Online Appendix: Bayesian versus maximum likelihood estimation of treatment effects in bivariate probit instrumental variable models

A. STAN CODE

```
// STAN code for Bayesian bivariate model
// based on code by Bob Carpenter
// http://tinyurl.com/horbxuu

functions { // define sum function
  int sum(int[,] a) {
    int s;
    s <- 0;
    for (i in 1:size(a))
      for (j in 1:size(a[i]))
        s <- s + a[i,j];
    return s;
  }
}
data {
  int<lower=1> K;/// number of independent variables, including intercept
  int<lower=1> D; //number of equations, in our case 2 (first and second stage)
  int<lower=0> N; // number of observations in dataset
  int<lower=0,upper=1> y[N,D]; // DVs in first and second stage, i.e. Outcome and Treatment
  vector[K] x[N]; // matrix of independent variables (including an intercept vector)
  real instr[N]; //instrument
}
transformed data { // counting how many in outcome and treatment vectors are 0 and 1s
//create matrix with indicator for each
  matrix[N,D] y2; //copy of y
  int<lower=0> N_pos;
  int<lower=1,upper=N> n_pos[sum(y)];
  int<lower=1,upper=D> d_pos[size(n_pos)];
  int<lower=0> N_neg;
  int<lower=1,upper=N> n_neg[(N * D) - size(n_pos)];
  int<lower=1,upper=D> d_neg[size(n_neg)];

  N_pos <- size(n_pos);
  N_neg <- size(n_neg);
  {
    int i;
    int j;
    i <- 1;
    j <- 1;
    for (n in 1:N) {
      for (d in 1:D) {
        if (y[n,d] == 1) {
          n_pos[i] <- n;
          d_pos[i] <- d;
          i <- i + 1;
        } else {
          n_neg[j] <- n;
          d_neg[j] <- d;
          j <- j + 1;
        }
      }
    }
  }
  y2 <- to_matrix(y);
}
parameters {
  matrix[D,K] beta; // coefficients for independent variables
  real beta2; // coefficient for treatment
  real pi; // coefficient for instrument
  cholesky_factor_corr[D] L_Omega; // cholesky decomposition of correlation matrix
  vector<lower=0>[N_pos] z_pos;
```

```
  // latent parameter of outcomes & treatment = 1, constrained so that minimum 0
  vector<upper=0>[N_neg] z_neg;
  // latent parameter of outcomes & treatment = 0, constrained so that maximum 0
}
transformed parameters {
  vector[D] z[N];
  // vector z is the transformed data for the normal draw, i.e. negative if y = 0, positive if y = 1
  for (n in 1:N_pos)
    z[n_pos[n], d_pos[n]] <- z_pos[n];
  for (n in 1:N_neg)
    z[n_neg[n], d_neg[n]] <- z_neg[n];
}
model {
  vector[N] treatment;
  treatment <- col(y2,1);
  // assigning treatment to new vector, so it can be used for calculation of second stage
  to_vector(beta) ~ normal(0, 2.5); // prior on coefficients on IVs
  beta2 ~ normal(0, 2.5); // prior on treatment effect
  pi ~ normal(0,2.5); // prior on instrument
  L_Omega ~ lkj_corr_cholesky(D); // prior on covariance decomp
  {
    vector[D] beta_x[N];
    //transposing the mean vector by hand so it can be used in the multi normal draw
    for(d in 1:D){
       for (n in 1:N){ //creating means for draw
            beta_x[n,1] <- row(beta,1) * x[n] + pi*instr[n]; // first stage
      beta_x[n,2] <- row(beta,2) * x[n] + treatment[n]*beta2; //second stage
         }
    }
    z ~ multi_normal_cholesky(beta_x, L_Omega); // draw from multinormal
  }
}
generated quantities {
  corr_matrix[D] Omega; //create correlation matrix from decomposition
  Omega <- multiply_lower_tri_self_transpose(L_Omega);
}
```

B. A BAYESIAN FRAMEWORK FOR BIVARIATE PROBIT MODELS

Since both the treatment and outcome are binary, the likelihood function is not directly tractable. We therefore use data augmentation, as proposed by Chib and Greenberg (1998), which to reformulates the model in terms of a latent, continuous variable for treatment and outcome $\mathbf{y}^* = (y^*, T^*)$. In addition, let $\mathbf{X_i} = \begin{pmatrix} \mathbf{x}_i' & T_i & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{x}_i' & z_i \end{pmatrix}$. For convenience, we also define $\mathbf{B} = [\boldsymbol{\gamma}_1', \pi, \boldsymbol{\gamma}_2', \beta,]$. This allows us to write the conditional distribution of the augmented data as

$$f(\mathbf{y_i}^*|\mathbf{X}, \mathbf{B}, \boldsymbol{\Omega}) \quad \propto \quad \exp[((\mathbf{y_i}^* - \mathbf{X_i B})'\Omega^{-1}(\mathbf{y_i}^* - \mathbf{X_i B})]., \tag{1}$$

The marginal posterior distributions for the augmented treatment data is:

$$T_i^*|\mathbf{X}, B, \Omega \propto \begin{cases} \phi(\mathbf{x}_i'\gamma_1 + z_i\pi + \rho(y_i^* - \mathbf{x}_i'\boldsymbol{\gamma}_2), 1-\rho^2); \mathrm{I}(-\infty, 0) & \text{if } T_i = 0 \\ \phi(\mathbf{x}_i'\boldsymbol{\gamma}_1 + z_i\pi + \rho(y_i^* - \mathbf{x}_i'\boldsymbol{\gamma}_2 - \beta), 1-\rho^2); \mathrm{I}(0, \infty) & \text{if } T_i = 1 \end{cases}, \tag{2}$$

where $\mathrm{I}(\cdot)$ indicates the allowable support of the truncated distributions. Likewise the marginal posterior distributions for the outcome is is:

$$y_i^*|\mathbf{X}, B, \Omega \propto \begin{cases} \phi(\mathbf{x}_i'\gamma_2 + T_i\beta + \rho(T_i^* - \mathbf{x}_i'\boldsymbol{\gamma}_1 - z_i\pi), 1-\rho^2); \mathrm{I}(-\infty, 0) & \text{if } y_i = 0 \\ \phi(\mathbf{x}_i'\boldsymbol{\gamma}_1 + T_i\beta + \rho(T_i^* - \mathbf{x}_i'\boldsymbol{\gamma}_1 - z_i\pi), 1-\rho^2); \mathrm{I}(0, \infty) & \text{if } y_i = 1 \end{cases} \tag{3}$$

