Estimating Latent Fields in Stochastic Dynamical Systems - A Case Study of COVID-19 in New Mexico

Jaideep Ray, Cosmin Safta & Wyatt Bridgman

Sandia National Laboratories, Livermore, CA

June 6th, 2023
Introduction

• **Aim:** Devise a method to infer a spatial quantity, the spread-rate of a disease, using limited data of epidemiological dynamics (case-count data)

• **Dataset:** COVID-19 case-counts in the counties of New Mexico

• **Why?**
  • Novel outbreaks are detected by analyzing (very noisy) case-count time-series; detection often delayed
    • Reporting errors, stochastic behavior in small populations (sparsely populated areas)
    • Outbreak detections (anomalous change in epidemiological dynamics) often uncertain; wait for case-counts to increase

• **Hypothesis:** Detect new outbreaks using the latent spread-rate of a disease, not case-counts

• **Technical challenges:**
  • How to infer the spread-rate field?
  • How to impose the spatial correlations seen in data? What kind of spatial structures do we have?
  • How to compute the spread-rate fast, in a parallel manner?
The practical problem – outbreak detection

• Two ways – temporal methods (SPC) & spatiotemporal method
  • Data used: case-counts of a disease, disaggregated in time & space

• Temporal methods: Fundamentally, anomaly detection
  • Using historical data, do a 2-week forecast of case-counts & uncertainty bounds (usually 95th percentile)
  • Wait for data; if 3 consecutive days > than 95th percentile, alarm!

• Spatiotemporal methods: Use historical & neighborhood data (autocorrelation) to make forecasts

• Shortcomings
  • Need long time-series data, prefer to be high-count / low variance
  • Not really feasible for novel diseases
Approach

• **Hypothesis:**
  • Use (latent) spread-rate to detect outbreaks, not case-counts directly
    • Not affected by reporting errors & only depends on human mixing patterns (behavior)

• **Inferring the spread-rate**
  • Pose and solve an inverse problem for the spread-rate in each NM county
  • Spread-rates in counties are auto-correlated. Devise a Gaussian Markov Random Field (GMRF) model to capture spatial pattern
  • Reformulate a spatiotemporal inverse problem for spread-rates in M counties. Use GMRF to impose autocorrelation
  • Solve with MCMC (for accuracy) and Variational Inference (VI; approximate, but fast); compare estimated spread-rates

• **Test:** Can disease detection be done with spread-rates, even the approximate VI one?
Formulating the temporal problem

• Assume $q(t; \theta)$, # of people infected on day $t$, in Area A
• $y_t^{(obs)}$: Case-counts from a location; $y_t(\theta)$: Predictions by model $M(t; \theta)$
• Convolve with incubation period for modeled cases
  • $y_t(\theta) = \int_{t_0}^{t} q(\tau - t_0; \theta)f_{inc}(t - \tau)d\tau$
• Infer $p(\theta|y_t^{(obs)})$ via Bayesian inference, using $y_t^{(obs)}$ & $y_t(\theta) = M(t; \theta)$
  • Provides (infers) the latent spread-rate curve

• Likelihood assumes Gaussian errors; parameter vector $\theta$ is 4-dimensional
• Inference can be done with MCMC, VI etc.
  • 4-dimension inference is easy
• Forecasting: $y_{t^*}, t^* > T$ conditioned on $p(\theta|y_t^{(obs)})$
Detecting change in epidemiological dynamics

• Model allows estimation of (past) infection-rate; forecasting with it assumes that it will not change drastically

• If forecasts are wrong, it implies a change in spread-rate (new variant, changes in human behavior etc.)

