# DISCUSSION PAPER SERIES

DP14790

## PANEL FORECASTS OF COUNTRY-LEVEL COVID-19 INFECTIONSLIU

Laura Liu, Hyungsik Roger Moon and Frank Schorfheide

**MACROECONOMICS AND GROWTH**

*ISSN 0265-8003*

# PANEL FORECASTS OF COUNTRY-LEVEL COVID-19 INFECTIONSLIU

*Laura Liu, Hyungsik Roger Moon and Frank Schorfheide*

Discussion Paper DP14790
Published 20 May 2020
Submitted 19 May 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Macroeconomics and Growth

# PANEL FORECASTS OF COUNTRY-LEVEL COVID-19 INFECTIONSLIU


## Abstract

We use dynamic panel data models to generate density forecasts for daily Covid-19 infections for a panel of countries/regions. At the core of our model is a specification that assumes that the growth rate of active infections can be represented by autoregressive fluctuations around a downward sloping deterministic trend function with a break. Our fully Bayesian approach allows us to flexibly estimate the cross-sectional distribution of heterogeneous coefficients and then implicitly use this distribution as prior to construct Bayes forecasts for the individual time series. According to our model, there is a lot of uncertainty about the evolution of infection rates, due to parameter uncertainty and the realization of future shocks. We find that over a one-week horizon the empirical coverage frequency of our interval forecasts is close to the nominal credible level. Weekly forecasts from our model are published at https://laurayuliu.com/covid19-panel-forecast/.



JEL Classification: C11, C23, C53

Keywords: Bayesian inference, COVID-19, Density forecasts, interval forecasts, panel data models, random effects, SIR model



Laura Liu - lauraliu@iu.edu
*Indiana University*

Hyungsik Roger Moon - moonr@usc.edu
*University of Southern California*

Frank Schorfheide - schorf@ssc.upenn.edu
*University of Pennsylvania and CEPR*



Acknowledgements
We thank the Johns Hopkins University Center for Systems Science and Engineering for making Covid-19 data publicly available on Github and Evan Chan for his help developing the website on which we publish our forecasts. Moon and Schorfheide gratefully acknowledge financial support from the National Science Foundation under Grants SES 1625586 and SES 1424843, respectively.

# Panel Forecasts of Country-Level Covid-19 Infections

Laura Liu
*Indiana University*

Hyungsik Roger Moon
*University of Southern California, Schaeffer Center, and Yonsei*

Frank Schorfheide*
*University of Pennsylvania CEPR, NBER, and PIER*

This Version: May 19, 2020

**Abstract**

We use dynamic panel data models to generate density forecasts for daily Covid-19 infections for a panel of countries/regions. At the core of our model is a specification that assumes that the growth rate of active infections can be represented by autoregressive fluctuations around a downward sloping deterministic trend function with a break. Our fully Bayesian approach allows us to flexibly estimate the cross-sectional distribution of heterogeneous coefficients and then implicitly use this distribution as prior to construct Bayes forecasts for the individual time series. According to our model, there is a lot of uncertainty about the evolution of infection rates, due to parameter uncertainty and the realization of future shocks. We find that over a one-week horizon the empirical coverage frequency of our interval forecasts is close to the nominal credible level. Weekly forecasts from our model are published at `https://laurayuliu.com/covid19-panel-forecast/`.

JEL CLASSIFICATION: C11, C23, C53

KEY WORDS: Bayesian inference, Covid-19, density forecasts, interval forecasts, panel data models, random effects, SIR model.

*Correspondence: L. Liu: Department of Economics, Indiana University, 100 S. Woodlawn Ave, Bloomington, IN 47405. Email: lauraliu@iu.edu. H.R. Moon: Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089. E-mail: moonr@usc.edu. F. Schorfheide: Department of Economics, 133 S. 36th Street, University of Pennsylvania, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu. We thank the Johns Hopkins University Center for Systems Science and Engineering for making Covid-19 data publicly available on Github and Evan Chan for his help developing the website on which we publish our forecasts. Moon and Schorfheide gratefully acknowledge financial support from the National Science Foundation under Grants SES 1625586 and SES 1424843, respectively.

# 1 Introduction

This paper contributes to the rapidly growing literature on generating forecasts related to the current Covid-19 pandemic. We are adapting forecasting techniques for panel data that we have recently developed for economic applications such as the prediction of bank profits, charge-off rates, and the growth (in terms of employment) of young firms; see Liu (2020), Liu, Moon, and Schorfheide (2020), and Liu, Moon, and Schorfheide (2019). We focus on the prediction of the smoothed daily number of active Covid-19 infections for a cross-section of approximately one hundred countries/regions. The data are obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. While we are currently focusing on country-level aggregates, our model could be easily modified to accommodate, say, state- or county-level data.

In economics, researchers distinguish, broadly speaking, between reduced-form and structural models. A reduced-form model summarizes spatial and temporal correlation structures among economic variables and can be used for predictive purposes assuming that the behavior of economic agents and policy makers over the prediction period is similar to the behavior during the estimation period. A structural model, on the other hand, attempts to identify causal relationships or parameters that characterize policy-invariant preferences of economic agents and production technologies. Structural economic models can be used to assess the effects of counterfactual policies during the estimation period or over the out-of-sample forecasting horizon.

The panel data model developed in this paper to generate forecasts of Covid-19 infections is a reduced-form model. It processes cross-sectional and time-series information about past infection levels and maps them into predictions of future infections. While the model specification is motivated by the time-path of infections generated by the workhorse compartmental model in the epidemiology literature, the so-called susceptible-infected-recovered (SIR) model, it is not designed to answer quantitative policy questions, e.g., about the impact of social-distancing measures on the path of future infection rates.

Building on a long tradition of econometric modeling dating back to Haavelmo (1944), our model is probabilistic. The growth rates of the infections are decomposed into a deterministic component which approximates the path predicted by a deterministic SIR model and a stochastic component that could be interpreted as either time-variation in the coefficients of an epidemiological model or deviations from such a model. We report interval and density forecasts of future infections that reflect two types of uncertainty: uncertainty

about model parameters and uncertainty about future shocks. We model the growth rate of active infections as autoregressive fluctuations around a deterministic trend function that is piecewise linear. The coefficients of this deterministic trend function are allowed to be heterogeneous across locations. The goal is not curve fitting – our model is distinctly less flexible in samples than some other models – but rather out-of-sample forecasts, which is why we prefer to project growth rates based on autoregressive fluctuations around a linear time trend.

A key feature of the Covid-19 pandemic is that the outbreaks did not take place simultaneously in all countries/regions. Thus, we can potentially learn from the speed of the spread of the disease and subsequent containment in country A, to make forecasts of what is likely to happen in country B, while simultaneously allowing for some heterogeneity across locations. In a panel data setting, one captures cross-sectional heterogeneity in the data with unit-specific parameters. The more precisely these heterogeneous coefficients are estimated, the more accurate are the forecasts. A natural way of disciplining the model is to assume that the heterogeneous coefficients are "drawn" from a common probability distribution. If this distribution has a large variance, then there is a lot of country-level heterogeneity in the evolution of Covid-19 infections. If instead, the distribution has a small variance, then the path of infections will be very similar across samples, and we can learn a lot from, say, China, that is relevant for predicting the path of the disease in South Korea or Germany.

Formally, the cross-sectional distribution of coefficients can be used as a so-called *a priori* distribution (prior) when making inference about country-specific coefficients. Using Bayesian inference, we combine the prior distribution with the unit-specific likelihood functions to compute *a posteriori* (posterior) distributions. This posterior distribution can then be used to generate density forecasts of future infections. Unfortunately, the cross-sectional distribution of heterogeneous coefficients is unknown. The key insight in the literature on Bayesian estimation of panel data models is that this distribution, which is called random effects distribution in the panel data model literature, can be extracted through simultaneous estimation from the cross-sectional dimension of the panel data set. There are several ways of implementing this basic idea. In this paper we will engage in a full Bayesian analysis by specifying a hyperprior for the distribution of heterogeneous coefficients and then constructing a joint posterior for the coefficients of this hyperprior as well as the actual unit-specific coefficients. Based on the posterior distribution, we simulate our panel model forward to generate density forecasts that reflect parameter uncertainty as well as uncertainty about shocks that capture deviations from the deterministic component of our forecasting model.

Our empirical analysis makes the following contributions. First, we present estimates of the random effects distribution as well as country-specific coefficients. Second, we document how density forecasts from our model have evolved over time, focusing on the forecasts for China, South Korea, and Germany for the origins of 2020-04-04 and 2020-04-18. We also examine the coverage frequencies of interval forecasts. Weekly forecasts are published on the companion website `https://laurayuliu.com/covid19-panel-forecast/`.

