Supplementary Appendix: Exploring the Dynamics of Latent Variable Models

Kevin Reuning

Miami University of Ohio

Michael R. Kenwick

The University of Pennsylvania

Christopher J. Fariss

The University of Michigan

Contents

Introduction to the Appendix

The supplementary material presented in this document provides additional details about the latent variable model developed in the article "Exploring the Dynamics of Latent Variable Models". The main article makes reference to the materials contained here.

Below in Appendix A, we first discuss the Student's t-distribution used in the robust dynamic latent variable model. In Appendix B, we discuss an alternative mixture latent variable model.

In Appendix C, we present a detailed discussion of the data generating process for the simulation and more information than we presented in the main manuscript. Specifically, in Appendix C.1, we provide an overview of the simulation analysis. In Appendix C.2, we discuss accuracy in time surrounding shocks to the latent trait. In Appendix C.3, we discuss accuracy in ranking observations. In Appendix C.4, we discuss within-unit rank correlations. In Appendix C.5, we discuss cross-validated accuracy. In Appendix C.6, we discuss differences between time periods. In Appendix C.7 and Appendix C.8, we provide simulation models that estimate the degrees of freedom parameter in the Student's t distribution.

Next, we discuss the effect of fixing the innovation variance at values that are too high or too low in Appendix D. Following this, we discuss the effect of inflated variance and directly estimate the variance in the Martin and Quinn (2002) data in Appendix E. We also provide additional posterior predictive checks for the comparison of models estimating the latent democracy variable in Appendix F. We cover convergence statistics for both replication models in Appendix G. Finally, we discuss extensions to WAIC in Appendix H.

In Appendix I, we briefly discuss the replication files that are publicly available. The estimates presented in this appendix along with the code necessary to implement the Bayesian models in STAN are publicly available at a dataverse archive here:
https://doi.org/10.7910/DVN/SSLCFF (Reuning, Kenwick and Fariss 2018).

A Student's t Parameterization

For the prior density function for the latent trait in the robust dynamic model, we employ the three parameter Student's t-distribution:

$$p(x|\nu,\mu,\sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}}\Big(1+\frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\Big)^{-\frac{\nu+1}{2}}$$

For $\nu > 2$:

$$\mathrm{E}(X) = \mu$$

$$\mathrm{var}(X) = \sigma^2\frac{\nu}{\nu-2}$$

As $\nu \to \infty$ then $\mathrm{var}(X) = \sigma^2$. It is possible to relate these values to the variance of the normal distribution, so that the Student's t has the same variance of the normal, but a different distribution of the probability mass (particularly at it's tails). When necessary, we set the variance of the normal and t-distributions to equivalent values using the following equation $\sigma_t^2 = \frac{\nu-2}{\nu}\sigma_{norm}^2$, where σ_{norm}^2 is the variance of the normal distribution.

B Mixture Model

In the main manuscript we consider three latent variable models: a static model, dynamic model, and robust dynamic model. We also considered a mixture model. For this model, we attempt to leverage the strengths of both the static and standard dynamic models by incorporating the two using a mixture modeling strategy. Here the items are assumed to be linked to the latent trait through two separate data generating processes, the first captured by the static model, and the second captured using the dynamic model. Consider again a punctuated equilibrium process for which we would expect that periods of stasis or slow-movement in θ will be captured using dynamic modeling, while periods of rapid change will be captured by the static alternative. The likelihood function for the mixture model is:

$$\mathcal{L} = \prod_{i,t=1}^{N,T} \prod_{k=1}^{K} \pi_{it} \Lambda(\alpha_k - \beta_k \theta_{it}^{dyn})^{y_{itk}} \big(1 - \Lambda(\alpha_k - \beta_k \theta_{it}^{dyn})\big)^{1-y_{itk}} + (1 - \pi_{it}) \Lambda(\alpha_k - \beta_k \theta_{it}^{stat})^{y_{itk}} \big(1 - \Lambda(\alpha_k - \beta_k \theta_{it}^{stat})\big)^{1-y_{itk}}$$

Where π_{it} is the mixture parameters for the static and dynamic models. We estimate these parameters for every unit-year in the data. The priors for the mixture model are detailed below. We place a prior of Beta(4,1) on the π_{it} parameter. This places a stronger prior on the estimate being drawn from the dynamic process. The expected value for the probability of the dynamic is 0.80, and the static is 0.20. This is because we expect periods of stasis to be more common than periods of volatility in most time series.

Mixture Model Priors

$$
\begin{aligned}
\theta_{it}^{stat} &\sim \mathrm{N}(0,1) \quad \forall i = 1,\ldots,N \quad \& \quad \forall t = 1,\ldots,T \\
\theta_{i1}^{dyn} &\sim \mathrm{N}(0,1) \quad \forall i = 1,\ldots,N \\
\theta_{it}^{dyn} &\sim \mathrm{N}(\pi_{i(t-1)}\theta_{i(t-1)}^{dyn} + (1-\pi_{i(t-1)})\theta_{i(t-1)}^{stat}, \sigma) \quad \forall i = 1,\ldots,N \quad \& \quad \forall t = 2,\ldots,T \\
\sigma &\sim \mathrm{HN}(0,3) \\
\pi_{it} &\sim \mathrm{Beta}(4,1)
\end{aligned}
$$

We evaluate the performance of this model relative to the static, dynamic, and robust dynamic model in the additional simulation analyses described in Appendix C.

C Simulation Analysis

C.1 Data Generating Process

Data for the simulations are generated using the process detailed in Algorithm 1. The initial $\theta_{n,1}$ is drawn from a standard normal distribution. Future values from that unit are then drawn using random walk assumptions where $\theta_{n,t}$ is drawn from a normal distribution centered on the previous value $\theta_{n,(t-1)}$. At each time point, we also allow the possibility for a "shock" where the auto-correlation across θs are broken.We implement this by drawing from a Bernoulli distribution with parameter p_{shock} at each time period. If a 1 is drawn then $\theta_{n,t}$ is drawn from the standard normal distribution, not from a normal distribution centered around the previous value. If $p_{shock} = 0$, the assumptions of the standard dynamic approach hold exactly.

When $p_{shock} > 0$ our simulation assumptions are closest to the assumptions of a mixture model, as the data are generated from a static process with probability $p_{shock} > 0$, and a dynamic model with probability $1 - p_{shock} > 0$. The mixture parameter in this case is analogous to the probability that a shock will occur. Note that the data generating process is different from that assumed by the robust model, which accommodates shocks through the increased tail density of the Students-t distribution rather than modeling the probability of a shock as an additional model parameter. This makes for a conservative evaluation of the robust model's performance.

We generate separate simulated data sets under a variety plausible vales for p_{shock}, the probability that a unit will experience a shock in each time period, and σ, the standard deviation of the time-series innovations. p_{shock} is set to values of 0, 0.01, and 0.1 and σ set to 0.01, 0.05, and 0.1 – we choose these relatively low values because we believe this best captures typical data generating processes for social science constructs.[15] For each

[15]In addition, existing latent variable models have often produced innovation standard deviation estimates that are close to these values. See Schnakenberg and Fariss (2014) for

combination of p_{shock} and σ values we generate 50 simulated data sets. In all simulations, we generate data for 50 units observed over 50 time periods with 5 items.

Algorithm 1 Simulation Process

1: Set number of units $N = 50$, number of time periods $T = 50$, number of items $K = 5$, probability of shock p_{shock}, innovation standard deviation σ.
2: Draw initial θ for each unit: $\theta_{n,1} \sim \mathrm{N}(0, 1)$.
3: Let S be an $(N-1)$ by T matrix where $s_{n,t} \sim \text{Bernouli}(p_{shock})$.
4: Draw remaining θ such that: $\theta_{n,t} \sim \begin{cases} \theta_{n,t-1} + \mathrm{N}(0, \sigma) & \text{If} \quad s_{n,t} = 0 \\ \mathrm{N}(0, 1) & \text{If} \quad s_{n,t} = 1 \end{cases}$
5: Draw parameters to create K items: $\alpha_k \sim \mathrm{U}(-3, 3)$ and $\beta_k \sim \mathrm{U}(0, 2)$.
6: Estimate k items: $y_{n,t,k} \sim \text{Bernoulli}\left(\Phi(2(\alpha_k + \beta_k \theta_{n,t}))\right)$ where Φ is the cumulative distribution function of the standard normal distribution.
7: Estimate posterior distributions for all four models using the same set of data.
8: Calculate relevant model statistics.
9: Repeat steps 2 to 8.

Computation is performed in R using Stan, a program for Bayesian analytics (Carpenter et al. 2016). Two parallel chains were run for 5,000 iterations, with 2,500 discarded as burn-in. To speed convergence the dynamic and robust models were estimated using a non-centered parameterization.[16] Conventional diagnostics are consistent with convergence.

C.2 Accuracy in Time Surrounding Shocks to the Latent Trait

We evaluate model performance using four metrics to capture how well each model fits the data. The first performance metric is model accuracy, defined as the proportion of observations for which the true value of the latent variable is contained within the 95 percent credible intervals generated by each model. Because we are interested in model performance in the time surrounding sudden changes in the latent variable, we group observations based

one estimated example and Martin and Quinn (2002) for one set by the researchers.

[16]Specifically, the draw of θ was separated into first drawing a raw value from a standard normal distribution, and then shifting and shrinking it based on the previous value and the estimated innovation parameters.

on the amount of time before or following shocks in the latent trait, and report accuracy accordingly.[17] These results demonstrate the degree of bias (rather than efficiency) in the estimates at time periods surrounding a shock.

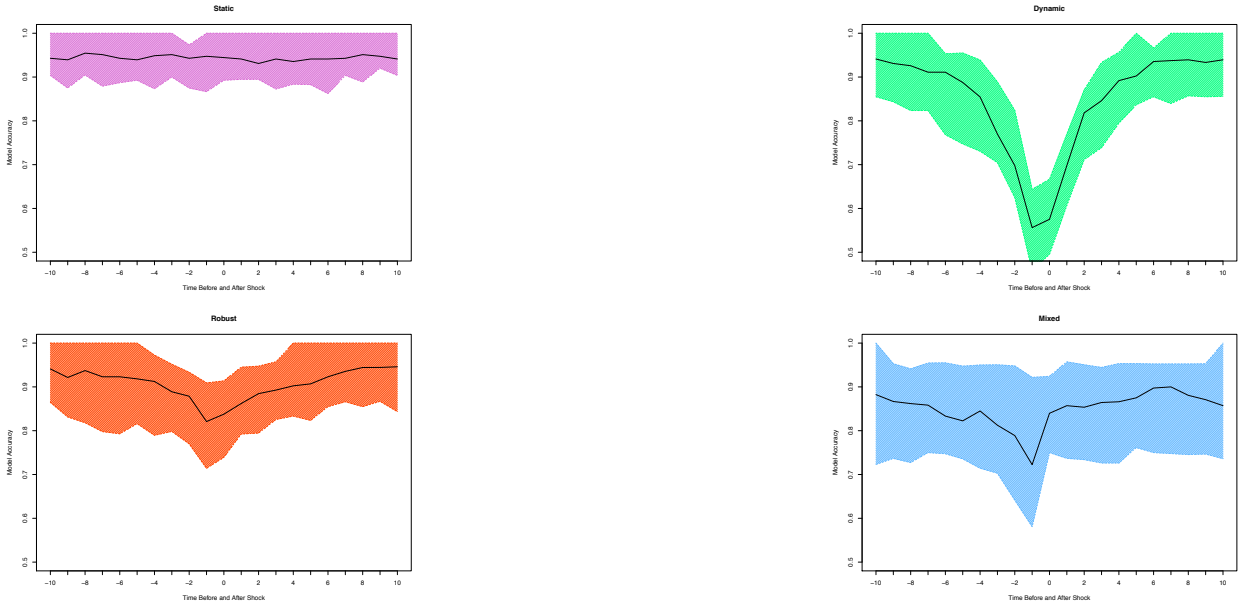
Figure 2 in the main manuscript reports results for each model (we reproduce this in Figure 6 here with the additional information from the mixture model now displayed). For this set of assessments, we selected a standard deviation on the innovation parameter of 0.05 and a shock probability of 0.01 for this set of simulations. Note that the static model generally produces the most accurate estimates, recovering the true latent estimate for about 95% of observations regardless of the proximity to a a shock. This is unsurprising, given the low-bias, low-efficiency features of the the static model, it easily accommodates shocks but generally produces large credible intervals. The performance of the dynamic model highlights the bias introduced by the over-smoothing of estimates across time. While the dynamic model generally recovers the true estimate for 90 to 95% of observations when shocks are not proximate, this value drops to about 60% for the periods immediately surrounding shocks. While neither the robust, nor mixture models completely reduce bias in the time surrounding shocks, both significantly improve the dynamic model. For the robust model, the median accuracy never dips below 85% and for the mixture model the median never dips below 75%. Thus, while neither model completely eliminates the bias caused by over-smoothing, both significantly improve upon the conventional approaches.

C.3 Accuracy in Ranking Observations

We now evaluate each model's ability to accurately rank observations on the values of the latent trait. We do so by ranking observations based on each model's estimates of the latent trait and then comparing this to the true ranking of the latent trait. As previously stated,

[17] If an observation is between two shocks, we place them with respect to the more proximate shock. We exclude observations from this analysis if they are equidistant between two shocks or are fewer than four periods away from two shocks.

