Bayesian Network Regularized Regression for Crime Modeling

Luis Carvalho
Joint work with Liz Upton
Dept. of Mathematics and Statistics
Boston University
lecarval@math.bu.edu

BU-Keio Workshop, August 2016
A motivating example: residential burglary in Boston
Introduction

A motivating example: residential burglary in Boston

Potential goals:
- Understanding crime rates: covariates? predictions?
A motivating example: residential burglary in Boston

Potential goals:
- Understanding crime rates: covariates? predictions?
- Identifying “hot zones” for intervention
Residential Burglary

- Data description: ~ 7K crimes occurring from July 2012 to October 2015 in Boston, provided by data.cityofboston.gov
- Reported occurrences are pooled in time and by intersection
- Covariates: averaged tax income, district type, distance to nearest police station
- Network with ~ 13K nodes, provided by boston.opendata.arcgis.com
Residential Burglary

- Data description: $\sim 7K$ crimes occurring from July 2012 to October 2015 in Boston, provided by data.cityofboston.gov
- Reported occurrences are pooled in time and by intersection
- Covariates: averaged tax income, district type, distance to nearest police station
- Network with $\sim 13K$ nodes, provided by boston.opendata.arcgis.com

First take: for $v$ in a network (undirected simple graph) $G$,

$$Y_v \overset{iid}{\sim} \text{Po} \left[ \exp \left( x_v^T \beta \right) \right]$$
Residential Burglary

- Data description: \( \sim 7K \) crimes occurring from July 2012 to October 2015 in Boston, provided by data.cityofboston.gov

- Reported occurrences are pooled in time and by intersection

- Covariates: averaged tax income, district type, distance to nearest police station

- Network with \( \sim 13K \) nodes, provided by boston.opendata.arcgis.com

- First take: for \( v \) in a network (undirected simple graph) \( G \),

\[
Y_v \overset{iid}{\sim} \text{Po}
\left[ \exp \left( \mathbf{x}_v^T \beta \right) \right]
\]

but:

- Crime rates are not spatially homogeneous
Residential Burglary

- Data description: \( \sim 7K \) crimes occurring from July 2012 to October 2015 in Boston, provided by data.cityofboston.gov

- Reported occurrences are pooled in time and by intersection

- Covariates: averaged tax income, district type, distance to nearest police station

- Network with \( \sim 13K \) nodes, provided by boston.opendata.arcgis.com

- First take: for \( v \) in a network (undirected simple graph) \( G \),

\[
Y_v \overset{iid}{\sim} \text{Po} \left[ \exp \left( x_v^T \beta \right) \right]
\]

but:

- Crime rates are not spatially homogeneous
- Crime rates can vary sharply
Network Regularized Regression

- Addressing the first issue,

\[ Y_v \sim \text{Po} \left[ \exp \left( x_v^T \beta(v) \right) \right] \]

where \( \beta \) is now network indexed
Network Regularized Regression

- Addressing the first issue,

  \[ Y_v \overset{\text{ind}}{\sim} \text{Po} \left( \exp \left( x_v^T \beta(v) \right) \right) \]

  where \( \beta \) is now network indexed

- To avoid overfitting, we impose smoothness on \( \beta \), e.g., under a single intercept model,

  \[
  \hat{\beta} := \arg \min_{\beta} \| Y - \beta \|_2^2 + \lambda \| M \beta \|_2^2 \\
  = \arg \min_{\beta} D(Y; \beta) + \lambda \beta^T M^T M \beta
  \]

  where \( M \) is a differential operator and \( \lambda \) is a roughness penalty
Network Regularized Regression

- Addressing the first issue,

\[ Y_v \overset{\text{ind}}{\sim} \text{Po} \left[ \exp \left( x_v^\top \beta(v) \right) \right] \]

where \( \beta \) is now network indexed

- To avoid overfitting, we impose smoothness on \( \beta \), e.g., under a single intercept model,

\[ \hat{\beta} := \arg\min_{\beta} ||Y - \beta||_2^2 + \lambda ||M\beta||_2^2 \]

\[ = \arg\min_{\beta} D(Y; \beta) + \lambda \beta^\top M^\top M \beta \]

where \( M \) is a differential operator and \( \lambda \) is a roughness penalty

