Sherif_analysis

Yvette Baurne, Frederic Delmar, and Jonas Wallin

2025-09-02

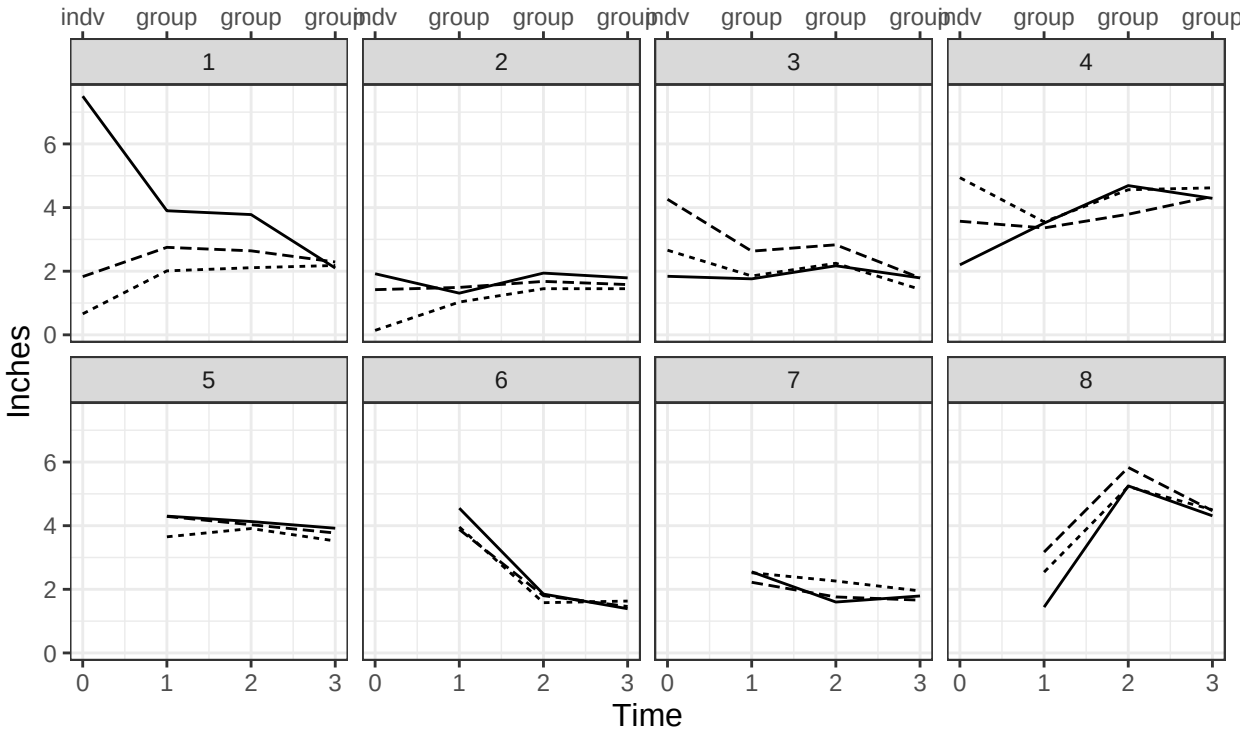
Visualsing the data

We begin by loading the Sherif dataset and plotting the raw measurements. At the first time point, each subject makes an individual estimate of the autokinetic effect. At the next two time points, they report in groups, and at the final time point they return to individual estimates. For this analysis, we drop that last individual measurement to focus on the group dynamics.

```
data("sherifdat")
sherifdat$time <- sherifdat$time + 1
sherifdat <- subset(sherifdat, time <= 3)
```

```
fig1 <- ggplot(data=sherifdat,
               aes(x=time,y=y,
               linetype = factor(person,
                         labels = c("Subject 1", "Subject 2", "Subject 3")))) +
  geom_line(linewidth=0.5) + xlab("Time") + ylab("Inches") +
  guides(linetype=guide_legend(title="Subjects within each group")) +
  scale_x_continuous(sec.axis = sec_axis(~.*1,
                              labels =  c("indv", "group", "group", "group")))

fig1 <- fig1 +  facet_wrap(~group, ncol = 4) +
  # labs(title = "Sherif (1935) autokinetic data") +
  theme_bw() +
  theme(legend.position="bottom") +
  theme(legend.text=element_text(size=12),
        legend.title=element_text(size=12),
        #axis.text.x = element_text(size = 11),
        #axis.text.y = element_text(size = 11),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))
print(fig1)
```



