

On the usage of the LongitudinalEvaluate package

Andreas Kristian Pedersen, Afsaneh M. Nejad, Anna Mejldal
Kristian Debrabant and Sören Möller.

Contents

1	Introduction	1
2	Example - The regional lockdown	1
2.1	Time invariant confounder	3
2.2	Time varying confounder	3
2.3	Comparison of the confounder trajectories	9

1 Introduction

This vignette shows how to utilize the three methods presented in this package. The first method assesses whether a time-invariant unmeasured confounder could explain the observed causal association. In particular, it constructs sensitivity curves showing the strength of the exposure-confounder and outcome-confounder association to explain the association. The second and third methods create continuous trajectories of a confounder process U_t capable of explaining the discrepancy between observed and true causal association between a binary exposure X_t and outcome Y_t .

2 Example - The regional lockdown

To illustrate the methods implemented in the LongitudinalEvaluate package we investigate whether the effect of the regional lockdown in Northern Jutland had on SARS COVID-19 incidence can be explained by unmeasured confounders.

During the Covid-19 pandemic and development of the vaccine, a variant of SARS Covid-19 was found on mink farms in Northern Jutland. To prevent spillover infections, a lockdown was imposed on six out of eleven municipalities in Northern Jutland. During the same period, however, mink was culled at the farms to decrease the risk of getting infected by SARS Covid-19, particularly in the lockdown area. The culling of the mink materialized faster in the lockdown area compared to the others implying this risk is an unmeasured time-varying confounder capable of explaining our proposed causal association. It is therefore possible to use both *SDEconfounderNonparametric* and *SDEconfounderYuima1d* to simulate the trajectory of a confounder capable of explaining the observed association.

Another potential confounder is population density as the non-lockdown areas, consisting of urban areas, have a much higher density than the primarily rural lockdown areas. However during the 1 month period of the lockdown this confounder remained stable, but still capable of explaining the observed effect. Therefore the shiny app at <https://andreaskpedersen.shinyapps.io/LongitudinalEvaluate/> can be used to assess whether the time-invariant confounder is capable of explaining the observed effect. A depiction of the temporal average treatment effect between the exposure and outcome over time can be seen in Figure 1.

