Dynamic Models for Time Series of Counts with Application to Transportation Safety

Volodymyr Serhiyenko
Department of Statistics,
University of Connecticut, Storrs

June 6, 2013

Ongoing project with Nalini Ravishanker, Dept. of Statistics, John Ivan and Md Saidul Islam, Dept. of Civil & Env. Eng.
Outline

- Description of Pedestrian Crash/Accident Data;
- MV Compositional Model for Different Severity Types;
- Static and Dynamic Models for Time Series of Crash Counts;
- Summary and Future Work.
Our Project Discusses:

1. Fit/Predict rates/proportions of crashes by severity type;
2. Fit/Predict total crash counts using:
   - Static GLM;
   - Dynamic GLM;
3. Use (1) and (2) to fit/predict crash counts by severity;
4. Separate DGLM’s for crash counts by severity;
Description of Highway Safety Data

- Crash Severity is defined as the highest injury severity experienced by any individual involved:
  - K: fatal;
  - A: severe injury;
  - B: evident minor injury;
  - C: non-evident possible injury;
  - O: property damage only.

- For each month, we used aggregated crash counts at each severity level, classified by area type (rural or urban area) as well as monthly aggregated Vehicle Miles Traveled (VMT) based on an average daily VMT for each year by area type (in CT from Jan. 1995 to Dec. 2009, data is available from ConnDOT).
MV Compositional Model for Different Severity Types

Let \( c_t = (C_{t1}, \cdots, C_{tG}) \): crash counts in month \( t \)
\( C_{tj} \geq 0, j = 1, \cdots, G \) (K,A,B,CO)
Rates/proportions:
\( X_{tj} = C_{tj}/\sum_{j=1}^{G} C_{tj}, j = 1, \cdots, G. \)
\( x_t \in S_t^g, g = G - 1 \)
Now we have \( G \)-variate compositional time series: \( (X_{t1}, \cdots, X_{tG}) \), such that for each \( t \) (Aitchison, 1982):

- \( X_{tj} \geq 0 \): non-negativity;
- \( X_{t1} + \cdots + X_{tG} = 1 \): unit sum constraint.

The composition is completely defined by any of \( g = G - 1 \) components and we choose C+O crashes as our baseline. Let \( x_t = (X_{t1}, \cdots, X_{tg}) \)
MV Compositional Model for Different Severity Types

(Box & Cox, 1964):

\[ Y_{tj} = \begin{cases} 
  \left( \frac{X_{tj}}{X_{tG}} \right)^{\lambda} - 1 & \text{if } \lambda \neq 0 \\
  \frac{X_{tj}}{X_{tG}} & \text{if } \lambda = 0 
\end{cases} \]

\[ \lambda \in \mathcal{R} \] is Box-Cox parameter,
\[ t = 1, \cdots, T, \quad j = 1, \cdots, g; \quad g = G - 1. \]

Let \( y_t = (Y_{t1}, \cdots, Y_{tg}) \) for \( t = 1, \cdots, T. \)
Multivariate VAR(p) model:

\[ y_t = \sum_{k=1}^{p} \Phi_k y_{t-k} + \eta u_t + \gamma t + S\delta + w_t \]

- \( \Phi_k \) is a 3 × 3 AR matrix;
- \( u_t \) is the exogenous process \( \log(VMT) \);
- \( \eta \) is a 3 × 1 parameter vector;
- \( \gamma \) is a coefficient vector corresponding to the time trend;
- \( S \) is a 3 × 12 parameter matrix for the seasonal part;
- \( \delta \) is an indicator vector corresponding to the seasonal part;
- \( w_t \) is \( g \)-variate iid \( N(0, \Sigma) \) errors.
MV Compositional Model for Different Severity Types

Data were divided into three parts:

- Fitting portion \((t = 1, \cdots, 168)\);
- 6 months Hold-out data \((t = 169, \cdots, 174)\);
- 12 months Hold-out data \((t = 169, \cdots, 180)\).