Further, placing a conjugate prior of the structural parameters, $\pi(\mathbf{B}) = \phi(\mathbf{B}_0, \boldsymbol{\Sigma}_{B_0})$, we get the usual multivariate Bayesian regression result, $\boldsymbol{B} \sim N(\hat{\mathbf{B}}, \hat{\boldsymbol{\Sigma}}_B)$, where $\hat{\boldsymbol{\Sigma}}_B = (\boldsymbol{\Sigma}_{B_0}^{-1} + \sum \mathbf{X_i'}\boldsymbol{\Omega}^{-1}\mathbf{X_i})^{-1}$ and $\hat{\mathbf{B}} = \hat{\boldsymbol{\Sigma}_B}(\boldsymbol{\Sigma}_{B_0}^{-1}\mathbf{B}_0 + \sum \mathbf{X_i'}\boldsymbol{\Omega}^{-1}\mathbf{y_i^*})$.

To complete the model, it is necessary to specify priors for and calculate the covariance matrix $\boldsymbol{\Omega}$, which must be positive definite. In this case, no closed-form conditional distribution is available. Chib and Greenberg (1998) and Chib (2003) recommend using a Metropolis-Hastings step with a tailored non-central t-distribution as the proposal density

for a transformation of the parameter ρ. However, we found the mixing in this model to be somewhat slow. Instead, we redefine the parameters in the model such that $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is a lower triangular matrix that results from the Cholensky decomposition of $\boldsymbol{\Omega}$. We then place a Lewandowski, Kurowicka and Joe (2009) prior over this triangular matrix, which ensures a positive definite correlation matrix.

C. LOCAL AVERAGE TREATMENT EFFECTS

The local average treatment effect is defined as

$$\Delta^{(LATE)} = \frac{E[y|z=1] - E[y|z=0]}{E[T|z=1] - E[T|z=0]}$$

For each individual, this is equivalent to

$$\frac{[\Pr(y_i=1,T_i=1|z_i=1) + \Pr(y_i=1,T_i=0|z_i=1)] - [[\Pr(y_i=1,T_i=1|z_i=0) + \Pr(y_i=1,T_i=0|z_i=0)]]}{Pr(T_i=1|z=1) - Pr(T_i=1|z=0)}.$$

Let $-\mathbf{\Omega}$ be equivalent to $\mathbf{\Omega}$ except that the off-diagonal elements are multiplied by -1. We can define the interim quantity

$$\Delta_i^{(LATE)} = \frac{M - O}{\Phi(\mathbf{x}_i'\boldsymbol{\gamma}_1 + \pi) - \Phi(\mathbf{x}_i'\boldsymbol{\gamma}_1)}$$

where

$$\begin{aligned} M &= \Phi^{(2)}\left(\begin{pmatrix} x_i\gamma_1 + \pi \\ x_i\gamma_2 + \beta \end{pmatrix}, \mathbf{\Omega}\right) + \Phi^{(2)}\left(\begin{pmatrix} -(x_i\gamma_1 + \pi) \\ x_i\gamma_2 \end{pmatrix}, -\mathbf{\Omega}\right) \\ O &= \Phi^{(2)}\left(\begin{pmatrix} x_i\gamma_1 \\ x_i\gamma_2 + \beta \end{pmatrix}, \mathbf{\Omega}\right) + \Phi^{(2)}\left(\begin{pmatrix} -x_i\gamma_1 \\ x_i\gamma_2 \end{pmatrix}, -\mathbf{\Omega}\right) \end{aligned}$$

This is again calculated for each observation at th observed values of x_i. Thus, $\Delta^{(LATE)} = \sum_{i=1}^{n} \Delta_i^{(LATE)}/n$

D. ADDITIONAL DETAILS FOR SIMULATION STUDY

In the simulations in the main text, coefficients have Gaussian priors with mean zero and standard deviation 2.5, which we found provided generally more accurate estimates while not overly ''shrinking'' the data towards zero in a probit setting. We estimated each `Stan` model using 3,000 iterations on three chains and discarded the first 500 iterations. We sporadically checked convergence using Gelman-Rubin statistics, but discovered no issues. Note that we keep `Stan`'s default settings. However, we generally suggests increasing of the target Metropolis acceptance rate (δ) for small samples. Adjusting this rate should allow researchers to achieve even better results than presented in the main text. In addition to the simulations in the main text, we also estimated all Bayesian models with Gaussian priors with standard deviations of 1.5 and 3.5, and an improper noninformative uniform prior (viz. no prior specified in `Stan` at all). We show a comparison for the models with different priors in the next section.

The Bayesian model estimated in `Stan` provides results for all 120,000 simulated data sets, while the ML estimator does not provide results for a small number of the simulated data sets. Due to the random creation of the data in the first and second stage, some of the estimations suffer from separation. This occurs relatively rarely and mostly when $N = 50$. Since `Stata` throws an error when either the first or second stage suffers from separation, it does not provide any results. Additionally in some instances the `Stata` ML algorithm does not converge. Thus, there were a number of data sets for which `Stata` does not provide any estimation results or for which no bootstrap of confidence intervals for the ATE or LATE are possible. The results for these 2,093 simulated data sets, i.e. less than 2% of the total, were not analyzed. In addition, in rare cases `Stata` converges to widely inaccurate estimates. When evaluating the ML procedure, therefore, we drop any simulated data sets when the absolute error of the estimated average treatment effect is larger than 2. Note that this occurs for zero cases in the Bayesian models, but for 1047 cases in the `Stata` ML estimation.