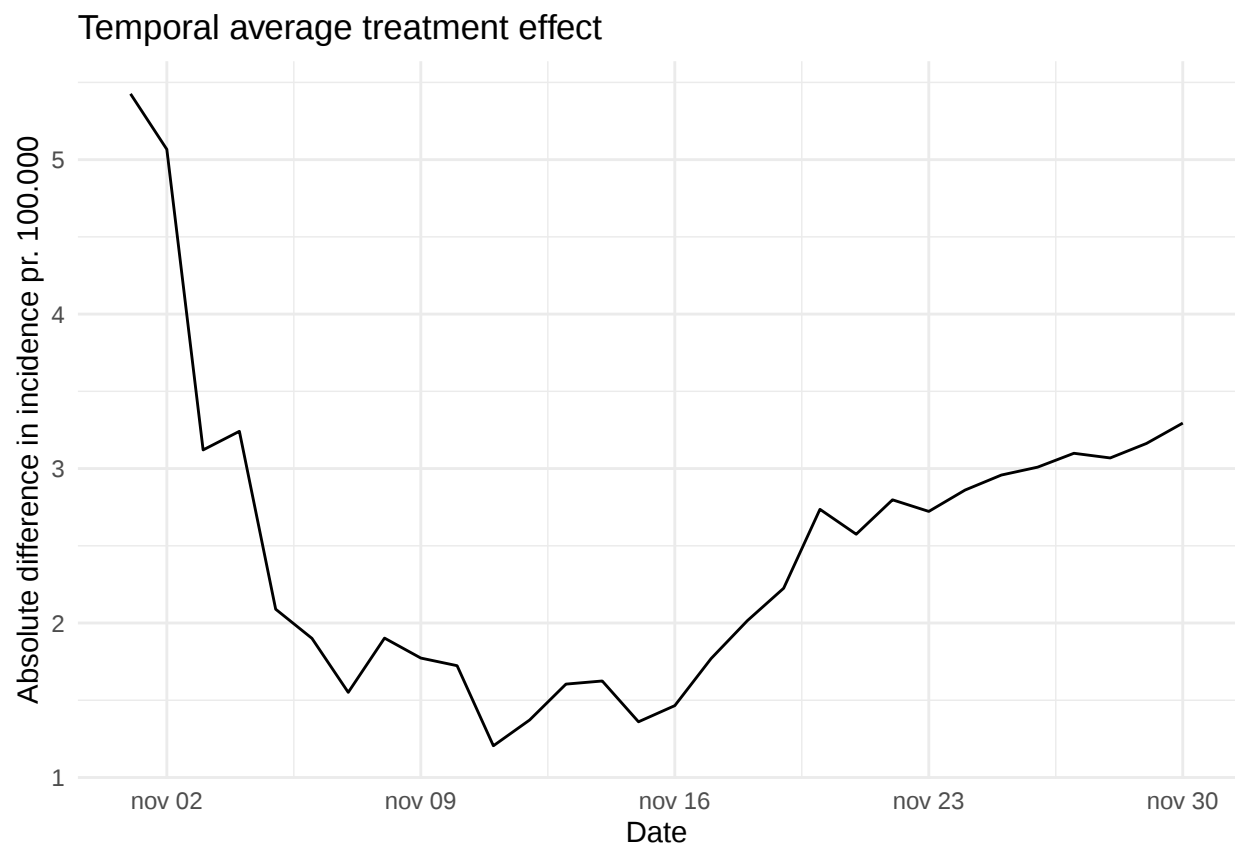


Figure 1: The temporal average treatment effect (TATE) over time between the exposure regional lockdown and outcome incidence of SARS Covid-19

2.1 Time invariant confounder

The Shiny application requires only the observed values of the temporal average treatment effect (TATE), which should be entered in the *Effect Sizes* field as a comma-separated sequence. To determine the order of the TATE process, change the plot type to *Difference-in-Difference*. This produces the plot shown in Figure 2.

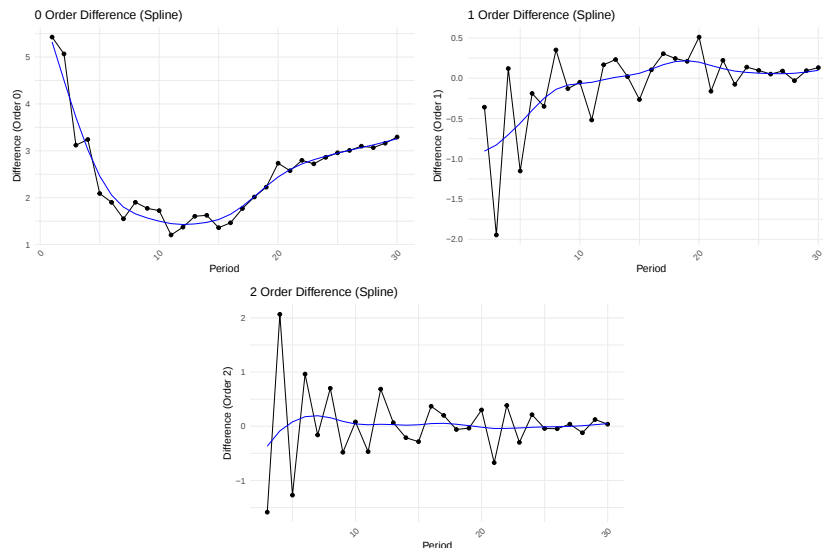


Figure 2: Difference in difference plots, stating the differencing order such that the TATE process becomes stationary. From the plots stationarity is achieved when the difference in difference operator is used twice, as the spline (blue line) reflects no temporal dependence, when the order is 2

From the difference-in-difference plot shown in Figure 2, it can be seen that the TATE process has order 2, as the fitted spline exhibits no systematic fluctuations over time. Consequently, the value entered in the *Order* field of the Shiny application should be set to 2. This produces the sensitivity curves shown in Figure 3.

The sensitivity curves can be used to assess the strength of association a confounding is required for explaining the observed effect. More precisely, for a given confounder–exposure association, the curves show the corresponding confounder–outcome association needed to explain away the estimated effect.

For example, suppose that population density is five times higher in the lockdown area than in the non-lockdown area. In this case, the sensitivity curve indicates that the confounder–outcome association at time 27 must be approximately 30 on the temporal average treatment effect scale. This confounder–outcome association corresponds to an average increase of approximately 30 SARS-CoV-2 cases per 100,000 individuals per day up to 27 November when the confounder increases with one unit-level.

