Variational Bayesian Inference for Spatial Negative Binomial Count Data Models with Unobserved Heterogeneity

Prateek Bansal\textsuperscript{1,*}, Rico Krueger\textsuperscript{2,*}, Michel Bierlaire\textsuperscript{2}, Daniel J. Graham\textsuperscript{1}

\textsuperscript{1}Imperial College London, \textsuperscript{2}Ecole Polytechnique Fédérale de Lausanne

* Equal contribution

20\textsuperscript{th} Swiss Transport Research Conference
Motivation

Spatial count data models are used to explain and predict frequencies of aggregate phenomena such as traffic accidents in geographically distinct entities (e.g. census tracts).

Spatial count data may exhibit …

- **spatial dependence** due to systematic correlation in unobserved factors.
- **spatial heterogeneity** due to spatially varying effects of covariates on dependent variable.
Motivation ii

Challenges

- Accounting for spatial effects complicates model estimation.
- Datasets are growing in size and are becoming available on a streaming basis.
Markov chain Monte Carlo (MCMC)

- Approximate posterior $P(\theta|y)$ numerically through samples from a Markov chain.
- Use Metropolis-Hastings algorithm and Gibbs sampling to construct Markov chain.
- Issues:
  - computation times,
  - storage of posterior draws,
  - convergence assessment,
  - serial correlation.
Variational Bayes (VB)

- Recast Bayesian estimation as an optimisation problem.
- Approximate posterior $P(\theta | y)$ analytically through a parametric variational distribution $q(\theta | \nu)$.
- Advantages:
  - Reduced storage requirements
  - Straightforward convergence assessment
  - Serial correlation no longer a concern
  - Stochastic optimisation
Research objective

Develop VB method for fast estimation of negative binomial model with spatial dependence and heterogeneity.
Model formulation

Negative binomial likelihood

\[ y_i \sim \text{NB}(r, p_i), \quad p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}, \quad \psi_i = X_i^\top \beta_i + \phi_i, \quad i = 1, \ldots, N \]

Hierarchical prior for link function parameters to accommodate spatial heterogeneity

\[ \beta_i \sim \text{Normal}(\mu, \Sigma), \quad i = 1, \ldots, N \]

Matrix exponential spatial specification (MESS; LeSage and Pace, 2007) of spatial dependence

\[ S\phi = \exp(\tau W)\phi = \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma^2 I_N) \]
Pólya-Gamma data augmentation (Polson et al., 2013; Zhou et al., 2012)

**Issue**
- Negative binomial distribution does not have a conjugate prior.
- Thus, the conditional distributions of the link function parameters do not constitute known distributions.

**Remedy**
- Introduce Pólya-Gamma-distributed auxiliary variables
  \[ \omega_i \sim \text{PG}(y_i + r, 0), \, i = 1, \ldots, N. \]
- Conditional on \( \omega \) and \( r \), the likelihood of observed counts is translated into a heteroskedastic Gaussian likelihood.
VB seeks to minimise the KL divergence (probability distance) between an approximating variational distribution $q(\theta)$ and the posterior of interest $P(\theta|y)$:

$$q^*(\theta) = \arg\min_q \left\{ \text{KL}(q(\theta)||P(\theta|y)) \right\}.$$ 

$q$ must be selected by the analyst. Its expressiveness determines the quality of the variational approximation and the complexity of the estimation problem.
Mean-field variational Bayes (MFVB)

- Impose posterior independence between parameter blocks $\Theta_1, \ldots, \Theta_J$:
  
  $$q(\Theta) = \prod_{j=1}^{J} q(\Theta_j)$$

- For conditionally-conjugate models, MFVB leads to a simple coordinate ascent algorithm.

- $\tau$ and $\sigma^2$ are recovered poorly under MFVB assumption.
Integrated nonfactorised variational Bayes (INFVB; Han et al., 2013; Wu, 2018)

- Decompose parameters $\Theta$ into two disjoint subsets $\{\Theta_c, \Theta_d\}$ to specify more flexible variational distribution:

$$q_{\text{INFVB}}(\Theta) = q(\Theta_c | \Theta_d)q(\Theta_d)$$

- Direct minimisation of KL divergence between $q_{\text{INFVB}}(\Theta)$ and posterior is challenging.

- Conditional on $\Theta_d$, INFVB involves a simple coordinate ascent algorithm.

- Define a grid with points $\Theta_d^{(g)}$ and run MFVB separately for each grid point. Then, compute weight of each grid point using the Boltzmann distribution.

- Select $\Theta_d = \{\tau, \sigma^{-2}\}$. 

Bansal, Krueger, Bierlaire, Graham (ICL, EPFL)
Data

- Benchmark INFVB against MCMC.
- Youth pedestrian injury counts in 603 census tracts of the Bronx and Manhattan from 2005–14 (Morris et al., 2019).
Sample description (N = 603)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth pedestrian injury count, 2005-14</td>
<td>9.69</td>
<td>8.35</td>
<td>0.00</td>
<td>44.00</td>
</tr>
<tr>
<td>Social fragmentation index</td>
<td>2.02</td>
<td>2.73</td>
<td>-4.50</td>
<td>18.67</td>
</tr>
<tr>
<td>Avg. annual daily traffic (AADT) in 1000, 2015</td>
<td>44.48</td>
<td>46.79</td>
<td>2.09</td>
<td>276.48</td>
</tr>
<tr>
<td>Private vehicle commute mode share, 2010-14</td>
<td>0.19</td>
<td>0.15</td>
<td>0.00</td>
<td>0.76</td>
</tr>
</tbody>
</table>
### Results: Estimation summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimation time [s]</th>
<th>LPPD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCMC</td>
<td>1733.77</td>
<td>-1713.24</td>
<td>3.79</td>
</tr>
<tr>
<td>INFVB</td>
<td>20.24</td>
<td>-1704.40</td>
<td>3.88</td>
</tr>
</tbody>
</table>

Note: LPPD = log-posterior predictive density. RMSE = root mean square error.
Results: Marginal posterior distributions

- **Intercept**
- **Priv. veh. commute mode share**
- **Social fragmentation index - mean**
- **Social fragmentation index - std. dev.**
- **log(AADT) - mean**
- **log(AADT) - std. dev.**

**References**

Bansal, Krueger, Bierlaire, Graham (ICL, EPFL)
Conclusion

Key points

- In case study, INFVB is more than 80 times faster than MCMC, while offering similar estimation accuracy.

Next steps

- Further testing
- Extension to models with spatio-temporal dependencies
- Online estimation
Thank you
References I