This paper is connected to several strands of the literature. The panel data forecasting approach is closely related to work by Gu and Koenker (2017a,b) and our own work in Liu (2020), Liu, Moon, and Schorfheide (2020), Liu, Moon, and Schorfheide (2019). All five papers focus on the estimation of the heterogeneous coefficients in linear panel data models. The forecasting model for the Covid-19 infections is very similar to the parametric benchmark model considered in Liu (2020). The approach has several desirable theoretical properties. For instance, Liu, Moon, and Schorfheide (2020), building on Brown and Greenshtein (2009), show that an empirical Bayes implementation of the forecasting approach based on Tweedie's formula can asymptotically (as the cross-sectional dimension tends to infinity) lead to forecasts that are as accurate as the so-called oracle forecasts. Here the oracle forecast is an infeasible benchmark that assumes that the distribution of the heterogeneous coefficients is known to the forecaster. Liu (2020) shows that the density forecast obtained from the full Bayesian analysis converges strongly to the oracle's density forecast as the cross-section gets large.

The piecewise linear conditional mean function for the infection growth rate resembles a spline; see de Boor (1990) for an introduction to spline approximation. Unlike a typical spline approximation in which the knot locations are free parameters and some continuity of smoothness restrictions are imposed, the knot placement in our setting is closely tied to the first component of the spline, and we do not impose continuity. However, going forward, it might become desirable to introduce additional knots in the deterministic trend component of infection growth rates and consider continuity restrictions. Smith and Kohn (1996) and Denison, Mallick, and Smith (1998) developed Bayesian approaches to automated knot selection. Alternatively, one could adopt a multiple-change-point approach as in Chib (1998), Giordani and Kohn (2008), and Koop and Potter (2009).

A growing number of researchers with backgrounds in epidemiology, biostatistics, machine learning, economics, and econometrics are engaged in modeling and forecasting aspects of the Covid-19 pandemic. Because this is a rapidly expanding and diverse field, we do not attempt to provide a meaningful survey at this moment. Instead, we simply provide a few

pointers. The paper by Avery, Bossert, Clark, Ellison, and Fisher Ellison (2020) cites a compilation of publicly available simulation models in footnote 15. The Center for Disease Control (CDC)[1] publishes forecasts from several different models and Nicholas Reich created a website[2] that combines Covid-19 forecasts from a variety of models. Murray (2020) and his team from the Institute for Health Metrics and Evaluation (IHME)[3] publish forecasts for Covid-19 related hospital demands and deaths. Fernandez-Villaverde and Jones (2020) generate forecasts from a variant of the SIR model.[4] Other forecasts are published by the Georgia State University School of Public Health[5] and independent data analysts, e.g., Youyang Gu.[6].

The remainder of this paper is organized as follows. Section 2 provides a brief survey of epidemiological models with a particular emphasis on the SIR model. The specification of our panel data model is presented in Section 3. The empirical analysis is conducted in Section 4. Finally, Section 5 concludes.

# 2 Modeling Epidemics

There is a long history of modeling epidemics. A recent survey of modeling approaches is provided by Bertozzi, Franco, Mohler, Short, and Sledge (2020). The authors distinguish three types of macroscopic models:[7] (i) the exponential growth model; (ii) self-exciting point processes / branching processes; (iii) compartmental models, most notably the SIR model that divides a population into susceptible ($S_t$), infected ($I_t$), and resistant ($R_t$) individuals. Our subsequent discussion will focus on the exponential growth model and the SIR model. While epidemiological models are often specified in continuous time, we will consider a discrete-time specification in this paper because it is more convenient for econometric inference.

The exponential model takes the form $I_t = I_0 \exp(\gamma_0 t)$. The number of infected individuals will grow exponentially at the constant rate $\gamma_0$. This is a reasonable assumption to describe the outbreak of a disease, but not the subsequent dynamics because the growth rate

[1] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html
[2] https://reichlab.io/covid19-forecast-hub/
[3] http://covid19.healthdata.org/
[4] https://web.stanford.edu/~chadj/Covid/Dashboard.html
[5] https://publichealth.gsu.edu/research/coronavirus/
[6] https://covid19-projections.com/
[7] As opposed to micro-simulation or agent-based models.

will typically fall over time and eventually turn negative as more and more people become resistant to the disease. The SIR model dates back to Kermack and McKendrick (1927). In its most elementary version it can be written in discrete-time as follows:

$$\begin{aligned} S_t &= S_{t-1} - \beta S_{t-1}(I_{t-1}/N) \\ I_t &= I_{t-1} + \beta S_{t-1}(I_{t-1}/N) - \gamma I_{t-1} \\ R_t &= R_{t-1} + \gamma I_{t-1}, \end{aligned} \tag{1}$$

where $N$ is the (fixed) size of the population, $\beta$ is the average number of contacts per person per time, and $\gamma$ is the rate of recovery or mortality. The model could be made stochastic by assuming that $\beta$ and $\gamma$ vary over time, e.g.,

$$\ln \beta_t = (1-\rho_\beta)\ln\beta + \rho_\beta \ln \beta_{t-1} + \epsilon_{\beta,t}, \quad \ln \gamma_t = (1-\rho_\gamma)\ln\gamma + \rho_\gamma \ln \gamma_{t-1} + \epsilon_{\gamma,t}.$$

In response to the recent Covid-19 pandemic, several introductory treatments of SIR models have been written for economists, e.g., Avery, Bossert, Clark, Ellison, and Fisher Ellison (2020) and Stock (2020). Moreover, there is a growing literature that combines compartmental models with economic components. In these models, economic agents account for the possibility of contracting a disease when making their decisions about market participation. This creates a link between infection rates and economic activity through the frequency of interactions. Examples of this work in macroeconomics include Eichenbaum, Rebelo, and Trabandt (2020), Glover, Heathcote, Krueger, and Rios-Rull (2020), and Krueger, Uhlig, and Xie (2020). The advantage of models that link health status to economic activity is that they can be used to assess the economic impact of, say, social distancing measures.

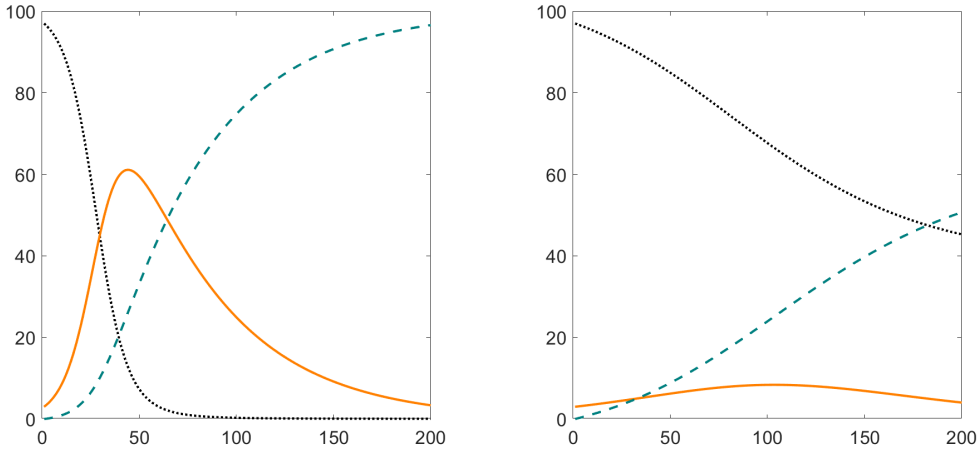
We now simulate the constant-coefficient SIR model in (1) under two different parameterizations for $(\beta, \gamma)$ that are unrelated to the current Covid-19 pandemic. The top panels of Figure 1 depict hypothetical time paths of $S_t$, $I_t$, and $R_t$. The size of the population is normalized to $N = 100$ and the outbreak of the disease is triggered by the initial condition $[S_0, I_0, R_0] = [97, 3, 0]$.

Under the first parameterization (left panels), the transmission rate $\beta = 0.15$ is very high and the recovery rate $\gamma = 0.02$ is relatively small. This leads to a fast rise in the number of infected individuals, which peaks at $I_{t_*} \approx 60$ in period $t_* \approx 50$. After the peak, the number of infections decreases, but more slowly than it increased during the initial outbreak. The bottom left panel shows the growth rate of the infections $100 \cdot \ln(I_t/I_{t-1})$ implied by the SIR

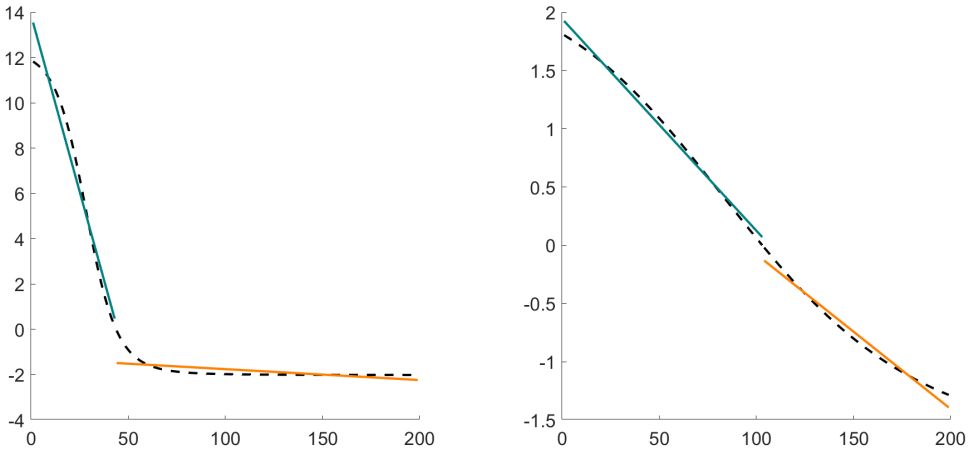
Figure 1: SIR Model Simulations

$\beta = 0.15$ and $\gamma = 0.02$ $\beta = 0.06$ and $\gamma = 0.04$

Levels of $S_t$ (black dotted), $I_t$ (orange solid), and $R_t$ (teal dashed)