2.2 Time varying confounder

2.2.1 Parametric approach and constant association

The parametric approach *SDEconfounderYuima1d* requires the following arguments

- *fit*: The fit object from the *qmle*-function from the *yuima* package
- *EFXU*: The association between the confounder and exposure expressed as a scalar.
- *EFUY*: The association between the confounder and outcome expressed as a scalar.
- *time*: The time scale of the confounder trajectory.
- *initial.value*: The initial value of the confounder.

Therefore when using the parametric approach, the first step is to fit a stochastic differential equation to the observed TATE process. This can be accomplished by the *yuima*-package and the *qmle* function. Note that

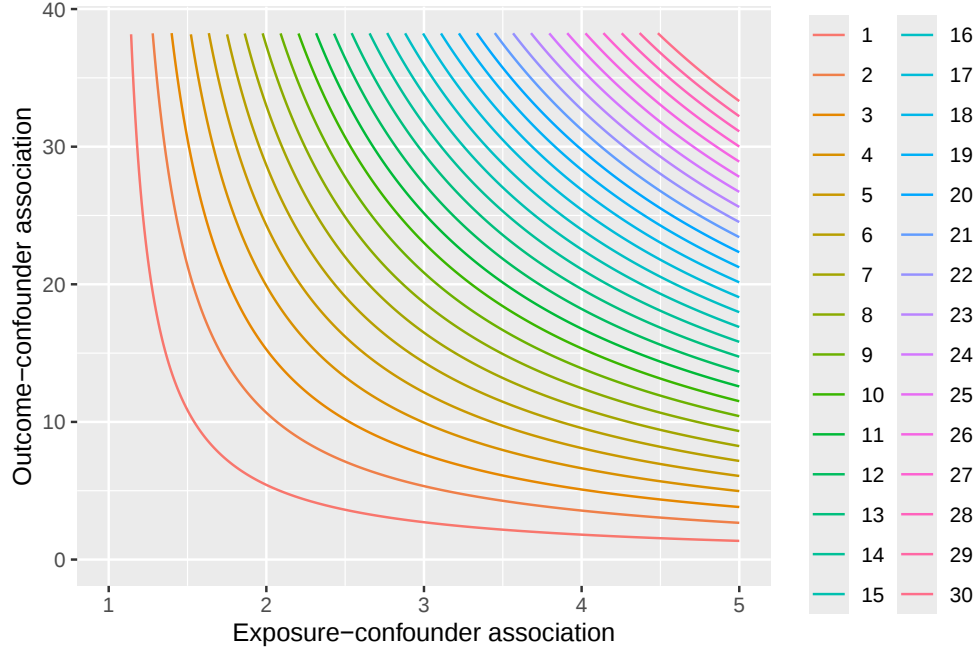


Figure 3: The confounder plot, stating the hypothetical association population density needs to have with the regional lockdown and incidence to explain the TATE between the regional lock down and incidence.

the *qmle* assumes that the drift and diffusion function is only dependent on the state variable. An example of how to fit stochastic differential equation using the *qmle* function can be seen below

```
Delta<-1/3
Covidexample$t<-1:30
mod<-setModel(drift="a/TATE+c*TATE", diffusion="sigma", state.var="TATE", time.var = "t",
  solve.var = "TATE",xinit = 5.425966)
model<-setYuima(model=mod, data=setData(zoo(Covidexample$TATE, order.by = Covidexample$t),
  delta=Delta))
fit <- qmle(model,start = list(a=-5,c = 5,sigma = 0.2),
  lower = list(a=-10,c = -10,sigma = 0),
  upper = list(a=0,c = 10,sigma =1))
```

The next step is to use the fitted object returned by *qmle* together with hypothetical values for the confounder–exposure and confounder–outcome associations. In this example, we assume that the confounder–exposure association corresponds to a 20% higher risk of getting infected through mink in the non-lockdown group compared to the lockdown group, corresponding to $RR_{U_t X_t} = 1.2$ based on a ratio scale. Furthermore, we assume that a one-percentage-point increase in infection risk yield an absolute increase in the incidence by 1.5 pr. 100000, corresponding to $EF_{U_t Y_t} = 1.5$ as it is on an absolute scale.