Figure 6: Model Accuracy in Time Surrounding Shocks to the Latent Variable



Note: The percentage of times the true latent variable is within 95% credible intervals. The horizontal axis is the distance away from the nearest shock – if two shocks are within 3 time periods away then the value was ignored. A distribution of values is estimated by generating 50 different simulated datasets – the bounds show the 20^{th} and 80^{th} percentile.

dynamic models are generally used to enhance efficiency, not necessarily to reduce bias in parameter estimates. By comparing the rankings generated by each model, however, we ensure that we do not evaluate models by efficiency gains alone, but also by their ability to accurately reproduce the relative ordering of units along the latent space, an issue important to applied researchers. This is also a means of assessing concurrent validity[18] – in this case, each is being compared on its ability to accurately categorize units into a known rank-ordering.

We compute these correlations with respect to the cross-sectional rank-ordering of all units (N) at each time period t. For each model, we iterate over each posterior draw $\mathbf{D} = \{1, \dots D\}$, and calculate cross-sectional ranking at each time period t, producing the

[18]Concurrent validity is a type of construct validity designed to assess the ability of a variable's operationalization to distinguish between groups that it should be able to distinguish between.

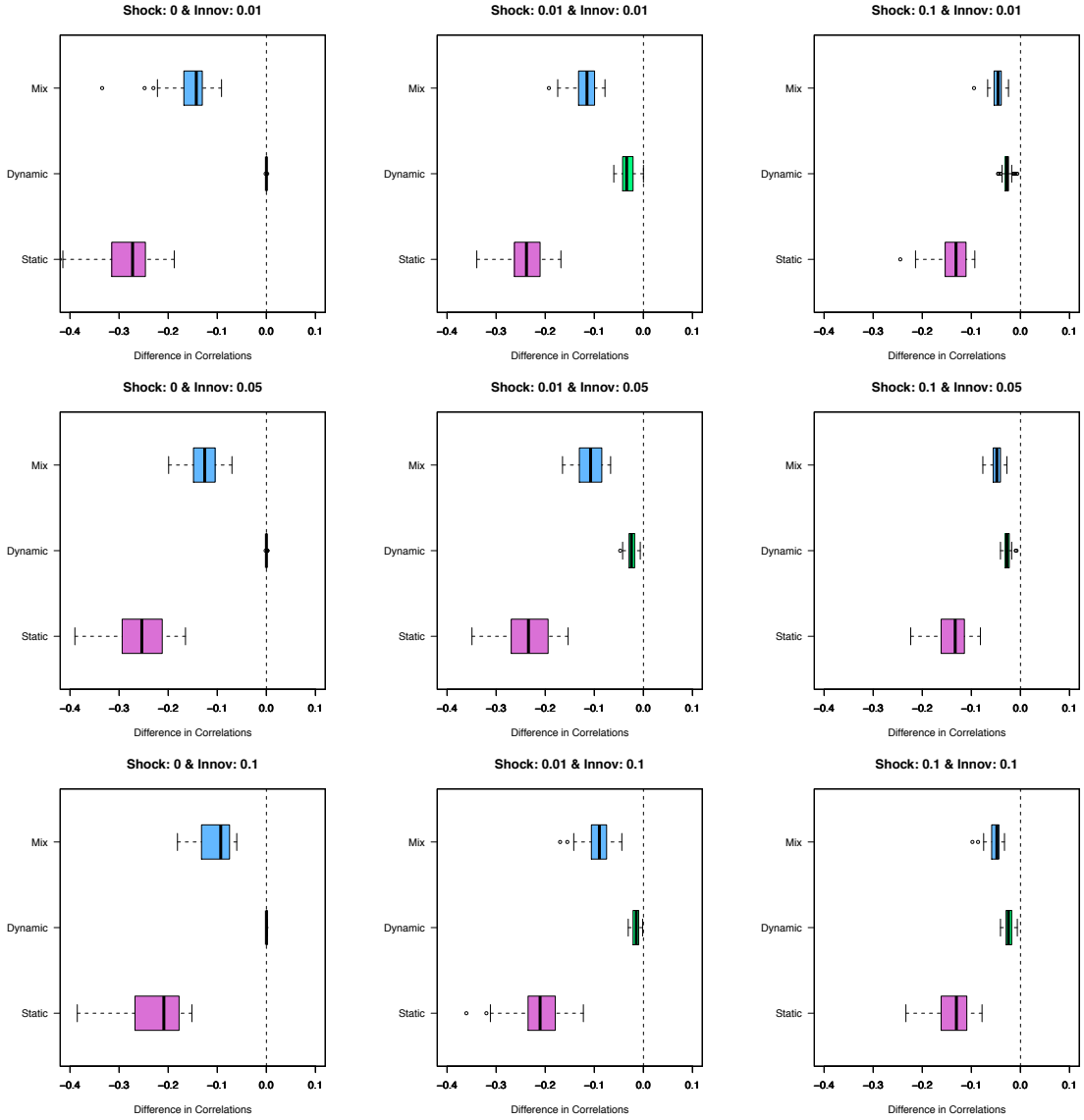
estimated ranking $\hat{\Theta}_{td} = (\hat{\theta}_{1,t,d}, \hat{\theta}_{2,t,d}, \ldots \hat{\theta}_{N-1,t,d}, \hat{\theta}_{N,t,d})$. We then compare this estimated ranking to the true ranking using the true latent trait $\Theta_t = (\theta_{1,t}, \theta_{2,t}, \ldots \theta_{N-1,t}, \theta_{N,t})$ by computing the median correlation (Kendall's Tau) value for each time period and across all posterior draws, such that

$$\text{Corr}_{\text{intra}} = \text{Median}(\tau(\Theta_t, \hat{\Theta}_{td})) \quad \forall t \in \mathbf{T}, \forall d \in \mathbf{D}. \tag{1}$$

Figure 7 displays the results of this analysis. The horizontal axis in each panel reports the difference in rank correlation produced by each modeling strategy using the robust model as the reference category. Negative values indicate the robust model is outperforming the alternative modeling strategy, while positive values indicate the opposite. The first column reports results when the true data generating process does not experience shocks, directly mirroring the assumptions made in the standard dynamic modeling strategy. When this is the case, we find that the robust model performs no worse than the dynamic model and is superior to both the static and mixture models across all values of the innovation parameter. In other words, the robust model performs no worse than the dynamic model even under conditions ideally suited to the assumptions made in for the dynamic model. The second and third column report results when the probability of a shock is set to 0.01 and 0.1, respectively. Among these cases, the robust model outperforms all other modeling strategies.

Two patterns are noteworthy. First, we find that while the mixture model often outperforms the static model, it is generally outperformed by the dynamic model. It therefore appears as though the mixture model is failing to adequately leverage the strengths of the static and dynamic models. It is possible that the modeling complexity of this strategy demands too much from the data, making it unable to effectively distinguish between instances where the dynamic and static modeling strategy is superior. Second, we find the robust model generally outperforms all other competing models under all conditions. These gains are the strongest when the probability of a shock is relatively high, but even when this

Figure 7: Comparing Robust to Alternative Strategies—Inter-Unit Rank Correlations



Note: Figure reports differences in the rank correlations between each model's estimates and the true values of the latent trait. To compare models, we take the difference between the rank correlation of the robust model with each alternative modeling strategy, such that negative values indicate the robust model is outperforming the alternative. Boxes indicate the distribution of these values across 50 independent data simulations – median values are reported with a black bar.

is not the case, the rankings produced by the robust model yield at least modest gains over the remaining models. This is strong evidence that the robust model is an improvement to current modeling techniques with little added complexity.

C.4 Within-Unit Rank Correlations

In the previous subsection, we report each model's ability to recover the cross-sectional ranking of each unit according to the value of its latent trait. Here, we report a second set of rank correlations that pertain to the models' ability to recover the rankings of the latent trait within each unit's time-series. The precise process by which this is done is detailed in the algorithm below, which mirrors the process we used to compute cross-sectional rankings.

Algorithm 2 Intra-Correlation Calculation

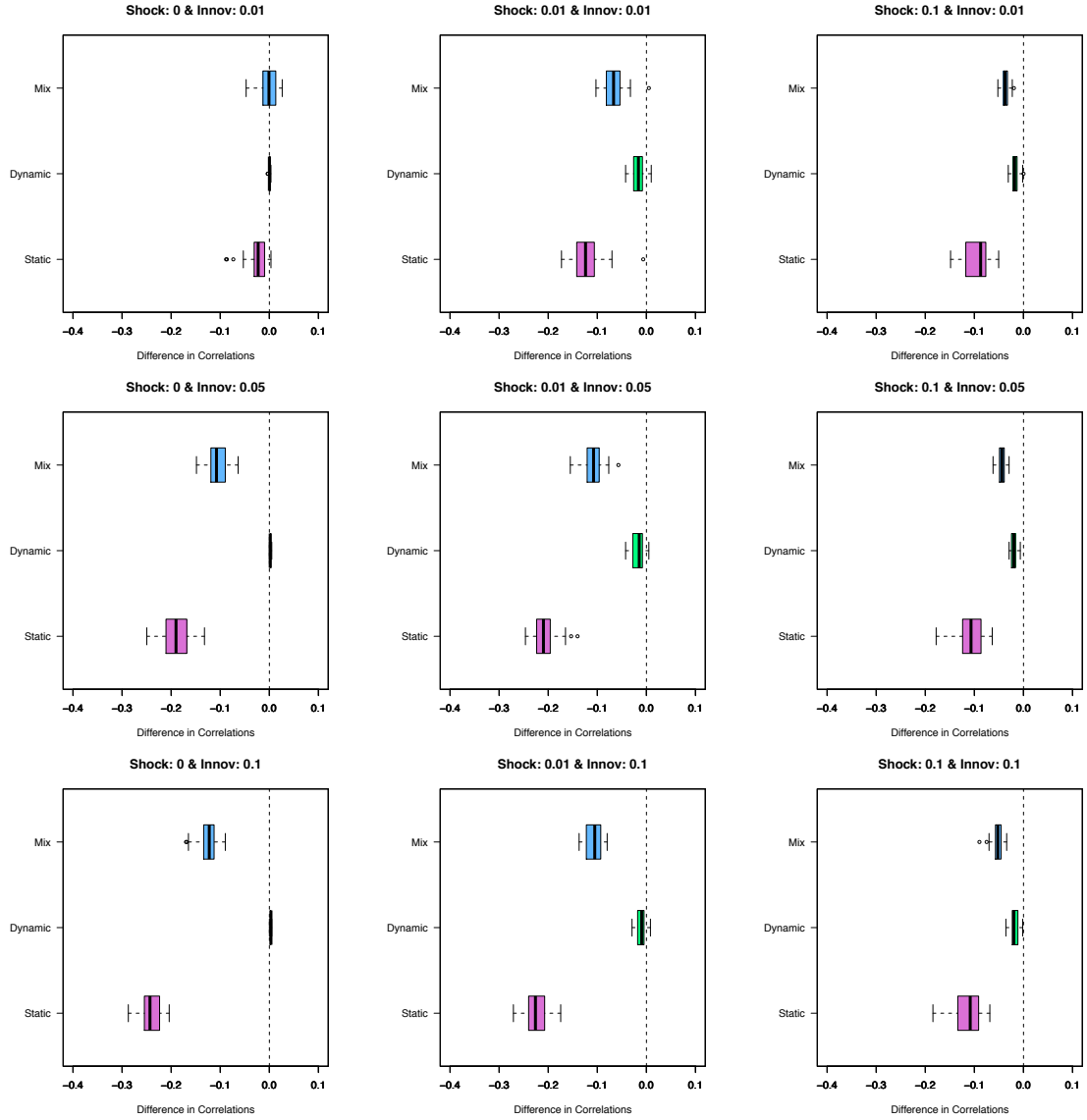
1: $\mathbf{D} = \{1, \dots D\}$ *Draws from the posterior*
2: $\mathbf{N} = \{1, \dots N\}$ *Units*
3: $\Theta_n = (\theta_{n,1}, \theta_{n,2}, \dots \theta_{n,T-1}, \theta_{n,T})$
4: $\hat{\Theta}_{nd} = (\hat{\theta}_{n,1,d}, \hat{\theta}_{n,2,d}, \dots \hat{\theta}_{n,T-1,d}, \hat{\theta}_{n,T,d})$
5: $\text{Corr}_{\text{intra}} = \text{Median}(\tau(\Theta_n, \hat{\Theta}_{nd})) \quad \forall n \in \mathbf{N}, \forall d \in \mathbf{D}$

Figure 8 reports the results from this analysis. As before, the horizontal axis in each panel reports the difference in rank correlation produced by each modeling strategy, using the robust model as the reference category. Negative values indicate the robust model is performing best, while positive values indicate that an alternative modeling strategy is producing superior results. When there is no probability of a shock occurring and the innovation parameter is set to 0.01, we find little difference between each of the four modeling strategies. This contrasts with the results we obtained from the cross-sectional rank correlation analysis, but is likely driven by the fact that there is less variation in the values of the latent trait within the time series relative to the cross-section. Under all other conditions, we find that the robust model generally outperforms the static, dynamic, and mixture models.

C.5 Cross-Validated Accuracy

We also perform a cross validation analysis to assess out-of-sample model performance. To do this, we simulate an additional manifest indicator of the latent trait that is left out from estimation of the measurement model. We then perform 10-fold cross-validation, partitioning

Figure 8: Comparing Robust to Alternative Strategies—Intra-Unit Rank Correlations



Note: The output of this figure pertains each model's ability to reproduce the true ranking of the latent trait's value within each simulated unit. The horizontal axis reports the difference in the rank correlations produced by the mixture, dynamic, and static models with that produced by the robust model. Negative values indicate the robust model is out performing the alternative. Rank correlations are computed on 1,000 draws from the posterior, and then taking the median value of these correlations. The distributions of these simulations are reported with box plots.

the simulated dataset into ten randomly assigned groups and running a regression model that predicts the held out indicator using the latent trait estimates as an independent variable. We then generate out-of-sample predictions for the held out partition of the data. This process is repeated ten times so that predictions are generated for the entire dataset. The

mean accuracy of the 10 folds is then used to evaluate how well we are able to predict the new indicator with different estimates of the latent variable.