- Similar works: network kernel-based regression (Smola and Kondor, 2003; Kolaczyk, 2009), and, more generally, functional data analysis (Ramsay and Silverman, 1996)
For network indexed coefficients, with $M$ the oriented weighted incidence matrix:

$$\beta^\top M^\top M \beta := \beta^\top L_w \beta = \sum_{(u,v) \in E(G)} w_{uv} (\beta(u) - \beta(v))^2$$

i.e., $L_w$ is weighted Laplacian
Network Regularized Regression

- For network indexed coefficients, with $M$ the oriented weighted incidence matrix:

$$\beta^\top M^\top M \beta := \beta^\top L_w \beta = \sum_{(u,v) \in E(G)} w_{uv} (\beta(u) - \beta(v))^2$$

i.e., $L_w$ is weighted Laplacian

- With $L_w := \Phi \Xi \Phi^\top$, $\Xi := \text{Diag}_{i=1,...,|V(G)|}(\xi_i)$, we adopt a basis expansion for $\beta$, $\beta = \Phi_{1:k} \theta$, $k \leq |V(G)|$, so the penalty becomes:

$$\beta^\top L_w \beta = \theta^\top \Phi_{1:k} \Phi \Xi \Phi^\top \Phi_{1:k} \theta = \theta^\top \text{Diag}_{i=1,...,k}(\xi_i) \theta$$
Network Regularized Regression

- For network indexed coefficients, with $M$ the oriented weighted incidence matrix:

$$\beta^\top M^\top M \beta := \beta^\top L_w \beta = \sum_{(u,v) \in E(G)} w_{uv} (\beta(u) - \beta(v))^2$$

i.e., $L_w$ is weighted Laplacian

- With $L_w := \Phi \Xi \Phi^\top$, $\Xi := \text{Diag}_{i=1,\ldots,|V(G)|}(\xi_i)$, we adopt a basis expansion for $\beta$, $\beta = \Phi_{1:k} \theta$, $k \leq |V(G)|$, so the penalty becomes:

$$\beta^\top L_w \beta = \theta^\top \Phi_{1:k}^\top \Phi \Xi \Phi^\top \Phi_{1:k} \theta = \theta^\top \text{Diag}_{i=1,\ldots,k}(\xi_i) \theta$$

- Under a Bayesian formulation, $\hat{\beta}$ is the posterior mode when

$$Y_v \mid \theta \stackrel{\text{ind}}{\sim} \text{Po}\left[ \exp\left(\phi_{kv}^\top \theta\right) \right]$$

$$\theta \sim N\left(0, \text{Diag}_{i=1,\ldots,k}\left\{ (\lambda \xi_i)^{-1} \right\} \right)$$
Toy example: $\mathbf{Y} = (10, 2, 3, 4)$, vertex 1 connected to triangle with vertices 2, 3, and 4, $w(u, v) \propto \exp\{-d(u, v)/2\}I[d(u, v) > 0]$, and

$$D = \begin{bmatrix} 0 & 10 & 0 & 0 \\ 10 & 0 & 5 & 3 \\ 0 & 5 & 0 & 2 \\ 0 & 3 & 2 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0.02 & -0.02 & 0 & 0 \\ -0.02 & 0.85 & -0.22 & -0.61 \\ 0 & -0.22 & 1.22 & -1 \\ 0 & -0.61 & -1 & 1.61 \end{bmatrix}$$
Bayesian Network Regularized Regression

- Addressing the issue of abrupt rate changes,

\[ Y_v | \zeta, \beta, Z_v \overset{\text{ind}}{\sim} \text{Po}\left[ \exp \left( Z_v \zeta + (1 - Z_v) \mathbf{x}_v^T \beta(v) \right) \right] \]

\[ Z_v | \gamma \overset{\text{ind}}{\sim} \text{Bern}\left[ \logit^{-1} \left( \mathbf{u}_v^T \gamma(v) \right) \right] \]

where:

- \( \zeta \) is the “background” crime rate
- \( Z_v \) codes for \( v \) being in a “hot zone”, also varying smoothly
- Both \( \beta \) and \( \gamma \) are network indexed and assume a basis expansion as before
Bayesian Network Regularized Regression

- Addressing the issue of abrupt rate changes,

\[
Y_v \mid \zeta, \beta, Z_v \overset{\text{ind}}{\sim} \text{Po} \left[ \exp \left( Z_v \zeta + (1 - Z_v) x_v^\top \beta(v) \right) \right]
\]

\[
Z_v \mid \gamma \overset{\text{ind}}{\sim} \text{Bern} \left[ \logit^{-1} \left( u_v^\top \gamma(v) \right) \right]
\]

where:

- \( \zeta \) is the “background” crime rate
- \( Z_v \) codes for \( v \) being in a “hot zone”, also varying smoothly
- Both \( \beta \) and \( \gamma \) are network indexed and assume a basis expansion as before

- Using basis coefficients, \( x_v^\top \beta(v) \to D_X(v)^\top \theta \) and \( u_v^\top \gamma(v) \to D_U(v)^\top \omega \),

\[
Y_v \mid \zeta, \theta, Z_v \overset{\text{ind}}{\sim} \text{Po} \left[ \exp \left( Z_v \zeta + (1 - Z_v) D_X(v)^\top \theta \right) \right]
\]

\[
Z_v \mid \omega \overset{\text{ind}}{\sim} \text{Bern} \left[ \logit^{-1} \left( D_U(v)^\top \omega \right) \right]
\]
Bayesian Network Regularized Regression

- Quick methodological recap:
  - Network regularized regression as a building block,
    \[ Y_v | \theta \sim F \left( g^{-1}(DX(v)^\top \theta) \right), \quad \theta \sim N(0, \lambda^{-1}_\theta \Omega(X, L_w(G))^{-}) \]
  - Change regions using latent network-indexed indicators \( Z \) and conditional responses
    \[ Z_v | \omega \sim \text{Bern} \left( \text{logit}^{-1}(DU(v)^\top \omega) \right), \quad \omega \sim N(0, \lambda^{-1}_\omega \Omega(U, L_w(G))^{-}) \]
Bayesian Network Regularized Regression

- Quick methodological recap:
  - Network regularized regression as a building block,
    \[ Y_v \mid \theta \sim \mathcal{F} \left[ g^{-1} \left( D_X(v) \top \theta \right) \right], \quad \theta \sim N \left( 0, \lambda_\theta^{-1} \Omega(X, L_w(G)) \right) \]
  - Change regions using latent network-indexed indicators \( Z \) and conditional responses
    \[ Z_v \mid \omega \sim \text{Bern} \left[ \logit^{-1} \left( D_U(v) \top \omega \right) \right], \quad \omega \sim N \left( 0, \lambda_\omega^{-1} \Omega(U, L_w(G)) \right) \]
  - There are now two main practical problems:
    - How to define the hyper-parameters controlling the smoothness of \( \beta \) and \( \gamma \)?
Bayesian Network Regularized Regression

- Quick methodological recap:
  - Network regularized regression as a building block,
    \[ Y_v \mid \theta \sim \mathcal{F} \left[ g^{-1} \left( D_X(v)^\top \theta \right) \right], \quad \theta \sim \mathcal{N} \left( 0, \lambda^{-1}_\theta \Omega(X, L_w(G))^{-} \right) \]
  - Change regions using latent network-indexed indicators \( Z \) and conditional responses
    \[ Z_v \mid \omega \sim \text{Bern} \left[ \logit^{-1} \left( D_U(v)^\top \omega \right) \right], \quad \omega \sim \mathcal{N} \left( 0, \lambda^{-1}_\omega \Omega(U, L_w(G))^{-} \right) \]
  - There are now two main practical problems:
    - How to define the hyper-parameters controlling the smoothness of \( \beta \) and \( \gamma \)?
    - How to fit this model efficiently for large scale datasets?
• Three main sets of hyper-parameters: $\theta \sim \mathcal{N}(0, \lambda^{-1} \Omega(X, L_w(G))^{-1})$, where $\Omega(X, L_w(G)) := D_X^\top L_w D_X$ and $D_X$ depends on $\Phi_{1:k}$.
Three main sets of hyper-parameters: \( \theta \sim N\left(0, \lambda^{-1}\Omega(X, L_w(G))^{-1}\right) \), where \( \Omega(X, L_w(G)) := D_X^\top L_w D_X \) and \( D_X \) depends on \( \Phi_{1:k} \).

