Chunyi Zhao
Two Six Capital
April 15, 2016
Overview

1. Pareto/NBD

2. Maximum Likelihood Estimation

3. Markov Chain Monte Carlo
   - The Bayesian point view
   - Metropolis-Hastings algorithm
   - Gibbs sampler
Assumption
- \( [x | \lambda, \tau, T] \sim \text{poisson}(\lambda t), t = \min(\tau, T) \)
- \( [\tau | \mu] \sim \text{exp}(\mu) \)

Choice of Prior
- \( [\lambda | r, \alpha] \sim \text{gamma}(r, \alpha) \)
- \( [\mu | s, \beta] \sim \text{gamma}(s, \beta) \)

Nice choice of prior leads to close-form likelihood and other convenient consequences.
Pareto/NBD Data structure

- **Data**
  - Dimension: customer $i$ for $i \in \{1 \cdots N\}$
  - Individual level summary statistics: $x_i, tx_i, T_i$

- **Latent variables**
  - Purchase rate: $\lambda_i$
  - Lifetime: $\mu_i$

- **Heterogeneity Parameters**
  - $r, \alpha$
  - $a, \beta$

- **DAG**
## Likelihood $\mathcal{L}(\theta|D)$

Likelihood is probability of observing the given data as a function of $\theta$.

$\mathcal{L}(\theta|D) = \mathbb{P}(D|\theta)$

## MLE In English

Assume $\theta, \theta^*, \mathcal{L}(\theta^*|D) > \mathcal{L}(\theta|D)$, this means given the observed data $\theta^*$ is more ”likely” to be the values for model parameters.

## Estimation

- Maximize likelihood $\equiv$ Maximize log-likelihood $\equiv$ Minimize $-(\text{log-likelihood})$
- The problem of local vs global minimum
- Choice of minimizer
Latent variable level likelihood: $P(x, tx, T|\lambda, \mu)$

- $\tau > T$:
  \[
  \mathcal{L}(\lambda, \mu|\tau > T, x, tx) = \frac{\lambda^x \cdot tx^{x-1} \cdot e^{-\lambda tx}}{\Gamma(x)} \cdot e^{\lambda(T-tx)} \cdot e^{-\mu T}
  \]

- $tx < \tau < T$:
  \[
  \mathcal{L}(\lambda, \mu|\tau < T, x, tx) = \frac{\lambda^x \cdot tx^{x-1} \cdot e^{-\lambda tx}}{\Gamma(x)} \cdot e^{\lambda(\tau-tx)} \cdot \mu e^{-\mu \tau}
  \]

Heterogeneity parameter level likelihood, i.e. the probability of individual’s transaction given $r, \alpha, s, \beta$ for a random customer

\[
P[X = x|r, \alpha, s, \beta, T] = P[X = x|r, \alpha, \tau > T]P[\tau > T|s, \beta] 
+ \int_0^T P[X = x|r, \alpha, \tau > t]f(t|s, \beta)dt
\]

"Nice" heterogeneity params distribution $\Rightarrow$ close-form solution.
The Bayesian point view

- Frequentist: one answer, Bayesian: a set of answers with different weights
- Bayes Theorem

\[
P(\theta, \phi | D) = \frac{\text{likelihood} \cdot \text{prior} \cdot \text{hyper}}{\text{marginal}}
\]

- Likelihood encodes our belief on the relationship among the parameters and the data.
- Prior encodes our belief on the parameters. ”Nice” = Conjugate.
- Hyperparameter controls latent variables, usually set flat and known.
- Posterior indicates the how well the reality fits the model.
- Goal: simulate posterior so that we can learn more from the model.
Markov Process

Mathematically, given a sequence of random variables \( \{X_1, X_2, \cdots, X_T\} \) representing a stochastic process, a process has the Markov property if:

\[
P(X_{t+1} \mid X_1, X_2, X_3, \cdots X_t) = P(X_{t+1} \mid X_t)
\]

Get posterior with MCMC

By constructing a MCMC whose stationary state is the desired distribution, i.e. the posterior, we are effectively drawing each iteration from the posterior once MCMC converges.

How to implement MCMC

Metropolis-Hastings algorithm, and its special case Gibbs sampler
Metropolis algorithm

- A random walk that uses an acceptance/rejection rule to converge to the specified target distribution.

- What we need
  - Posterior distribution \( P(\theta^* | D), P(\theta^{t-1} | D) \)
  - Proposal distribution/jumping distribution \( J_t(\theta^* | \theta^{t-1}) \)
  - Proposal distribution needs to be symmetric!