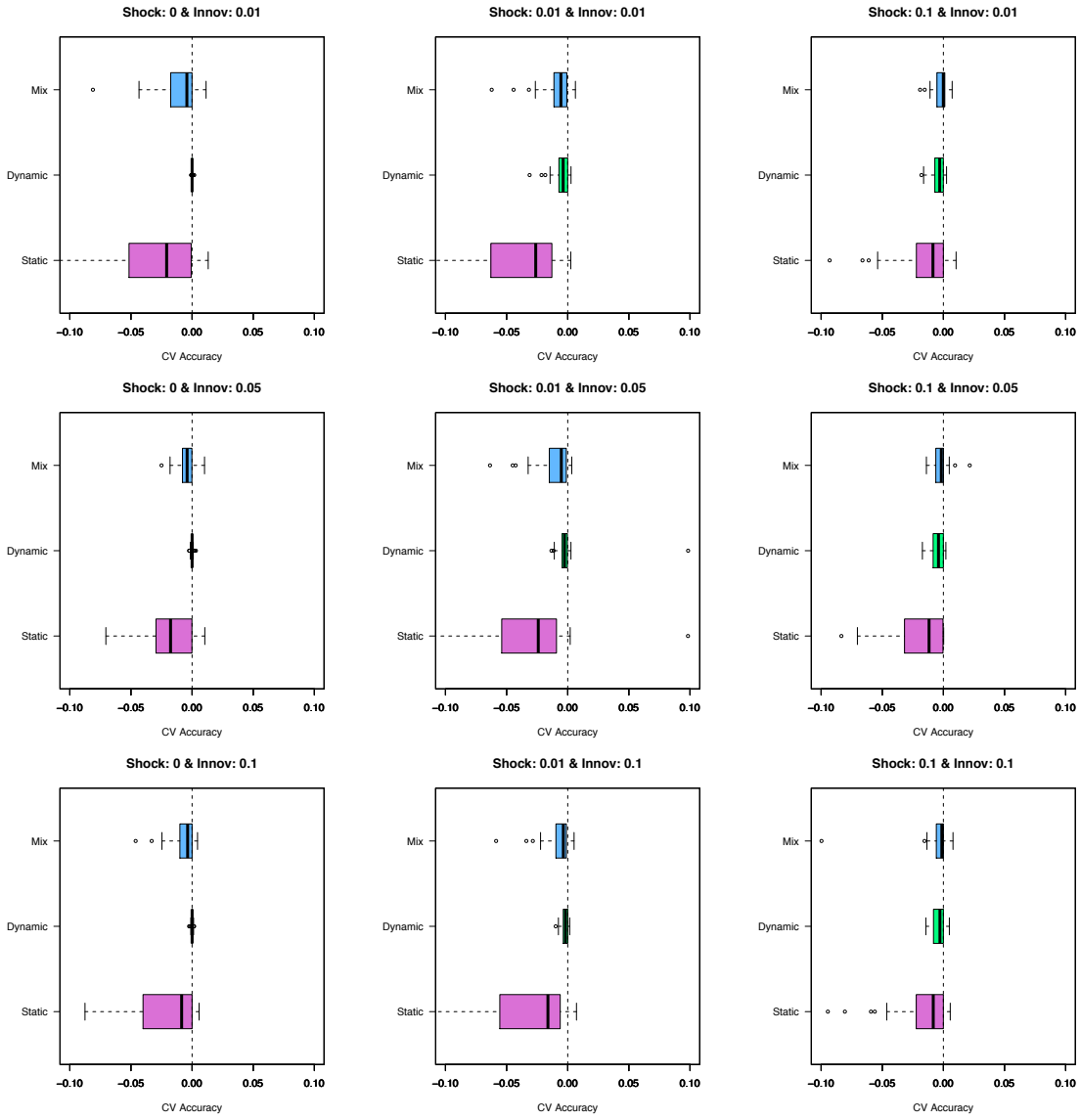
Algorithm 3 Cross Validated Accuracy

1: $y_{n,t,k} = \mathcal{I}\Big[\Lambda^{-1}\big(\alpha_k + \beta_k\theta_{n,t} + \mathrm{N}(0,1)\big)\Big] \quad k \in \{1-6\}$
2: Let $\hat{\theta}_{nt}$ be the median of the posterior distribution estimated with $Y_{n,t,1}, Y_{n,t,2}, \ldots Y_{n,t,5}$
3: Let Y_6^m by a random partition of $Y_{n,t,6}$, such that $Y_6^1 \cap Y_6^2 \cap \ldots Y_6^9 \cap Y_6^{10} = \varnothing$ and $Y_6^1 \cup Y_6^2 \cup \ldots Y_6^9 \cup Y_6^{10} = Y_{n,t,6}$
4: **for all** $m \in \{1-10\}$ **do**
5: Let $\bar{Y}_6^m = Y_6 \cap Y_6^m$ & $\bar{\Theta}^m$ is the set of related $\hat{\theta}_{ij}$
6: Estimate $\hat{\alpha}, \hat{\beta}$, where $\bar{Y}_6^m = \Lambda^{-1}\big(\alpha + \beta\hat{\theta} + \epsilon\big)$
7: $\hat{Y}_6^m = \mathcal{I}\Big[\Lambda^{-1}\big(\hat{\alpha} + \hat{\beta}\bar{\Theta}^m\big)\Big]$
8: $P_m = (\hat{Y}_6^m == Y_6^m)/\mathrm{Length}(Y_6^m)$
9: **end for**
10: $\mathrm{CV} = \frac{1}{10}\sum_{m=1}^{10} P_m$

Figure 9 reports the results of this analysis. The horizontal axis reports the difference in the percentage of observations correctly predicted by each model with the robust model again serving as the reference category. As before, model performance is reported across data sets with the shock and innovation standard deviation parameters set to a variety of reasonable values. When there is no probability of a shock, the robust model outperforms the static and mixture models and performs no worse than the dynamic model. As the values of the shock parameter is increased (columns two and three), the robust model continues to outperform the static and mixture model, and now produces slightly more accurate estimates than the dynamic model.

We perform an additional test to evaluate how well the models perform out-of-sample on observations that have recently experienced shocks. To do this, we partition the simulated data into units that occur during or within two time periods of a shock and those that are not proximate to a shock. We then train a model on the group that did not experience shocks (again using the latent variable estimates as the independent variable and held out items as the dependent variable) and generate out-of-sample predictions for the group proximate

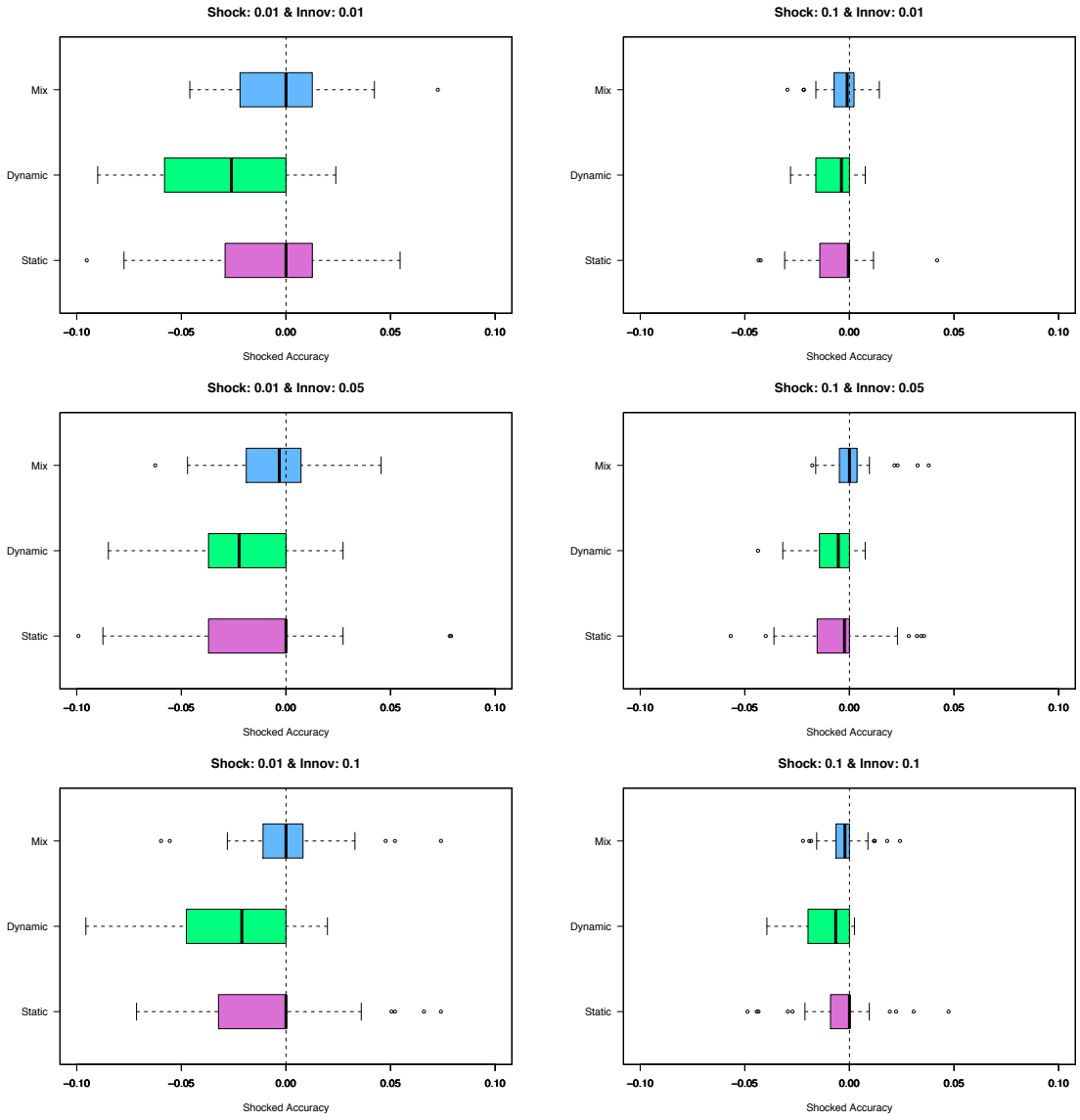
Figure 9: Difference in Cross Validated Accuracy



Note: Difference in 10-fold cross validated error between models, using the robust as the baseline. Negative values indicate that the robust model is out performing other models. Cross validate error is calculated on an item that was held out from estimation of the latent variable. Distributions are estimated by generating 50 independent sets of data.

to shocks. Model performance is reported in Figure 10. In all cases the robust model performs as well, or better than the other three models. When shocks are relatively unlikely, the difference between the robust model and the dynamic model and static model is at its most extreme. Interestingly, the mixture model performs relatively well in this test, and the median accuracy for the mixture model approaches the median accuracy for the robust

Figure 10: Difference in Cross Validated Accuracy for Time Around Units Shocks



Note: Difference in 10-fold cross validated error between models, using the robust as the baseline. Negative values indicate that the robust model is out performing other models. Cross validate error is calculated on an item that was held out from estimation of the latent variable. Distributions are estimated by generating 50 independent sets of data.

model.

Overall, we find that the robust model tends to produce slightly more accurate out-of-sample predictions than the alternative modeling strategies. While these differences are often modest, they nevertheless indicate that the results uncovered elsewhere in this section are not driven by over-fitting.