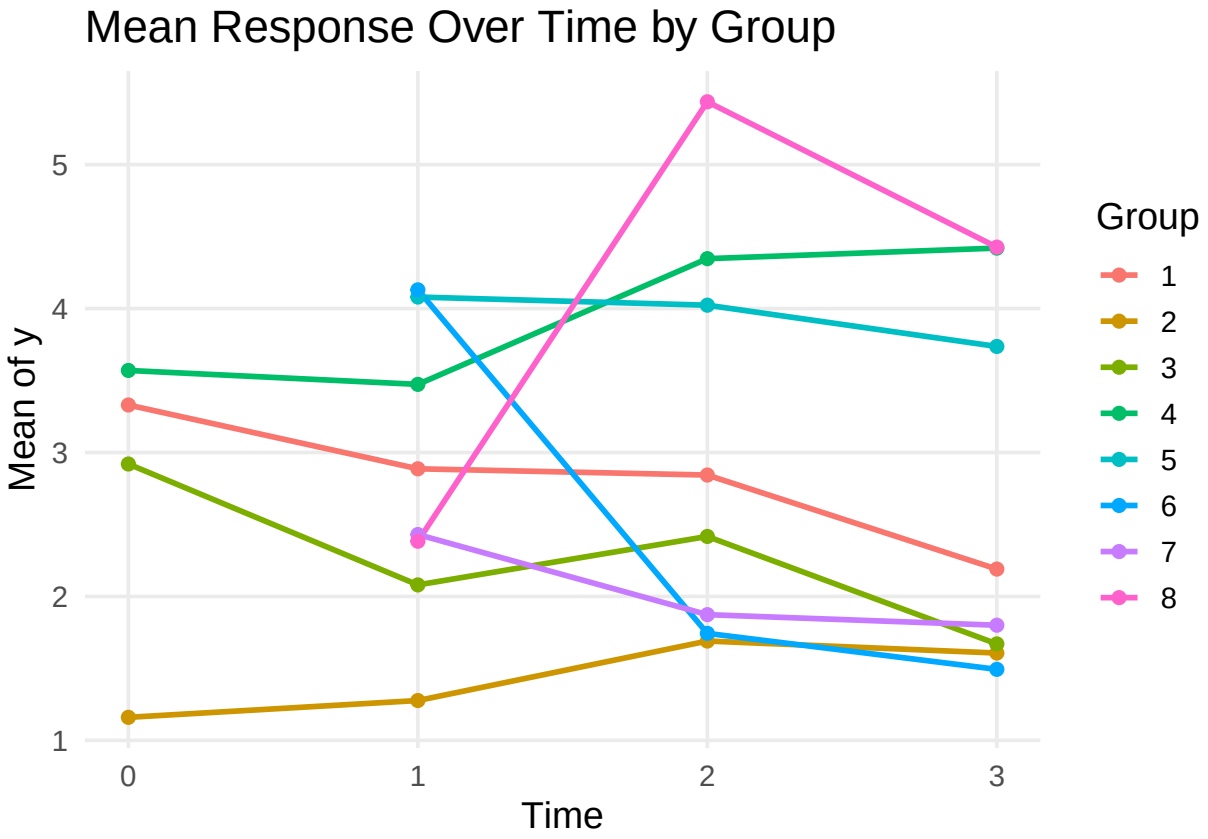
Subjects within each group — Subject 1 ···· Subject 2 - - - Subject 3

