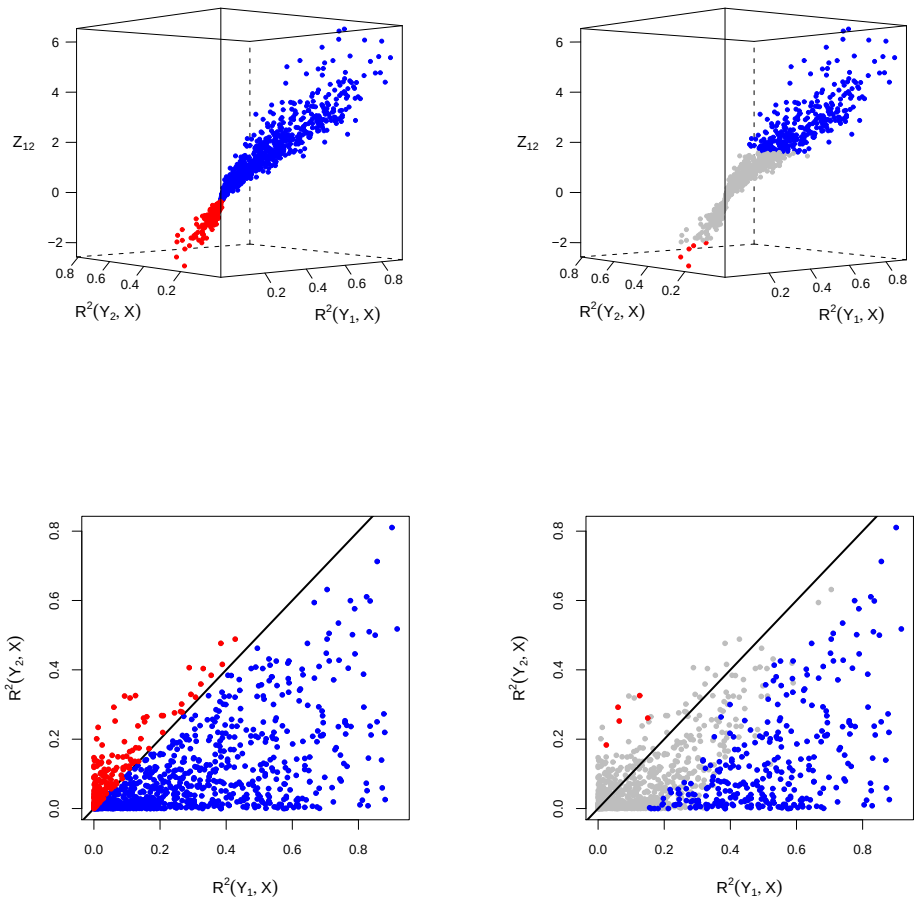

Figure S2 Model selection via log-likelihood ratio versus Vuong's test.

Figure S2 illustrates how Vuong's test works. We generated 1,000 data-sets from the model $X \to Y_1 \to Y_2$ and applied Vuong's test to the comparison of models $M_1 : X \to Y_1 \to Y_2$ against $M_2 : X \to Y_2 \to Y_1$. The top panels present 3D scatter plots of the test statistics Z_{12} against the R^2 values of the regression of Y_1 on X, $R^2(Y_1, X)$, and the R^2 values of the regression of Y_2 on X, $R^2(Y_2, X)$. The data points are color coded as blue, red and grey, representing, respectively, M_1, M_2 and "no calls". Note that because model M_1 corresponds to the true model, we have that the a M_1 call is always correct, whereas

a M_2 call is always incorrect in this example. Therefore, blue and red points represent, respectively, correct and incorrect calls. The bottom panels follow the same color coding and show the projections of the 3D scatter plots into the $R^2(Y_1, X)$ by $R^2(Y_2, X)$ plane.

The left panels of Figure S2 show the model selection results based on the log-likelihood ratio (LR) criterium, where positive $L\hat{R}_{12}$ values support M_1 and negative $L\hat{R}_{12}$ values support M_2 (note that we actually use the Z_{12} test statistics, instead of $L\hat{R}_{12}$ statistics, but the results are equivalent). Because we generate the data from model $X \rightarrow Y_1 \rightarrow Y_2$, it will usually be the case that X explains a greater proportion of the variability of Y_1 than of Y_2. In other words, $R^2(Y_1, X)$ will tend to be higher than $R^2(Y_2, X)$. However, some of the data-sets show the opposite trend due to random noise on the data. The bottom left panel shows that the log-likelihood criterium tends to make incorrect calls when $R^2(Y_1, X) < R^2(Y_2, X)$.

The right panels of Figure S2 show the model selection results derived from Vuong's test. Now we see that most of the incorrect calls made by the log-likelihood criterium (red points) are not significant (grey points) according to Vuong's test, that requires that $Z_{12} \leq -1.64$ or $Z_{12} \geq 1.64$ for statistical significance at a 5% level. The drawback is the reduction in power to detect the correct calls, since not only red dots are replaced by grey dots, but many of the blue dots are turned into grey, as well. These figures illustrate how Vuong's test trade an increase in precision for a reduction in statistical power to detect true positives.