Growth rate $100 \cdot \ln(I_t/I_{t-1})$, actual (black dashed) and Fitted (colored solid)




*Notes:* We normalize the size of the population to $N = 100$ and set the initial conditions to $S_0 = 97$, $I_0 = 3$, and $R_0 = 0$.

model. It is a monotonically decreasing function of time that we approximate by fitting a piecewise linear least-squares regression line with a break point at $t_*$ which is the point in time when the infections peak and the growth rate transitions from being positive to being negative. Under the second parameterization the transmission rate $\beta = 0.06$ is much lower and the recovery rate is slightly faster. This leads to an almost bell-curve shaped path of

infections. While the resulting growth rate of the infections is not exactly a linear function of time $t$, the break at $t_*$ is much less pronounced. While the piecewise-linear regression functions do not fit perfectly, they capture the general time-dependence of the growth-rate path implied by the SIR model. In particular, they allow for a potentially much slower change in the growth rate of infections after the peak.

We use these simulations as a motivation for the subsequent specification of our empirical model.[8] This model assumes that the growth rate of infections is a decreasing piecewise-linear function of time with a break when the growth rates cross zero and the infections peak. This deterministic component is augmented by a stochastic component that follows a first-order autoregressive, AR(1), process.

# 3 A Bayesian Panel Data Model

We now describe our empirical model in more detail. We begin with the specification of a regression model for the growth rate of infections in Section 3.1. Our model features location-specific regression coefficients and heteroskedasticity. The prior distribution for the Bayesian analysis is summarized in Section 3.2. Posterior inference is implemented through a Gibbs sampler that is outlined in Section 3.3. The algorithm to obtain simulated infection paths from the posterior predictive distribution is outlined in Section 3.4.

## 3.1 Panel Regression Specification

We specify a panel data model for infection growth rates $y_{it} = \Delta \ln I_{it}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. We assume that

$$y_{it} = \gamma_i' x_t + \delta_i' x_t \mathbb{I}\{t > t_i^*\} + u_{it}, \quad u_{it} = \rho u_{it-1} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_i^2), \tag{2}$$

where $\gamma_i = [\gamma_{0i}, \gamma_{1i}]'$ is a $2 \times 1$ vector of heterogeneous coefficients and $x_t = [1, t]'$. $\mathbb{I}\{t > t_*\}$ is the indicator function that is equal to one if $t > t_i^*$ and zero if $t \leq t_i^*$. The $2 \times 1$ vector $\delta_i = [\delta_{0i}, \delta_{1i}]'$ captures the size of the break in the regression coefficients at $t = t_i^*$. The deterministic part of $y_{it}$ corresponds to the piecewise-linear regression functions fitted to the infection growth paths simulated from the SIR in Figure 1.

[8] For forecasts generated directly from an enriched version of the SIR model see, for instance, Fernandez-Villaverde and Jones (2020).

The serially-correlated process $u_{it}$ generates stochastic deviations from the deterministic path $\gamma_i' x_t$ of the infection growth rate. The $u_{it}$ shocks may capture time variation in the $(\beta, \gamma)$ parameters of the SIR model or, alternatively, model misspecification. In Section 2 the break point $t_i^*$ was given by the peak of the infection path. Abstracting from a potential discontinuity at the kink, we define $t_i^*$ as

$$t_i^* = -\gamma_{0i}/\gamma_{1i}, \tag{3}$$

which implies that $\mathbb{E}[y_{it}|t = t_i^*] = 0$. Because of the AR(1) process $u_{it}$, $t_i^*$ is not the peak of the observed sample path, nor is it an unbiased or consistent estimate of the period in which the infections peak. For $\delta_i = 0$, the model reduces to

$$y_{it} = \gamma_i' x_t + u_{it}, \tag{4}$$

Note that the break date $t_i^*$ is identified in this model even if $\delta_i = 0$, because we assume the break occurs when the deterministic component of the growth rate falls below zero.

To construct a likelihood function we define the quasi-difference operator $\Delta_\rho = 1 - \rho L$ such that $\Delta_\rho u_{it} = \epsilon_{it}$. Thus, we can rewrite (2) as follows

$$y_{it} = \rho y_{it-1} + \gamma_i' \Delta_\rho x_t + \delta_i' \Delta_\rho x_t \mathbb{I}\{t > t_i^*\} + \epsilon_{it}. \tag{5}$$

Now let $\lambda_i = [\gamma_i', \delta_i]'$ and $n_\lambda$ be the dimension of $\lambda$. The parameters of the panel data model are $(\rho, \lambda_{1:N}, \sigma^2_{1:N})$. Here, we use the notation $Z_{1:L}$ to denote the sequence $z_1, \ldots, z_L$. Using this notation, we denote the panel observations by $Y_{1:N,1:T}$. We will subsequently condition on $Y_{1:N,0}$ to initialize conditional likelihood function. Finally, from the growth-rates $y_{it}$ we can easily recover the level of active infections as

$$I_{it} = I_{i0} \exp\left[\sum_{\tau=1}^{t} y_{i\tau}\right]. \tag{6}$$

## 3.2 Prior Distribution

To conduct Bayesian inference, we need to specify a prior distribution for $(\rho, \lambda_{1:N}, \sigma^2_{1:N})$. We do so conditional on a vector of hyperparameters $\xi$ that do not enter the likelihood function.

Our prior distribution has the following factorization:

$$p\big(\rho, \lambda_{1:N}, \sigma^2_{1:N}, \xi\big) \propto p(\rho) \left( \prod_{i=1}^{N} p(\lambda_i|\xi) f(\lambda_i) \right) \left( \prod_{i=1}^{N} p(\sigma_i^2|\xi) \right) p(\xi), \tag{7}$$

where $\propto$ denotes proportionality and $f(\cdot)$ is an indicator function that we will use to impose the following sign restrictions on the elements of $\lambda_i$:

$$f(\lambda_i) = \mathbb{I}\{\gamma_{1i} < 0\} \cdot \mathbb{I}\{\delta_{0i} < 0\} \cdot \mathbb{I}\{\delta_{1i} > 0\} \cdot \mathbb{I}\{\gamma_{1i} + \delta_{1i} < 0\}.$$

The restriction $\gamma_{1i} < 0$ ensures that the growth rates are falling over time. After the break point the rate of decline decreases ($\delta_{1i} > 0$), but stays negative ($\gamma_{1i} + \delta_{1i} < 0$). In addition we assume that the decrease in the rate of decline is associated with a downward shift, i.e., $\delta_{0i} < 0$, of the intercept as shown in the SIR simulation.

Because of the presence of the indicator function $f(\cdot)$ the right-hand side of (7) is not a properly normalized density. In view of the indicator function $f(\cdot)$ we define the random effects distribution of $\lambda_i$ given $\xi$ as

$$\pi(\lambda_i|\xi) = \frac{1}{C(\xi)} p(\lambda_i|\xi) f(\lambda_i), \quad C(\xi) = \int p(\lambda_i|\xi) f(\lambda_i) \lambda_i. \tag{8}$$

In turn, the marginal prior distribution of the hyperparameters is given by

$$\pi(\xi) = p(\xi)[C(\xi)]^N. \tag{9}$$

Building on Liu (2020), we use the following densities $p(\cdot)$ in (7) for $\rho$, $\lambda_i$, and $\sigma_i^2$:

$$\rho \sim N(0.5, 1)\mathbb{I}\{0 \leq \rho \leq 0.99\}, \quad \lambda_i \sim N(\mu, \Sigma), \quad \sigma_i^2 \sim IG(a, b). \tag{10}$$

Thus, the vector of hyperparameters is $\xi = (\mu, \Sigma, a, b)$. We decompose $p(\xi) = p(\mu, \Sigma)p(a, b)$. The density $p(\mu, \Sigma)$ is constructed as follows:

$$\mu|\Sigma \sim N(0, \Sigma), \quad \Sigma \sim IW(W_0, \nu). \tag{11}$$

The degrees of freedom for the Inverse Wishart distribution is set to

$$\nu = (2n_\lambda + 1)(n_\lambda - 1) + 1 = 28.$$

The shape matrix $W_0$ is diagonal with elements

$$W_{0,kk} = \frac{(\nu - n_\lambda - 1)\hat{\mathbb{V}}^i\big(\hat{\mathbb{E}}_i^t[y_{it}]\big)}{n_\lambda(\hat{\mathbb{E}}[x_{k,it}])^2}, \quad k = 1, \ldots, n_\lambda.$$

Here, $\hat{\mathbb{E}}_i^t[z_{it}]$ is the sample mean of the time series $z_{it}$, $t = 0, \ldots, T$, $\hat{\mathbb{V}}[z_i]$ is the cross-sectional sample variance of $z_i$, $i = 1, \ldots, N$, and $\hat{\mathbb{E}}[z_{it}]$ is a sample average of $z_{it}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The matrix $W_0$ is constructed to align the scale of the variance of $\mu_i$ with the cross-sectional variance of the data, adjusting for the average magnitudes of the regressors that multiply the $\lambda_i$ elements.