In total, we dropped 2,093 + 1047 = 3140 of the 120,000 simulated data sets when evaluating the ML estimator and 2,093 of the simulations for the Bayesian estimate (those for which the Stata MLE did not converge). We did not have to drop any additional simulations for the Bayesian models. However, dropping these estimations for the `Stata` estimator biases our results in *favor* of the ML estimator in comparison to the correctly specified Bayesian model, especially considering that we leave out cases where the ML estimate is very inaccurate.[1]

[1]When comparing models based on the LATE we additionally drop observations where the absolute error of the true LATE and estimated LATE is larger than 2. This leads us to drop 220 simulation results for the Bayesian model (with prior SD = 2.5), 739 simulation results for the 2SLS model, and 1157 simulations the STATA ML model.

E. COMPARISON TO TWO-STAGE LEAST SQUARES

In this section we compare the estimates of the local average treatment effect (LATE) for the each simulated data set for the Bayesian model (prior SD 2.5), the STATA MLE estimator, and the STATA 2SLS estimator. Figure E.1 shows the mean squared error for each of the three estimators for varying instrument strength and sample sizes. As one can easily see, the 2SLS estimator performs by far the worst on these simulated data sets.

Again, we can compare coverage for the Bayesian, Stata MLE, and 2SLS estimator. Figure E.2 shows the coverage rates for all three estimators. As one can clearly see, the MLE and 2SLS estimator are widely off the mark and under-cover significantly. Surprisingly, the coverage of both the MLE and 2SLS estimators actually becomes worse with larger samples and stronger instruments. We believe this to be the case because the standard errors decrease even though the estimate is often false.

To further compare these methods, in Table E.1 we show the estimates for our application (Sondheimer and Green 2010) in the main text, when applying standard IV methods where the outcome is assumed to be continuous. Note that many of these estimates are *prima facia* implausible. For instance, in the STAR experiment, 2SLS estimates the effect of education as leading to a 90% increase in the predicted probability of voting in a federal election. Note also, that the implied confidence interval falls well outside of the possible $[-1, 1]$ interval for both the STAR and IHAD experiments.

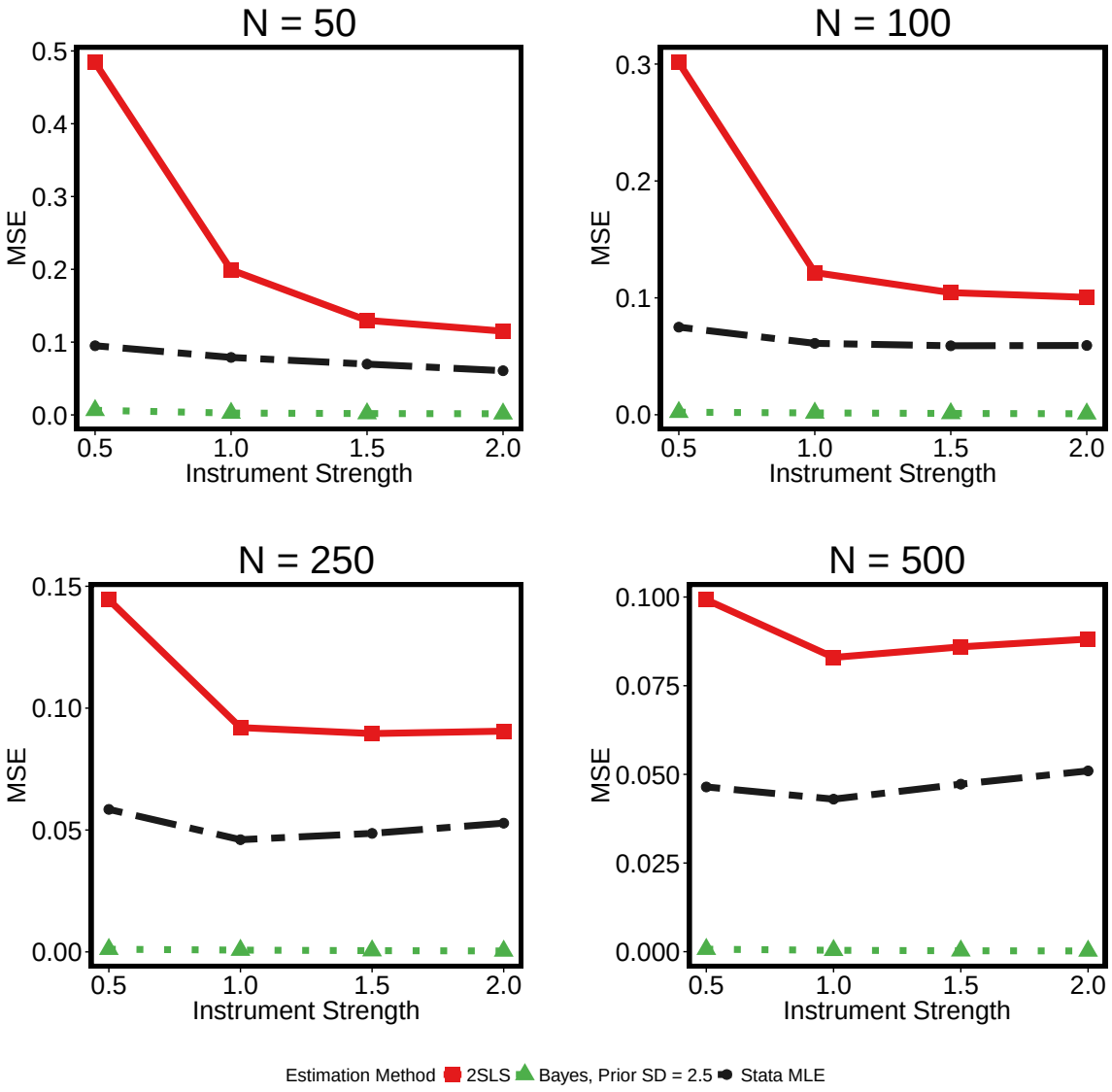

Figure E.1. This Figure compares the mean squared error (MSE) for the Bayesian (prior SD = 2.5), ML estimates, and the 2SLS estimates for subsets of the simulated data sets. The MSE is calculated based on the estimated and true LATE. The top-left plot shows the MSEs for data sets including 50 observations and different strength-levels of the instrument. Going from the top left to the bottom right the sample size increases for each plot. Both the Bayesian and ML specifications outperform the 2SLS estimate in terms of the MSE across *all* parameter settings.