We next examine group-level dynamics by averaging each group's responses over time and plotting these means together. The trajectories hint at some autoregressive pattern. We also observe that between-group variability appears to grow over the measurement occasions, but with only eight groups this trend should be interpreted cautiously.

```
group_time_summary <- sherifdat %>%
  group_by(group, time) %>%
  summarise(
    y_mean       = mean(y),
    .groups = "drop"
  )

fig <- ggplot(group_time_summary, aes(x = time, y = y_mean,
                                      group = factor(group),
                                      color = factor(group))) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Mean Response Over Time by Group",
    x     = "Time",
    y     = "Mean of y",
    color = "Group"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "right",
    panel.grid.minor = element_blank()
  )
var_df <- group_time_summary %>%
```

```
  group_by(time) %>%
  summarise(var_y_mean = var(y_mean), .groups = "drop")
print(fig)
```




```
print(var_df)
```

```
## # A tibble: 4 x 2
##    time var_y_mean
##   <dbl>      <dbl>
## 1     0       1.19
## 2     1       1.00
## 3     2       1.96
## 4     3       1.68
```

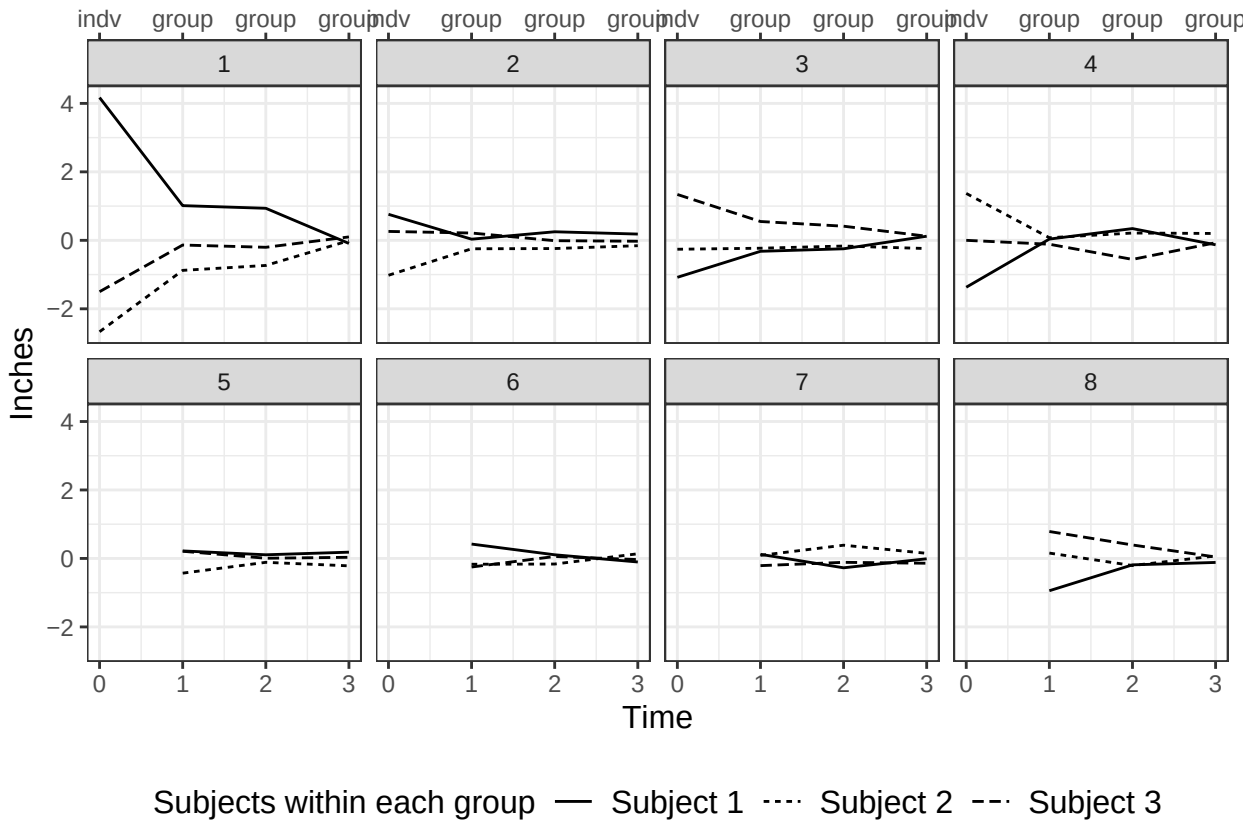
We also examine individual deviations after subtracting each group's mean. The resulting plots reveal a strong, non-stationary "consensus emergence," where subjects converge toward the group average over time. There is some suggestion of autoregressive dynamics in these deviations, but the dominant convergence effect makes it difficult to distinguish clearly.

```
sherifdat_cell_demeaned <- sherifdat %>%
  group_by(group, time) %>%
  mutate(

    y_ct_mean   = mean(y),              # cell mean
    y_demeaned  = y - y_ct_mean         # subtract it
  ) %>%
  ungroup()
```

```r
fig1 <- ggplot(data=sherifdat_cell_demeaned,
               aes(x=time,y=y_demeaned, linetype = factor(person,
                           labels = c("Subject 1", "Subject 2", "Subject 3")))) +
  geom_line(linewidth=0.5) + xlab("Time") + ylab("Inches") +
  guides(linetype=guide_legend(title="Subjects within each group")) +
  scale_x_continuous(sec.axis = sec_axis(~.*1,
                              labels = c("indv", "group", "group", "group" )))

fig1 <- fig1 +  facet_wrap(~group, ncol = 4) +
  # labs(title = "Sherif (1935) autokinetic data") +
  theme_bw() +
  theme(legend.position="bottom") +
  theme(legend.text=element_text(size=12),
        legend.title=element_text(size=12),
        #axis.text.x = element_text(size = 11),
        #axis.text.y = element_text(size = 11),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))
print(fig1)
```




```r
var_time_df <- sherifdat_cell_demeaned %>%
  group_by(time) %>%
  summarise(
    var_y = var(y_demeaned),
    .groups = "drop"
  )
print(var_time_df)
```

```
## # A tibble: 4 x 2
##    time  var_y
##   <dbl>  <dbl>
## 1     0 3.20
## 2     1 0.198
## 3     2 0.124
## 4     3 0.0174
```

