Reasoning about Uncertainty in the Stan Modeling Language

John Haman
Thomas Johnson

10 April 2019
What is Stan?

- Stan is a Bayesian model specification (and compilation) language
  - Similar to BUGS and other model languages, but more flexible and easier to use
- Started at Columbia University in 2012
- Open source
- Under active development
Why do we use Stan at IDA?

- Model building in Stan is transparent and structured
  - Model assumptions are specified at the beginning of the model process
- Inferential tasks are generic
  - Uncertainty quantification is done by calculating summary statistics from posterior samples
- Stan is fast, memory efficient
- Available as an R package (rstan)
What Stan buys us

- Access to Bayesian statistical models
  - High level modeling language
  - Easy to create statistical models
- Uncertainty intervals for all model parameters
- Flexibility to fit a large array of models
Ballistic Data (notional)
The Notional Experiment

Old

Far

New

Near
Circular Error Probable (CEP)

Want to define a circle in which we predict the next shot will land with probability 50%.

**Challenge:** Want CEP radius to change with experimental factors
Circular Error Probable (CEP)

$CEP_{50}$ should be something like this.

- Green: Near / Old
- Blue: Far / Old
- Red: Near / New
- Black: Far / New
Constant variance model is inadequate

Equal radii = bad!

- [Legend]
  - Green: Near / Old
  - Blue: Far / Old
  - Red: Near / New
  - Black: Far / New
Proposed Model*

\[ Y = \begin{pmatrix} x \\ y \end{pmatrix} \sim \text{Normal} \left( \begin{pmatrix} M \\ 0 \\ 0 \\ M \end{pmatrix} \begin{pmatrix} \rho \\ \zeta \end{pmatrix}, \Sigma \right) \]

\[ \Sigma \text{ is a matrix that supplies different variances depending on factor settings in } M. \]

This is just a complicated linear model,

\[ Y = X\beta + \epsilon \]

*JMP calls this a log-linear variance model
The Traditional Way

- Fit model with `nlme`, or `statmod` in R
- Code custom approximate F-statistic
- Doable, but the math is hard
- Not easy to extend
The Stan way – more transparent

All Stan programs start with three “blocks”

Sampling statement

Sampling statement

9 April 2019
Stan Output: Posterior Samples

Each column is a vector of samples

<table>
<thead>
<tr>
<th>Sample #/Term</th>
<th>$\sigma_{\text{Far}}$</th>
<th>$\sigma_{\text{Near}}$</th>
<th>$\rho_{\text{Weapon}}$</th>
<th>$\rho_{\text{Range}}$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.75</td>
<td>2.31</td>
<td>3.01</td>
<td>5.39</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>9.29</td>
<td>2.49</td>
<td>2.45</td>
<td>4.31</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>9.73</td>
<td>2.45</td>
<td>2.71</td>
<td>4.39</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>7.86</td>
<td>2.48</td>
<td>2.39</td>
<td>5.66</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>8.24</td>
<td>2.26</td>
<td>2.67</td>
<td>4.32</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

All inference done by computing column standard deviations or quantiles.
Posterior Distributions of All Model Parameters

Different standard deviations!!

Distribution of x direction parameters

Distribution of y direction parameters
Generate New Data from the Stan model

x-Direction Miss Distance (ft)

y-Direction Miss Distance (ft)

Only “Far / New” data shown.

Average

Observed data
Size the Circle (CEP)

50% of generated points inside circle
Calculate the Desired CEP for each Factor
Pros:
- Reliable uncertainty quantification
- Uncertainty intervals for all parameters, no bootstrapping required
- Extensible
- Model checking available

Cons:
- More work upfront
- Have to learn something new
IDA Resources

mc-stan.org

- Tutorials
- Videos
- Reference Manual
- Case Studies