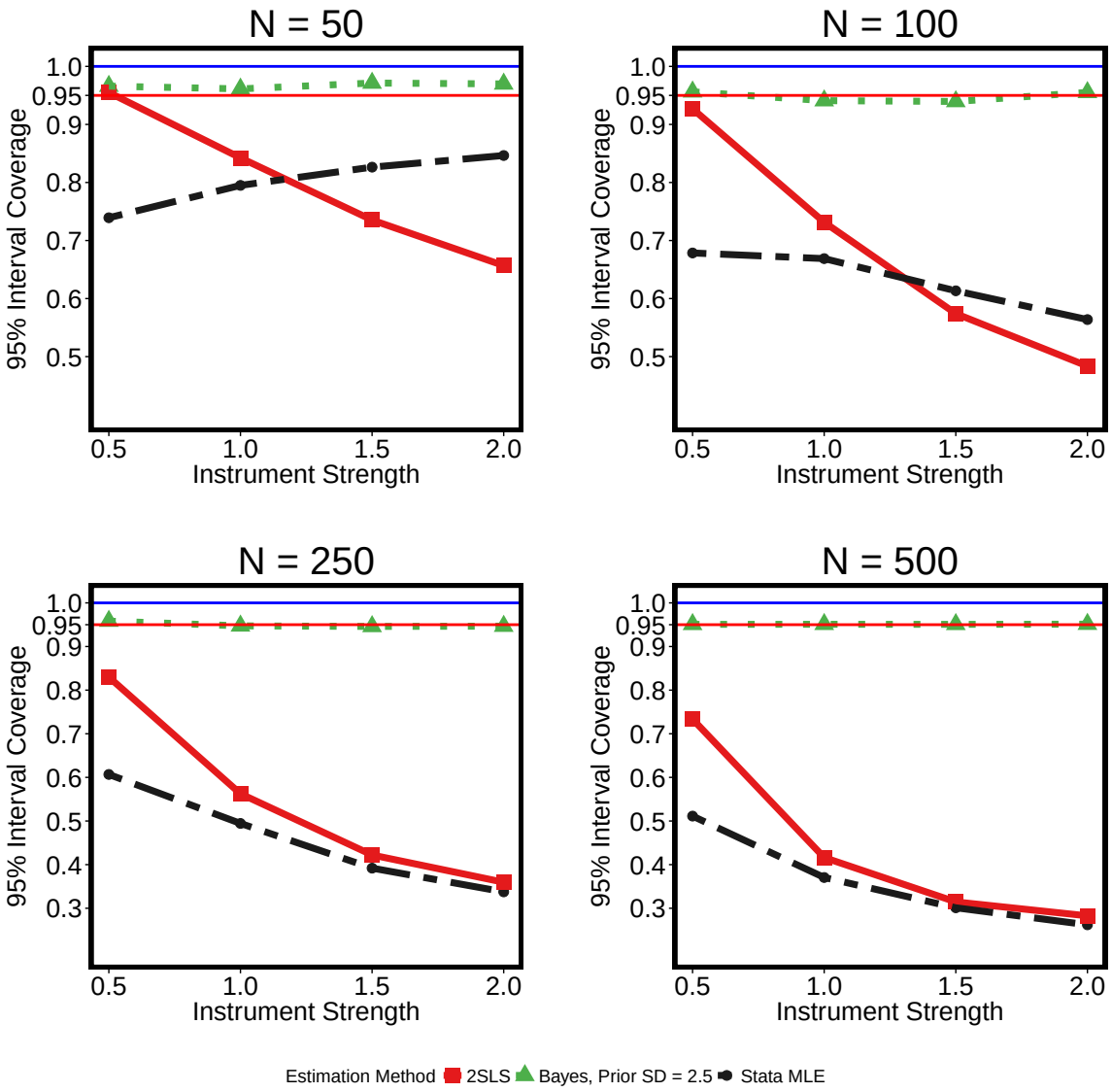

Figure E.2. This Figure compares the 95% coverage probabilities for the Bayesian, Stata MLE, and 2SLS estimators for the LATE estimate, which should be about 0.95 for a well-calibrated model. The Bayesian estimator is much closer to this value for nearly all parameter combinations, although it is somewhat conservative. The coverage probabilities for the MLE and 2SLS model are too small, especially with strong instruments or large sample sizes.

Table E.1. Estimating treatment effects for schooling experiments using 2SLS and LIML instrumental variables models

	Perry		IHAD		STAR	
	2SLS	*Fuller LIML*	*2SLS*	*Fuller LIML*	*2SLS*	*Fuller LIML*
HS Graduation Level						
Intercept	0.44	0.44	0.62	0.62	0.85	0.85
	(0.06)	(0.06)	(0.08)	(0.08)	(0.01)	(0.01)
Treatment	0.21	0.21	0.17	0.17	0.05	0.05
	(0.09)	(0.09)	(0.13)	(0.13)	(0.03)	(0.03)
Voter Turnout Level						
Intercept	0.01	0.02	0.02	0.15	-0.35	-0.23
	(0.18)	(0.16)	(0.59)	(0.44)	(0.70)	(0.60)
HS Graduation	0.27	0.25	0.50	0.31	0.90	0.76
	(0.32)	(0.29)	(0.86)	(0.86)	(0.80)	(0.69)
N	123	123	58	58	811	811

Coefficients are 2SLS/Fuller-corrected LIML estimates with standard errors in parentheses. Note that the estimates of the effect of graduation on turnout are implausibly large for the STAR and IHAD experiments. In the Fuller-corrected LIML models, we set a = 1.