We then fit a baseline model to these data.

```
GP.h <- ce(y ~ 1,
           ~ 1 | person,
           ~ 1 | group,
           emergence = ~ 1,
           method = "GP",
           method.team = "OU.homeostasis",
           time = "time",
           data = sherifdat)
```

And examine how the variability changes over time on group and individual level, we can see that the fitted model has similar dynamics as the raw data, i.e. the variance of group effects increases over time, while the variance of individual effects decreases over time.

```
Covs_inital.fit <- get.Cov(GP.h$covariances,GP.h$object)
print('Variance for team effect over time')
```

```
## [1] "Variance for team effect over time"
```

```
print(diag(Covs_inital.fit$SigmaT))
```

```
##         1         4         7        10
## 0.8631948 1.0271348 1.2222149 1.4543504
```

```
Covs_inital.fit <- get.Cov(GP.h$covariances,GP.h$object)
print('Variance for individual effect over time')
```

```
## [1] "Variance for individual effect over time"
```

```
print(diag(Covs_inital.fit$SigmaI))
```

```
##           1           4           7          10
## 2.976048809 0.353802018 0.042061373 0.005000701
```

We now also explore if a linear random effect model for individuals could be better.

```
#random effects for indivuals
CEI2.h <- ce(y ~ 1,
             ~ 1 | person,
             ~ 1 | group,
             emergence = ~ -1 + time,
             time = "time",
             method = "CEI2",
             method.team = "OU.homeostasis",
             data = sherifdat)


GP <- ce(y ~ 1+time,
         ~ 1 | person,
         ~ 1 + time | group,
```

```
          emergence = ~ 1,
          method = "GP",
          time = "time",
          data = sherifdat)
```