• Our insight: This could be formalized into a rigorous outbreak detector / change in epidemiological dynamics

Flattening CA’s curve; first lockdown in March 2020
The spatiotemporal problem

- **Temporal estimation problem**: The posterior distribution

\[
p\left(\theta \mid y_t^{(obs)}\right) \propto \left(y_t^{(obs)} - M(t; \theta)\right)^T \Gamma^{-\frac{1}{2}} \left(y_t^{(obs)} - M(t; \theta)\right) p_{prior}(\theta), \quad \Gamma = \text{diag}(\sigma_A + \sigma_M y_t^{(obs)})
\]

- \(\theta\) is 4-dimensional; the inversion is 6 dimensional

- **The spatiotemporal estimation problem**: 
  - \(y_t^{(obs)}\) contains case-counts for all times till \(t\), from all areas \(A_j, j = 1 \cdots J\)
  - \(\Gamma\) spans over all time \(t\), and all \(A_j\) and must enforce all spatial autocorrelations. What is it?

- **Modeling the spatial problem**: 
  - Is there any spatial correlation? What form does it take?
  - What does \(\Gamma\) look like in a spatiotemporal inversion problem?
Spatial modeling

• Created a simple regression model for case-counts in NM
  • $Y = w_0 + \sum_k w_k \phi_k + \epsilon, \epsilon \sim N(0, \xi^2)$
  • $\phi_k$: exogenous covariates of epidemiology/risk factors (population, socioeconomic conditions, transport connectivity etc.)
  • $\epsilon$ shows spatial correlations in epidemiological dynamics not explained by exogenous covariates

• Clear spatial pattern
  • Rio Grande valley (inhabited; blue) shows similar $\epsilon$
  • Further out, red counties have similar behavior
  • Northwest / Southeast counties show max $\epsilon$

• To do:
  • Clearly, clustered, but need to get significance via a statistical test
  • Need to capture this pattern in a GMRF model
$\Gamma$ for GMRF

- Existence of clusters determined by Moran’s I test
- How far does autocorrelation extend in the (large) counties of NM?
  - Also determined by Moran’s I test, computed with 1-hop and 2-hop neighborhoods
  - **Finding**: autocorrelation is only between nearest neighbors
- **Precision matrix** $\Gamma^{-1} = \frac{1}{t^2}\left[I - \lambda W\right]$, $W$ is the nearest-neighbor connectivity matrix, $\lambda$ is the strength of spatial autocorrelation
- **Posterior**:
  - $p (\Theta | Y^{(obs)}_t) \propto \Pi_t \frac{\left(\psi_t^{(obs)} - M(t; \Theta)\right)^T \Gamma^{-\frac{1}{2}} \left(\psi_t^{(obs)} - M(t; \Theta)\right)}{|\Gamma|^\frac{1}{2}} p_{prior}(\Theta), \Theta = \{\theta_j\}, j = 1 \ldots J$
  - $\psi_t = M(t; \Theta)$ predicts case counts on Day $t$
  - $\Theta$ contains $4 \times J$ parameters to infer, along with $(\tau, \lambda)$; high-dimension even for $J = 3$
3- county results using MCMC

- Estimation with 3 counties
- Provides infection-rate curve too
Speeding up with VI

• ** Curse of dimensionality:** Dimensionality of the inverse problem grows as $\sim 4J$, $J = \# \text{ of areal units}$
  • For NM, $J = 33$. Too high-dimensional for MCMC

• **Solution:** mean-field variational inference
  • Approximate $p\left(\theta \mid Y_t^{obs}\right)$ as a multivariate Gaussian with a diagonal covariance
  • Estimation now implies estimating $(\bar{\theta}_k, \text{Var}(\theta_k)), \ k = 1...K (=4J)$
  • Test on Santa Fe county

• **Mathematical development**
  • Objective function (likelihood) to be maximized to estimate $(\bar{\theta}_k, \text{Var}(\theta_k))$
  • Parallel iterative methods to optimize (Adams)

• **Effect of approximation:** VI underestimates uncertainty
  • Much faster & already parallelized
3-county results using VI

- Estimation with 3 counties
9-county inference with VI

Bernalillo, Santa Fe, Valencia, Sandoval, McKinley, San Juan, Rio Arriba, Chaves, Dona Ana
Detecting the fall wave, 2020

• Detect the arrival of the Fall wave of 2020 in Bernalillo county

• Process:
  • Infer spread-rate using data till Sept 15th; forecast ahead w/ 95th percentile; detect outliers
  • Redo with negative binomial fit (RKI; Hohle & Paul, 2008)