F. COMPARISON OF DIFFERENT PRIORS

In this section we provide comparison of the different prior specifications for the Bayesian model and the Stata ML estimate. As one can see in Figure F.3, all Bayesian specifications clearly outperform the ML estimates in terms of the MSE for the ATE. The Bayesian model with the uniform prior performs very similar to the Stata ML model, but slightly better. In general, the model with the Gaussian prior with standard deviation 2.5 performs best, except for very large sample sizes.

Figure F.4 provides the evaluation of the credible (confidence) intervals across the different prior specifications of the Bayesian models and the ML estimates. As one can see, when the prior on the coefficient is too tight, the Bayesian estimates may provide too much shrinkage and the credible intervals are too tight (especially for small samples). Essentially, the estimates of the Bayesian model with a prior standard deviation of 1.5 provide coefficient estimates that are very conservative (closer to zero). This is especially visible in Figure F.5, which provides a comparison of the different MSEs for different ''true'' treatment effects. Based on our experience and the simulation results, we would therefore recommend researchers to use prior standard deviations of around 2.5 unless sample sizes are extremely large. On the other hand, tight priors do give conservative estimates and may therefore be used by researchers. As to be expected, the Bayesian model with completely uninformative priors does not perform very well, which indicates the importance and benefit of specifying reasonable priors. Indeed, without a prior, the Bayesian method must function entirely off of the likelihood and behaves poorly for the reasons discussed in the main text.

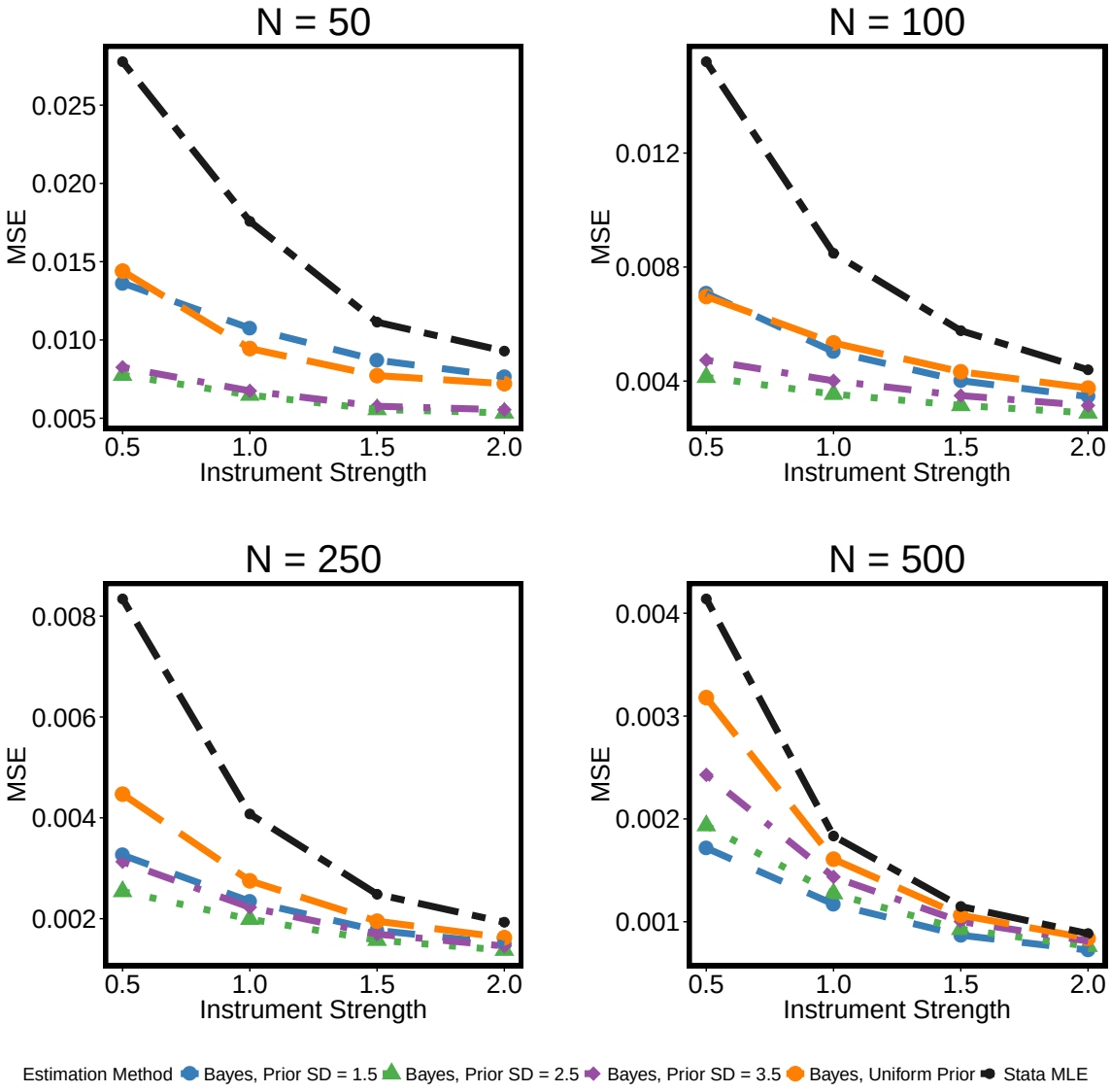

Figure F.3. This Figure compares the mean squared error (MSE) on the ATE for the Bayesian (all prior specifications) and ML estimates for subsets of the simulated datasets. The top-left plot shows the MSEs for data sets including 50 observations and different strength-levels of the instrument. Going from the top left to the bottom right the sample size increases for each plot. All Bayesian specifications with informative priors outperform the ML estimate in terms of the MSE across *all* parameter settings and this pattern is particularly clear with small samples and weak instruments. However, in terms of MSE, the prior specification with medium standard deviation (2.5) generally performs the best, except for very large sample sizes. The Bayesian specification with the noninformative uniform prior performs similar to the Stata estimate.