Finally, we want to evaluate whether the confounder can explain the observed effect, hence the fit object is based directly on the TATE-process. Therefore the confounder trajectory can be simulated as followed:

```
U_t<-SDEconfounderYuima1d(fit, EFXU=1.2, EFUY=1.5,time=Covidexample$t,
  initial.value = 5.425966/(1.2*1.5))
```

The object `U_t` contains a dataframe with four variables: The confounder value, time, and the upper and lower limit of the confounders values from the simulation of the confounder. These values can be used to create plots as seen in figure 4, which the following code does

```
plot(U_t)
```

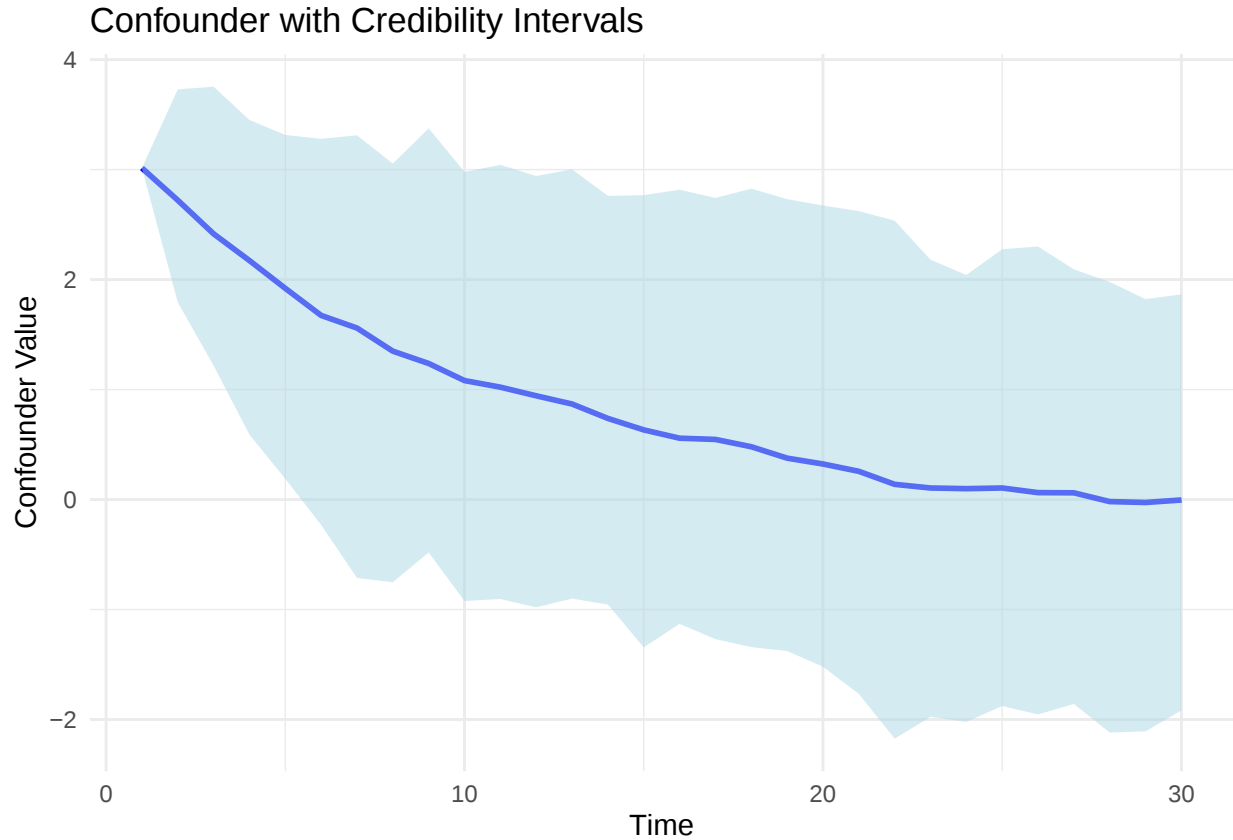


Figure 4: The trajectory based on `SDEconfounderYuima1d` of an unmeasured confounder capable of explaining the association between regional lockdown and incidence. The association between the exposure and confounder was set to 1.2 on the risk ratio scale and the association between the outcome and confounder was set to 1.5 on the absolute scale.

The plot in figure 4 is based on the `ggplot` function and suggest that a confounder with the given hypothetical parameters would have to decline exponentially over time and close in on zero at the end of the study period in order to explain our observed association between the regional lockdown and incidence. The trajectory of the confounder depicted in 4 aligns with the presumed unmeasured confounder, risk of getting infected with SARS-Covid 19 due to the mink. This suggests the unmeasured confounder can potentially explain the given association between the regional lockdown and the incidence, under the assumption of the *qmle*-function and `SDEconfounderYuima1d`.

