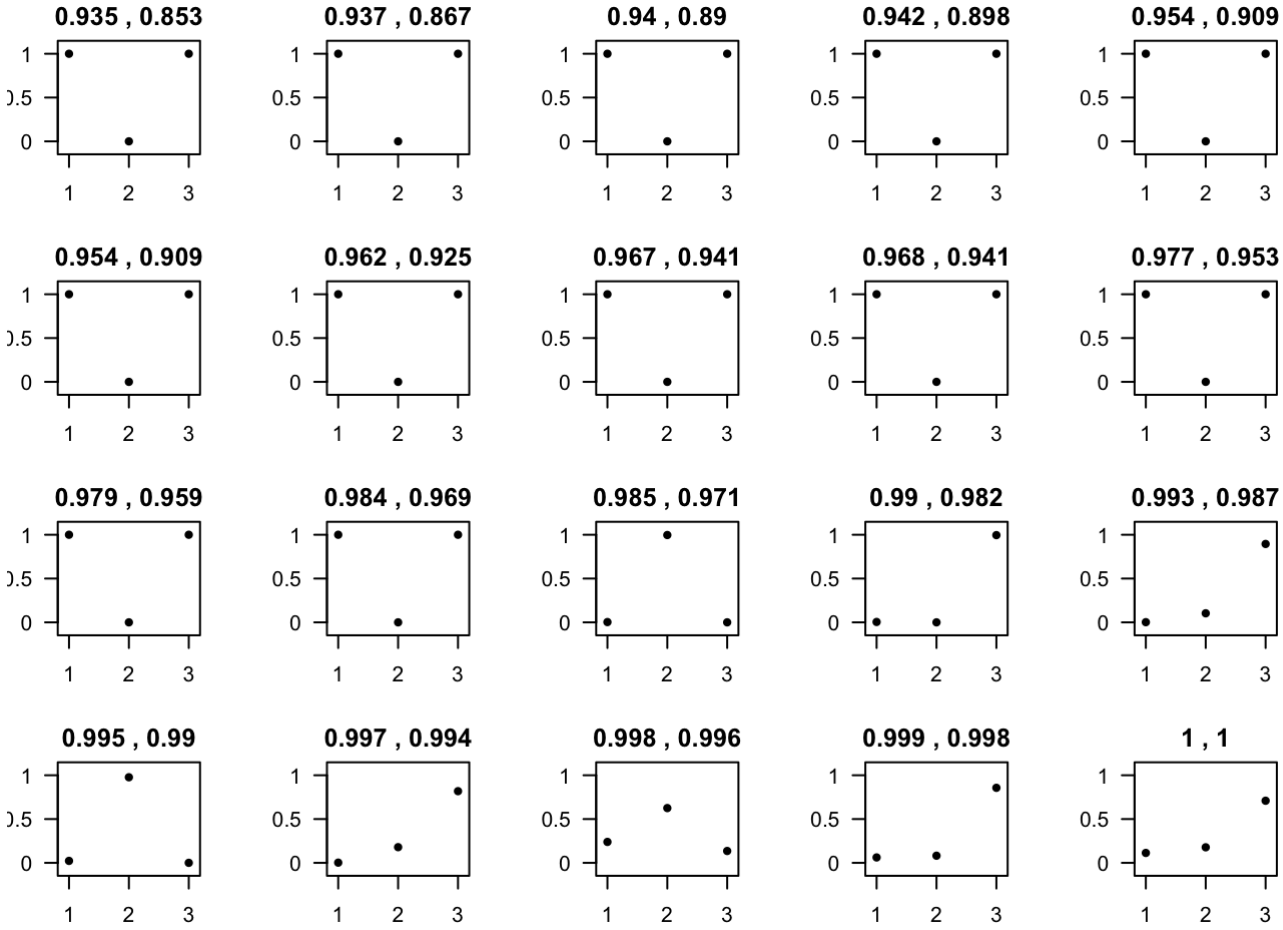
SuSiE Correlation Experiment

Stuart Brabbs

The basic model of this experiment is a situation where some variable Y depends on variables X1 and X3, which are each highly correlated with a third variable X2, but only correlated with each other as an artifact of their correlation with X2. The true model is thus $Y = \beta_1 X_1 + \beta_3 X_3$. However, competing models involving X2 are $Y = \beta_2 X_2$, $Y = \beta_1 X_1 + \beta_2 X_2$, and $Y = \beta_2 X_2 + \beta_3 X_3$. For the sake of easing the simulation, we assume that X2 has a causal effect on X1 and X3, but there is no causal effect between X1 and X3, i.e. X2 is a confounder. We simulate X2 as a random sample of 1000 from $N(0, 1)$, and then simulate both X1 and X3 as X2 plus a random value from $N(0, \sigma^2)$, where the standard deviation σ is varied to adjust correlation.