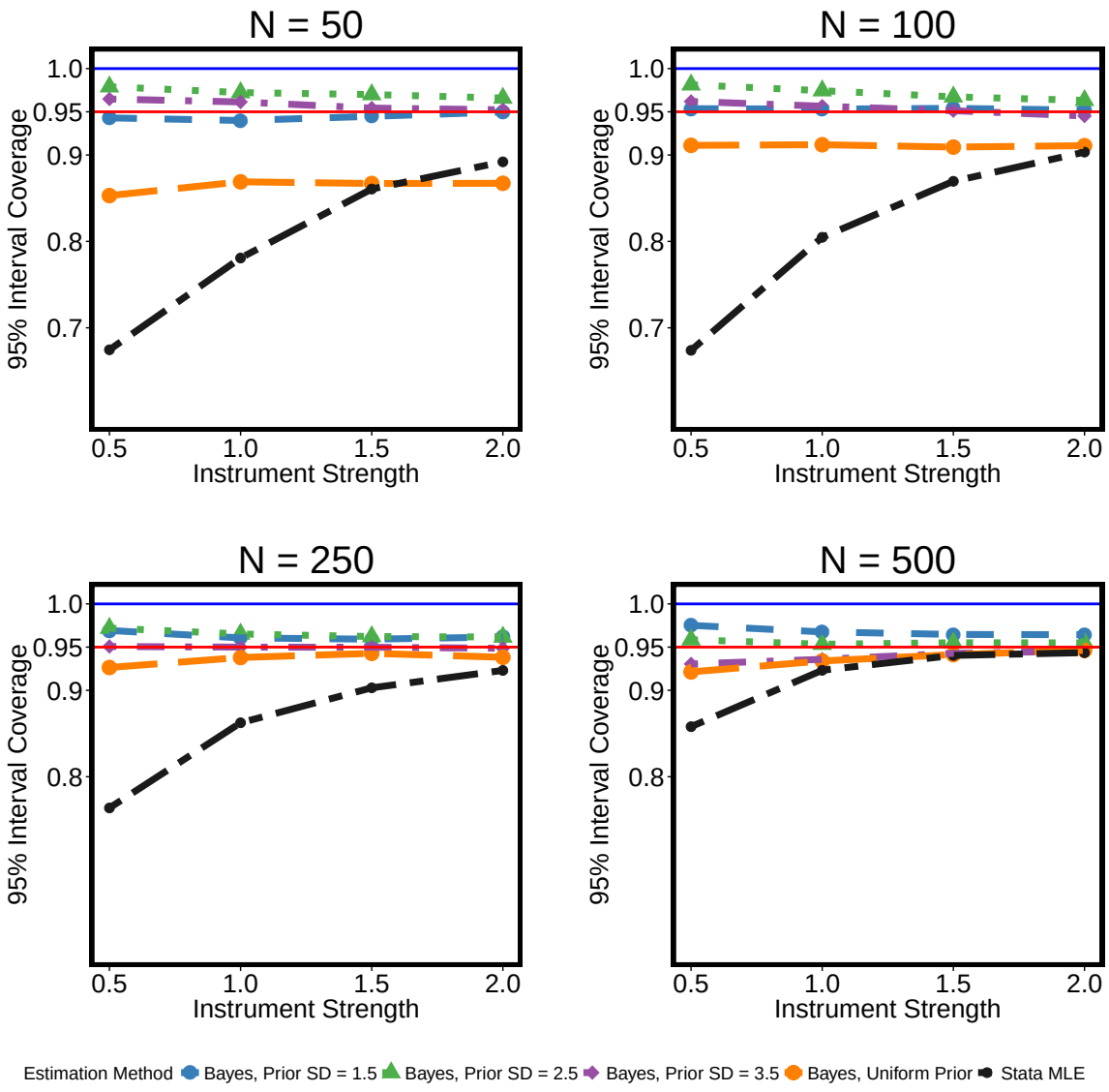

Figure F.4. This Figure compares the 95% coverage probabilities for the Bayesian and ML estimators (including all prior specifications), which should be about 0.95 for a well-calibrated model. The Bayesian estimator is much closer to this value for nearly all parameter combinations and most prior specifications, although it is somewhat conservative. The coverage probabilities for the ML model are too small, especially with weak instruments or small sample sizes. Similarly, if the standard deviation on the Gaussian prior is too tight, the shrinkage in the Bayesian model becomes too strong.