C.6 Difference Between Time Periods

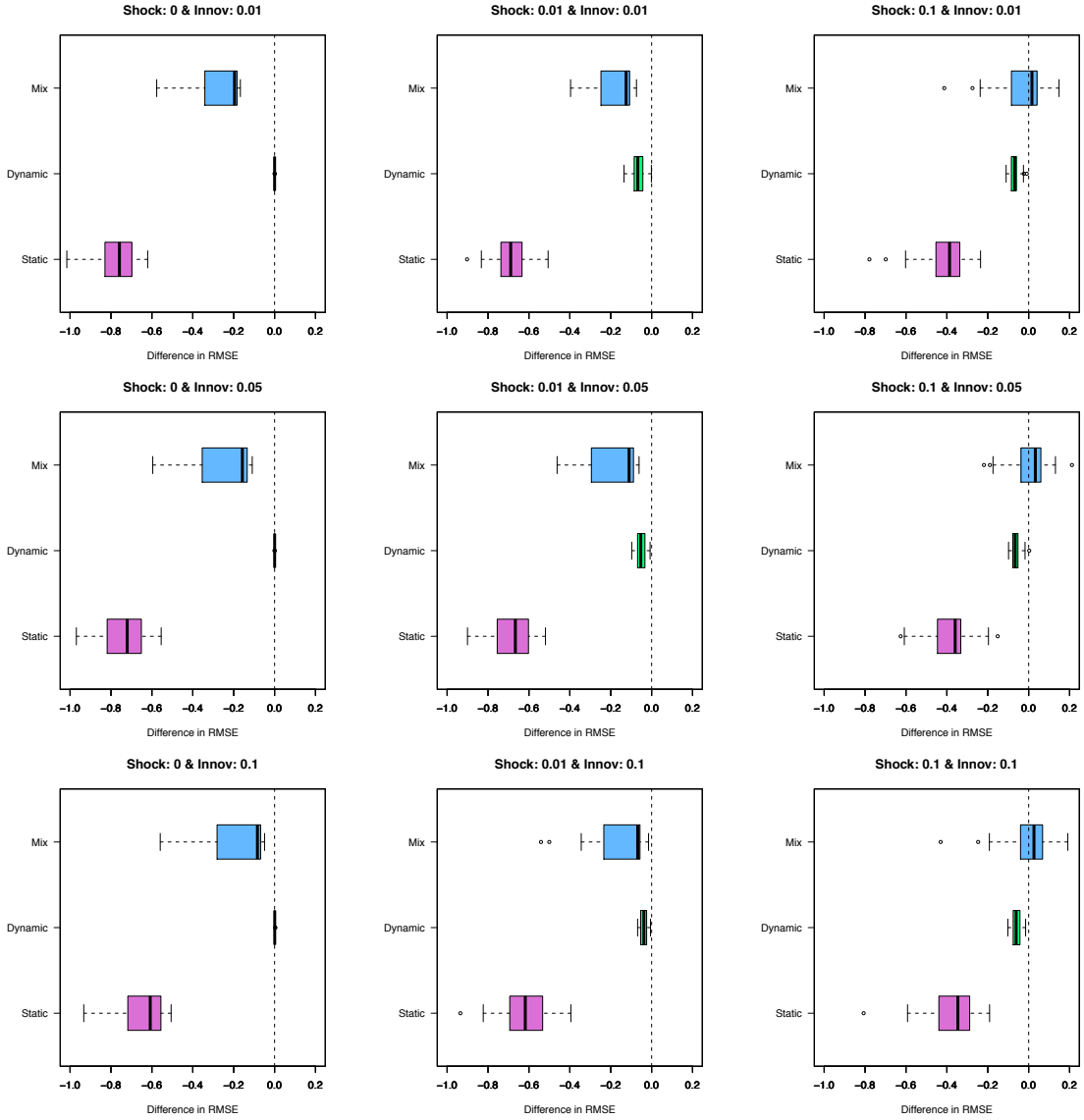
Finally, we examine the ability of the different models to estimate temporal changes in the latent trait. We calculate the change between time periods for each unit's true latent value ($\delta_{n,t-1} = \theta_{n,t} - \theta_{n,t-1}$) and the change between time periods for the estimated latent variables from each of the four proposed models ($\hat{\delta}_{n,t-1} = \hat{\theta}_{n,t} - \hat{\theta}_{n,t-1}$). A model is preferred if the true change ($\delta_{n,t-1}$) is most similar to the change estimated by the latent variable model ($\hat{\delta}_{n,t-1}$). We use Root Mean Squared Error (RMSE) to asses the degree of similarity: RMSE = $\sqrt{\frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{t} (\hat{\delta}_{n,t} - \delta_{n,t})^2}$. A lower RMSE statistic indicates increased accuracy in the ability for the latent variable to estimate overtime change.

Results are reported in Figure 11. We again report differences in RMSE so that negative values indicate that the robust model is outperforming the comparison model.[19] Not surprisingly, the static models do not fare well at this type of posterior prediction. By minimizing bias, the static model increases uncertainty around each unit estimate which makes it difficult to predict changes when they occur. All of the other models fare better than the static model. When the probability of a shock is 0, the robust and dynamic versions of the model both fare equally well at predicting the temporal changes. However, as the probability of a shock to the system increases, the robust model begins to perform better than the dynamic version of the model. In some cases the mixture model actually outperforms the robust model, but even here the difference is marginal.

In sum, the simulated results indicate that the robust model generally outperforms standard modeling strategies across a variety of metrics when considering data that is generated by a dynamic process that experiences shocks. Even in instances where units do not experience shocks, we find that the robust model performs no worse than the dynamic model suggesting that there is no loss of performance even when the data are ideally suited to the

[19] To accomplish this we subtracted the comparison model from the robust model instead of subtract the robust model from the comparison model as we had previously done.

Figure 11: Difference in Root Mean Square Error



Note: Difference in root mean squared error (RMSE) for the estimated median change between time periods and the real change between time periods. Negative values indicate that the robust model is outperforming an alternative modeling strategy; positive values indicate the opposite.

assumptions of the dynamic model.

C.7 Estimating Degrees of Freedom

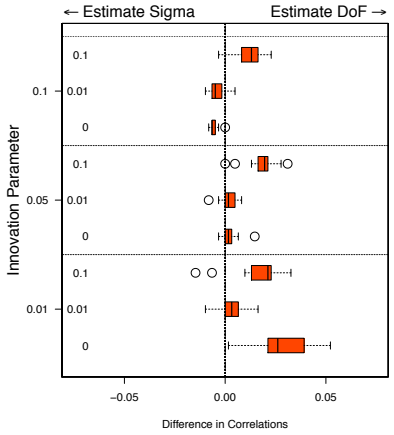
In addition to estimating models with a free σ we also investigated the possibility of estimating ν directly after fixing σ to a predetermined value. We fixed σ at 0.05 (which is in the middle of our true σ) and put a Gamma$(2, 0.1)$ prior on ν bounded at $\nu > 1$. In Figure 12 we compare model output to estimating ν versus estimating σ on the same generated datasets. For these simulations and the next in the subsection we only estimate the results on 25 simulations instead of 50 as in the simulations presented in the earlier part of this section.

Figure 12 does not provide evidence to prefer any single model. In some cases estimating ν improves model accuracy while for other data generating processes the model estimating σ is preferred. In almost all cases the differences are minimal. The only case where this is not true is validation using Root Mean Squared Error (RMSE) of the change between time points. For parameter values, estimating σ significantly improves model fit.

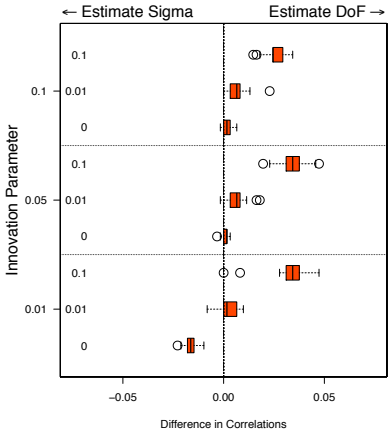
It is important to note also that estimating ν is difficult. As ν approaches 1 the Student's t-distribution approaches the Cauchy distribution which has an undefined variance and expected value. This can compound problems with rotational identification when a researcher has to identify a latent variable by providing tight priors on a set of actors (as Martin and Quinn (2002) do). As estimates of the degrees of freedom approaches 0, then the innovation variance increases and so it becomes likely for a unit to jump from positive values to negative values across a time period.

Figure 12: Differences in Estimate σ and ν

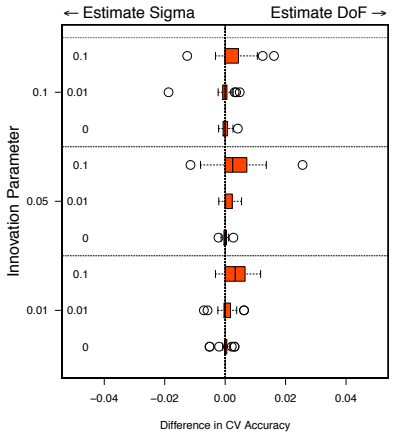
(a) Intra Unit Correlations



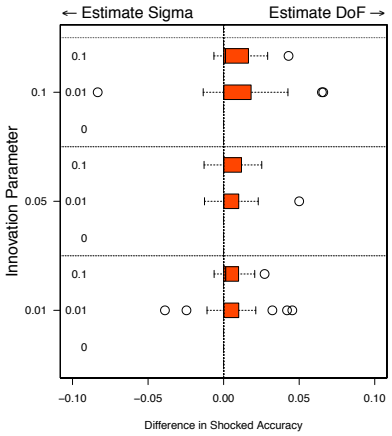
(b) Intra Time Correlations



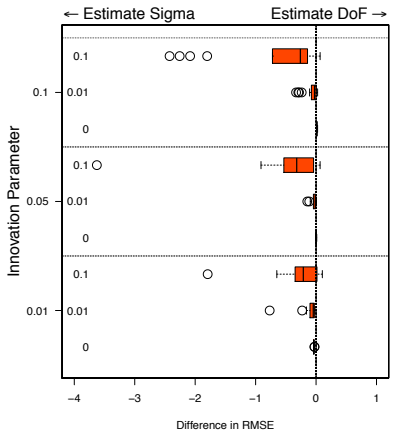
(c) Cross Validated Predictions



(d) Predictions to Shocked Units



(e) RMSE



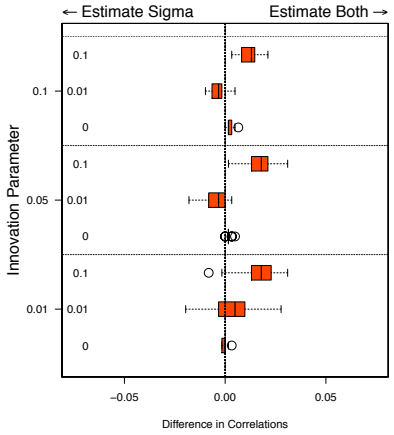
Note: Differences in variety of measures of statistical accuracy for models estimated with free σ versus those estimated with a free ν. Differences are taken so that negative values indicate that estimating σ performs better. The number on the Y axis indicate what σ was set to in the data generating process, the numbers just to the left of the boxplots indicate what the probability of shock was.

C.8 Estimating Both Degrees of Freedom and Scale Parameter

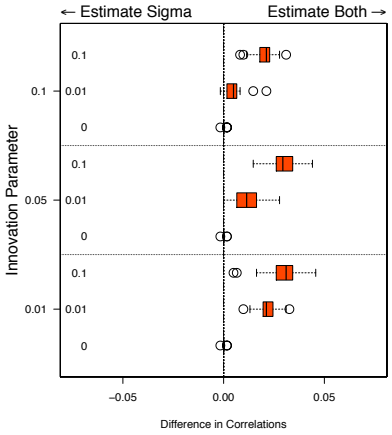
In addition to estimating models with a free σ or a free ν, we also investigated the possibility of estimating both σ and ν. These results are consistent with those obtained from models comparing the estimation of a a free σ and a fixed ν to the estimation of a fixed σ and a free ν.

Figure 13: Differences Between Estimating only σ and Estimating σ and ν

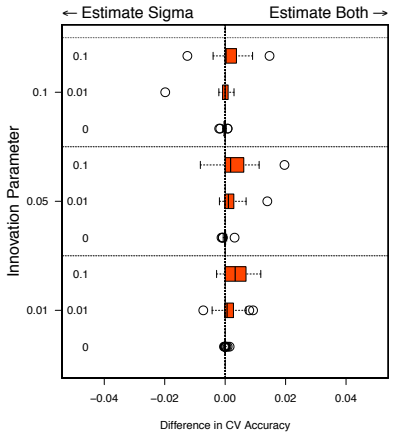
(a) Intra Unit Correlations



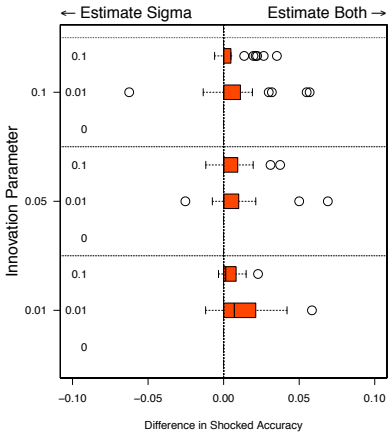
(b) Intra Time Correlations



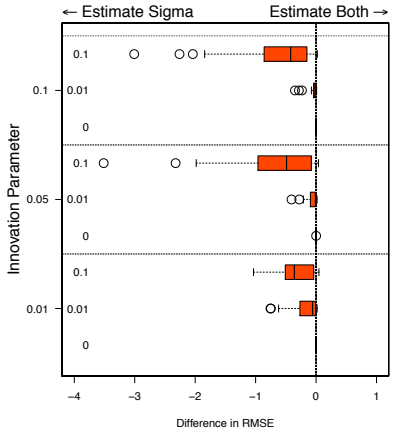
(c) Cross Validated Predictions



(d) Predictions to Shocked Units



(e) RMSE



Note: Differences in variety of measures of statistical accuracy for models estimated with free σ versus those estimated with a free ν and free σ. Differences are taken so that negative values indicate that estimating only σ performs better. The number on the Y axis indicate what σ was set to in the data generating process, the numbers just to the left of the boxplots indicate what the probability of shock was. See Appendix C for more details on the different estimates of model accuracy.

D Fixing the Innovation Variance

As discussed in the main manuscript, sparse data structures sometimes lead researchers to set the innovation variance (σ) to a fixed value instead of estimating this quantity from the data. Whenever this is the case, there is a risk that researchers may assign a value for σ that is either too low or too high. Setting very small values to σ can be particularly problematic in a time-series setting, as model estimates of the latent trait become increasingly time invariant as σ decreases towards 0. Setting σ to high values can be equally problematic. As $\sigma \to \infty$ the model begins to generate parameter estimates independently for each time period. This is problematic if there is not sufficient information at each time period to identify the model – indeed it is this issue which lead Martin and Quinn (2002) to fix the value of σ in the first place. In short, setting σ to a value that is too low risks biased model estimates, while setting a value that is too high may result in a model that is not identified. The latter issue will be apparent to practitioners – an unidentified model can often be diagnosed through poor convergence diagnostics – but the former risks biases in model estimates that might otherwise be overlooked by researchers if σ is not chosen with care. In the remainder of this section we conduct a simulation analysis to demonstrate how researchers may identify and correct these biases.