• Result:
  • Our method detects the start of the fall wave; RKI method fails
  • RKI’s time-series method needs long training data (>2 months)
  • We exploit knowledge of incubation period & parameterized infection-rate profile
Conclusions

• We have developed a VI method to infer a latent field, give indirect observations
  • Our case: latent infection-rate or spread-rate field, from case-count data
  • Requires a forward problem (epidemiological problem); spread-rate is smooth in space-time
• Algorithmic innovations: Estimation is high-dimensional; MCMC not up-to-the-task
  • Requires a Gaussian Markov Random Field model to spatially regularize (enforce spatial auto-correlation)
  • Estimation performed using Variational Inference
  • Tested on the counties of New Mexico, COVID-19 data
• Final use: Detect arrival of Fall wave in NM, posing it as an anomalous epidemiological behavior
  • Detect better than conventional detectors that employ case-counts natively
    • Often conventional detectors are not robust – can get better performance with smaller training data
  • Better detection artefact of exploiting a smooth infection-rate, unaffected by reporting errors etc.
Acknowledgements

Backup slides: Scalability to more counties, approximation error
### Comparison of MCMC and VI

#### Bernalillo

<table>
<thead>
<tr>
<th></th>
<th>MCMC</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>(-6.5e+00 3.5e+00)</td>
<td>(-1.1e+01 3.2e-01)</td>
</tr>
<tr>
<td>$N$</td>
<td>(4.4e+03 1.3e+02)</td>
<td>(1.54e+02 4.81e+00)</td>
</tr>
<tr>
<td>$k$</td>
<td>(4.4e+00 8.3e-01)</td>
<td>(6.6e+00 8.7e-02)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>(1.1e+01 1.4e+00)</td>
<td>(9.0e+00 1.30e-01)</td>
</tr>
<tr>
<td>RMSE</td>
<td>9.77</td>
<td>43.16</td>
</tr>
</tbody>
</table>

#### Valencia

<table>
<thead>
<tr>
<th></th>
<th>MCMC</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>(-1.8e+01 5.6e+00)</td>
<td>(-7.0e+00 3.3e-01)</td>
</tr>
<tr>
<td>$N$</td>
<td>(4.9e+02 1.1e+01)</td>
<td>(1.6e+02 5.1e+0)</td>
</tr>
<tr>
<td>$k$</td>
<td>(8.3e+00 1.7e+00)</td>
<td>(7.6e+00 9.9e-02)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>(7.5e+00 9.7e-01)</td>
<td>(7.8e+00 1.1e-01)</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.08</td>
<td>5.6</td>
</tr>
</tbody>
</table>

#### Santa Fe

<table>
<thead>
<tr>
<th></th>
<th>MCMC</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>(-1.5e+01 4.9e+00)</td>
<td>(-1.1e+01 3.6e-01)</td>
</tr>
<tr>
<td>$N$</td>
<td>(8.2e+02 3.3e+01)</td>
<td>(1.7e+02 6.2e+00)</td>
</tr>
<tr>
<td>$k$</td>
<td>(4.6e+00 8.6e-01)</td>
<td>(5.0e+00 9.7e-02)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>(1.5e+01 2.3e+00)</td>
<td>(1.5e+01 2.6e-01)</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.52</td>
<td>2.4</td>
</tr>
</tbody>
</table>

#### Regional parameters

<table>
<thead>
<tr>
<th></th>
<th>MCMC</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^2$</td>
<td>(2.0e+01 6.3e+00)</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>(5.3e-01 3.0e-02)</td>
<td></td>
</tr>
<tr>
<td>$\sigma_a$</td>
<td>(2.1e-06 3.8e-07)</td>
<td></td>
</tr>
<tr>
<td>$\sigma_M$</td>
<td>(5.8e-03 3.0e-03)</td>
<td></td>
</tr>
</tbody>
</table>

- VI uncertainties (std dev) are always under-estimated. In some cases, the parameter means also disagree significantly.
- Prediction RMSEs always larger than MCMC, but expected (approximate method).