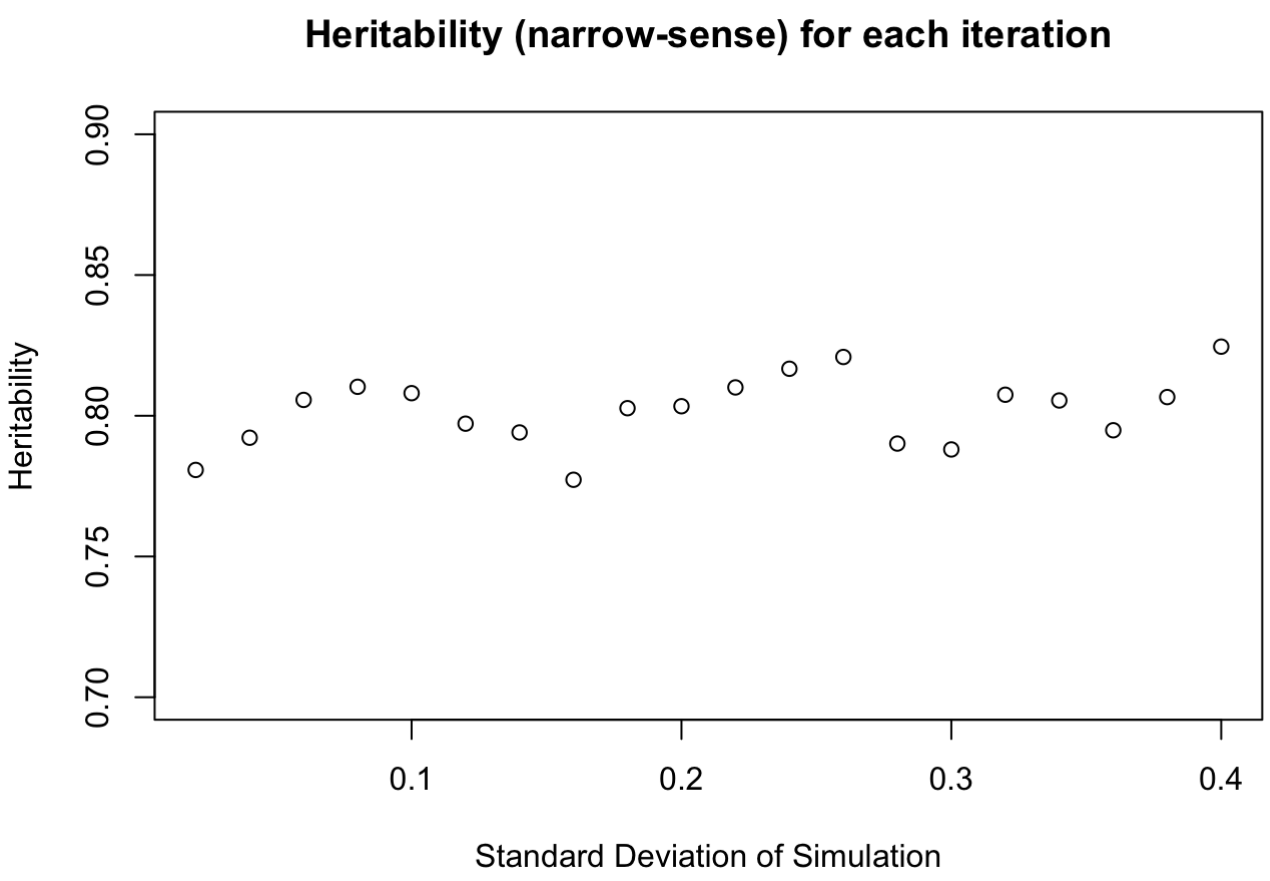
We first use a set seed and vary the standard deviation for the X1 and X3 simulations from 0.4 to 0.02. We then plot the posterior inclusion probabilities (PIPs) for X1, X2, X3 from SuSiE on each dataset. Above each plot we include first the average of the partial correlation between X1 - X2 and X2 - X3, and second the partial correlation between X1 - X3:

```
set.seed(100)
r2all <- c()
sds <- sort(seq(0.02, 0.4, by=0.02), decreasing = TRUE)
cors <- cbind(rep(0,20), rep(0,20))
par(mfrow = c(4,5))
for (i in c(1:20)) {
  x2 <- rnorm(1000)
  x1 <- x2 + rnorm(1000, sd = sds[i])
  x3 <- x2 + rnorm(1000, sd = sds[i])
  X <- cbind(x1, x2, x3)
  y <- x1 + x3 + rnorm(1000)

  pcor1 <- round(mean(cor(x1, x2), cor(x2,x3)), digits = 3)
  pcor2 <- round(cor(x1,x3), digits = 3)
  cors[i,1] <- pcor1
  cors[i,2] <- pcor2

  res <- susie(X, y, L= 2)
  r2 <- 1 - (res$sigma2)/var(y)
  r2all <- c(r2all, r2)
  nam <- paste0("res", sds[i])
  assign(nam, res)

  par(mar = c(3,2,2,3))
  title <- paste(pcor1, pcor2, sep = " , ")
  plot(res$pip, xlab = "Predictor", ylab = "PIP", main = title, pch = 20, xlim = c(0.9,
3.1), ylim = c(-0.1, 1.1), xaxt = "n", yaxt = "n")
  axis(side = 2, at = c(0,0.5,1), labels = c("0", "0.5", "1"), las = 1)
  axis(side = 1, at = c(1,2,3), labels = c("1", "2", "3"))
}
```