To obtain the density $p(a, b)$, we follow Llera and Beckmann (2016) and let

$$b \sim G(\underline{\alpha}_b, \underline{\beta}_b), \quad p(a|b) \propto \frac{\underline{\alpha}_a^{-(1+a)} b^{a\underline{\gamma}_a}}{\Gamma(a)^{\underline{\beta}_a}}. \tag{12}$$

The parameters $(\underline{\alpha}_a, \underline{\beta}_a, \underline{\gamma}_a, \underline{\alpha}_b, \underline{\beta}_b)$ need to be chosen by the researcher. We use $\underline{\alpha}_a = 1$, $\underline{\beta}_a = \underline{\gamma}_a = \underline{\alpha}_b = \underline{\beta}_b = 0.01$, which specifies relatively uninformative priors for hyperparameters $a$ and $b$.

## 3.3 (Approximate) Posterior Inference

Posterior inference is based on an application of Bayes Theorem. Let $p(Y_{1:N,1:T}|\lambda_{1:N}, \sigma^2_{1:N})$ denote the likelihood function (for notational convenience we dropped $Y_{1:N,0}$ from the conditioning set). Then the posterior density is proportional to

$$p(\rho, \lambda_{1:N}, \sigma^2_{1:N}, \xi|Y_{1:N,0:T}) \propto p(Y_{1:N,1:T}|\lambda_{1:N}, \sigma^2_{1:N}) p(\rho) p\big(\rho, \lambda_{1:N}, \sigma^2_{1:N}, \xi\big), \tag{13}$$

where the prior was given in (7). To generate draws from the posterior distribution we use a Gibbs sampler that iterates over the conditional posterior distributions

$$\begin{aligned} &\lambda_{1:N}|(Y_{1:N,0:T}, \rho, \sigma^2_{1:N}, \xi), \quad \rho|(Y_{1:N,0:T}, \lambda_{1:N}, \sigma^2_{1:N}, \xi), \\ &\sigma^2_{1:N}|(Y_{1:N,0:T}, \lambda_{1:N}, \rho, \xi), \quad \xi|(Y_{1:N,0:T}, \lambda_{1:N}, \sigma^2_{1:N}, \xi). \end{aligned} \tag{14}$$

The Gibbs sampler generates a sequence of draws $\big(\rho^s, \lambda^s_{1:N}, (\sigma^2_{1:N})^s, \xi^s\big)$, $s = 1, \ldots, N_{sim}$, from the posterior distribution. The implementation of the Gibbs sampler closely follows Liu (2020).

For the Gibbs sampler to be efficient, it is desirable to have a model specification in which it is possible to directly sample from the conditional posterior distributions in (14). Unfortunately, the exact likelihood function leads to a non-standard conditional posterior distribution for $\lambda_{1:N}|(Y_{1:N,0:T}, \rho, \sigma^2_{1:N}, \xi)$ because $\gamma_i$ enters the indicator function in (2) through the definition of $t^*_i$. Thus, rather than using the exact likelihood function, we will use a limited-information likelihood function of the form

$$p_l(Y_{1:N,1:T}|\lambda_{1:N}, \sigma^2_{1:N}) = \prod_{i=1}^{N} p_l(Y_{i,1:T}|\lambda_i, \sigma^2_i). \tag{15}$$

The densities $p_l(Y_{i,1:T}|\lambda_i, \sigma^2_i)$ are constructed as follows. Let $\Delta$ be some positive number, e.g., three or five time periods. Given a sample $(Y_{i,1:T}, \ln I_{i,1:T})$ we define

$$t_{i,max} = \text{argmax}_{1\leq t\leq T} \ \ln I_{i,1:T}.$$

If $t_{i,max} = T$, then it is likely that $t^*_i \geq T$. On the other hand, if $t_{i,max} < T$, then it is likely that $t_* = t_{i,max}$. Thus, we distinguish two cases:

*Case 1:* Suppose $t_{i,max} = T$: we drop observations $Y_{i,T-\Delta+1:T}$ and define

$$p_l(Y_{i,1:T}|\gamma_i, \delta_i, \sigma^2_i) = p(Y_{i,1:T-\Delta}|\gamma_i, \rho, \sigma^2_i).$$

Because $\delta_i$ does not enter the likelihood function, its posterior is $p(\delta_i|Y_{i,1:T-\Delta}, \gamma_i, \rho) = p(\delta_i|\gamma_i)$.

*Case 2:* Suppose $t_{i,max} < T$: we drop observations $Y_{i,t_{i,max}-\Delta+1:t_{i,max}+\Delta-1}$ and define

$$p_l(Y_{i,1:T}|\gamma_i, \delta_i, \sigma^2_i) = p(Y_{i,1:t_{i,max}-\Delta}, Y_{i,t_{i,max}+\Delta:T}|\gamma_i, \delta_i, \rho, \sigma^2_i).$$

Now $\delta_i$ does enter the likelihood function and its prior gets updated in view of the data.

## 3.4 Forecasting Infection Rates

Bayesian forecasts reflect parameter and shock uncertainty. We simulate trajectories of infection growth rates from the posterior predictive distribution using the Algorithm 1. The simulated growth rates can be converted into simulated trajectories for active infections using (6).

**Algorithm 1 (Simulating from the Posterior Predictive Distribution)**

1. *For* $s = 1, \ldots, N_{sim}$
   (a) *Use parameter draw* $s$ *from the posterior distribution:* $\big(\rho^s, \lambda^s_{1:N}, (\sigma^2_{1:N})^s\big)$.
   (b) *For* $i = 1, \ldots, N$*:*
      i. *Compute* $t_i^{*s} = -\gamma^s_{i0}/\gamma^s_{i1}$.
      ii. *Generate a sequence of draws* $\epsilon_{it} \sim N\big(0, (\sigma_i^2)^s\big)$, $t = T+1, \ldots, T+H$.
      iii. *Iterate (5) forward for* $t = T+1, \ldots, T+H$ *to obtain* $Y^s_{i,T+1:T+H}$.
      iv. *Compute* $I^s_{iT+h} = I_{iT} \exp\big[\sum_{l=1}^{h} y^s_{iT+l}\big]$, $h = 1, \ldots, H$.
2. *Based on the simulated paths* $I^s_{1:N,T+1:T+H}$, $s = 1, \ldots, N_{sim}$, *compute point, interval, and density forecasts for each period* $t = T+1, \ldots, T+H$.

# 4 Empirical Analysis

The data set used in the empirical analysis is described in Section 4.1. We discuss the posterior estimates in Section 4.2. Finally, we present density and interval forecasts in Section 4.3.

## 4.1 Data

The data set is obtained from CSSE at Johns Hopkins University.[9] We define the total number of active infections in location $i$ and period $t$ as the number of confirmed cases minus the number of recovered cases and deaths. We understand that infections are measured with error because there is evidence that a significant number of infected individuals are asymptomatic and hence not captured in the official statistics. Moreover, determining the precise number of Covid-19 related deaths is non-trivial (dying with versus dying of Covid-19). The goal of our modeling effort is to predict the number of active infections as recorded in the CSSE data set.

Throughout our study we use country-level aggregates. The time period $t$ corresponds to a day and we fit our model to one-sided three-day rolling averages to smooth out noise

[9] `https://github.com/CSSEGISandData/COVID-19`

generate by the timing of the reporting. In a slight abuse of notation, the time subscript $t$ in (2) is meant to be event time and hence is specific on the location $i$. The event time is initialized once the number of confirmed cases in a location reaches 100.[10] For each location, we let the time series of infections end at the same calendar time. As a result, the panel is unbalanced.

Our empirical analysis is based on a cross-section of approximately 100 countries/regions. We start out from 185 locations and eliminate a subset of locations according to the following rules: (i) we eliminate locations that have not reached 100 active infections. (ii) We eliminate locations for which $t_{i,max} - \Delta < 0$. This guarantees that we have at least one observation in the limited-information likelihood function to extract information about $\gamma_i$. (iii) For each location $i$ we regress the growth rates from period $t = 0$ to $t = T$ on a time trend and an intercept and eliminate locations where the OLS estimate of the time-trend coefficient is positive because the SIR model implies a decreasing growth rate. The resulting cross-sectional dimension of our panel is $N = 110$.