Model VAR(1) had the smallest AIC value among all fitted VAR(p) models for \(p = 1, \cdots, 10\). We iterated \(\lambda\) from \(-2\) to 2 with step 0.1 and recorded MAE (Mean Absolute Error) on transformed proportions \((X_{tj})\).

\[
MAE = \frac{1}{n} \sum_t | X_{tj} - \hat{X}_{tj} |
\]

\(\lambda\) was fixed at 0.7 for VAR(1) model. Note: R function VAR in package(vars) was used for model fitting. 10\(^{-5}\) was added to rates for log to be well defined.
MV Compositional Modelling

Stacked Plot of Fitted Proportions of Compositional Model

Volodymyr Serhiyenko
Department of Statistics, University of Connecticut, Storrs
QPRC, June 5-7, 2013
MV Compositional Model for Different Severity Types

- Proportion of K crashes stays approximately at the same level throughout the time period;
- Proportion of A level pedestrian crashes is decreasing, with increase in the proportion of B level crashes (over time we observe less severe injuries);
- C+O rates stays approximately at the same level throughout the time period;

Note: We want to know how well we are predicting pedestrian crash counts, which we computed as:

\[(\text{Comp. Rates}) \times (\text{Total Counts}) = (\text{Crash Counts by severity type})^{*}\]

*Later we will show MAE results*
Static Model for Crash Counts

- Time Series Regression Model for Crash Counts:

\[ y_t | \lambda_t \sim \text{Poisson}(\lambda_t) \]
\[ \lambda_t = \alpha \times V_t \times e^{\gamma_1 t + \gamma_2 t^2 + s' \delta} \]

or:
\[ \log(\lambda_t) = \log(\alpha) + \log(V_t) + \gamma_1 t + \gamma_2 t^2 + s' \delta \]

where \( \alpha \) is a static intercept coefficient; \( \log(V_t) \) is log of Vehicle Miles Traveled (VMT); \( \gamma_1 \) and \( \gamma_2 \) represent time trend coefficients; \( s' \) is a \( 1 \times 12 \) parameter vector corresponding to the seasonal part; \( \delta \) is a \( 12 \times 1 \) indicator vector corresponding to the seasonal part.
Dynamic Model for Crash Counts

Dynamic Linear Models, Shumway and Stoffer:

- Obs. eqn: \( y_t = F'_t x_t + v_t \)
- State/System eqn.: \( x_t = G_t x_{t-1} + w_t \)

where \( v_t \sim \text{iid } N(0, V_t) \) \( w_t \sim \text{iid } N(0, W_t) \)

- \( y_t \) time sequence of scalar obs;
- \( x_t \) sequence of state (latent) parameters;
- \( F'_t \) vector of explanatory variables;
- \( G_t \) matrix describing the state evolution.
Dynamic Model for Crash Counts

DGLM: Gamerman, Bmka (1998):

- $y_t$: univariate count time series, $t = 1, \ldots, n$;
- $\lambda_t = E(y_t | \tilde{\theta}_t)$;

Observation equation and system/state equation:

$$p(y_t | \tilde{\theta}_t) \propto \exp \left\{ \frac{y_t \tilde{\theta}_t - b(\tilde{\theta}_t)}{\phi_t} \right\}$$

$$g(\lambda_t) = \eta_t = F'_t \beta_t$$

$$\beta_t = G_t \beta_{t-1} + w_t$$

Example: Poisson or NegBin DGLM, with log link function.

Note: MCMC methods required for Bayesian inference.
Dynamic Model for Crash Counts

Shan Hu et al (2012):

- Observation Equation:

\[ y_t | \lambda_t \sim \text{Poisson}(\lambda_t) \]
\[ \lambda_t = \alpha_t \times V_t \times e^{S_t} \]

or: \[ \log(\lambda_t) = \log(\alpha_t) + \log(V_t) + S_t \]

where \( \beta_{0t} = \log(\alpha_t) \) is the state parameter (dynamic coefficient), \( \log(V_t) \) is log of Vehicle Miles Traveled (VMT), \( S_t \) is seasonal component.

- System/State Equations (AR(1)):

\[ \beta_{0t} = \rho \beta_{0,t-1} + w_{0,t} \]
\[ S_t = -(S_{t-1} + \ldots + S_{t-11}) + w_{1,t} \]

where \( w_{0,t} \sim N(0, W_0) \) and \( w_{1,t} \sim N(0, W_1) \).
Dynamic Model for Crash Counts

Approximate Bayesian Inference

- Integrated Nested Laplace approximations (INLA): Rue et al. (2009);
- This approach enables very fast approximate computation of posterior distributions and facilitates posterior inference;
- In the case of Gaussian observations, the INLA method is exact up to integration error;
- Extensive comparison between INLA and MCMC approaches have been performed in Rue(2009), Eidsvik(2010) and others showing the advantages of using the approximate method;
- Simulated examples and case studies for large class of statistical models are available in R-INLA, see http://www.r-inla.org/.
### Comparison of the Models for Counts for Pedestrian Crashes by Severity Levels

(* denotes minimum value)