We can also plot the narrow-sense heritability for each iteration, which is equivalent to the R^2 value for each model fit:

```
plot(sds, r2all, main = "Heritability (narrow-sense) for each iteration", xlab = "Standa
rd Deviation of Simulation", ylab = "Heritability", ylim = c(0.7, 0.9))
```



We see that while the heritability varies, there does not seem to be a significant change to it as the correlations vary. Note also that an easy way to calculate heritability with the SuSiE output is to use $h^2 = 1 - sigma2/Var(Y)$, where sigma2 is the residual variance.

We then output the partial correlation values at which the PIPs first change from assigning 1 to X1 and X3:

```
for (i in c(1:20)) {
  res <- paste0("res", sds[i])
  if(get(res)$pip[1] > 0.95 || get(res)$pip[2] > 0.95){}
  else {
    ourcor <- cors[i,]
    break
  }
}
ourcor
```

```
## [1] 0.990 0.982
```

We thus see that for this simulation, there is not an identifiability issue until the X1-X2/X2-X3 correlations are at 0.990, and the corresponding X1-X3 correlation reaches 0.982.

We now run the same simulation from above but 100 times, without any set seed, and for each simulation find the partial correlation values at which the PIPs first change from assigning 1 to X1 and X3. We then output the mean of the X1 - X2/X2 - X3 partial correlations and the mean of the X1 - X3 partial correlation, and plot the mean X1 -

X2/X2 - X3 correlation and X1 - X3 correlation for each simulation, including the mean of the simulations as a red line:

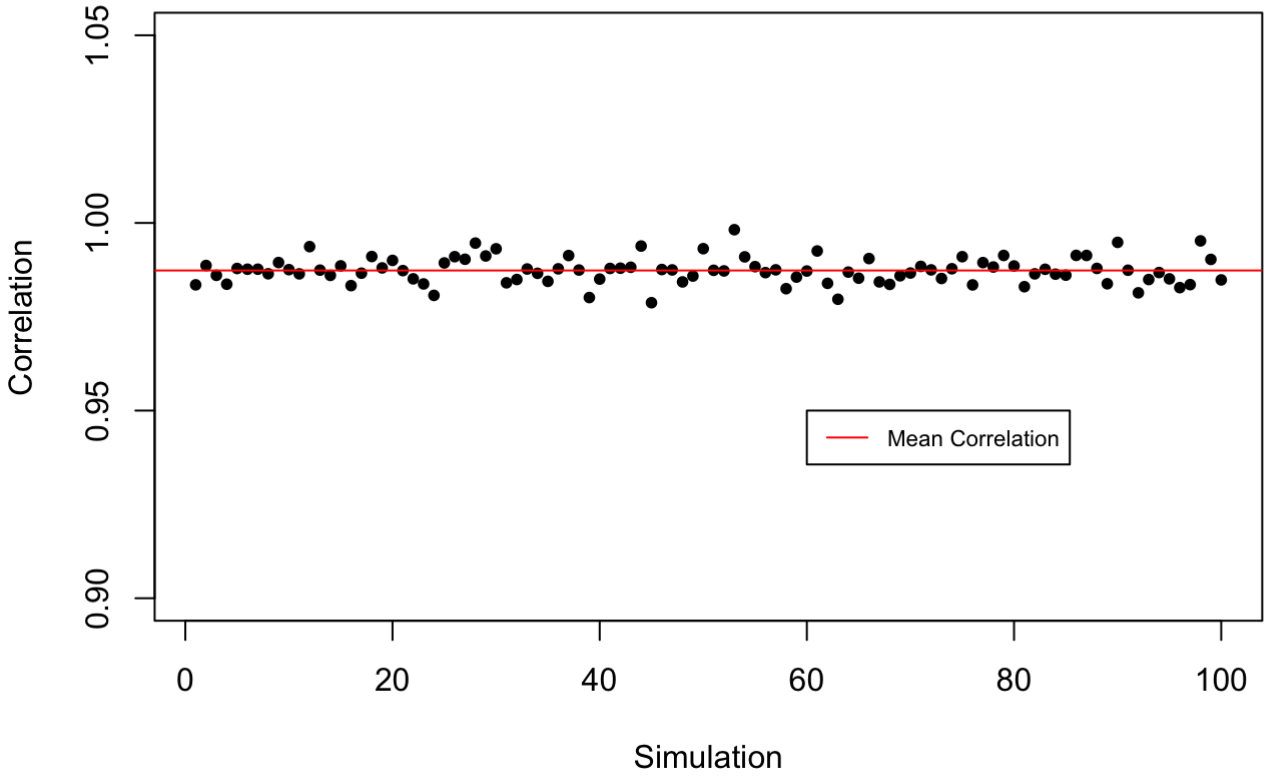
```
manycors1 <- c()
manycors2 <- c()

for (val in c(1:100)) {
  sds <- sort(seq(0.02, 0.4, by=0.02), decreasing = TRUE)
  cors <- cbind(rep(0,20), rep(0,20))
  for (i in c(1:20)) {
    x2 <- rnorm(1000)
    x1 <- x2 + rnorm(1000, sd = sds[i])
    x3 <- x2 + rnorm(1000, sd = sds[i])
    X <- cbind(x1, x2, x3)
    y <- x1 + x3 + rnorm(1000)

    pcor1 <- mean(cor(x1, x2), cor(x2,x3))
    pcor2 <- cor(x1,x3)
    cors[i,1] <- pcor1
    cors[i,2] <- pcor2

    res <- susie(X, y, L= 2)
    nam <- paste0("res", sds[i])
    assign(nam, res)
  }
  for (i in c(1:20)) {
  res <- paste0("res", sds[i])
  if(get(res)$pip[1] > 0.95 || get(res)$pip[2] > 0.95){}
  else {
    ourcor1 <- cors[i,1]
    ourcor2 <- cors[i,2]
    break
  }
  }
  manycors1 <- c(manycors1, ourcor1)
  manycors2 <- c(manycors2, ourcor2)
}
mean(manycors1)
```

```
## [1] 0.9873337
```