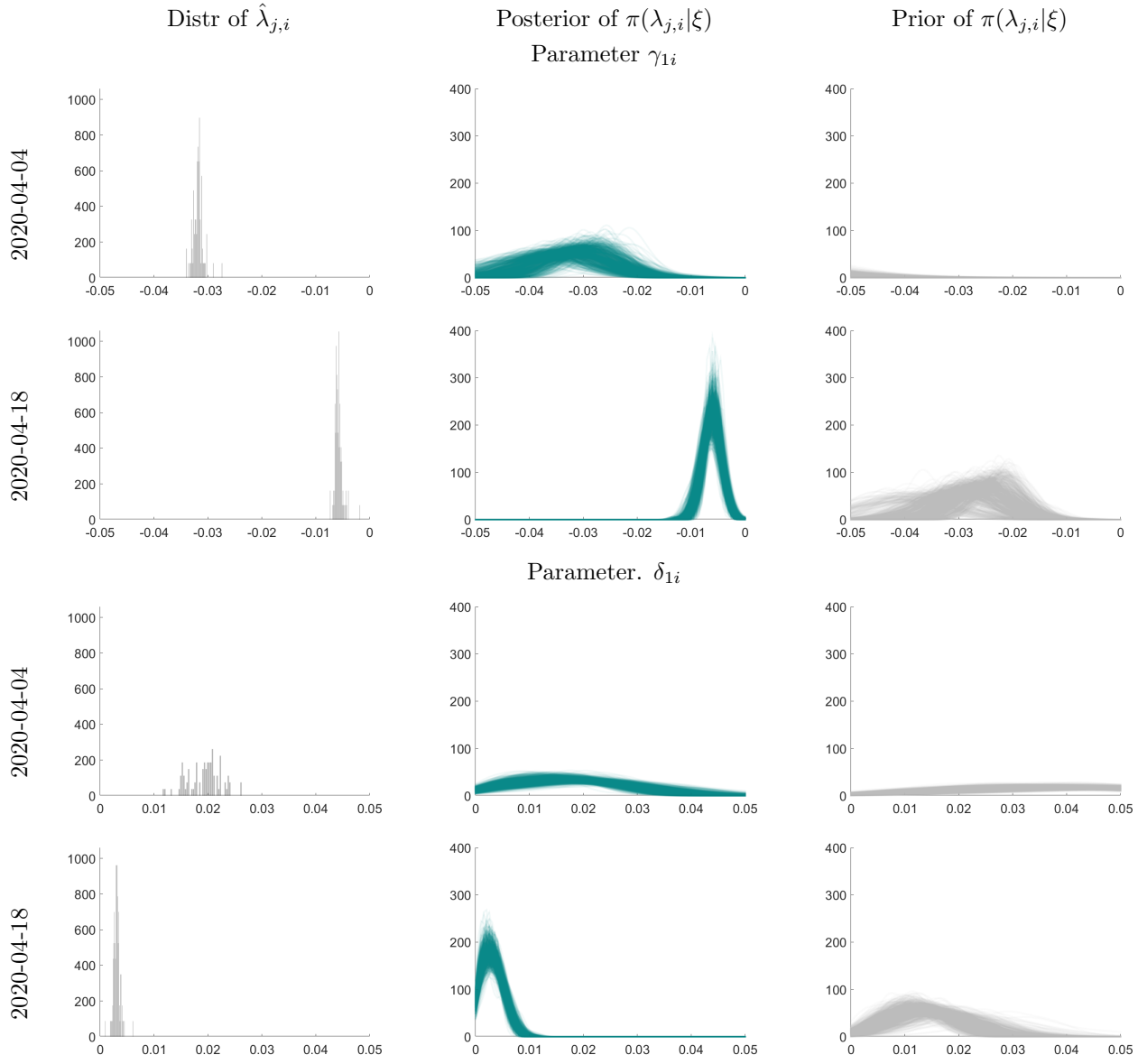
## 4.2 Parameter Estimates

Before discussing the forecasts, we will examine the parameter estimates. Throughout this subsection we focus on two estimation samples. For each location $i$, the first observation included in both samples is determined by the point in time in which the number of infections reaches 100. The last observation for each location is determined by calendar time. The first estimation sample ends on 2020-04-04. At this point only seven countries/regions in our panel have reached the peak level of infections. The second estimation sample ends two weeks later on 2020-04-18 when 36 locations have moved beyond the peak in terms of the number of active infections.

Our Gibbs sampler generates draws from the joint posterior of $(\rho, \lambda_{1:N}, \sigma^2_{1:N}, \xi)|Y_{1:N,0:T}$. We begin with a discussion of the estimates of $\gamma_{1i}$ and $\delta_{1i}$, which affect the speed at which the growth rates is expected to change on a daily basis. $\gamma_{1i}$ measures the average daily decline in the growth rate of active infections. For instance, suppose the at the beginning of the outbreak, in event time $t = 0$, the growth rate $\ln(I_t/I_{t-1}) = 0.2$, i.e., approximately 20%. A value of $\gamma_{1i} = -0.02$ implies that, on average, the growth rate declines by 0.02, meaning that

[10] In calendar time, let $\tau_0 = \min_\tau$ s.t. $I_\tau > 100$. Using $I_{\tau_0}, I_{\tau_0+1}, \ldots$, we take log differences to compute growth rates $\ln(I_{\tau_0+1}/I_{\tau_0})$, $\ln(I_{\tau_0+2}/I_{\tau_0+1})$, .... In the estimation we need one growth rate observation to initialize lags. Thus, in event time, period $\tau_0$ corresponds to $t = -1$.

Figure 2: Heterogeneous Coefficients Estimates and Random Effects Distributions




*Notes:* Point estimator $\hat{\lambda}_{j,i}$ is posterior mean of $\gamma_{1i}$ or $\delta_{1i}$, respectively.

after 10 days it is expected to reach zero and turn negative subsequently. A positive value of $\delta_{1i} = 0.01$ implies that after the growth rate becomes negative, its decline is reduced (in absolute value) to $\gamma_{1i} + \delta_{1i} = -0.01$.

In the panels in the first column of Figure 2 we plot the cross-sectional distributions of posterior mean estimates $\hat{\gamma}_{1i}$ and $\hat{\delta}_{1i}$. Between 2020-04-04 and 2020-04-18 the distribution of the estimates $\hat{\gamma}_{1i}$ shifts to the right. While in the early sample the growth rate of the

infections appears to fall quickly over time ($\hat{\gamma}_{1i} \approx -0.032$), two weeks later the estimate has fallen (in absolute value) to approximately -0.005. The estimates $\hat{\delta}_{1i}$ show a similar shift from approximately 0.02 to below 0.005. The additional two weeks of data have led to a more concentrated cross-sectional distribution of estimates, indicating that the deterministic component of the infection growth rates is becoming more similar as countries/regions move beyond the early stages of the infections.

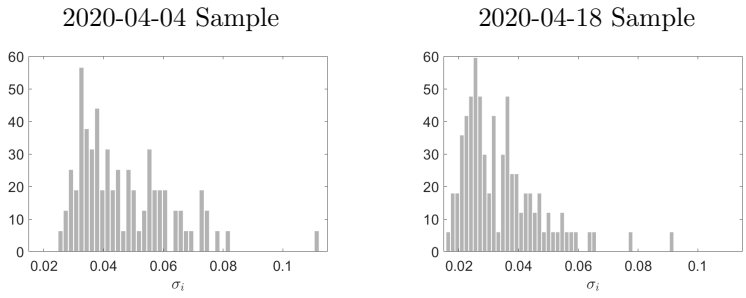
An important component of our model is the random effects distribution $\pi(\lambda_i|\xi)$ defined in (8). Prior and posterior uncertainty with respect to the hyperparameters $\xi$ generate uncertainty about the random effects distribution. In the remaining panels of Figure 2 we plot draw from the posterior (center column) and prior (right column) distribution of the random effects density $\pi(\lambda_i|\xi)$. Each draw is represented by a hairline. Because the normalization constant $C(\xi)$ of $\pi(\lambda_i|\xi)$ is difficult to compute due to the truncation of a joint Normal distribution, we show kernel density estimates obtained from draws from $\pi(\lambda_i|\xi)$.

The random effects densities drawn from the posterior approximately peak around values of $\gamma_{1i}$ and $\delta_{1i}$ for which the histograms on the left are peaking. Thus, the estimates of the densities cohere with the estimates of the heterogeneous coefficients. The histograms also show the increase in information between the 2020-04-04 and 2020-04-18 samples. The precise relationship between the hairlines that represent draws from the distribution of the random effects densities and the posterior point estimates are discussed in more detail in Liu, Moon, and Schorfheide (2019). The random effects densities are generally more diffuse than the distributions of the point estimates represented by the histograms because the random effects densities can be viewed as priors of $\lambda_i$ whereas the point estimates combine information from these priors and the time series $Y_{i,1:T}$.

The random effects densities drawn from the prior distribution of $\xi$ are fairly flat. Because of the truncation, the means implied by the RE densities for $\gamma_{1i}$ are negative, whereas the means implied by the densities for $\delta_{1i}$ are positive. The priors for the random effects densities are dependent on the sample because the overall prior is indexed by data-dependent tuning parameters; see Section 3.2.