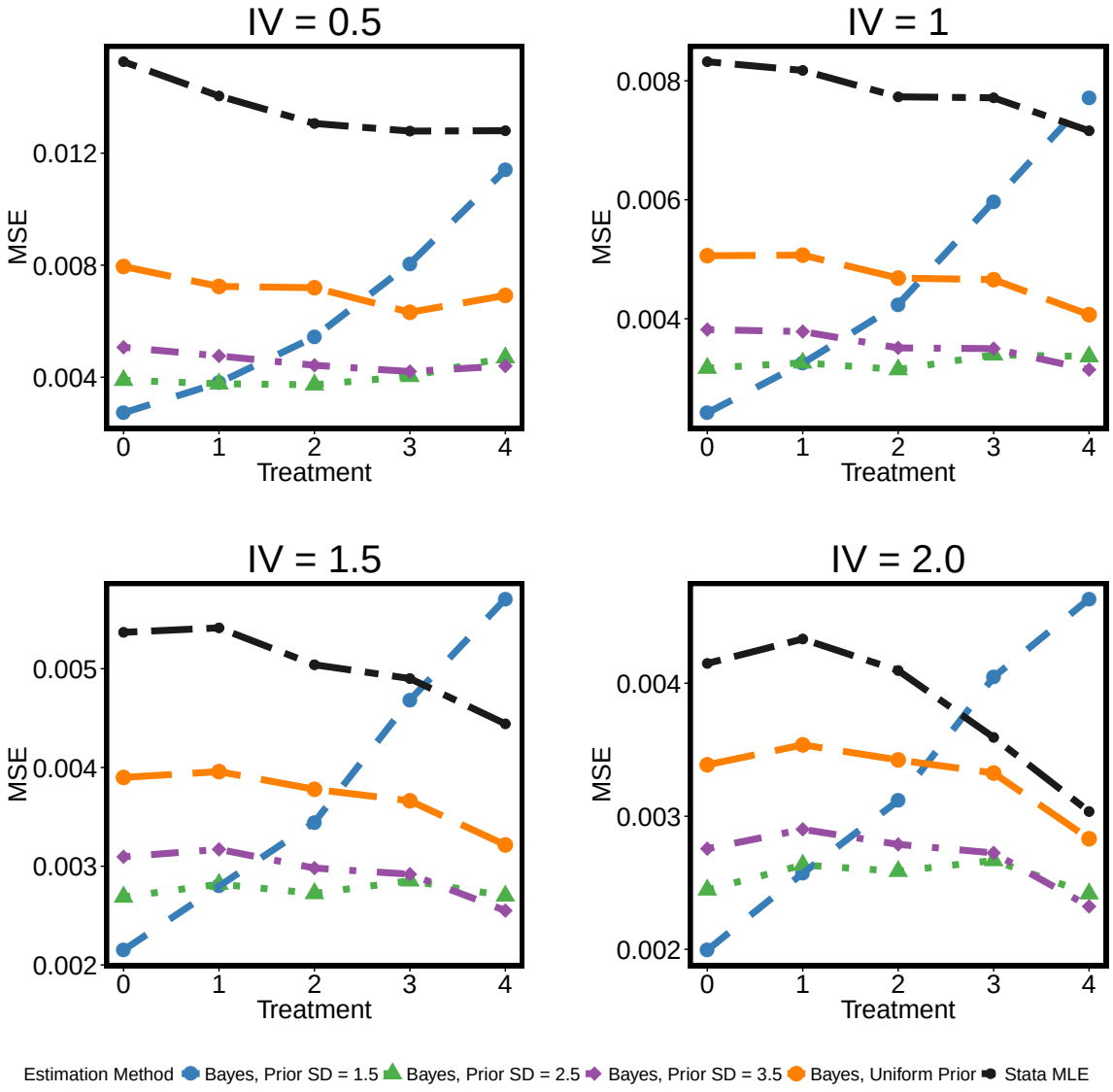

Figure F.5. This Figure compares the mean squared error (MSE) on the ATE for the Bayesian (all prior specifications) and ML estimates for subsets of the simulated data sets. The top-left plot shows the MSEs for data sets with the true IV coefficient equal to 0.5 and different treatment strength. Going from the top left to the bottom right the strength of the instrument increases for each plot. Again, in general the Bayesian specifications with informative priors outperform the ML estimate in terms of the MSE. This pattern is particularly clear with weak instruments and treatments. However, as one can clearly see, when the treatment is particularly strong, the Bayesian specification with a tight prior variance shrinks the estimate too far towards zero and thus suffers in terms of the MSE.