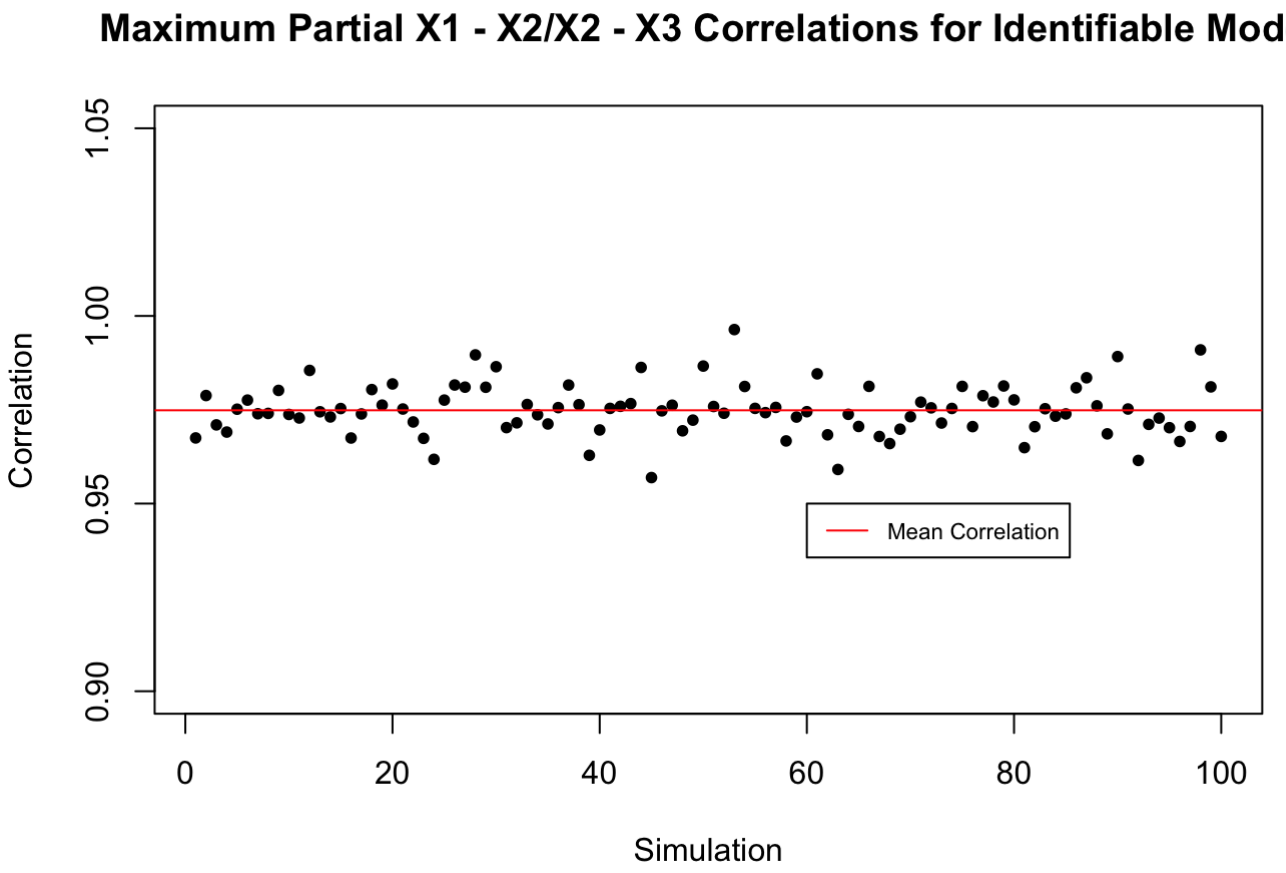
```
mean(manycors2)
```

```
## [1] 0.9748582
```

```
plot(manycors1, main = "Maximum Partial X1 - X2/X2 - X3 Correlations for Identifiable Mo
del", xlab = "Simulation", ylab = "Correlation", ylim = c(0.9,1.05), pch = 20)
abline(h = mean(manycors1), col = "red")
legend(60, 0.95, legend = "Mean Correlation", col = "red", lty = 1:2, cex = 0.7)
```




```
plot(manycors2, main = "Maximum Partial X1 - X2/X2 - X3 Correlations for Identifiable Mo
del", xlab = "Simulation", ylab = "Correlation", ylim = c(0.9,1.05), pch = 20)
abline(h = mean(manycors2), col = "red")
legend(60, 0.95, legend = "Mean Correlation", col = "red", lty = 1:2, cex = 0.7)
```