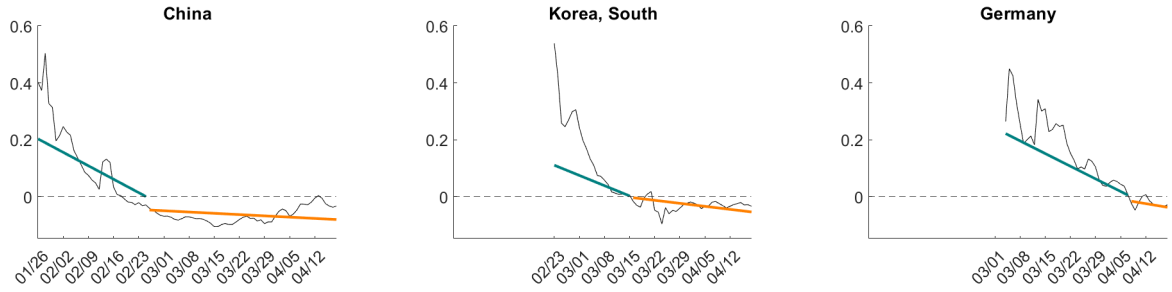
Our posterior sampler also generates estimates for the homogeneous autoregressive coefficient $\rho$. The estimates are $\hat{\rho} = 0.9898$ for the 2020-04-04 sample and $\hat{\rho} = 0.7849$ for the 2020-04-18 sample. In Figure 3 we show histograms of the cross-sectional distribution of $\hat{\sigma}_i$. Overall, the fit of the panel data model appears to improve as time progresses: the

Figure 3: Cross-sectional Dispersion of Innovation Variances




*Notes:* Histogram of posterior mean estimates $\hat{\sigma}_i$.

Figure 4: Fitted Regression Lines for Daily Infection Growth Rates



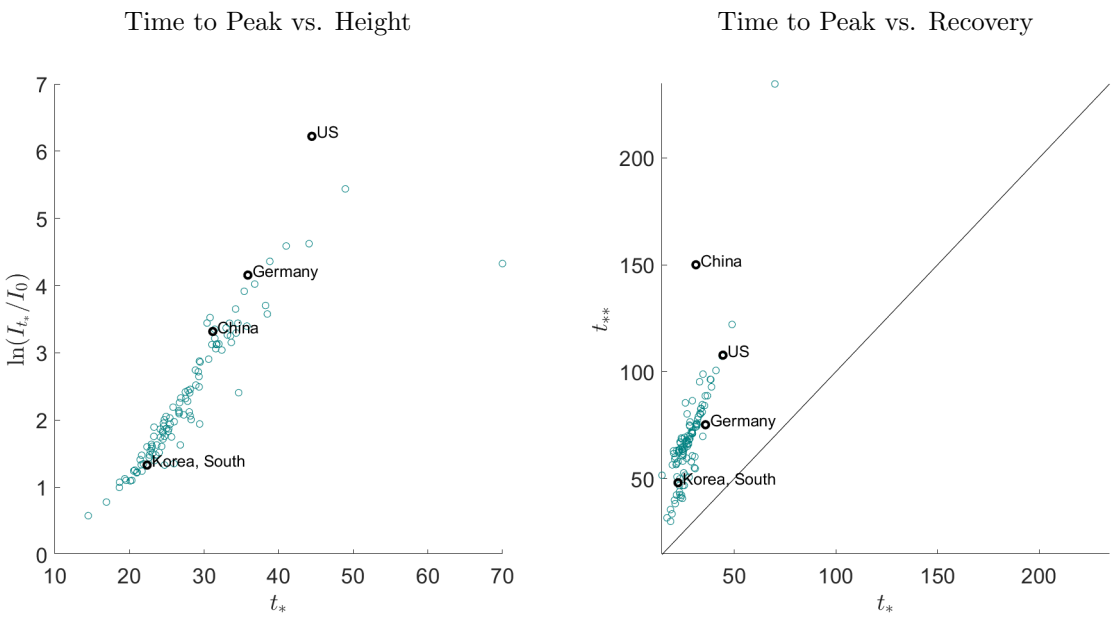

*Notes:* Estimation sample ends in 2020-04-18.

autocorrelation $\rho$ of the shock process $u_{it}$ falls and the distribution of $\hat{\sigma}_i$ shifts to the left and becomes a bit more concentrated.

After examining the cross-sectional distribution of the $\hat{\gamma}_{1i}$ and $\hat{\delta}_{1i}$ estimates, we will now examine the implied regression functions that capture the deterministic component of the infection growth rates for three specific countries: China, South Korea, and Germany. These three countries experienced the outbreak at different points in time. The posterior median estimates from which the regression lines depicted in Figure 4 are constructed, reflect the prior information from the random effects distributions depicted in Figure 2 and the time series information for each country. By construction, the regression lines are piecewise linear, and the break occurs at the point in time when the deterministic component implies a zero growth rate. The fitted regression line for South Korea reflects a fair amount of shrinkage induced by the prior distribution, because the initial rapid decline in the growth rate is unusual according to the estimated cross-sectional random effects distribution.

Because the coefficients $\gamma_i$ and $\delta_i$ cannot be directly interpreted in terms of the speed and

Figure 5: Parameter Transformations $t^*$, $\ln(I_t/I_*)$, and $t^{**}$




*Notes:* The results are based on the 2020-04-18 sample. Results are in event time. $t_0$ is the period in which the number of infections exceeds 100 for the first time.

the severity of the outbreak, we are transforming the $\lambda_i$s as follows (omitting the $i$ subscripts): First, we use the definition of $t^* = -\gamma_0/\gamma_1$ from (3). Note that $t^*$ is not restricted to be an integer. Second, according to the deterministic part of the growth rate model, the log level of infections at the peak, relative to the starting point is approximately

$$\ln(I_{t^*}/I_0) = \int_0^{t^*} (\gamma_0 + \gamma_1 t)dt = -\frac{\gamma_0^2}{2\gamma_1}. \tag{16}$$

Third, after the break at $t = t_*$ the growth rate continues to decline according to $(\gamma_0 + \delta_0) + (\gamma_1 + \delta_1)t$. We define the time $t^{**}$, i.e., the time it takes to return to the initial level $I_0$, as the solution to

$$\int_0^{t^{**}} [\gamma_0 + \delta_0 + (\gamma_1 + \delta_1)t]dt - \frac{\gamma_0^2}{2\gamma_1} = 0 \tag{17}$$

Note that $(t^*, \ln(I_{t^*}/I_0), t^{**})$ is a nonlinear transformation of $(\gamma_0, \gamma_1, \delta_0, \delta_1)$. The triplet does not measure the actual or expected time to peak, height of the peak, time to recover.

Pairwise scatter plots of $\left(t^*, \ln(I_{t^*}/I_0), t^{**}\right)$ are depicted in the two panels of Figure 5. Each dot is generated as follows: for each MCMC draw $s = 1, \ldots, N_{sim}$ we transform $(\gamma_i, \delta_i)^s$ into $\left(t_i^*, \ln(I_{i,t_i^*}/I_{i0}), t_i^{**}\right)^s$. We then compute medians of the transformed objects. We indicate

the values for China, South Korea, Germany, and the U.S. According to the first panel, there is a strong positive correlation between time to peak $t_*$ and height of peak $\ln(I_{t^*}/I_0)$. The relationship is remarkably linear across locations. The second panel shows that the time to recovery $t^{**}$ is (a lot) larger than the time to peak $t^*$. Here China is an outlier. The actual time to recover from the epidemic was a lot shorter, which is due to favorable shocks $u_{it}$ in the model.