- Algorithm
  - Initialize the chain with \( \theta^0 \).
  - For \( t = 1, 2, \cdots \)
    - Sample \( \theta^* \) from \( J_t(\theta^* | \theta^{t-1}) \).
    - Calculate the ratio \( r = \frac{P(\theta^* | D)}{P(\theta^{t-1} | D)} \)
    - Accept probability \( \alpha = \min(1, r) \)
    - Implementation: generate \( u \sim \text{uniform}(0, 1) \). If \( r < u \), then accept \( \theta^* \), else keep \( \theta^{t-1} \)
The goal of MCMC is to construct a Markov chain such that its unique stationary distribution is the posterior distribution.

Stationary distribution of a Markov chain: the Markov chain converges to a state that $\theta_t$ for $\forall t$ has the same distribution.

Goal: prove $P(\theta^t = \theta^*) = P(\theta^{t-1} = \theta^*)$

Assume $\theta_a, \theta_b$ s.t $P(\theta_b | D) > P(\theta_a | D)$.

$P(\theta^{t-1} = \theta^a, \theta^t = \theta^b) = P(\theta_a | D) \cdot J_t(\theta_b | \theta_a) \cdot r, r = 1$ due to our assumption

$P(\theta^{t-1} = \theta^b, \theta^t = \theta^a) = P(\theta_b | D) \cdot J_t(\theta_a | \theta_b) \cdot r, r = \left( \frac{P(\theta_a | D)}{P(\theta_b | D)} \right)$

By doing some math we will have

$P(\theta^{t-1} = \theta^b, \theta^t = \theta^a) = P(\theta_a | D) \ast J_t(\theta_a | \theta_b) = P(\theta^{t-1} = \theta^a, \theta^t = \theta^b)$, since the proposal density $J_t(\cdot | \cdot)$ is symmetric.

$P(\theta | D)$ is the same despite the choice of $\theta \Rightarrow$ the posterior distribution $P(\theta | D)$ is the stationary distribution of the Markov Chain of $\theta$. 
Metropolis-Hastings algorithm

- Metropolis-Hastings algorithm relaxes the constrain on the proposal distribution, so that \( J(\cdot|\cdot) \) is not required to be symmetric.
- Instead, change \( r = \frac{\mathbb{P}(\theta^*|D)/J_t(\theta^*|\theta^{t-1})}{\mathbb{P}(\theta^{t-1}|D)/J_t(\theta^{t-1}|\theta^*)} \). The rest of the algorithm remains the same.
- See appendix for why MH algorithm works.
- Gibbs sampler is a special case of MH, where \( r = 1 \).

\[
J(\theta^*|\theta^{t-1}) = \mathbb{P}(\theta^*_j|\theta^*_{-j}, D), \theta^*_{-j} = \theta_{-j}^{t-1}
\]

\[
r = \frac{\mathbb{P}(\theta^*|D)/J_t(\theta^*|\theta^{t-1})}{\mathbb{P}(\theta^{t-1}|D)/J_t(\theta^{t-1}|\theta^*)} = \frac{\mathbb{P}(\theta^*_j, \theta^*_{-j}|D)/\mathbb{P}(\theta^*_j|\theta^*_{-j}, D)}{\mathbb{P}(\theta^{t-1}_j, \theta^{t-1}_{-j}|D)/\mathbb{P}(\theta^{t-1}_j|\theta^{t-1}_{-j}, D)}
\]

\[
= \frac{\mathbb{P}(\theta^*_j|\theta^*_{-j}, D)\mathbb{P}(\theta^{t-1}_{-j}|D)/\mathbb{P}(\theta^*_j|\theta^*_{-j}, D)}{\mathbb{P}(\theta^{t-1}_j|\theta^*_{-j}, D)\mathbb{P}(\theta^{t-1}_{-j}|D)/\mathbb{P}(\theta^{t-1}_j|\theta^*_{-j}, D)} = 1
\]
Gibbs Sampler

- Update certain parameter given rest of parameters according to conditional posterior distribution.
- When to use Gibbs sampler
  - When conditional posterior is well defined and easy to sample from.
  - Conjugate Prior
- What we need
  - \( \theta = \{ \theta_1, \cdots, \theta_J \} \)
  - \( [\theta_j | \theta_{-j}^{t-1}, D] \Rightarrow \) conditional posterior distribution of one parameter given the rest of parameters at \( t-1 \).
  - Notice, at \( t \), the above distribution is calculated with parameters that are updated already at iteration \( t \) and parameters that are not yet updated at \( t-1 \).
- Algorithm
  - initialize \( \theta^{(0)} \)
  - For \( t = 1, 2, \cdots \), update \( \theta_1^{(t)}, \theta_2^{(t)}, \cdots \theta_J^{(t)} \) in turn.
The result of MCMC algorithm is a sample of posterior distribution.