Figure 14 displays the results of a simulation analysis conducted to demonstrate the types of modeling biases likely to result when one assigns values to σ that are too low.[20] The simulation constructs a latent trait, x that follows a random walk with drift, increasing in expectation by 0.15 every time period with standard deviation of $\sqrt{0.05}$ (formally $x_t = x_{t-1} + \text{N}(0, \sqrt{0.05})$) over 100 observation periods. Four manifest indicators, Y_t^j, are generated from the latent trait where $Y_t^j = \alpha_j + \beta_j X_t + \text{N}(0, \epsilon_j)$. We fix α_1 to 0 and β_1 to 1, while $\beta_j \sim \text{U}(0, 2)$ and $\alpha_j \sim \text{U}(-2, 2)$ for the remaining three items. This is done so that we can re-

[20] We are grateful to an anonymous reviewer for suggesting this simulation and providing an example of it.

estimate the latent trait on the same scale as the true latent trait. For all items $\epsilon_j \sim \text{U}(.5, 2)$. We use these indicators to generate estimates of the latent variable using both the dynamic model priors and the robust model priors. For each of these, we fix the innovation standard deviation at either the true standard deviation, a fifth of the true standard deviation or five times the true standard deviation. In Figure 14, we plot these estimates with a 95% credible interval (displayed in blue along) with the true values (displayed in points).

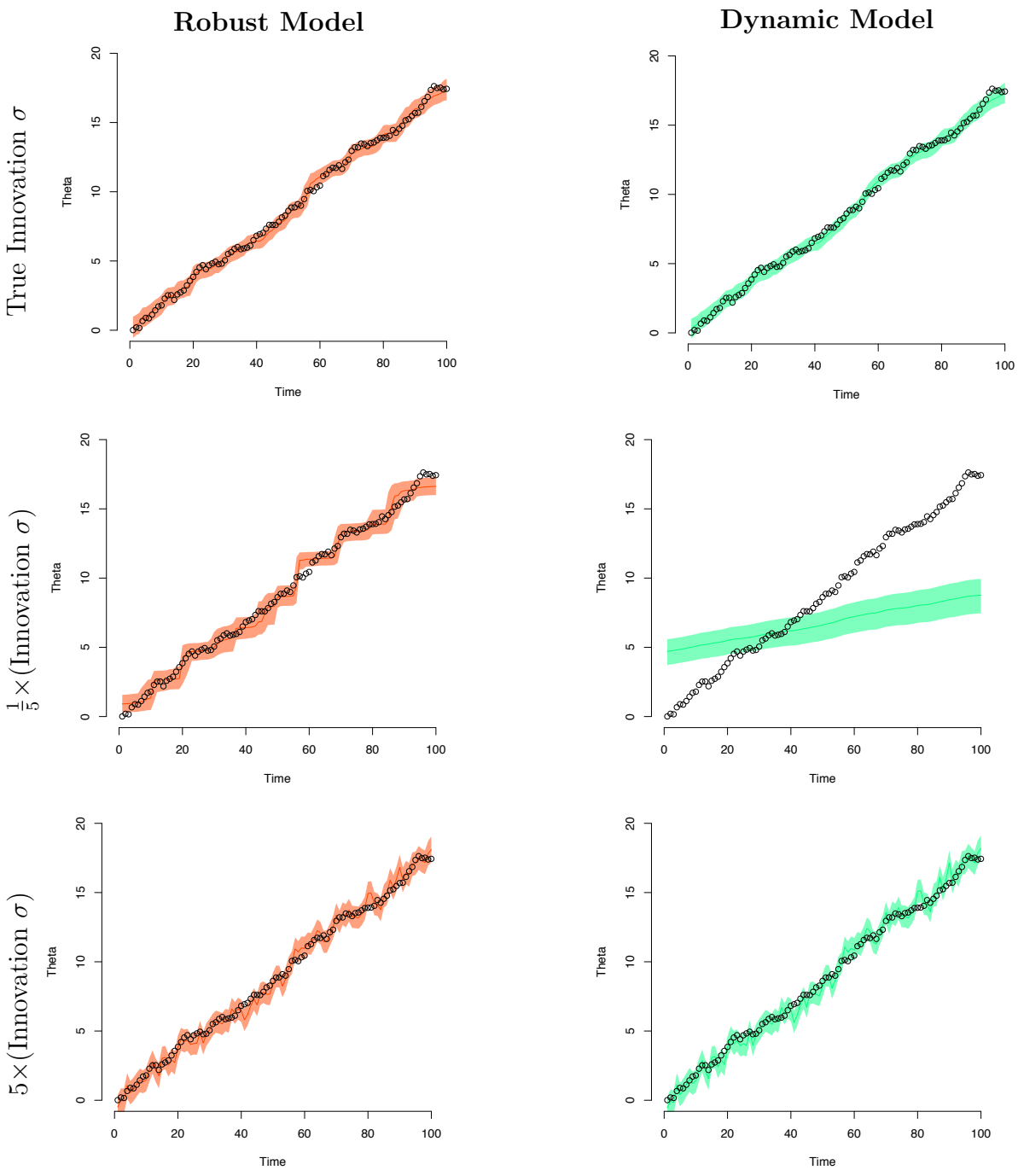
The first row in Figure 14 shows the estimates when the innovation standard deviation is set to the correct value. Both the robust model and the dynamic model behave similarly and the true latent variable is almost always contained within the credible intervals obtained from both models. In the second row we show the estimates when the innovation standard deviation is assumed to be much smaller than it is–in this case when it is fixed to a fifth the truth standard deviation. Under these conditions the robust and dynamic models perform very differently. For almost all cases, the credible intervals for the robust model contain the true value of the latent trait, but instead of fitting a smooth linear trend, the model fits an approximately stepwise function to the latent trait. In applied work, without further investigation, an applied researcher might conclude that the latent trait experiences a series of temporal shocks when this is not, in truth, the case. The results of the dynamic model are equally problematic. Here, model estimates of the latent trait follow a smooth linear trend, but one that is far more shallow than the true, underlying process. As a result, the dynamic model's credible intervals rarely contain the true values of the latent trait. The applied researcher might therefore incorrectly conclude that there is little temporal variation in the latent trait.

Finally the last row shows what happens when the innovation standard deviation is assumed to be much larger than it is in truth. Here there appears to be more variation overtime with a lot of change between each time period. This happens in both the case of the dynamic model and the robust model and reflects the fact that the models are not taking sufficient temporal information into account. As stated above, however, setting σ to a value

that is even high might result in a lack of model identification when modeling sparse data structures.

To unpack this issue further, we conduct a second simulation similar to those conducted in Appendix C. Again, we estimate a dataset of 50 units over 50 time periods with a true innovation standard deviation of $\sqrt{0.1}$. Next we estimated a dynamic model and robust model with fixed innovation standard deviation for each with the overall innovation standard deviation fixed at $\sqrt{0.001}$. Figure 15 plots a unit drawn from these datasets that we thought was illustrative. In the left column we show estimates using the misspecified variance and the right column shows it with the correct variance. When the innovation variance is specified to be smaller than it is we again get estimates that do not adequately capture the true path of the latent variable. In the case of the robust model we have two regimes with no change with a shock between them. In the dynamic model we see very little change overtime. These problems vanish in both cases when the correct innovation variance is assigned.

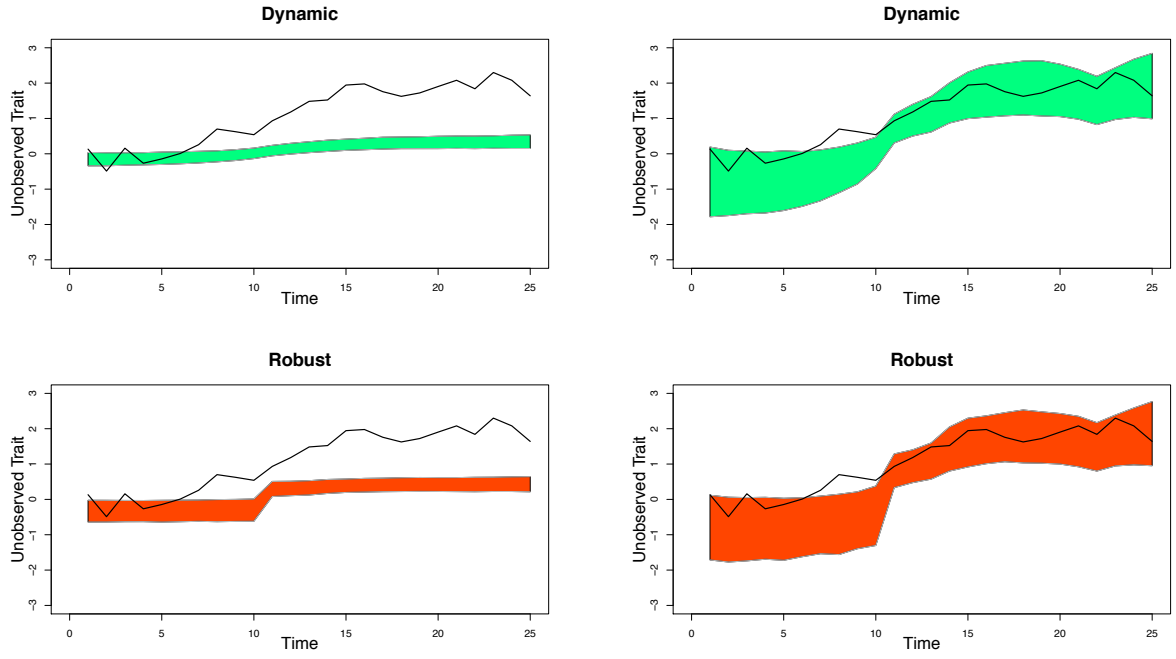
In sum, as we note in the main text, practitioners that fix the value of the innovation variance of latent variable models should take additional steps to validate their models. Like Martin and Quinn (2002), we recommend that practitioners evaluate model performance with the innovation variance set to multiple values to ensure that model estimates are not significantly biased. Again, these issues are unlikely to manifest when users estimate the innovation parameter directly from the data and so we recommend doing so whenever possible.



Note: Simulated sparse data structure that requires the researcher to set the innovation standard deviation (σ) to a fixed value instead of estimating this quantity from the data. Whenever this is the case, there is a risk that researchers may assign a value for σ that is either too low or too high. In both cases, the robust model and the dynamic model behave similarly, properly recovering the latent variable when the innovation variance is set to the correct value (top row). This is not the case for either model when the innovation standard deviation is set too low (middle row) or too high (bottom row). Like Martin and Quinn (2002), we recommend that practitioners evaluate model performance with the innovation standard deviation set to multiple values to ensure that model estimates are unbiased.

Figure 15: True and Estimated Latent Scores for Different Fixed Variance

(a) $\sigma = \sqrt{0.001}$ *(b)* $\sigma = \sqrt{0.1}$



Note: Estimated latent variables for a single unit when fixing the total innovation variance. The true innovations standard deviation is $\sqrt{0.1}$ and so the column on the left assumes a significantly smaller variance.

E Innovation Variance Check for Martin and Quinn

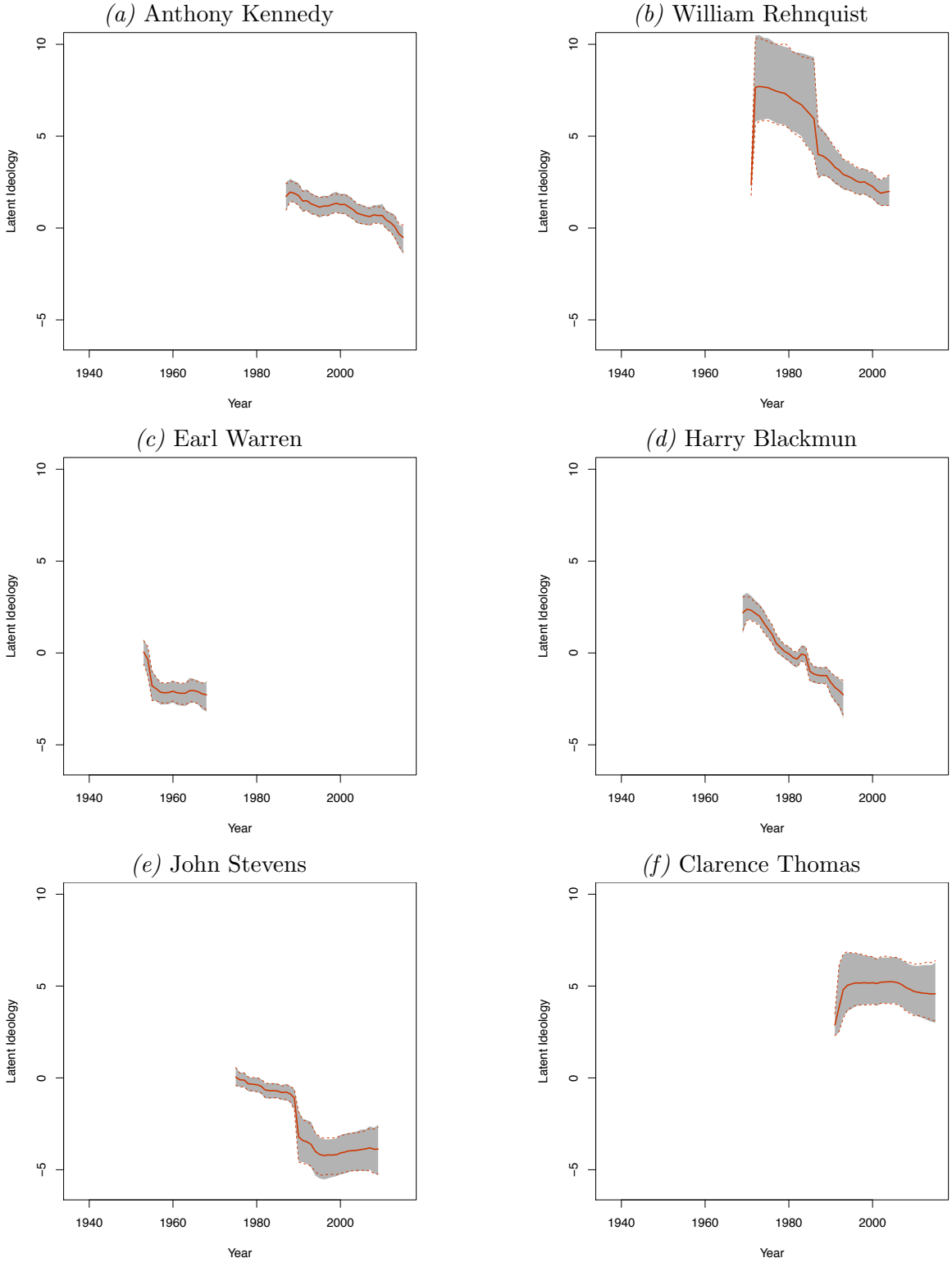
As discussed above, it is particularly important to validate models where the innovation variance has been fixed to a specific value. To do this in the case of Martin and Quinn (2002) we followed two tracks. First, we re-estimated both the dynamic and robust model with the innovation variance set to higher values than those provided by Martin and Quinn to examine whether it produced substantive changes to our results. Next we used a relatively restrictive prior to estimate a version of Martin and Quinn (2002) where we directly estimated the innovation variance. We continue to find that the robust model outperforms the dynamic model for each of these modeling choices.