The random effect on the group level is much worse, compared to fitting Gaussian processes on two levels. The random effect on individual level is basically equivalent to the Gaussian process. While a random effect model for groups is clearly much worse.

```
cat('AIC_GP_GP = ', GP.h$AIC,'\n')
```

```
## AIC_GP_GP =  179.3234
```

```
cat('AIC_CEI_GP = ', CEI2.h$AIC,'\n')
```

```
## AIC_CEI_GP =  179.3599
```

```
cat('AIC_GP_L = ', GP$AIC,'\n')
```

```
## AIC_GP_L =  224.3147
```

We can now examine how the variance components evolve over time in our best-fitting model. Early on, individual-level variability dominates, but by the final measurement occasion the group-level variance takes over. Overall, the group effect starts small yet steadily increases, measurement error remains small, and individual variability declines. These patterns confirm a pronounced non-stationary process at both the group and individual levels.

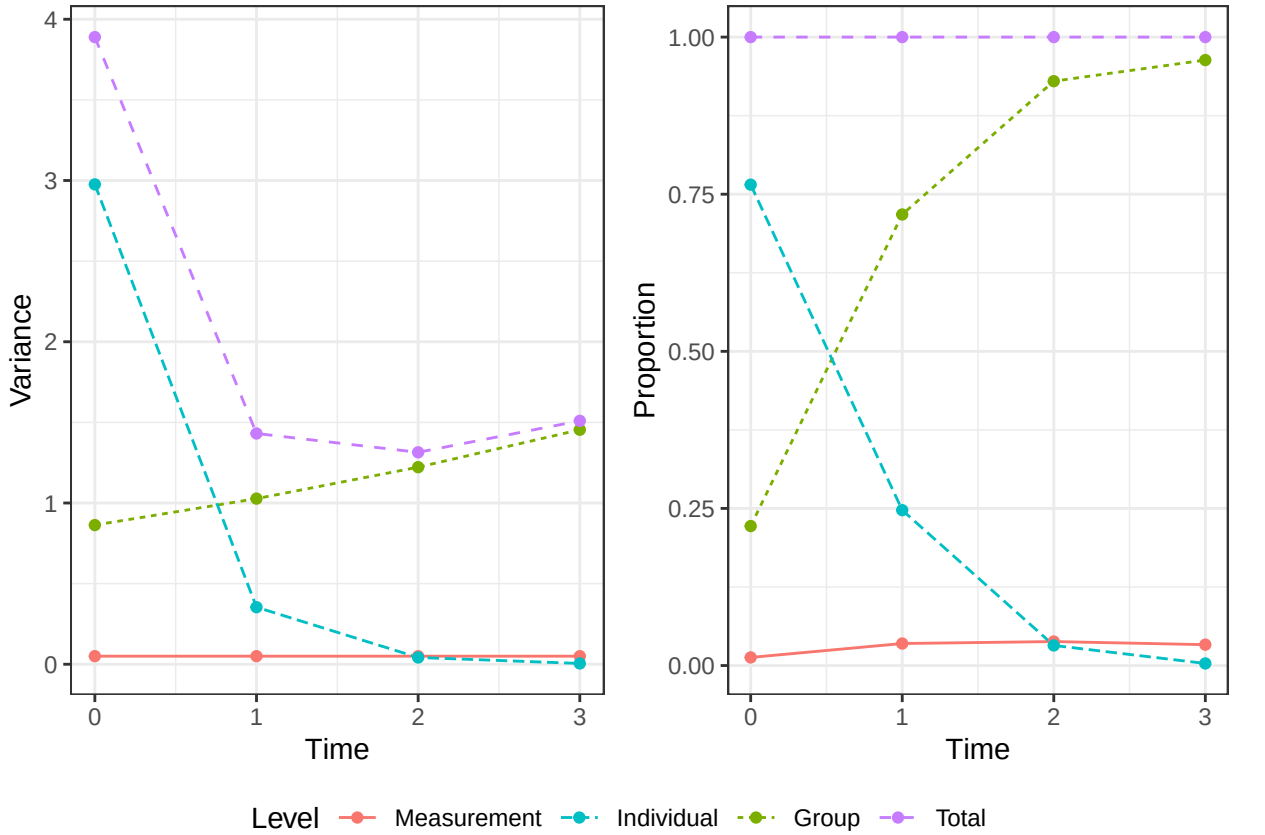
```
Covs <- get.Cov(GP.h$covariances,GP.h$object)

dat <- data.frame(t = 0:3,
                  VeY = diag(Covs$SigmaE),
                  VP = diag(Covs$SigmaI),
                  VG = diag(Covs$SigmaT))
dat$tot <- rowSums(dat[,2:4])
dat$PVeY <- dat$VeY/dat$tot
dat$PVP <- dat$VP/dat$tot
dat$PVG <- dat$VG/dat$tot
dat$PVY <- dat$tot/dat$tot
dat$VY <- dat$tot


pvar <- dat %>%
  pivot_longer(c(VeY,VP,VG,VY),names_to = "var",values_to = "value") %>%
  ggplot(aes(x=t,y=value,col=var,linetype=var)) +
  geom_line() +
  geom_point() +
  theme_bw()+
  ylab("Variance") +
  xlab("Time") +
  scale_color_discrete(name="Level",labels =
    c("VeY" = "Measurement","VP" = "Individual","VG" = "Group","VY" = "Total"),
    breaks=c("VeY","VP","VG","VY"))+
  scale_linetype_discrete(name="Level",labels =
    c("VeY" = "Measurement","VP" = "Individual","VG" = "Group","VY" = "Total"),
              breaks=c("VeY","VP","VG","VY"))+
  theme(legend.position = "none")
```

```
pperc <- dat %>%
  pivot_longer(c(PVeY,PVP,PVG,PVY),names_to = "var",values_to = "value") %>%
  ggplot(aes(x=t,y=value,col=var,linetype=var)) +
  geom_line() +
  geom_point() +
  theme_bw()+
  ylab("Proportion") +
  xlab("Time") +
  scale_color_discrete(name="Level",
  labels = c("PVeY" = "Measurement","PVP" = "Individual",
             "PVG" = "Group","PVY" = "Total"),
  breaks=c("PVeY","PVP","PVG","PVY"))+
  scale_linetype_discrete(name="Level",
labels = c("PVeY" = "Measurement","PVP" = "Individual",
           "PVG" = "Group","PVY" = "Total"),
breaks=c("PVeY","PVP","PVG","PVY"))+
  theme(legend.position = "right")

Figure4 <- ggpubr::ggarrange(pvar, pperc,
                      #labels = c("A", "B"),
                      ncol = 2, nrow = 1,
                      common.legend = T,
                      legend="bottom")
print(Figure4)
```