2.2.2 Nonparametric approach and constant association

In the nonparametric approach the inputs are similar to those of the parametric approach, except the observed temporal average treatment effect process is supplied directly instead of a fitted object of the *qmle* function. The drift and diffusion function is estimated by way of kernel regression. The predictions from these kernel regressions are then used together with the numerical schemes for stochastic differential equations. The schemes are employed to incorporate variability into our function to acquire simulation-based uncertainty intervals or credibility intervals, if priors are specified. The priors are based on the two association linked with the confounder. The following code generates the plot in figure 5 as in the previous section using the Euler-Maruyama scheme.

```
U_t<-SDEconfounderNonparametric(Covidexample$TATE, ATE=Covidexample$ATE,Covidexample$t,
                                EFXU=1.2, EFUY=1.5, scheme = "Euler-Maruyama")
plot(U_t)
```

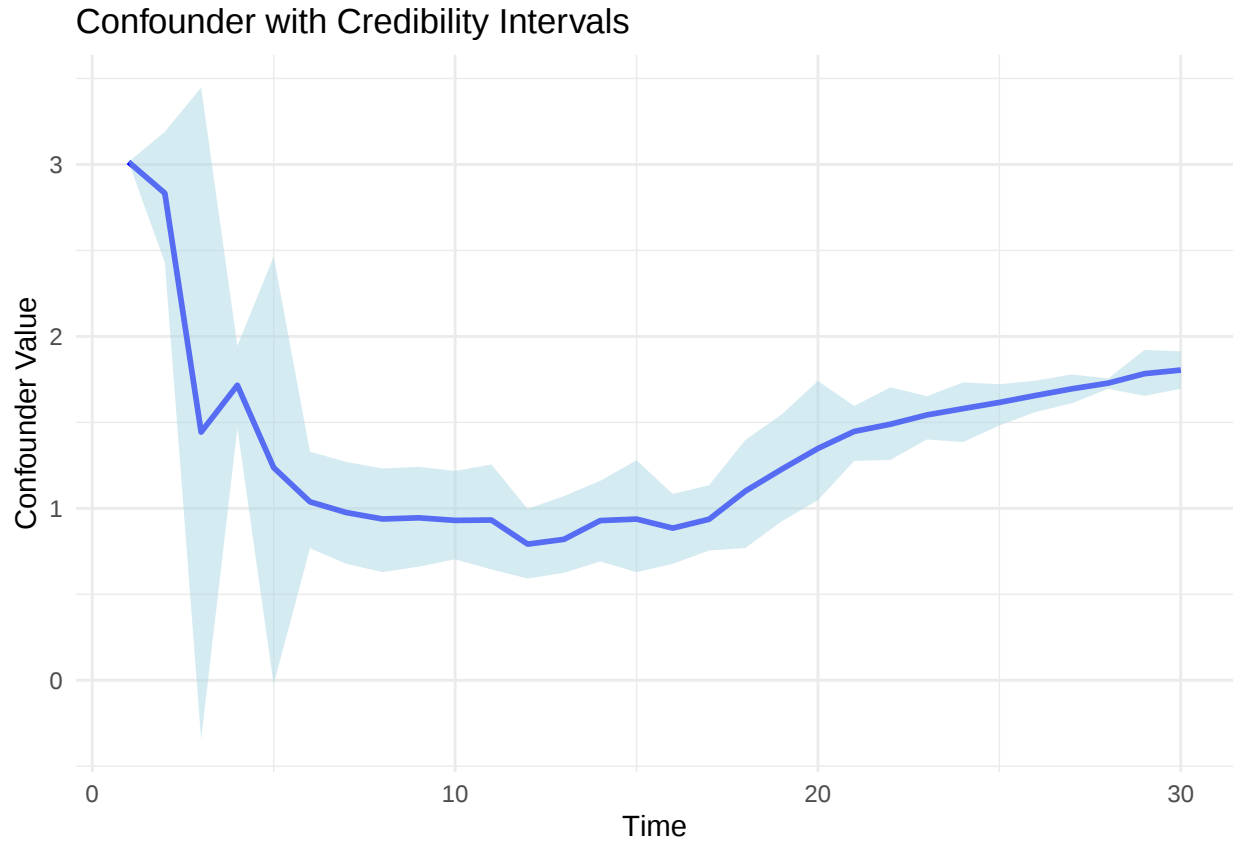


Figure 5: The trajectory based on `SDEconfounderNonparametric` of an unmeasured confounder capable of explaining the association between regional lockdown and incidence. The association between the exposure and confounder was set to 1.2 on the risk ratio scale and the association between the outcome and confounder was set to 1.5 on the absolute scale.