E.1 Increasing The Innovation Variance

In the main manuscript, we compare the robust model to a standard dynamic model with the total innovation variance set to 0.10. Here, we compare these models with the total innovation variances increased to 0.15. The WAIC statistic for these models is comparable to the models presented in the manuscript with the robust model continuing to outperform the dynamic model (a difference of 185 with a standard error of 15.3). The posterior predictions for the models in the manuscript perform similarly to posterior predictions from these additional models as well. The accuracy for the robust model with the inflated variance was 72.97% and the dynamic was 72.84%, which was as a slight improvement compared to what was presented in the manuscript.

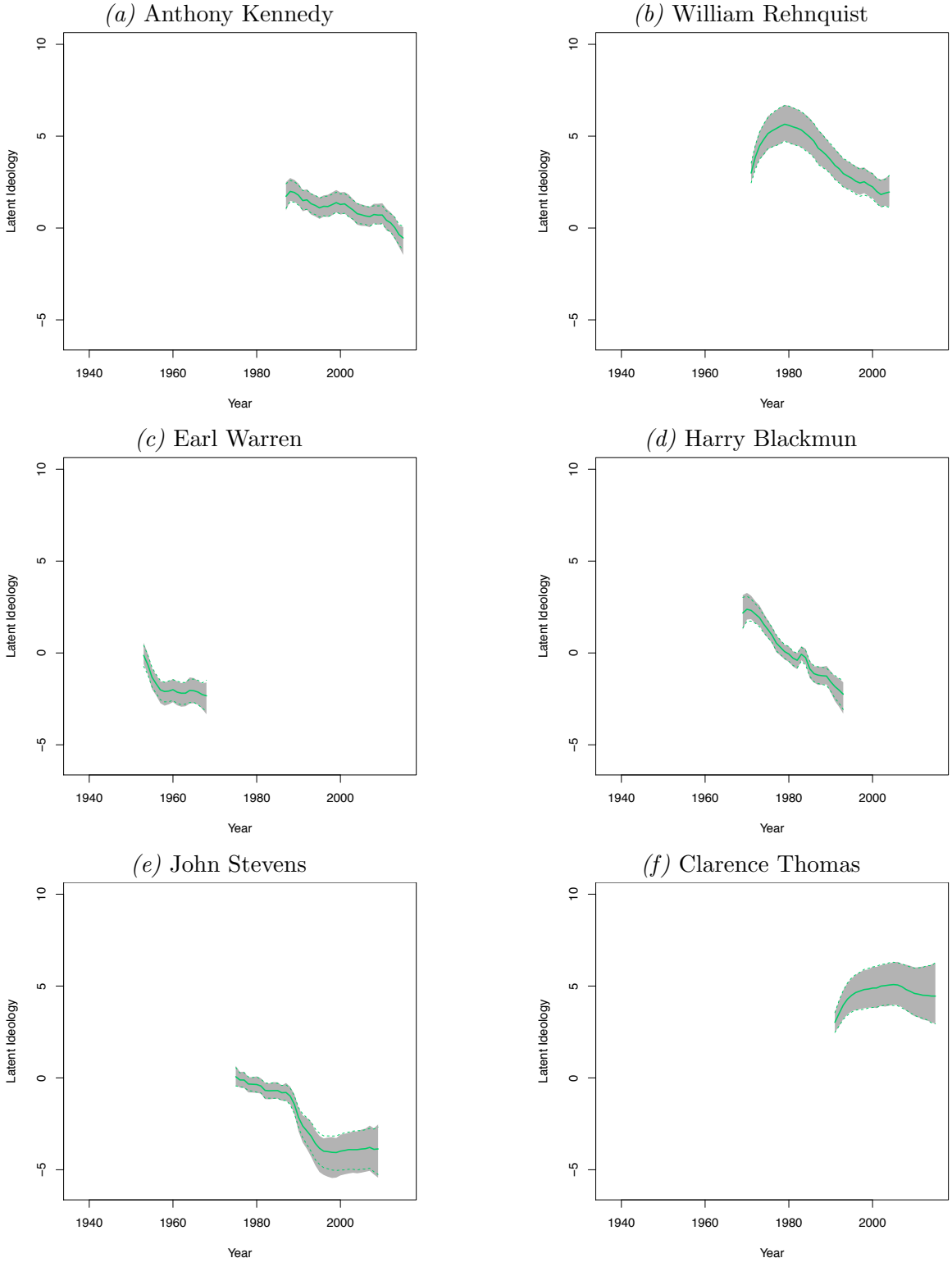
Finally, and more importantly, the observed temporal dynamics of the estimated latent variables are very similar. The inflated variance estimates correlate to the original estimates at 0.9998 for both models. In Figure 16 we plot the robust latent variables presented in the main text with the inflated variance version in gray. A similar pattern is revealed in all cases. In Figure 17 we do the same but the with the dynamic model, again we find a similar pattern. Overall, the results are consisteint between models with the total innovation variances increased from 0.10 to 0.15.

Figure: 16: Robust with Fixed Variance vs. Robust with Inflated Fixed Variance



Note: Two sets of judicial ideology scores, which are estimated using two different versions of the latent variable model (both robust). The robust model with fixed variance set to the Martin and Quinn (2002) value is represented by the solid red line for the median value and the dashed lines for the 2.5 and 97.5 credible interval. The robust model with fixed variance set to a larger value is represented by the shaded area, which is the 2.5 to 97.5 credible interval.

Figure: 17: Dynamic with Fixed Variance vs. Dynamic with Inflated Fixed Variance



Note: Two sets of judicial ideology scores, which are estimated using two different versions of the latent variable model (both dynamic). The dynamic model with fixed variance set to the Martin and Quinn (2002) value is represented by the solid green line for the median value and the dashed lines for the 2.5 and 97.5 credible interval. The dynamic model with fixed variance set to a larger value is represented by the shaded area, which is the 2.5 to 97.5 credible interval.

E.2 Estimated Variance

Finally, as an additional check we estimated the σ innovation parameter directly. To accomplish this we set a relative tight prior on σ of Beta(1,2) which constrains σ to be between 0 and 1, and placing half of the probability mass at $\sigma \leq .25$. We continue to fix a very tight variance on Douglas (0.0001) because of his unique voting pattern. The total innovation variance estimated in both the dynamic and the robust models are larger than those set by Martin and Quinn (2002) but do not change the substantive conclusions drawn in the main text.

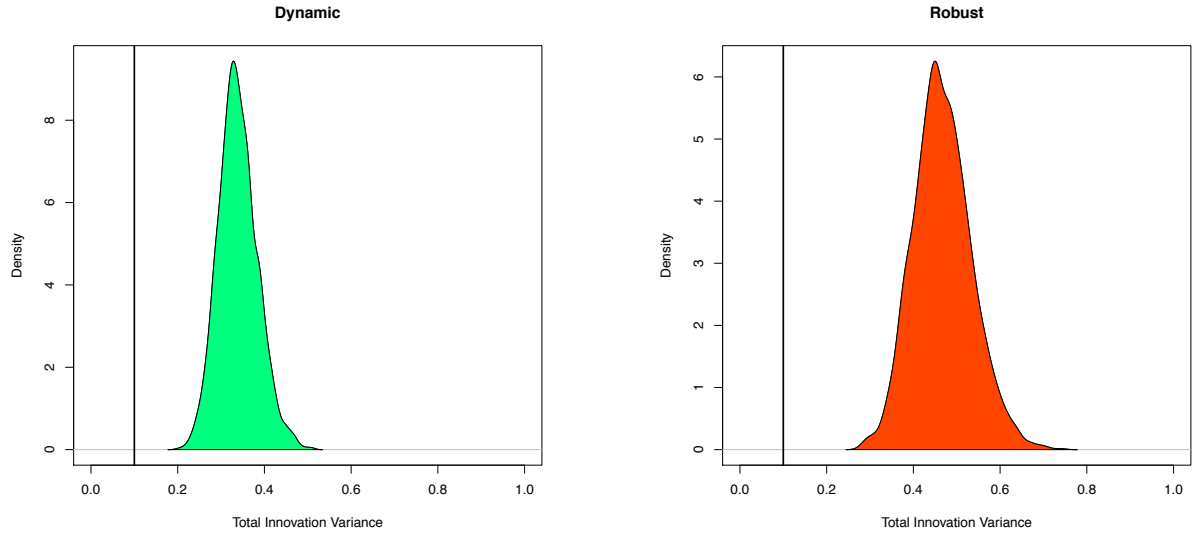
Before discussing the ideology patterns we note that using WAIC and posterior predictive checks we continue to find that the robust model performs better. The difference in WAIC is 261 with a standard error of 16.3, with the robust model out performing the dynamic. The posterior predictive accuracy of the robust model is 73.22% while for the dynamic it is 73.01%, the largest predictive accuracy found in any of the models presented here.

In Figure 18 we plot the innovation variance estimates for the two models. In the dynamic model this equates to squaring the estimated σ while in the robust we square it and then multiply it by 2 to account for the 4 degrees of freedom. The horizontal line is the variance set by Martin and Quinn (2002). In both cases the estimated variance is substantively larger than what was used by Martin and Quinn (2002).

The larger variance does not substantively change the patterns found in the main text. The latent variables with the estimated variance correlate at 0.995 (robust) and 0.997 (dynamic) with the original models. To demonstrate this we plot the fixed variance model's estimates of the latent variable over top of the estimated variance models in Figure 19 and 20. Although the estimated latent variables do not track perfectly compared to what is presented in the main text, they remain very similar.

Finally, to emphasize this points we recreate the figures in the main text but this time using just the estimated variance models. This is Figure 21. There are still important differences for Rehnquist and Thomas when comparing the robust to the dynamic model

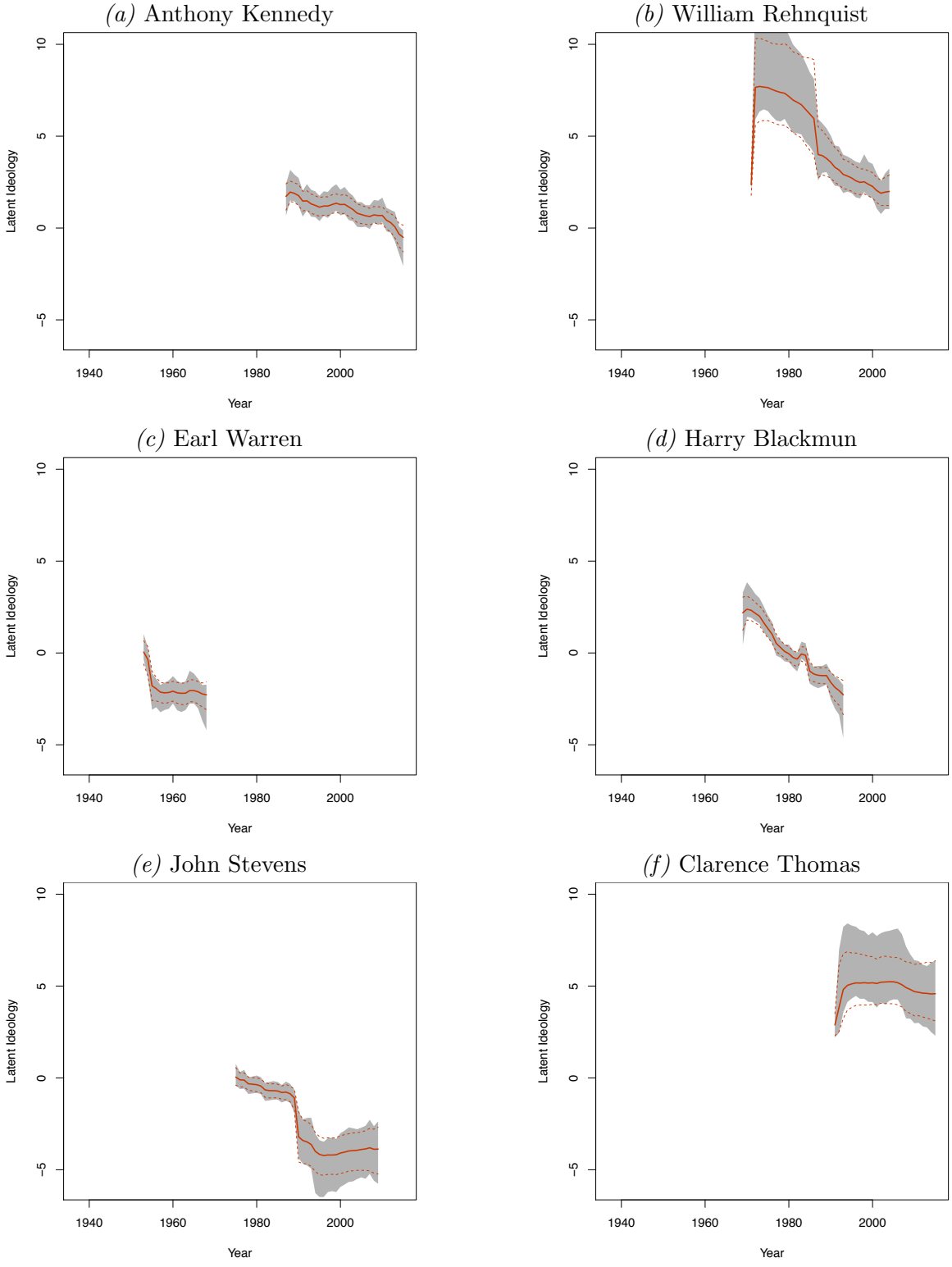
Figure 18: Estimated Total Innovation Variance



Note: Estimated total innovation variance for each model. For the dynamic model this is calculated by taking $\hat{\sigma}^2$ for the robust model it is calculated by taking $\hat{\sigma}^2 \frac{\nu}{\nu-2}$ where $\nu = 2$. The horizontal line is the original variance set by Martin and Quinn (2002).