The mean X1-X2/X2-X3 partial correlation where the identifiability issue arises is:

```
mean(manycors1)
```

```
## [1] 0.9873337
```

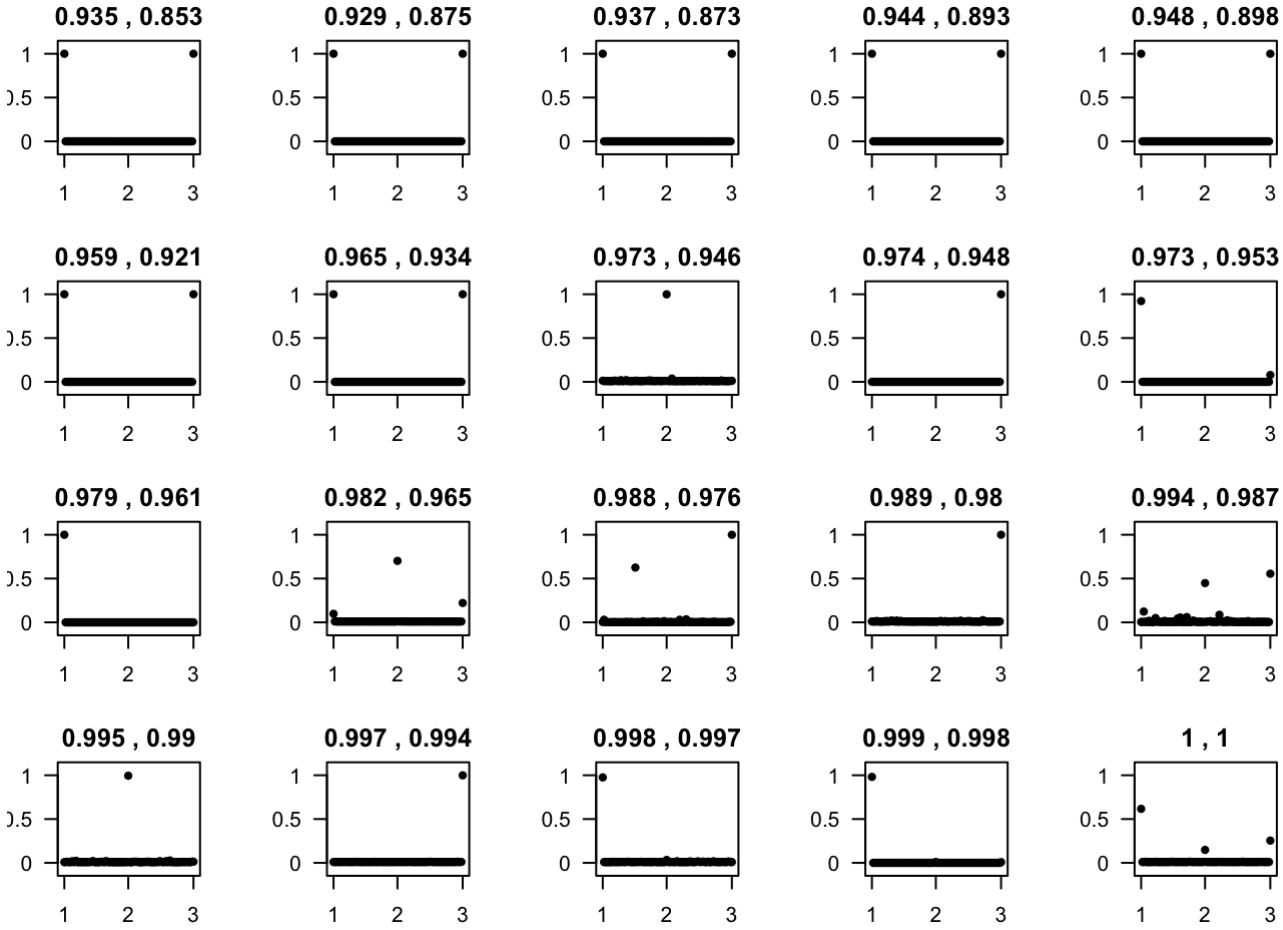
And the mean X1-X3 partial correlation is where the identifiability issue arises is:

```
mean(manycors2)
```

```
## [1] 0.9748582
```