The plot is again based on `ggplot2` package and suggests that a confounder with the given hypothetical parameters shall decline exponentially until the 16 of November and then increase slightly from the 16 and onward to explain away our proposed causal association. This does not realign with the unmeasured time-varying confounder, risk of getting infected of SARS Covid-19 through mink, as the culling would make this confounder decreases monotonically.

2.2.3 Time varying associations

Using the non-parametric method one can simulate a time-varying confounder with time-varying associations to the exposure and outcome. First one tries to simulate the observed process using the *SDEconfounderNonparametric*. This can be done by setting *ATE* equal to zero and the two associations related directly to the confounder *EFXU* and *EFUY* equal to 1. This code can be seen below

```
U_t<-SDEconfounderNonparametric(Covidexample$TATE, ATE=Covidexample$ATE,Covidexample$time,
                                EFXU=1, EFUY=1, scheme = "Euler-Maruyama")
```

U_t contains a dataframe called `trajectories` with each simulated trajectory of the observed TATE. Thus, one can update each of these trajectories with time varying associations and use the `MatrixStats` package together

with ggplot to get a similar figure as in the previous two sections. Under an SIR-type model, one possible specification of the time-varying associations are

$$EF_{U_{t,1}Y_t} \approx \frac{1}{t}(U_{t,1} + \alpha t)e^{U_{t,1} + (\alpha - \gamma)t},$$

$$f'(U_{t,0}) = \beta U_{t,0}^{\beta-1}$$

where the first equation relates to the association between the confounder and outcome $EF_{U_{t,1}Y_t}$ and the second equation relates to the association between the confounder and exposure $f'(U_{t,0})$. Therefore one can simulate the unmeasured confounder recursively using the following code giving figure 6.

```

beta<-1.2
alpha<-0.3
gamma<-0.5

U_mat <- matrix(NA_real_, nrow = 30, ncol = 100)

for (k in 2:100){

  E <- U_t@trajectories[, k]

  U <- numeric(30)

  U[1] <- E[1] * 0.27

  for (t in 1:29) {
    denom <- (beta * U[t]^(beta - 1)) *
      ((1/t) * (U[t]^beta + alpha*t) *
        exp(U[t]^beta + (alpha - gamma)*t))

    U[t + 1] <- (E[t + 1] - E[t]) / denom + U[t]
  }
  U_mat[, k] <- U

}

alpha_level<-0.05

U_mat<-t(as.matrix(U_mat))
U_mat<-U_mat[2:100,]
Mean <- colMeans2(U_mat, na.rm = TRUE)
CIll <- colQuantiles(U_mat, probs = alpha_level/2, na.rm = TRUE)
CIul <- colQuantiles(U_mat, probs = 1 - alpha_level/2, na.rm = TRUE)

Confounderdataset<-data.frame(
  Confounder = Mean,
  time = Covidexample$t,
  ConfounderCIll = CIll,
  ConfounderCIul = CIul

```

```

)

names(Confounderdataset)<-c("mean_Confounder","time", "CIll", "CIul")

ggplot(Confounderdataset, aes(x = time, y = mean_Confounder)) +
  geom_line(color = "blue", linewidth = 1) +
  geom_ribbon(
    aes(ymin = CIll, ymax = CIul),
    fill = "lightblue",
    alpha = 0.5
  ) +
  labs(
    x = "Time",
    y = "Confounder Value",
    title = "Confounder with Credibility Intervals"
  ) +
  theme_minimal()