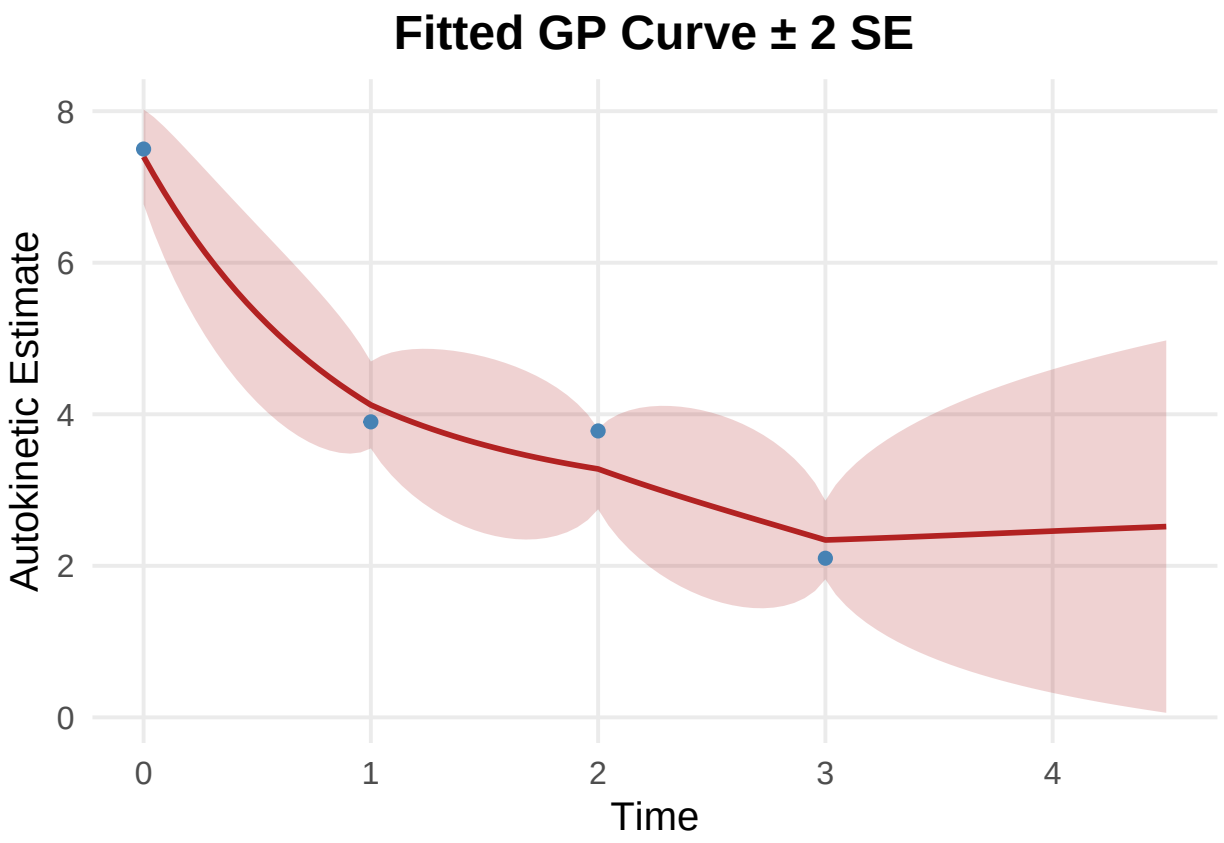
Finally we also show how a prediction of a individual future response also what was the likely response between observations. We can see that the uncertainty increases as at the end this due to the increasing