<table>
<thead>
<tr>
<th>Model</th>
<th>Within sample (fitting portion)</th>
<th>6 months ahead prediction</th>
<th>12 months ahead prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamic Setup</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K severity crash level</td>
<td>1.0048*</td>
<td>0.7342*</td>
<td>0.970</td>
</tr>
<tr>
<td>A severity crash level</td>
<td>2.8424*</td>
<td>2.4912</td>
<td>2.2976</td>
</tr>
<tr>
<td>B severity crash level</td>
<td>3.6964*</td>
<td>2.9794*</td>
<td>2.9172</td>
</tr>
<tr>
<td>CO severity crash level</td>
<td>2.9075*</td>
<td>2.9281</td>
<td>2.2499*</td>
</tr>
<tr>
<td><strong>Static Setup</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K severity crash level</td>
<td>1.0094</td>
<td>0.7563</td>
<td>0.9703*</td>
</tr>
<tr>
<td>A severity crash level</td>
<td>2.8711</td>
<td>2.4064*</td>
<td>2.2467*</td>
</tr>
<tr>
<td>B severity crash level</td>
<td>3.8180</td>
<td>3.0461</td>
<td>2.6980*</td>
</tr>
<tr>
<td>CO severity crash level</td>
<td>3.0773</td>
<td>2.8506*</td>
<td>2.3918</td>
</tr>
</tbody>
</table>
Static and Dynamic Models for Crash Counts

Pedestrian Crash Counts with the Model Fits

Observed counts
Fitted counts from the Dynamic Model
Fitted counts from the Static Model

Pedestrian Accidents


Volodymyr Serhiyenko
Department of Statistics, University of Connecticut, Storrs
QPRC, June 5-7, 2013
Posterior Mean for Intercept Alpha

- Total number of crashes
- K level crashes
- A level crashes
- B level crashes
- CO level crashes

Volodymyr Serhiyenko
Department of Statistics, University of Connecticut, Storrs
QPRC, June 5-7, 2013
Summary and Future Work

Results from UV Dynamic Modeling:

- Dynamic model gives better fits than Static Model in terms of Mean Absolute Error and enables us to discover temporal trends (suitable for pedestrian crashes data);
- There is a decreasing trend in all pedestrian crash counts before April 2005, followed by a noticeable increase after that which lasted until May 2007, and then a flattening out until the end of the fitting period;
- The behavior of all pedestrian crashes appears to be largely due to changes in the severity level A;
- There is a slight decrease of C+O and increase of B injury severity levels of pedestrian crashes;
- The level K crash counts stay approximately at the same level relatively to the other levels.
Summary and Future Work

Future work:

- Investigate the reasons for sudden increase in crash type A as well as build Suitable Dynamic Models for Rural Areas and compare results;
- Investigate Multivariate Dynamic Compositional Models for Proportions of crashes (possible use of MARSS R function);
- Apply finite mixtures of MV Poisson for different types of pedestrian crashes, which allows for overdispersion in marginal distributions and negative associations (Karlis & Meligkotsidou (2007) and Shan Hu (2012, PhD thesis)):

\[ p(\mathbf{Y} | \Phi) = \sum_{h=1}^{H} \pi_h MP_m(\mathbf{Y} | \lambda_h). \]
Thank you!

Questions?
In 2009, 4,092 pedestrians were killed and an estimated 59,000 were injured in traffic crashes in the United States; On average, a pedestrian was killed every two hours and injured every nine minutes in traffic crashes; In 2009, pedestrian deaths accounted for 12 percent of all traffic fatalities, and made up 3 percent of all the people injured in traffic crashes; The 4,092 pedestrian fatalities in 2009 were a decrease of 7 percent from 2008 and a decrease of 14 percent from 2000;

Given response $y$, latent Gaussian variables $x$, and hyperparameters $\theta$, posterior marginals of interest:

$$
\pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y)d\theta 
$$

$$
\pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j}
$$

The INLA approach constructs nested approximations:

$$
\tilde{\pi}(x_i|y) = \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y)d\theta 
$$

$$
\tilde{\pi}(\theta_j|y) = \int \tilde{\pi}(\theta|y)d\theta_{-j}
$$
The INLA approach consists of three steps:

- Approximate \( \pi(\theta|y) \) by a Laplace approximation.
- Approximate \( \pi(x_i|\theta,y) \) using either a Gaussian approximation, simplified Laplace approximation, or Laplace approximation.
- Use numeric integration

\[
\tilde{\pi}(x_i|y) = \sum_j \tilde{\pi}(x_i|\theta_j,y)\tilde{\pi}(\theta_j|y)\Delta_j
\]

Note: In Step 2, the Laplace approximation is preferred, but can be computationally expensive. We may use the Gaussian or simplified Laplace approximations if they give small Kullback - Leibler divergence between \( \pi(x_i|y) \) and \( \tilde{\pi}(x_i|y) \).