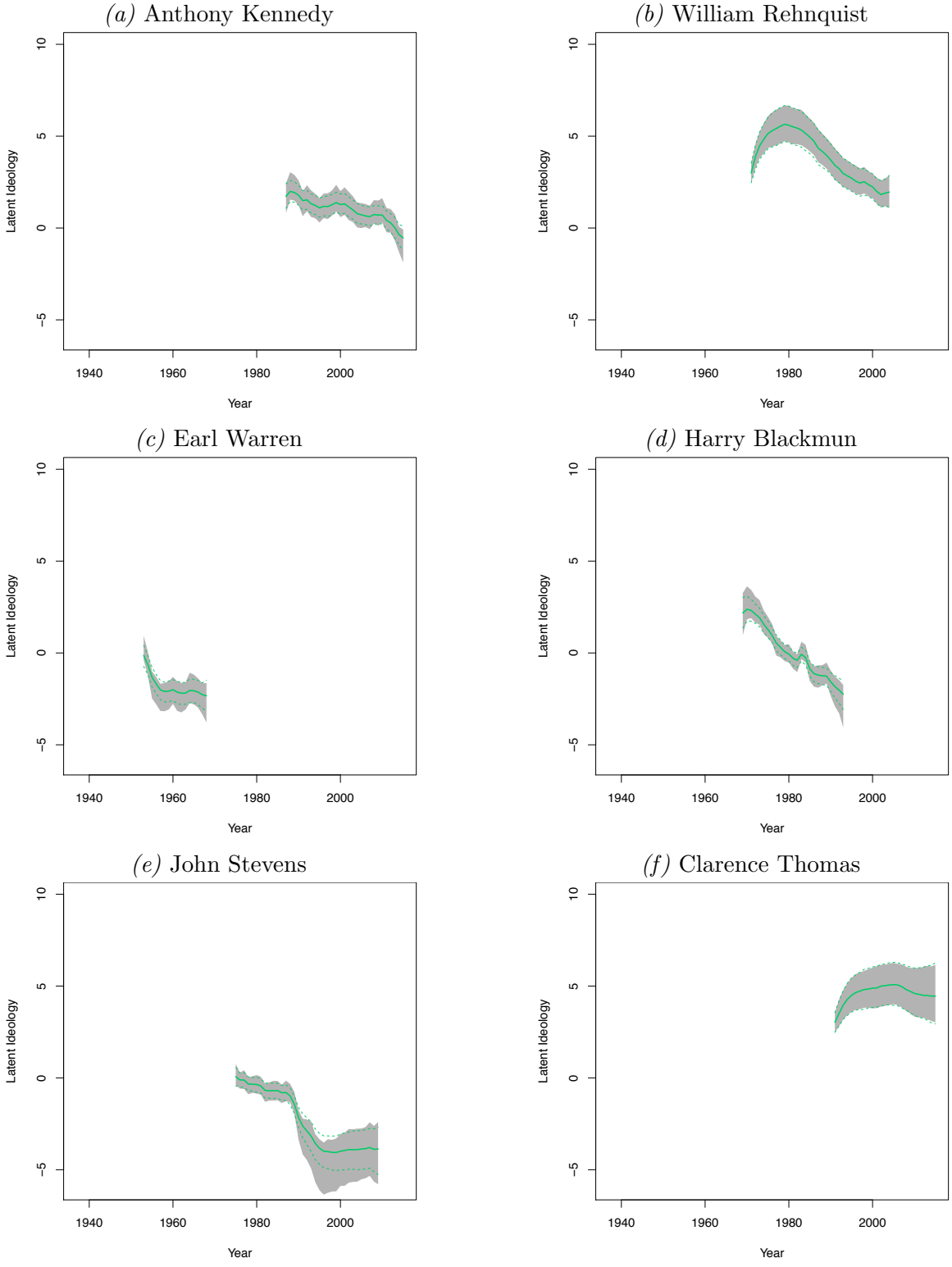
though there is now clear convergence for Warren between the robust and the dynamic models. The robust model then continues to show interesting patterns that might better illuminate patterns of judicial ideology for future scholars even when the innovation variance is directly estimated.

Figure: 19: Robust with Fixed Variance vs. Robust with Estimated Variance



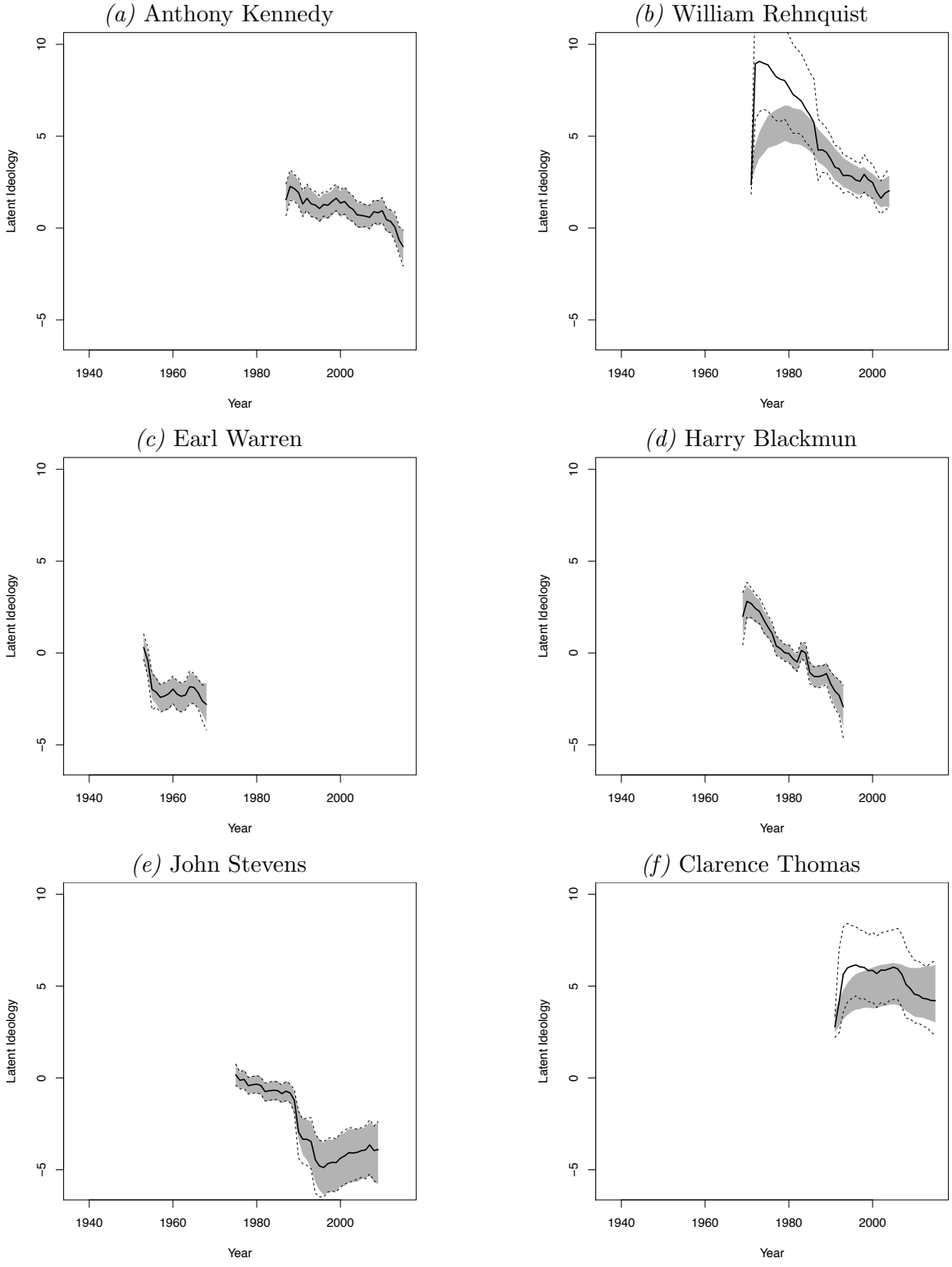
Note: Two sets of judicial ideology scores, which are estimated using two different versions of the latent variable model (both robust). The robust model with fixed variance is represented by the solid red line for the median value and the dashed lines for the 2.5 and 97.5 credible interval. The robust model with estimated variance is represented by the shaded area, which is the 2.5 to 97.5 credible interval.

Figure: 20: Dynamic with Fixed Variance vs. Dynamic with Estimated Variance



Note: Two sets of judicial ideology scores, which are estimated using two different versions of the latent variable model (both dynamic). The dynamic model with fixed variance is represented by the solid green line for the median value and the dashed lines for the 2.5 and 97.5 credible interval. The dynamic model with estimated variance is represented by the shaded area, which is the 2.5 to 97.5 credible interval.

Figure: 21: Robust with Estimated Variance vs. Dynamic with Estimated Variance



Note: Two sets of judicial ideology scores, which are estimated using two different versions of the latent variable model (robust and dynamic). The robust model with estimated variance is represented by the solid black line for the median value and the dashed lines for the 2.5 and 97.5 credible interval. The dynamic model with estimated variance is represented by the shaded area, which is the 2.5 to 97.5 credible interval.

F Additional Posterior Predictive Checks for Democracy

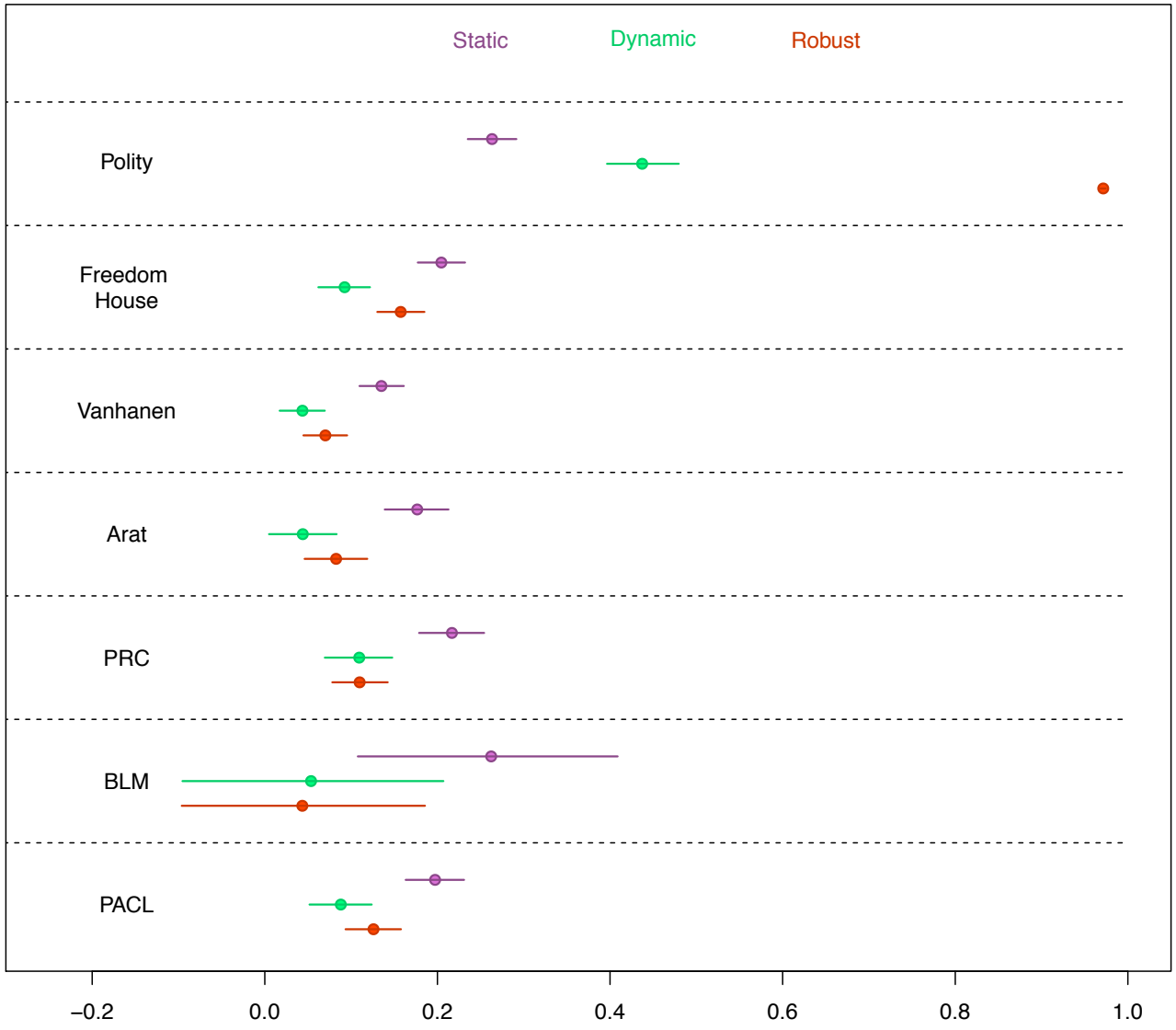
F.1 Correlation with Change in Indicators

As an additional way to compare the models, we examined how each predicted change in the indicators compared to the true change of the indicators. Using the posterior estimates, we calculated predictions for each annual indicator, and then the change between years. We then estimate the correlation of these predicted changes with the changes in the real indicators. We note that this is a hard test on difference in the model as change is relatively rare in indicators of democracy.