We now repeat the same steps as above, but introduce 97 other variables with no effect on the outcome. We first use a set seed:

```
set.seed(100)
r2all <- c()
sds <- sort(seq(0.02, 0.4, by=0.02), decreasing = TRUE)
cors <- cbind(rep(0,20), rep(0,20))
par(mfrow = c(4,5))
for (i in c(1:20)) {
  x2 <- rnorm(1000)
  x1 <- x2 + rnorm(1000, sd = sds[i])
  x3 <- x2 + rnorm(1000, sd = sds[i])
  X <- x1
  for (j in c(1:48)){
    xnew <- rnorm(1000)
    X <- cbind(X, xnew)
  }
  X <- cbind(X, x2)
  for (j in c(1:49)){
    xnew <- rnorm(1000)
    X <- cbind(X, xnew)
  }
  X <- cbind(X, x3)
  y <- x1 + x3 + rnorm(1000)

  pcor1 <- round(mean(cor(x1, x2), cor(x2,x3)), digits = 3)
  pcor2 <- round(cor(x1,x3), digits = 3)
  cors[i,1] <- pcor1
  cors[i,2] <- pcor2

  res <- susie(X, y, L= 2)
  r2 <- 1 - (res$sigma2)/var(y)
  r2all <- c(r2all, r2)
  nam <- paste0("res", sds[i])
  assign(nam, res)

  par(mar = c(3,2,2,3))
  title <- paste(pcor1, pcor2, sep = " , ")
  plot(res$pip, xlab = "Predictor", ylab = "PIP", main = title, pch = 20, xlim = c(0,101
), ylim = c(-0.1, 1.1), xaxt = "n", yaxt = "n")
  axis(side = 2, at = c(0,0.5,1), labels = c("0", "0.5", "1"), las = 1)
  axis(side = 1, at = c(1,50,100), labels = c("1", "2", "3"))
}
```



From this plot alone, it appears that we have an identifiability issue at a lower correlation than in the previous simulations. We then run 100 different simulations of this:

