Motivation

In a probabilistic program

\[ p(y|x) \propto p(x)p(y|x) \]

'usual' conditioning is deterministic: \( p(x|y = c) \).

Works when observations

\begin{itemize}
  \item are summarized or obfuscated
  \item are collected by multiple parties,
  \item are noisy and obtained online,
  \item reflect partial knowledge about future.
\end{itemize}

Won't work when observations

\begin{itemize}
  \item \( x \) is the output of stochastic code
  \item \( y \) are noisy and obtained online,
  \item \( x \) reflect partial knowledge about future.
\end{itemize}

Intuition

We know that \( z \sim \mathcal{N}(0, 1) \)

We want to infer \( x \) such that \( y \approx x + z \)

\[
\begin{array}{c|c|c|c}
  \text{Naive} & \text{Stochastic} & \text{Transformed} \\
  \hline
  z \sim \mathcal{N}(0, 1) & z \sim \mathcal{N}(0, 1) & z \sim \mathcal{N}(0, 1) \\
  x \sim \text{Gamma}(2, 2) & x \sim \text{Gamma}(2, 2) & x \sim \text{Gamma}(2, 2) \\
  y \mid x, z \sim \mathcal{N}(x + z, 1) & y \mid x, z \sim \mathcal{N}(x + z, 1) & y \mid x, z \sim \mathcal{N}(x, 1) \\
\end{array}
\]

Definition

Probabilistic program computes

\[ p(x, z) = p(x)p(z|x) \]

Our objective is to infer \( p(x|D_z) \) when \( z \sim D_z \).

Conditioning on \( D_z \equiv \text{conditioning on all values} \):

\[
p(x|D_z) \propto p(x|z, D_z) = p(x) \prod_{z \in \Omega_0} (p(z|D_z))^{p_{D_z}(z)dz} = \exp \left( \log p(x) + \int_{z \in \Omega_0} p_D(z) \log p(z|x)dz \right) \propto \exp (\log p(x) - \text{KL}[p_D(z)||p(z|x)])
\]

Population of New York

<table>
<thead>
<tr>
<th>Data</th>
<th>Population</th>
<th>Sample1</th>
<th>Sample2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N=804)</td>
<td>(n=100)</td>
<td>(n=100)</td>
<td></td>
</tr>
<tr>
<td>mean 17,135</td>
<td>19,667</td>
<td>38,505</td>
<td></td>
</tr>
<tr>
<td>sd 139,147</td>
<td>142,218</td>
<td>228,625</td>
<td></td>
</tr>
<tr>
<td>0% 19</td>
<td>164</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>5% 336</td>
<td>308</td>
<td>315</td>
<td></td>
</tr>
<tr>
<td>25% 800</td>
<td>891</td>
<td>863</td>
<td></td>
</tr>
<tr>
<td>50% 1,668</td>
<td>2,081</td>
<td>1,740</td>
<td></td>
</tr>
<tr>
<td>75% 5,050</td>
<td>6,049</td>
<td>5,239</td>
<td></td>
</tr>
<tr>
<td>95% 30,295</td>
<td>25,130</td>
<td>41,718</td>
<td></td>
</tr>
<tr>
<td>100% 2,627,319</td>
<td>1,424,815</td>
<td>1809578</td>
<td></td>
</tr>
</tbody>
</table>

Model

\[
\begin{align*}
  m & \sim \text{Normal}(\text{mean, } \frac{sd}{\sqrt{n}}) \\
  s^2 & \sim \text{InvGamma}(\frac{n}{2}, \frac{n}{2sd^2}) \\
  \sigma & = \sqrt{\log (s^2/n^2 + 1)} \\
  \mu & = \log m - \frac{\sigma^2}{2} \\
  z_{1...n} \mid m, s^2 & \sim \text{LogNormal}(\mu, \sigma)
\end{align*}
\]