G. MAE STATISTICS FOR SIMULATION

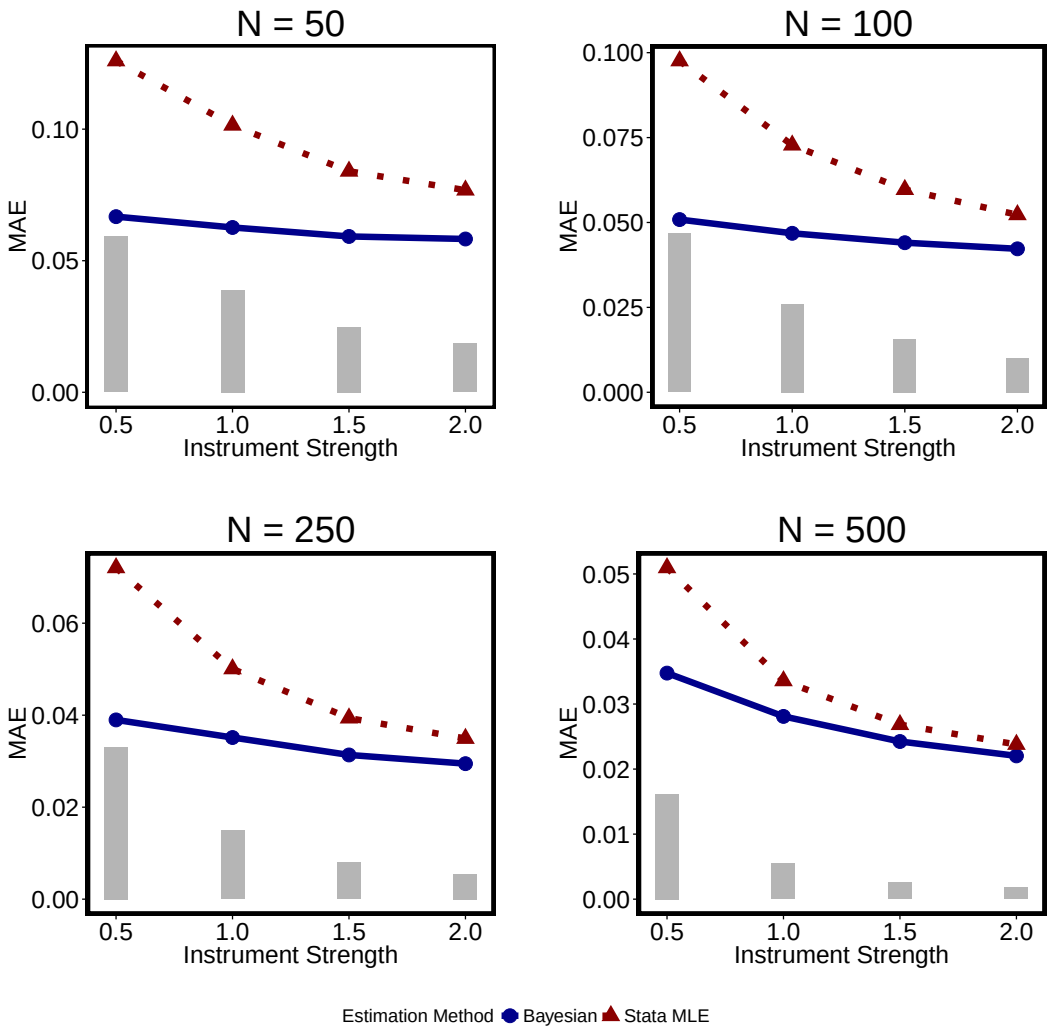

Figure G.6. This Figure shows the mean absolute error (MAE) for the ATE for the Bayesian estimate of the treatment effect in comparison to the MAE of the maximum likelihood estimate. As in Figure 1 in the main text, the top-left plot shows the MAEs for data sets with 50 observations. Going from the top left to the bottom right the sample size increases for each plot. Gray vertical plots show the difference between the two MAEs at each parameter value. It is immediately obvious that, again, across *all* parameter combinations displayed, the Bayesian estimator outperforms the ML estimate in terms of the MAE. With increasing sample size and stronger instruments, the difference decreases. Still, even at large sample sizes and with strong instruments the Bayesian estimate remains superior.

H. COMPARING SIMULATIONS NUMBERS

In general we generated 500 different data sets for each parameter combination, resulting in 120,000 different simulated data sets. In addition, for a subset of the parameter space we also created 2,000 simulated data sets per parameter combination. Here we provide a quick comparison of the main results for this subset of the parameter space. Specifically. for N = (50, 100, 250, and 500), $\rho = 0.75$, $\pi = 1.5$, and $\beta = 3$ we simulated an additional $2,000$ data sets.

Table H.2 shows a comparison of the mean absolute error for the parameter estimate of β for each model specification and parameter combination for the 500 and 2,000 simulated data sets.[2] The differences are very small and not substantively meaningful giving confidence that our 500 simulation runs per parameter combination are sufficient. Similarly, Table H.3 shows the variance of the absolute error for each model at the different parameter settings. Again, differences between 500 and 2,000 simulated data sets are very minor.

Table H.2. Mean of Absolute Error

N	Bayes (SD 1.5)	Bayes (SD 1.5)	Bayes (SD 1.5)	Bayes (SD Uniform)	STATA ML	# Simulations
50.000	0.948	0.517	0.874	2.243	1.532	500.000
50.000	0.939	0.506	0.873	2.302	1.578	2000.000
100.000	0.551	0.528	0.810	1.464	0.984	500.000
100.000	0.549	0.525	0.809	1.479	0.991	2000.000
250.000	0.363	0.427	0.526	0.685	0.523	500.000
250.000	0.357	0.422	0.526	0.688	0.519	2000.000
500.000	0.271	0.339	0.390	0.459	0.368	500.000
500.000	0.273	0.330	0.378	0.446	0.364	2000.000

[2]Again, we have dropped observations for which the absolute error ($|\beta - \hat{\beta}|$) is larger than 10.

Table H.3. Variance of Absolute Error

N	Bayes (SD 1.5)	Bayes (SD 1.5)	Bayes (SD 1.5)	Bayes (SD Uniform)	STATA ML	# Simulations
50.000	0.153	0.163	0.451	4.530	2.856	500.000
50.000	0.151	0.156	0.442	4.414	2.863	2000.000
100.000	0.127	0.162	0.390	1.547	0.946	500.000
100.000	0.117	0.162	0.401	1.629	0.972	2000.000
250.000	0.066	0.109	0.180	0.311	0.174	500.000
250.000	0.064	0.112	0.182	0.310	0.186	2000.000
500.000	0.043	0.074	0.098	0.134	0.089	500.000
500.000	0.040	0.069	0.090	0.125	0.082	2000.000

I. CALCULATING BOOTSTRAPPED CONFIDENCE INTERVALS

In the main text we compare the results from our Bayesian model with the ML model estimates produced in `Stata`. In particular, we are interested in comparing how each approach accurately recovers the ATE as defined in the main text. When comparing the accuracy of point estimates, this is relatively straightforward. However, to compare the coverage rates of the two estimation methods we need to derive confidence intervals of the ATE for the ML model, which is not directly available from the `Stata` output. To calculate standard errors for the derived causal quantities of interest, we rely on a parametric bootstrapping method. The purpose of this appendix is to briefly provide the details for how we calculating these quantities.

Broadly speaking our strategy is to simulate draws from the assumed asymptotic distribution of the parameters of the model and to base our calculations on these draws. Specifically, assume that $\boldsymbol{\Sigma}$ is the estimated variance covariance matrix provided by `Stata`. Further, let $\tilde{\rho}$ represent the arc-hyperbolic tangent of ρ from the main text, and $\boldsymbol{\tau} = (\gamma_1, \pi, \gamma_2, \beta, \tilde{\rho})'$ be the complete vector of parameters from the model. We assume that the parameters in model are distributed according to following multivariate normal, $MVN(\boldsymbol{\tau}, \boldsymbol{\Sigma})$.

Our general approach is to take a draw from this multivariate normal distribution and transform $\tilde{\rho}$ to ρ using the tangent function. We then calculate the ATE as specified in the main text. We then repeat this process 1,000 times. We use calculate the 95% bootstrapped confidence interval using the the 97.5 and 2.5 percentile estimates.[3]

[3]For the LATE calculations in Appendix E we use only 500 draws due to the computational demands of these calculations.

References

Chib, Siddhartha. 2003. On Inferring Effects of Binary Treatments with Unobserved Confounders. In *Bayesian Statistics*, ed. J.M. Bernardo, M.J. Bayarry, J.O. Berger and A.P. Dawid. Vol. 7 Oxford: Oxford University Press pp. 65--84.

Chib, Siddhartha and Edward Greenberg. 1998. ''Analysis of Multivariate Probit Models.'' *Biometrika* 85(2):347--361.

Lewandowski, Daniel, Dorota Kurowicka and Harry Joe. 2009. ''Generating random correlation matrices based on vines and extended onion method.'' *Journal of multivariate analysis* 100(9):1989--2001.

Sondheimer, Rachel Milstein and Donald P. Green. 2010. ''Using Experiments to Estimate the Effects of Education on Voter Turnout.'' *American Journal of Political Science* 54(1):174--189.