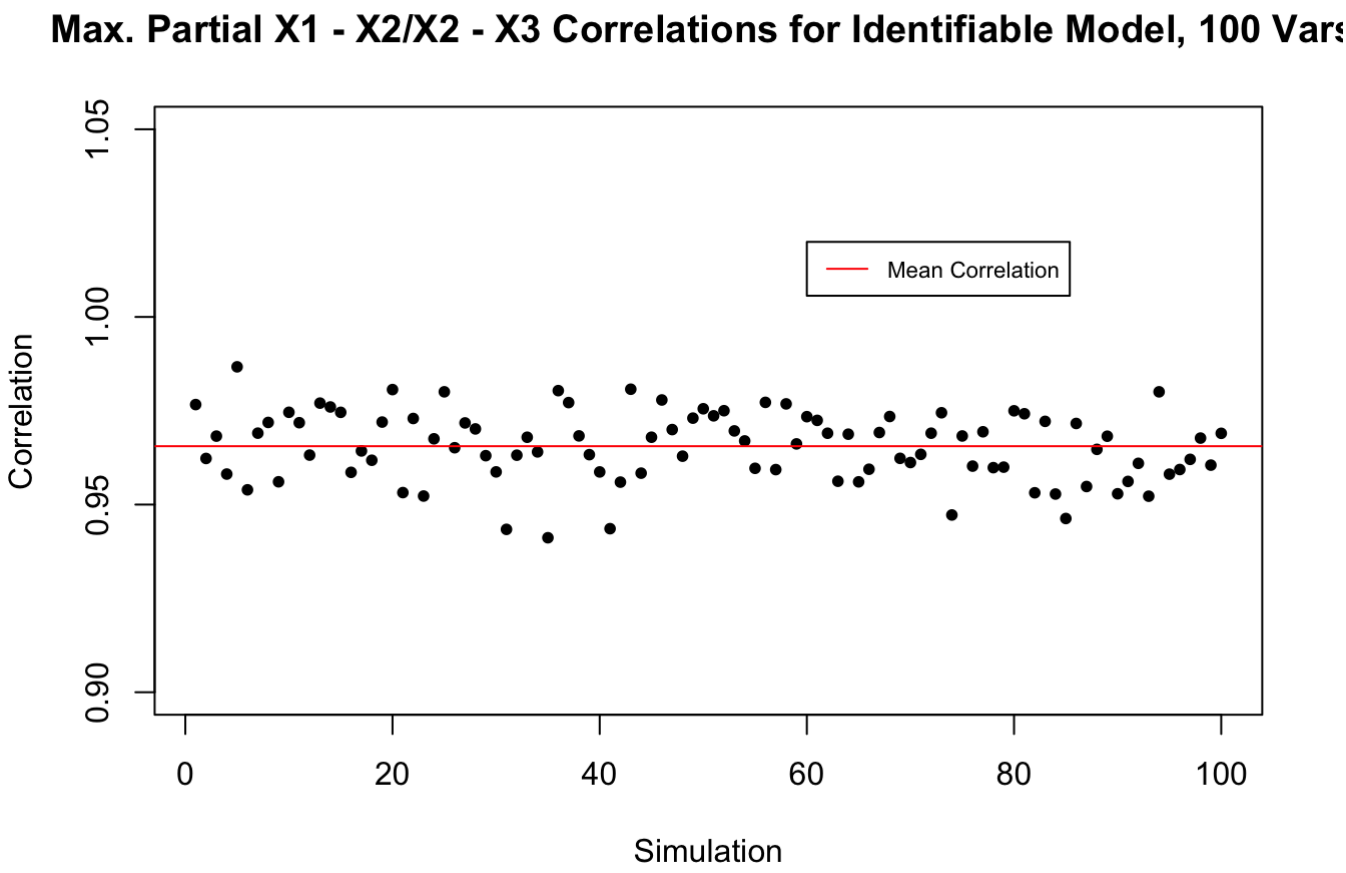
```
manycors1 <- c()
manycors2 <- c()

for (val in c(1:100)) {
  sds <- sort(seq(0.02, 0.4, by=0.02), decreasing = TRUE)
  cors <- cbind(rep(0,20), rep(0,20))
  for (i in c(1:20)) {
    x2 <- rnorm(1000)
    x1 <- x2 + rnorm(1000, sd = sds[i])
    x3 <- x2 + rnorm(1000, sd = sds[i])
    X <- x1
    for (j in c(1:48)){
      xnew <- rnorm(1000)
    X <- cbind(X, xnew)
    }
    X <- cbind(X, x2)
    for (j in c(1:49)){
      xnew <- rnorm(1000)
      X <- cbind(X, xnew)
    }
    X <- cbind(X, x3)
    y <- x1 + x3 + rnorm(1000)

    pcor1 <- mean(cor(x1, x2), cor(x2,x3))
    pcor2 <- cor(x1,x3)
    cors[i,1] <- pcor1
    cors[i,2] <- pcor2

    res <- susie(X, y, L= 2)
    nam <- paste0("res", sds[i])
    assign(nam, res)
  }
  for (i in c(1:20)) {
  res <- paste0("res", sds[i])
  if(get(res)$pip[1] > 0.95 || get(res)$pip[2] > 0.95){}
  else {
    ourcor1 <- cors[i,1]
    ourcor2 <- cors[i,2]
    break
  }
  }
  manycors1 <- c(manycors1, ourcor1)
  manycors2 <- c(manycors2, ourcor2)
}
mean(manycors1)
```

```
## [1] 0.9655569
```

