter -> Carex.bigelowii

In path analysis, we call the response variables (wtith 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 NAs 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
```

This 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 the path diagram corresponding to this model, and add the estimated path coefficients. 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).

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 hypothesised 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
```

```
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 first candidate model.

```r
library(lavaan)
```

```
## Warning: package 'lavaan' was built under R version 4.2.2

## This is lavaan 0.6-12
## lavaan is FREE software! Please report any bugs.
```

```r
library(semPlot)
```

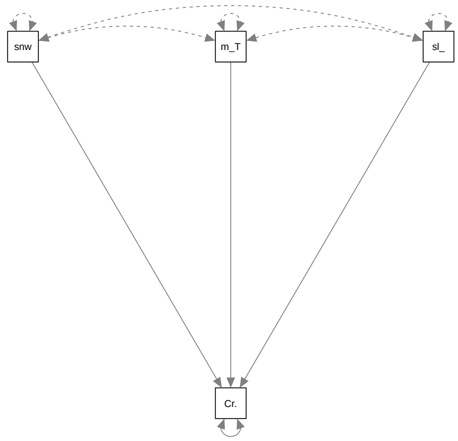
```
## Warning: package 'semPlot' was built under R version 4.2.2
```

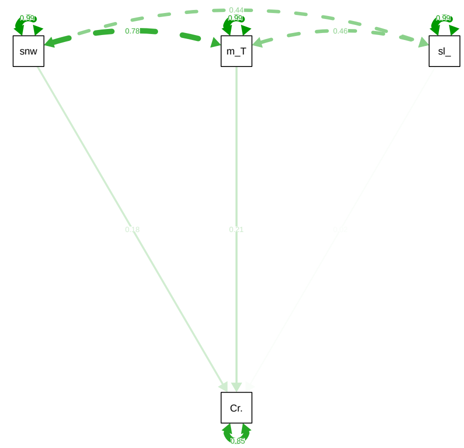
```r
mod = '
  Carex.bigelowii ~ snow + min_T_winter + soil_moist
'

lmod = sem(mod, data=plants)
summary(lmod)
```

```
## lavaan 0.6-12 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         4
##
##   Number of observations                            93
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                     Estimate  Std.Err  z-value  P(>|z|)
##   Carex.bigelowii ~
##     snow               0.177    0.159    1.112    0.266
##     min_T_winter       0.206    0.161    1.282    0.200
##     soil_moist         0.023    0.110    0.206    0.837
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Carex.bigelowi    0.850    0.125    6.819    0.000
```

```r
par(mfrow=c(1,2))
semPaths(lmod, what="diagram")
semPaths(lmod, what="est")
```





Similarly we can fit our alternative model.

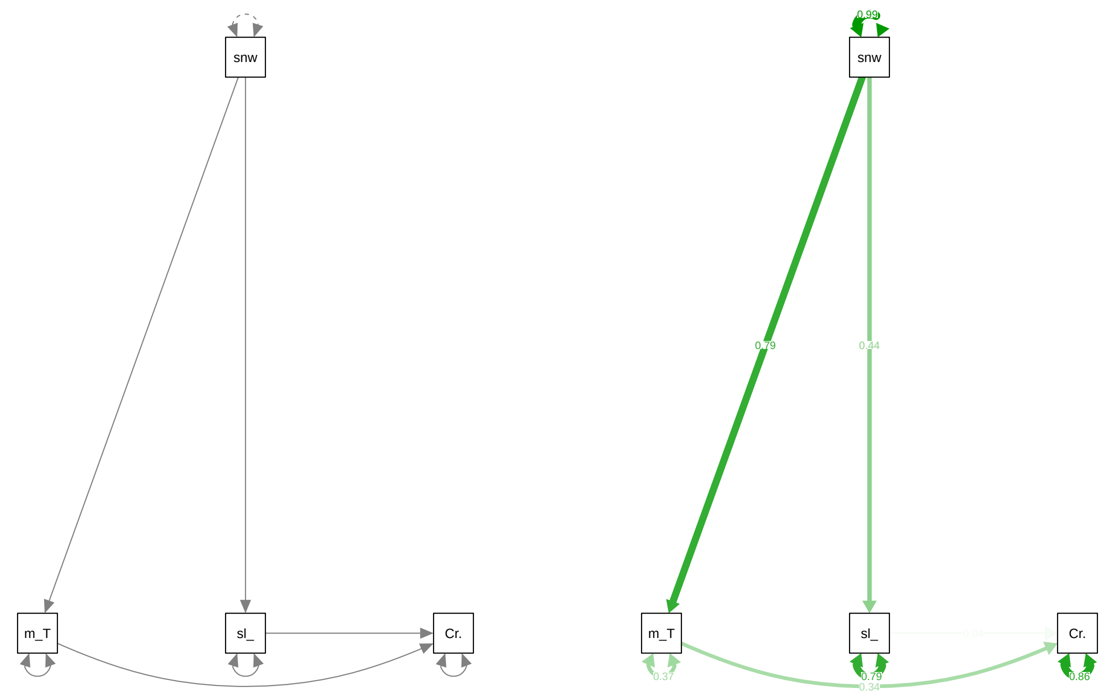
```
mod2 = '
  min_T_winter ~ snow
  soil_moist ~ snow
  Carex.bigelowii ~ min_T_winter + soil_moist
'

lmod2 = sem(mod2, data=plants)
summary(lmod2, fit.measures=F)
```

```
## lavaan 0.6-12 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         7
##
##   Number of observations                            93
##
## Model Test User Model:
##
##   Test statistic                                 5.137
##   Degrees of freedom                                 2
##   P-value (Chi-square)                           0.077
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   min_T_winter ~
##     snow              0.793    0.063   12.544    0.000
##   soil_moist ~
```

```
##     snow                 0.443    0.093    4.770    0.000
##   Carex.bigelowii ~
##     min_T_winter         0.339    0.103    3.278    0.001
##     soil_moist           0.040    0.103    0.386    0.700
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .min_T_winter      0.367    0.054    6.819    0.000
##    .soil_moist        0.795    0.117    6.819    0.000
##    .Carex.bigelowi    0.862    0.126    6.819    0.000
```

```
par(mfrow=c(1,2))
semPaths(lmod2, what="diagram")
semPaths(lmod2, what="est")
```



Note that for a SEM containing multiple components, the `lavaan` package provides a hypothesis test for the entire model. The interpretation of these 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.

Finally, there are several further extensions of structural equation modelling, allowing e.g the inclusion of unmeasured latent variables, and the use of more flexible link functions (through GLMs). One package for fitting such flexible models is `piecewiseSEM`

```r
library(piecewiseSEM)
```

```
## Warning: package 'piecewiseSEM' was built under R version 4.2.2

##
##   This is piecewiseSEM version 2.1.0.
##
##
##   Questions or bugs can be addressed to <LefcheckJ@si.edu>.
```

```r
model=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)

summary(model)
```

```
##   |                                                                              |

##
## Structural Equation Model of model
##
## 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(model)
```

The `piecewiseSEM` implements tests of so-called *directed separation*, which is a test for conditional non-independence of variables. Here, we test e.g. if *Carex* abundance is really conditionally independent of snow cover, i.e. after we have accounted for winter temperature and soil moisture. We also get an overall test for the model, which is the same as we got with our `lavaan` SEM model.