To define \( L_w \) we need a measure of similarity as weights in \( G \): in our application, we use \( w(u, v) \propto \exp\left\{ -d(u, v)/\psi \right\} \) and set the “network range” \( \psi \) such that median similarity is 0.8.
Three main sets of hyper-parameters: \( \theta \sim N\left(0, \lambda^{-1} \Omega(X, L_w(G))^\top \right) \), where \( \Omega(X, L_w(G)) := D_X^\top L_w D_X \) and \( D_X \) depends on \( \Phi_{1:k} \)

To define \( L_w \) we need a measure of similarity as weights in \( G \): in our application, we use \( w(u, v) \propto \exp\left\{ -d(u, v)/\psi \right\} \) and set the “network range” \( \psi \) such that median similarity is 0.8

Penalty \( \lambda \) and basis rank \( k \) can be defined jointly using leave-one-out cross-validation via PRESS working residuals (“LOOP”)

![Graph 1](image1.png)

![Graph 2](image2.png)
Model Fitting

Crime Counts: Predicted and Actual

Pearson Residuals

Latent Variable = 0
Latent Variable = 1
Model Fitting

Probability of Zero Crime

Latent Variable Values

- 0.00 - 0.19
- 0.19 - 0.48
- 0.48 - 0.75
- 0.75 - 0.95
- 0.95 - 1.00
Model Fitting

Parcel Gross Tax

Tax Effect

-2.33 - -0.01
-0.01 - 0.04
0.04 - 0.09
0.09 - 0.21
0.21 - 2.17
Conclusions and future work

- Summary
Conclusions and future work

- Summary
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
Conclusions and future work

• Summary
  • Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  • Methodology can be used as a building block for more elaborated models
Conclusions and future work

- Summary
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  - Methodology can be used as a building block for more elaborated models
  - Main concern: representative models and computational efficiency
Conclusions and future work

- Summary
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  - Methodology can be used as a building block for more elaborated models
  - Main concern: representative models and computational efficiency
- New challenges
Conclusions and future work

- **Summary**
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  - Methodology can be used as a building block for more elaborated models
  - Main concern: representative models and computational efficiency

- **New challenges**
  - Refinements and extensions: dynamic model, basis selection, covariance structure
Conclusions and future work

- **Summary**
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  - Methodology can be used as a building block for more elaborated models
  - Main concern: representative models and computational efficiency

- **New challenges**
  - Refinements and extensions: dynamic model, basis selection, covariance structure
  - Other applications in Biology, Epidemiology, and Engineering
Conclusions and future work

- **Summary**
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  - Methodology can be used as a building block for more elaborated models
  - Main concern: representative models and computational efficiency

- **New challenges**
  - Refinements and extensions: dynamic model, basis selection, covariance structure
  - Other applications in Biology, Epidemiology, and Engineering
  - Bayesian *network* regression (for topology inference)
Conclusions and future work

• **Summary**
  - Network regularization is useful in a number of applications and can be more suitable than other types of regularization
  - Methodology can be used as a building block for more elaborated models
  - Main concern: representative models and computational efficiency

• **New challenges**
  - Refinements and extensions: dynamic model, basis selection, covariance structure
  - Other applications in Biology, Epidemiology, and Engineering
  - Bayesian *network* regression (for topology inference)

Thank you!