The robust model continues to outperform the other two models in predicting change for Polity. The mean correlation for the robust model with change in Polity is 0.97 while for the dynamic model it is 0.43 and the Static it is 0.26. The robust model almost perfectly replicates change in the polity model.

In addition, by looking at change we see how the robust model can outperform the dynamic model. For all remaining annual indicators, the robust model performs as well or better than the dynamic model. The static model continues to perform well for all indicators except for Polity.

Figure 22: Posterior Predictive Checks of Item Change



Note: This plot displays each models performance in accurately predicting annual change of the seven annual democracy indicators used to generate model estimates. The horizontal axis reports the correlation between real change in each indicator and the change predicted from the posterior of the measurement model. Models are displayed along the vertical axis. Dots correspond to the median value from the set of predictions, while solid lines denote the 2.5 and 97.5 percentile values of each models performance. The static model estimates are colored purple, the dynamic model is colored green, and the robust model orange.

G Convergence Diagnostics

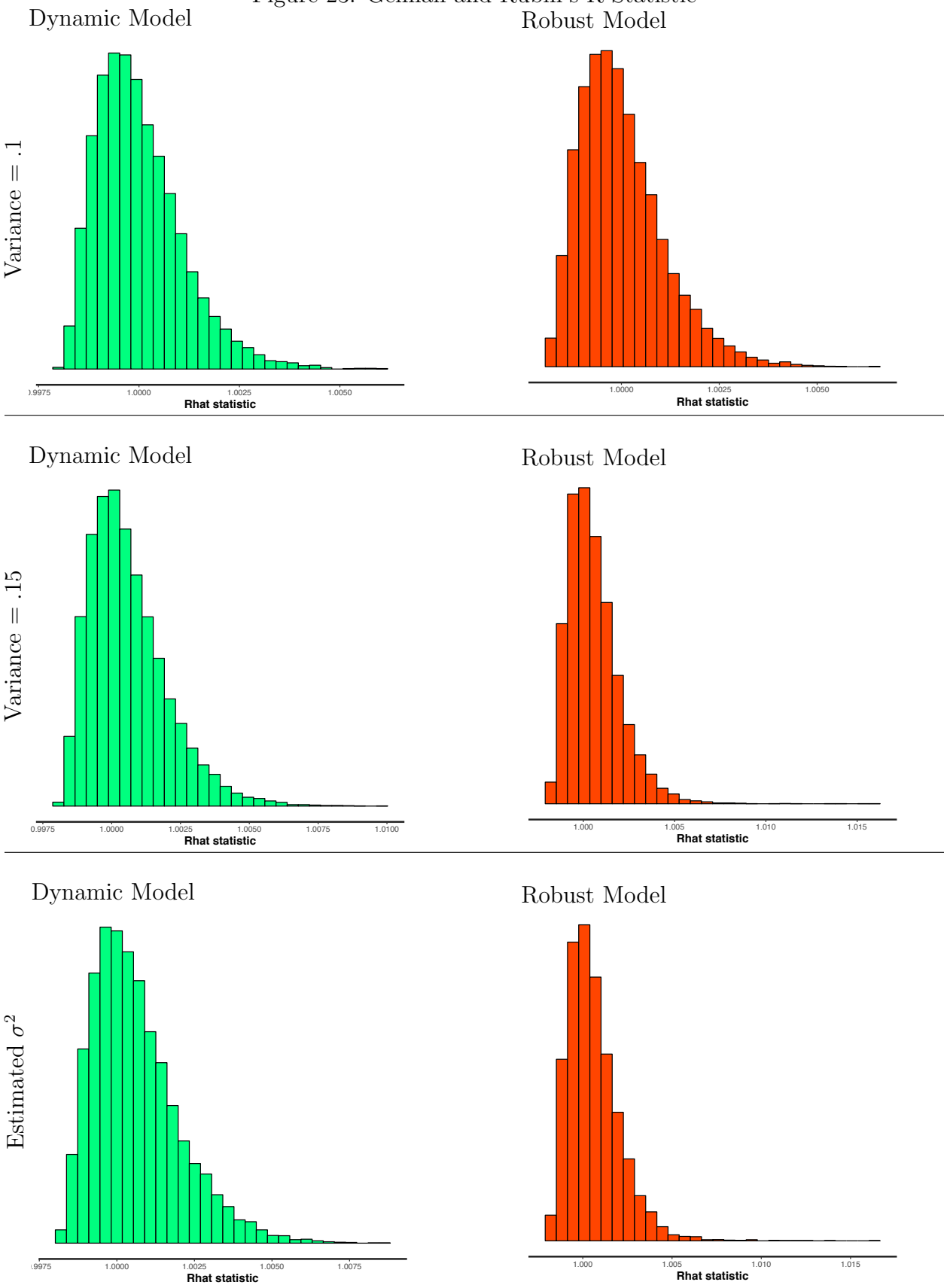
In this section, we display some of the convergence diagnostics for the substantive applications. We omit the statistics for the simulation analysis because of the sheer size of them (1,350 separate model estimates for the main three model types). We summarize model convergence using Gelman and Rubin's (1992) $\hat{R}$. This summary statistic is a measure of the ratio of the average variance within each change to the overall variance in all chains. A ratio close to 1 indicates convergence with estimates below 1.10 seen as generally indicative of convergence. The estimates for both sets of applied models demonstrate convergence.

G.1 Martin and Quinn (2002)

Figure 23 plots histograms of the $\hat{R}$ statistics for the dynamic and robust models presented in the text along with the inflated variance models discussed in Appendix E. For all four models $\hat{R}$ values are close to 1, with none greater than 1.10.[21]

[21]Note that because of the number of parameters estimated in Martin and Quinn (2002) we thinned the posterior by saving every second draw.

Figure 23: Gelman and Rubin's $\hat{R}$ Statistic

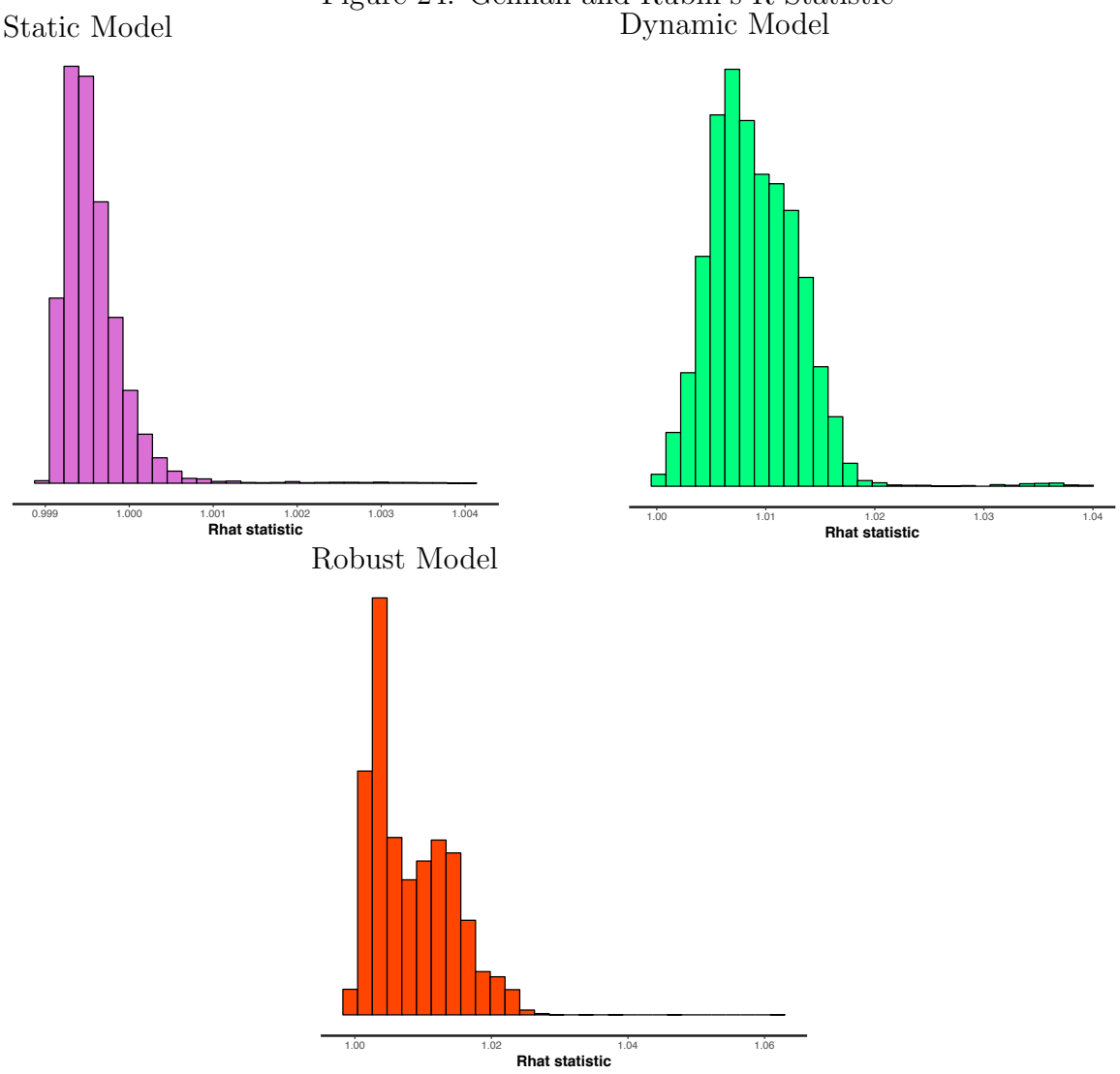


Note: Histogram of $\hat{R}$s from each model. All values are below 1.10, which indicates convergence.

G.2 Pemstein, Meserve and Melton (2010)

Figure 24 plots histograms of the $\hat{R}$ statistics for each model. In all cases, the vast majority of $\hat{R}$ statistics are close to 0, with almost all of them below 1.04 in all three models. None of the $\hat{R}$ statistics come close to approaching 1.10 indicating support for convergence.

Figure 24: Gelman and Rubin's $\hat{R}$ Statistic



Note: Histogram of $\hat{R}$s from each model. All values are below 1.10, which indicates convergence.

H WAIC for Hierarchical and IRT Models

WAIC (the Watanabe-Akaike or widely applicable information criterion) is currently one of the most preferred model diagnostics for Bayesian models (e.g., Gelman et al. 2014). However, several open research questions remain under-explored when using WAIC with hierarchical or IRT models like ours. Given the modeling structures presented in our main manuscript, these questions are important to consider in light of our dynamic models. In particular, as we discussed in the manuscript, there is an important question of what particular quantity or unit-structure is "left out" when validating models. That is, should individual items be left-out for all unit-time periods, for units from a panel, or all unit-years? Or should all the items be left out for one of these unit structures? Newly published research extends WAIC in cases in which items are clustered within an observation (Furr 2017) as well as other work incorporating time dynamics (Li et al. 2016). Another recent area of work are diagnostics, and best practices for WAIC and other models (Vehtari, Gelman and Gabry 2016). When there is concern over the validity of WAIC statistics it is useful to also estimate a K-fold cross validation. As a check we computed these for the models presented in the manuscript. The held-out log likelihood for Martin and Quinn dynamic model is -24,422 and for the robust model it is -24,375. The model with the log likelihood closest to 0 is preferred, and so the results are are substantively the same as the WAIC. The log likelihood for Pempstein et al static model is -47,702, for the dynamic it is -39,969, and for the robust it is -36,129. Again, the substantive results are the same.

We suggest that while this area of research continues, that authors should provide, as we have, multiple checks of model fit. As we argue in the main manuscript, posterior predictive checks are one very powerful way to test how well an IRT model fits data. In addition, when applicable, preforming K-fold crossvalidation can be useful, although again within Hierarchical or an IRT context, there are still concerns over what level of observation should be held out (e.g., individual items, unit-years, panels, etc). Overall, fit statistics, posterior

predictive checks, and visual analysis of the temporal patterns of well-known cases allow for the evaluation of competing models without relying on a single statistical tool.

I Replication Notes

All measurement models were estimated on an High Performance Computer (HPC). The models for the Unified Democracy Scale and Martin and Quinn take approximately 12 hours per chain. These models can be estimated on a laptop. If the user wishes to do so, we recommend using 2 cores, 2 chains, and to let each model run over night. The measurement models using simulated data take approximately 6 to 24 hours per chain. Because of the number of simulation, it is not feasible to run all simulations except on an HPC system. Please note that for some of the simulated results, the replication with the seed does not reproduce the exact simulated estimates. This is potentially an issue with differences in the random number generator (RNG) across platforms or Monte Carlo error or perhaps both. To exactly replicate the estimates, the system specification, operating system, software versions, and seed must all be identical, otherwise, some variation in the estimates will occur. We have successfully estimated the results across several platforms, operating systems, software versions, and seeds. We provide more detail about the file structure and computing requirement for reproducing all of the results presented in the paper and appendix at the dataverse repository available here: https://doi.org/10.7910/DVN/SSLCFF (Reuning, Kenwick and Fariss 2018).