## 4.3 Predictive Densities

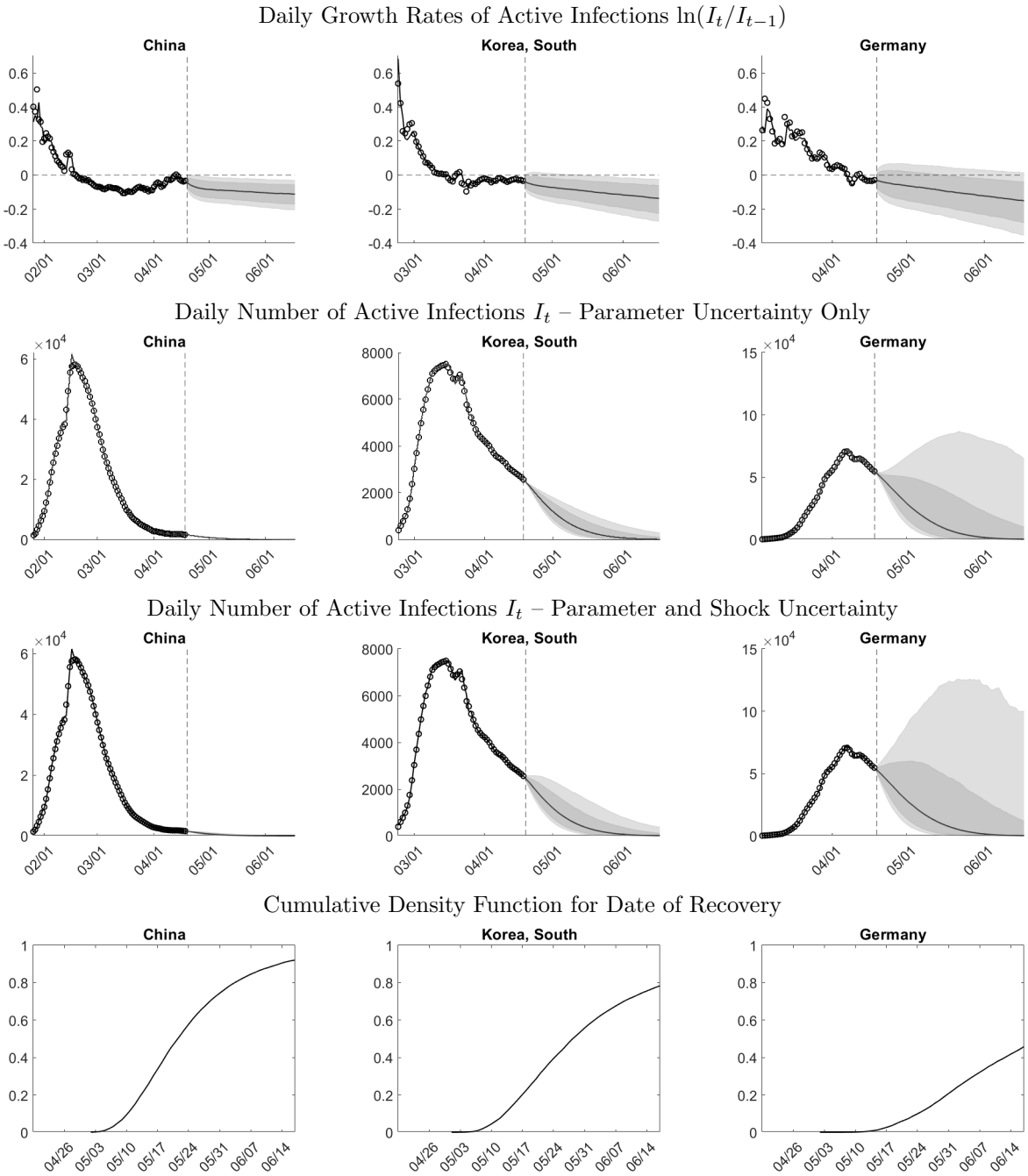
We now turn to density forecasts generated from the estimated panel data model. We use Algorithm 1 to simulate trajectories of infection growth rates which, conditional on observations of the initial levels $I_{iT}$, we convert into stocks of active infections. For each forecast horizon $h$ we use the values $y^s_{iT+h}$ and $I^s_{iT+h}$, $s = 1, \ldots, N_{sim}$ to approximate the predictive density. Strictly speaking, we are not reporting complete predictive densities. Instead, we plot medians and construct equal-tail-probability bands that capture the range between the 20-80% and 10-90% quantiles. The wider the bands, the greater the uncertainty. As in the estimation section, we consider two samples: one ends on 2020-04-04 and the other one on 2020-04-18. The end of the estimation sample is the origin of our forecasts.

Figure 6 shows density forecasts over 60 days for the growth rate, the level of active infections, and the recovery date in China, South Korea, and Germany based on 2020-04-18 data. The forecast origin is indicated by the vertical dashed line. At the forecast origin, the three countries are at different stages of the epidemic. In China the number of active infections has fallen from 58,000 to 1,600. In South Korea, the level of infections is 67 percent below its peak value. Finally, Germany has barely moved beyond the peak. Prior to the forecast origin we show the actual values and in-sample fitted values.[11] Additional density forecasts for more than 100 countries/regions are provided on the companion website `https://laurayuliu.com/covid19-panel-forecast/`.

The panels in the first row of Figure 6 show forecasts for the growth rate of active infections. At the forecast origin, the actual growth rates for all three countries are negative. The median forecast is driven by the deterministic trend component in our model for $y_{it}$; see (2) and Figure 4. The bands reflect both parameter uncertainty and stochastic fluctuations

[11] The fitted values are generated as follows: for each draw from the posterior distribution, we generate a one-step-ahead in-sample prediction for each country/region. Then we compute the median across these in-sample predictions for each location.

Figure 6: Forecasts for China, South Korea, and Germany, Origin is 2020-04-18




*Notes:* Rows 1 to 3: The vertical lines indicate the forecast origin. The circles indicate actual infections. The solid lines prior to the forecast origin represent in-sample one-step-ahead forecasts. The solid lines after the forecast origin represent medians of the posterior predictive distribution. The grey shaded bands indicate the 20%-80% (dark) and 10%-90% (light) interquantile ranges of the posterior predictive distribution. Bottom row: cumulative density function (associated with posterior predictive distribution) of of date of recovery defined as $\tau$ such that $I_\tau = I_0$.

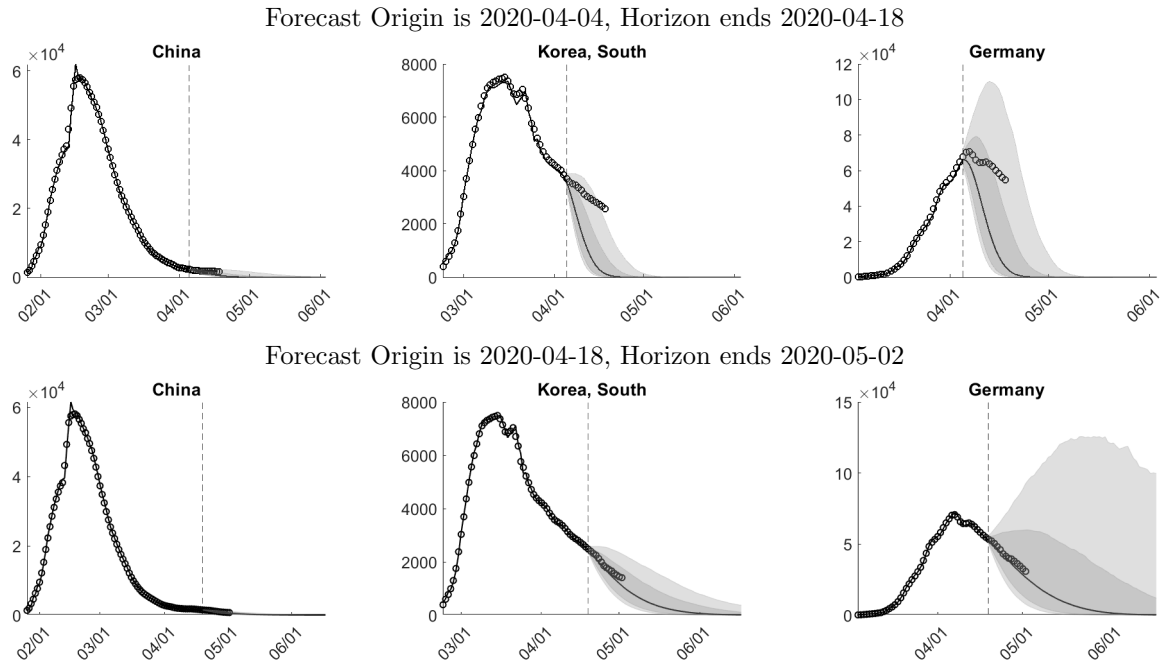
around the trend component generated by the autoregressive process $u_{it}$. The width of the bands is the smallest for China and the largest for Germany. Two factors contribute to the wider bands for Germany. First, the estimated innovation standard deviation $\hat{\sigma}_i$ is larger for Germany than for China and South Korea. Second, recall that at the peak, the parameters of the deterministic component of our model shift by $\delta_i$. The less time has passed since the peak, the fewer observations are available to estimate $\delta_i$, which increases the contribution of parameter uncertainty to the predictive distribution.

The second and third rows of Figure 6 depict predictions for the daily level of active infections. The path of active infections broadly resembles the paths simulated with the SIR model in Section 2. The rise of infections during the outbreak tends to be faster than the subsequent decline, which is a feature that is captured by the break in the conditional mean function of our model for the infection growth rate $y_{it}$ in (2). The difference between the bands depicted in the second and third rows is that the former reflects parameter uncertainty only (we set future shocks equal to zero), whereas the latter reflects parameter and shock uncertainty. In the case of Germany, shock uncertainty increases the width of the bands by approximately 50%. Due to the exponential transformation that is used to recover the levels, the predictive densities are highly skewed and exhibit a large upside risk. This is particularly evident for Germany. The growth rate prediction in the first row indicates that there is an approximately 20% probability of a positive infection growth rate. Converted into levels, temporarily positive growth rates of infections generate a "second wave" of infections in our model.

In the bottom row of Figure 6 we plot cumulative density function for the date of recovery, which we define as the first date when the infections fall below the initial level $I_{i0}$. The density function is calculated by examining each of the future trajectories $I^s_{iT+h}$ for $h = 1, \ldots, 60$ generated by Algorithm 1. For China the probability that the infection rate will fall below $I_{i0}$ over the two month period is greater than 90%, whereas for Germany the probability is slightly less than 50%.