variability of the group effect. It should be noted that the exact form of this variability is rather uncertain. We can also see that the 95% prediction intervals goes down next two the observations. This is since we have rather small measurement error we are rather certain of the value of the processes at the measurement locations.

```
n.time = 100
sherifdat.new <- data.frame(person = c(rep(1,n.time)),
                            group  = c(rep(1,n.time)),
                            time   = c(seq(0,4.5,length.out = n.time) ))
pred.data.new   <- predict.ce(GP.h, sherifdat.new)
# 1) Build a data frame of predictions + CIs
pred_df <- tibble(
  time = sherifdat.new$time,
  fit  = pred.data.new[,1],
  se   = sqrt(pred.data.new[,3])
) %>%
  mutate(
    lower = fit - 2*se,
    upper = fit + 2*se
  )

# 2) Extract the observed points for person 1, group 1
obs_df <- sherifdat %>%
  filter(person == 1, group == 1)

# 3) Plot with ribbon + line + points
fig <- ggplot(pred_df, aes(x = time)) +
  geom_ribbon(aes(ymin = lower, ymax = upper),
              fill = "firebrick", alpha = 0.2) +
  geom_line(aes(y = fit), color = "firebrick", linewidth = 1) +
  geom_point(data = obs_df, aes(x = time, y = y),
             color = "steelblue", size = 2) +
  labs(
    title = "Fitted GP Curve ± 2 SE",
    x     = "Time",
    y     = "Autokinetic Estimate"
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title       = element_text(face = "bold", hjust = 0.5),
    panel.grid.minor = element_blank()
  )
print(fig)
```


Fitted GP Curve ± 2 SE
Autokinetic Estimate
Time
0
2
4
6
8
0
1
2
3
4