A hierarchical model for extremes of ozone, NO and NO$_2$

Emma Eastoe

Department of Mathematics and Statistics
Lancaster University, UK

September 22, 2008
Outline

1. Aims
   - Data
   - Questions of interest

2. Background - univariate extremes
   - Stationary
   - With covariates

3. Hierarchical model
   - The model
   - Simulation
   - Inference

4. Results
   - Margins
   - Joint distribution of NO and NO₂

5. Comments
The data

- Daily maxima of hourly ozone, NO and NO$_2$ concentrations at a site in central Reading, UK
- NO and NO$_2$ are primary pollutants e.g. released during combustion of fossil fuels
- Ozone is a secondary pollutant, i.e. formed in the atmosphere
- NO and NO$_2$ are precursors for ozone
- Temperature, quantity of sunlight and wind speed are also important factors in determining concentrations
Notation

- \( \mathbf{Y} \): \( d \)-dimensional random variable, generally observe a sequence \( \mathbf{Y}_i \) of length \( n \)
- \( \mathbf{X} \): associated \( n \times p \) matrix of covariates
General questions for MV extremes

- The (univariate) marginal probability that one of the components is large
- The probability that all components of the vector are simultaneously large
- Conditional on a subset $A$ of the components being large, the probability distribution of the $d - |A|$ remaining components
- The probability that some function of (a subset of) the components is large.

Generally, ‘is large’ means exceeds some high threshold level.
Stationary sequence of random variables \( \{Y_t\} \)

Asymptotically motivated model

For high enough level \( u \) model the size and rate of (local maxima of) threshold exceedances

Size is modelled by a generalised Pareto distribution (GPD), for \( y > 0 \),

\[
\Pr[Y_t > u + y \mid Y_t > u] = \left[ 1 + \frac{\xi y}{\psi_u} \right]^{-1/\xi}_+, \\
\]

where \( \psi_u > 0 \) is threshold dependent.

Model the rate \( \phi_u = \Pr[Y_t > u] \) as a Bernoulli random variable.
Model (functions of the) parameters as functions of covariates, $x_t$

Use linear functions only

$$\log \psi_u(x_t) = \psi' x_t, \quad \xi(x_t) = \xi' x_t \quad \text{and} \quad \logit \phi_u(x_t) = \phi' x_t$$

where $\psi$, $\xi$ and $\phi$ are vectors of regression coefficients

Inference straightforward through maximum likelihood, which can be split into two parts - one for $(\psi, \xi)$ and one for $\phi$
First model aspects of the underlying process and use these to preprocess the data, producing an approximately stationary sequence.

In the absence of a better model, we suggest the following:

\[ Y_t^{\lambda(x_t)} - 1 \]
\[ \frac{\lambda(x_t)}{\lambda(x_t)} = \mu(x_t) + \sigma(x_t)Z_t \]

where \( \sigma(x)_t > 0 \) and \( Z_t \) is assumed to be stationary.

Model the extremes of \( Z_t \) using a threshold \( u_z \) and the \( \text{GPD}(\psi_u(x_t), \xi(x_t)) \) and rate \( \phi_u(x_t) \) model.

Functions of location, scale and Box-Cox parameters modelled as linear functions of covariates.
Suppose we can order the $d$-dimensional random variable in some meaningful way $Y_t = (Y_{1t}, \ldots, Y_{dt})$

Then model $Y_{it}$ conditionally on
- Covariates $X_t$, and
- Set $S_i = \{ Y_{jt} : j < i \}$, which is empty for $i = 1$

In our example, we model NO conditional on covariates, NO$_2$ conditional on NO and covariates and ozone conditional on NO, NO$_2$ and covariates
At each level we model the sequence \{Y_{it}\} using the pre-processing approach, allowing a model on both tails of each component:

1. Estimate the parameters \(\mu_i(x_t, S_{it})\), \(\sigma_i(x_t, S_{it})\) and \(\lambda_i(x_t, S_{it})\).
2. Transform \(Y_{it}\) to \(Z_{it}\).
3. Obtain upper and lower thresholds of \(Z_{it}\), denoted \(u^l_i\) (lower) and \(u^u_i\) (upper).
4. Estimate the GPD and rate parameters using the data above (below) the thresholds \(u^u_i\) (\(u^l_i\)): \(\phi^l_i(x_t, S_{it})\), \(\phi^u_i(x_t, S_{it})\), etc.
To estimate extreme probabilities, we use simulation

Denote simulated data by \((Y^*, X^*)\)

To simulate \(N\)-years of data requires \(L = 365N\) observations

To simulate covariates for day \(t \in \{1, \ldots, 365\}\) in any given year, since we have no model, resample across the observed years from values observed on day \(t\) only.
Simulate $\mathbf{Y}_t$ in the order $\mathbf{Y}_{1t}, \ldots, \mathbf{Y}_{dt}$. For each $i$, conditional on $(\mathbf{Y}_{1t}, \ldots, \mathbf{Y}_{(i-1)t}, \mathbf{X}^*_t)$,

1. Simulate a $U_t \sim \text{Uniform}(0, 1)$
2. If $U_t > \phi^u_i(x^*_t, S^*_it)$, then simulate $Z^*_it$ from the upper tail model
3. If $U_t < \phi^l_i(x^*_t, S^*_it)$, then simulate $Z^*_it$ from the lower tail model
4. If $\phi^l_i(x^*_t, S^*_it) \leq U_t \leq \phi^u_i(x^*_t, S^*_it)$, then resample $Z^*_it$ from \{ $Z_{it} : u^l_i \leq Z_{it} \leq u^u_i$ \}
5. Back transform to the original scale
- Adopt a Bayesian framework: posterior distributions, better able to quantify uncertainty
- Requires MCMC: Gibbs updates for location and rate parameters, Metropolis-Hastings random walk for the rest
Also include first order interactions and indicators for

- Weekend
- Year
- Season
Recalling that $Y_{1t}$ is NO, $Y_{2t}$ is NO$_2$ and $Y_{3t}$ is ozone, and using 90% thresholds:
QQ plots with 95% (99%) credibility regions. Simulated data sets of the same length as the observed data for each of the samples from the posterior. These were ordered and the median (quantiles) found.
Estimated 5-, 10- and 100-year return levels, with 95% credibility regions.

<table>
<thead>
<tr>
<th></th>
<th>5-year</th>
<th>10-year</th>
<th>100-year</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>706</td>
<td>819</td>
<td>1346</td>
</tr>
<tr>
<td></td>
<td>(596,910)</td>
<td>(670,1143)</td>
<td>(914,3381)</td>
</tr>
<tr>
<td>NO₂</td>
<td>156</td>
<td>167</td>
<td>210</td>
</tr>
<tr>
<td></td>
<td>(146,175)</td>
<td>(152,194)</td>
<td>(174,364)</td>
</tr>
<tr>
<td>O₃</td>
<td>218</td>
<td>234</td>
<td>289</td>
</tr>
<tr>
<td></td>
<td>(197,247)</td>
<td>(208,271)</td>
<td>(239,387)</td>
</tr>
</tbody>
</table>

- 5-year levels exceeded only 2 or 3 times for each series
- Only 10-year level for NO₂ exceeded (once)
- None of 100-year levels exceeded
Conditional joint distribution

Conditional on ozone achieving it’s 10-year marginal return level. This condition deceases the probability that both NO and NO$_2$ exceed their marginal threshold level, but increases the probability that at least one does.
Presented an approach for the prediction of the probabilities of) joint and marginal extreme events for a multivariate, non-stationary data set

Proposed method is a hierarchical extension to the pre-processing method previously described in the univariate case

Drawback - relies on the assumption that we can extrapolate the response-covariate regression technique

Method also relies on finding a sensible ‘hierarchy’

Future work - how well does the model capture levels of asymptotic (in)dependence?