In Figure 7 we overlay two weeks of actual infections onto density forecasts generated from the 2020-04-04 (top panels) and 2020-04-18 (bottom panels). On 2020-04-04 the model forecasts a fairly quick recovery from the pandemic. This "optimism" is consistent with Figure 2 which indicates that $|\hat{\gamma}_{1i} + \hat{\delta}_{1i}|$ is larger in the earlier sample. Comparing the predictive density to the actuals, indicate that while the actual realizations are still within the 10-90% bands, the longer horizon the further they are in the tails. Thus, the model overestimates the speed of recovery. Two weeks later, on 2020-04-18, the estimates have

Figure 7: Interval Forecasts and Actuals




*Notes:* The vertical lines indicate the forecast origins. The circles indicate actual infections. The solid lines prior to the forecast origin represent in-sample one-step-ahead forecasts. The solid lines after the forecast origin represent medians of the posterior predictive distribution. The grey shaded bands indicate the 20%-80% (dark) and 10%-90% (light) interquantile ranges of the posterior predictive distribution.

caught on to the slower decline, which translates into a more drawn-out recovery. While for South Korea the width of the bands associated with the short-run forecasts is smaller for the 2020-04-18 sample, the width for Germany increases. The 2020-04-18 predictions are remarkably accurate: between 2020-04-18 and 2020-05-02 the median forecasts are very close to the actuals for all three countries.

We now turn to a more systematic evaluation of the forecasts, focusing on the coverage probability of interval forecasts represented by the bands in Figures 6 and 7. Denote the interval forecasts represented by the bands by $C_{i,T+h|T}(Y_{1:N,0:T})$. In addition to the 20%-80% and 10%-90% intervals, we will also consider 25%-75% and 5%-95% intervals. Table 1 reports the cross-sectional empirical coverage frequency, defined as

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\{y_{iT+h} \in C_{i,T+h|T}(Y_{1:N,0:T})\},$$

for different forecast origins and targets. The empirical coverage frequencies can be compared

Table 1: Interval Forecast Performance

| Forecast Origin | Target | Quantile Range & Coverage of Interval Forecasts 0.25 to 0.75 Cov. 0.50 | 0.20 to 0.80 Cov. 0.60 | 0.10 to 0.90 Cov. 0.80 | 0.05 to 0.95 Cov. 0.90 |
|---|---|---|---|---|---|
| One Week Ahead | | | | | |
| Apr-25 | May-02 | 0.41 | 0.57 | 0.85 | 0.92 |
| Apr-18 | Apr-25 | 0.36 | 0.48 | 0.83 | 0.90 |
| Apr-11 | Apr-18 | 0.41 | 0.55 | 0.82 | 0.93 |
| Apr-04 | Apr-11 | 0.05 | 0.14 | 0.47 | 0.76 |
| Two Weeks Ahead | | | | | |
| Apr-18 | May-02 | 0.28 | 0.41 | 0.69 | 0.86 |
| Apr-11 | Apr-25 | 0.16 | 0.30 | 0.69 | 0.87 |
| Apr-04 | Apr-18 | 0.00 | 0.00 | 0.25 | 0.52 |
| Three Weeks Ahead | | | | | |
| Apr-11 | May-02 | 0.09 | 0.10 | 0.50 | 0.77 |
| Apr-04 | Apr-25 | 0.00 | 0.00 | 0.09 | 0.40 |
| Four Weeks Ahead | | | | | |
| Apr-04 | May-02 | 0.00 | 0.00 | 0.02 | 0.22 |

*Notes:* We report empirical coverage frequencies.

to the nominal credible level of the interval forecasts. However, this comparison is delicate. While in finite samples the two objects tend to differ, one can show that if the posterior distribution of $(\rho, \xi)$ concentrates around a limit point as $N \longrightarrow \infty$, then under suitable regularity conditions, the discrepancy between the empirical coverage frequency and the credible level will vanish.[12]

The results in Table 1. Over a one-week horizon, and starting with the 2020-04-11 forecasts, the empirical coverage frequency is fairly close to the nominal credible level. For the 2020-04-04 origin there is a larger discrepancy. At the two-week horizon, again starting with the 2020-04-11 forecast, the empirical coverage frequency is somewhat smaller than the nominal coverage level, but still relatively close.

For forecast horizons of more than two weeks, the empirical coverage frequency is unfortunately very low. A look at the full set of forecasts generated based on 2020-04-18 and plotted on the companion website `https://laurayuliu.com/covid19-panel-forecast/` provides some insights into the prediction errors. Large forecast errors occur for many countries/regions that have not yet reached the peak of the infections, e.g., Afghanistan or

[12]See Liu, Moon, and Schorfheide (2019) for a more detailed discussion.

Algeria. While the model predicts a likely downturn over the next two weeks, the actual number of infections in these countries steadily rises and eventually moves outside of the forecast bands. This problem was more severe in early April, because in very few locations the infections had peaked, resulting in the low coverage rate for the 2020-04-04 forecasts.

We will publish forecasts online at `https://laurayuliu.com/covid19-panel-forecast/` on a weekly basis and continue to monitor the empirical coverage frequencies.

# 5 Conclusion

We adopted a panel forecasting model initially developed for applications in economics to forecast active Covid-19 infections. A key feature of our model is that it exploits the experience of countries/regions in which the epidemic occurred early on, to sharpen forecasts and parameter estimates for locations in which the outbreak took place later in time. At the core of our model is a specification that assumes that the growth rate of active infections can be represented by autoregressive fluctuations around a downward sloping deterministic trend function with a break. Our specification is inspired by infection dynamics generated from a simple SIR model.

According to our model, there is a lot of uncertainty about the evolution of infection rates, due to parameter uncertainty and the realization of future shocks. Moreover, due to the inherent nonlinearities, predictive densities for the level of infections are highly skewed and exhibit substantial upside risk. Consequently, it is important to report density or interval forecasts, rather than point forecasts. We find that over a one-week horizon the empirical coverage frequency of our interval forecasts is close to the nominal credible level.

A natural extension of our model is to allow for additional, data-determined breaks in the deterministic trend function as the pandemic unfolds and countries/regions are adopting new policies that accelerate or decelerate the spread of the virus and as more and more people become resistant to the infection. It is also worthwhile to link the heterogeneous coefficient estimates (or transformations thereof) to country-specific variables that measure social norms and policies to fight the pandemic. This could be done in a second step through ex-post regressions with the heterogeneous coefficient estimates as left-hand-side variables or, more elegantly, in a correlated random effects framework.

# References


Avery, C., W. Bossert, A. T. Clark, G. Ellison, and S. Fisher Ellison (2020): "Policy Implications of Models of the Spread of Coronavirus: Perspectives and Opportunities for Economists," *Covid Economics, CEPR Press*, 12, 21–68.

Bertozzi, A. L., E. Franco, G. Mohler, M. B. Short, and D. Sledge (2020): "The Challenges of Modeling and Forecasting the Spread of COVID-19," *arXiv*, 2004.0474v1.

Brown, L. D., and E. Greenshtein (2009): "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-dimensional Vector of Normal Means," *The Annals of Statistics*, pp. 1685–1704.

Chib, S. (1998): "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86, 221–241.

de Boor, C. (1990): *Splinefunktionen*, vol. Lectures in Mathematrics, ETH Zürich. Birkhäuser Verlag, Basel.

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998): "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society B*, 60(2), 333–350.

Eichenbaum, M. S., S. Rebelo, and M. Trabandt (2020): "The Macroeconomics of Epidemics," *NBER Working Paper*, 26882.

Fernandez-Villaverde, J., and C. I. Jones (2020): "Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities," *Manuscript, University of Pennsylvania*.

Giordani, P., and R. Kohn (2008): "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models," *Journal of Business and Economics Statistics*, 26, 66–77.

Glover, A., J. Heathcote, D. Krueger, and J.-V. Rios-Rull (2020): "Health versus Wealth: On the Distributional Effects of Controlling a Pandemic," *Covid Economics, CEPR Press*, 6, 22–64.

Gu, J., and R. Koenker (2017a): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Economics*, 32(3), 575–599.

——— (2017b): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *Journal of Business & Economic Statistics*, 35(1), 1–16.

Haavelmo, T. (1944): "The Probability Approach in Econometrics," *Econometrica*, 12, 1–115.

Kermack, W. O., and A. G. McKendrick (1927): "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772), 700–721.

Koop, G., and S. M. Potter (2009): "Elicitation in Multiple Change-Point Models," *International Economic Review*, 50(3), 751–772.

Krueger, D., H. Uhlig, and T. Xie (2020): "Macroeconomic Dynamics and Reallocation in an Epidemic," *Manuscript, University of Pennsylvania*.

Liu, L. (2020): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *arXiv preprint arXiv:1805.04178*.

Liu, L., H. R. Moon, and F. Schorfheide (2019): "Forecasting with a Panel Tobit Models," *NBER Working Paper*, 26569.

——— (2020): "Forecasting with Dynamic Panel Data Models," *Econometrica*, 88(1), 171–201.

Llera, A., and C. Beckmann (2016): "Estimating an Inverse Gamma distribution," *arXiv preprint arXiv:1605.01019*.

Murray, C. J. (2020): "Forecasting the Impact of the First Wave of the COVID-19 Pandemic on Hospital Demand and Deaths for the USA and European Economic Area Countries," *medRxiv, https://doi.org/10.1101/2020.04.21.20074732*.

Smith, M., and R. Kohn (1996): "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343.

Stock, J. H. (2020): "Dealing with Data Gaps," *Covid Economics, CEPR Press*, 3, 1–11.