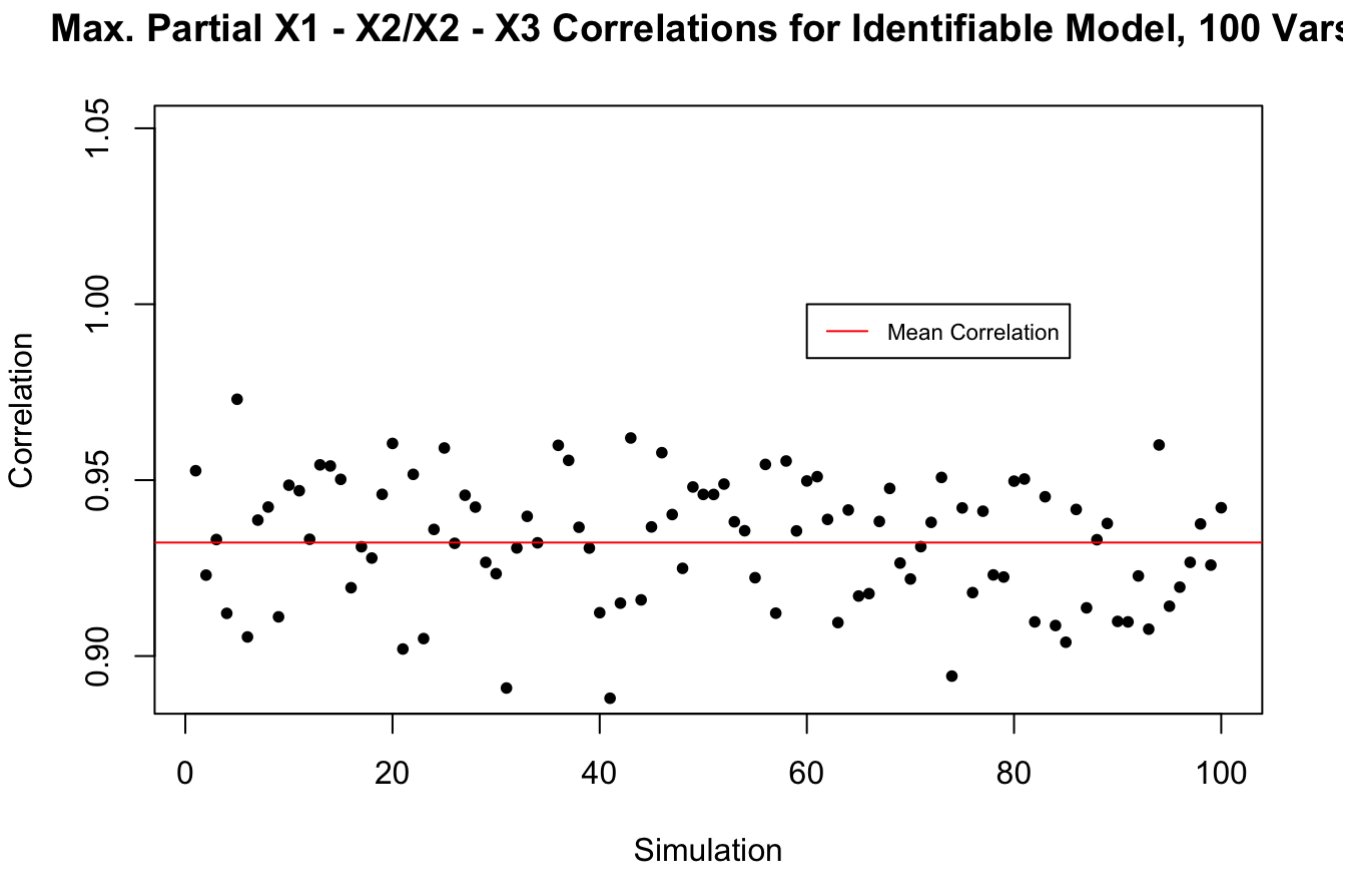
```
mean(manycors2)
```

```
## [1] 0.9322582
```

```
plot(manycors1, main = "Max. Partial X1 - X2/X2 - X3 Correlations for Identifiable Mode
l, 100 Vars.", xlab = "Simulation", ylab = "Correlation", ylim = c(0.9,1.05), pch = 20)
abline(h = mean(manycors1), col = "red")
legend(60, 1.02, legend = "Mean Correlation", col = "red", lty = 1:2, cex = 0.7)
```




```
plot(manycors2, main = "Max. Partial X1 - X2/X2 - X3 Correlations for Identifiable Mode
l, 100 Vars.", xlab = "Simulation", ylab = "Correlation", ylim = c(0.89,1.05), pch = 20)
abline(h = mean(manycors2), col = "red")
legend(60, 1, legend = "Mean Correlation", col = "red", lty = 1:2, cex = 0.7)
```



For the simulations with more extraneous variables, the mean X1-X2/X2-X3 partial correlation where the identifiability issue arises is:

```
mean(manycors1)
```

```
## [1] 0.9655569
```

And the mean X1-X3 partial correlation where the identifiability issue arises is:

```
mean(manycors2)
```

```
## [1] 0.9322582
```

We can thus see that with the inclusion of the 97 other extraneous variables, the mean X1-X2/X2-X3 correlation at which the model becomes unidentifiable drops about 0.02, and the corresponding mean X1-X3 correlation drops about 0.04.

From this, we can conclude that in this simple setting SuSiE can correctly identify the model up to very high correlations between the X1-X2 and X2-X3 pairs, though when applied to a situation with many more noise variables, we see that in extreme correlations, such as 0.96 and beyond, identifiability does become an issue, and the algorithm begins to switch between all four possible models. Even in these cases, however, SuSiE sometimes partially identifies the model, assigning significant PIP to the true causal variables, but there is no guarantee of this happening for a given dataset.