```

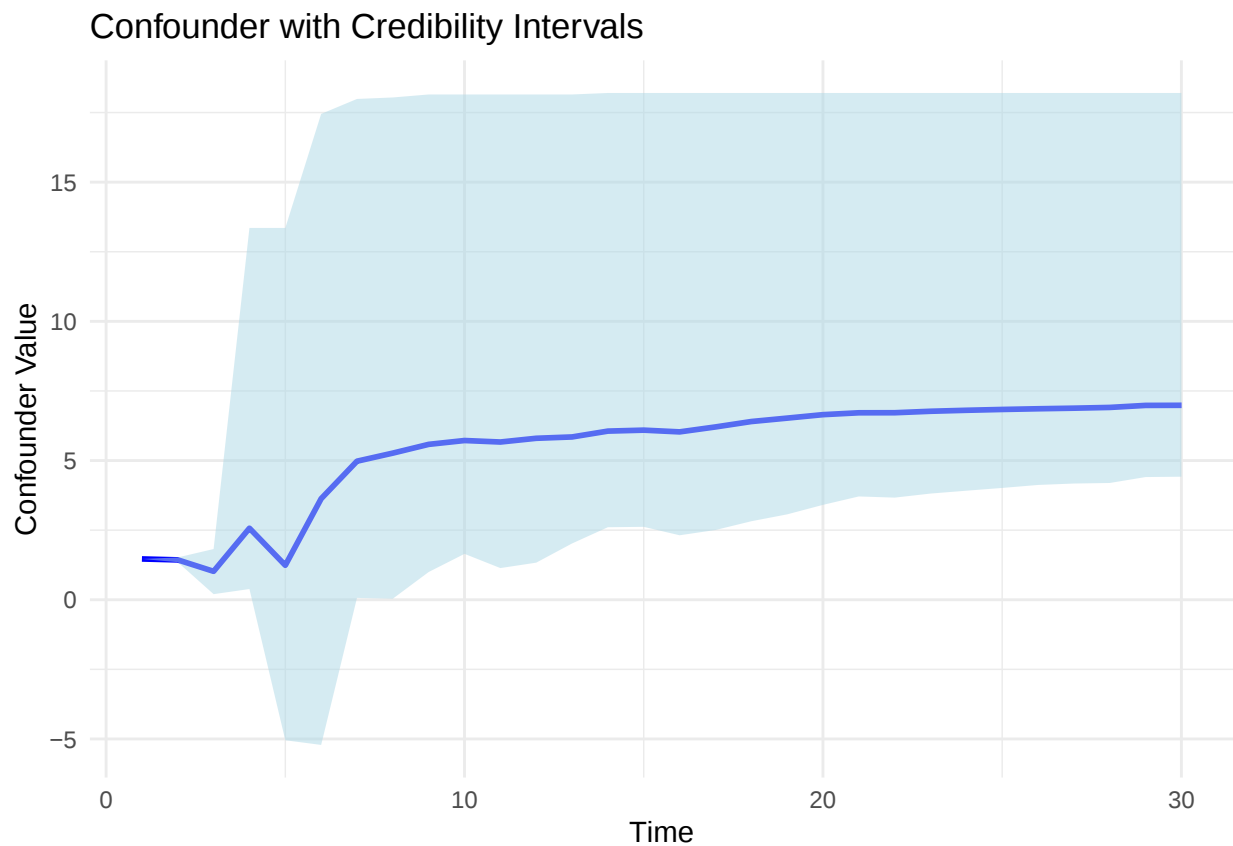


Figure 6: The trajectory based of an unmeasured confounder capable of explaining the association between regional lockdown and incidence, where the association from the confounder to the exposure and outcome is time-varying.

Figure 6 suggests that the risk of getting infected by mink should increase over the period in order to explain the observed association. Again this seems unlikely given the culling of the mink in the same period should

decrease the risk. Again we conclude that the confounder is not capable of explaining the association.

2.3 Comparison of the confounder trajectories

The differences between the three confounder trajectories can be attributed to how the observed data are used and the assumptions on the confounder–exposure and confounder–outcome associations.

The parametric approach requires fitting a stochastic differential equation to the observed TATE process. The estimates of the parameters from the fitted object are then combined with the hypothetical confounder-related parameters to construct the confounder trajectory with uncertainty intervals. However, the estimation is based on quasi-maximum likelihood, which assumes that both the drift and diffusion functions only depend on the state variable, restricting its use to these types of stochastic processes.

In contrast, the nonparametric approach does not impose this restriction, because of its use of kernel regression to estimate both the drift and diffusion functions that depends on the time component and state variable, simultaneously. These estimates are then used in numerical schemes to simulate the observed process. The resulting predictions are evaluated at the observed TATE values, since the sensitivity analysis is conditional on the data. Consequently, the nonparametric approach is more closely aligned with the observed data, particularly when the drift and diffusion functions are time-dependent.