Simple example, let’s look at the gamma data generating process.
\[
x_i \sim \text{gamma}(\alpha, \beta).
\]

- Histogram: bin the samples according to the value \( p \).
- Density: \[
\int_0^\infty f(x) = 1 \Rightarrow \sum_{i=0}^N [b(i) - b(i-1)]d(mid(i)), \text{ s.t } d(mid(i)) \]
  has the same ratio given by frequency
- Quantile: solve cdf \( F(x) = \alpha \) where \( \alpha \) is the quantile.

Credible interval: the subset of posterior parameter space \( \mathcal{C} \) s.t \[
\int_{\mathcal{C}} f(\theta|D)d\theta = 1 - \alpha.
\]

- If \( \theta_L^* \) is the \( \alpha/2 \) posterior quantile for \( \theta \), and \( \theta_U^* \) is the \( 1\alpha \) posterior quantile for \( \theta \), then \( (\theta_L^*, \theta_U^*) \) is a \( 100(1\alpha)\% \) credible interval for \( \theta \).

- Which means, given the observed data there is a \( 1 - \alpha \%) \) probability the true value of \( \theta \) falls in to the above interval.

HPD interval: what if the density is highly skewed?

- additional condition on \( \mathcal{C} = \{\theta : f(\theta|D) \geq k\} \), \( k \) is the horizontal line at \( 1 - \alpha \).
Convergence Diagnosis

- "Burn-in": discarding early iterations of the simulation runs ⇒ eliminating unrepresentative samples
- "Thinning": dependence of the iterations in each sequence (within-sequence correlation) ⇒ achieving the effect of random draws.

Visualization:
- Trace plot
- Running mean plot
- Auto-correlation plot: Auto-correlation \( pk \) is the correlation between a certain draw and its \( k^{th} \) lag.

Gelman and Rubin diagnostic
- Running multiple chain from over-dispersed starting points. Calculating within and between sequence variance (\( B, W \)) to approximate marginal posterior variance (\( \text{var}(\theta|D) \)).
- Scale reduction factor \( R = \sqrt{\frac{\text{var}(\theta|D)}{W}} \). If \( R > 1 \), then further simulations are needed.
A combined approach: Gibbs sampler for latent variables, and MH algorithm for $r, \alpha, s, \beta$

Original flavor: assumptions remain the same

- $[\lambda_i | \tau_i, x_i, T_i, r, \alpha] \sim \text{gamma}(r + x, \alpha + \min(\tau, T_i))$
- $[\mu_i | \tau_i, s, \beta] \sim \text{gamma}(s, \beta + T_i), \tau > T; \text{gamma}(s + 1, \beta + \tau_i), \text{o.w.}$
- $[\tau_i | tx_i, T_i, \lambda_i, \mu_i]$ two cases based on $\mathbb{P}(\tau_i > T_i | \lambda_i, \mu_i, D_i)$
  - Flip a coin with weight $p = \frac{1}{1+\mu_i/(\lambda_i+\mu_i)[\exp((\lambda_i+\mu_i)(T_i-tx_i))-1]}$ to decide whether still alive at $T_i$
  - If alive $\tau_i = T_i + r \exp(\mu_i)$ exponential is memoryless.
  - If dead $\tau_i \sim$ Double truncated exponential distribution with mean $1/(\lambda_i + \mu_i)$ in $[tx_i, T_i]$.
  - Derive the CDF of the above distribution.
  - Inverse sampling: generate $u \sim \text{uniform}(0, 1)$ solve $F(x) = u$, $x$ is the next step.
Conjugate priors for gamma parameters \((r, \alpha), (s, \beta)\) are not easy to sample directly from. \(\Rightarrow\) MH step

Let \(\phi = \{p, q, s, r\}\) be the set of hyperparameters and \(\theta = \{\lambda, \mu\}\) be the latent variables. The posterior

\[
P(\phi|D) \propto P(D|\phi) \cdot P(\phi) \propto P(D|\theta) \cdot P(\theta|\phi)
\]

- For a gamma process: data \(x_1, \cdots, x_n \overset{iid}{\sim} \text{gamma}(\alpha, \beta)\)

\[
P(X|\alpha, \beta) \propto \frac{P^{\alpha-1} \exp(-\beta S)}{(\Gamma(\alpha)^r \beta^{-\alpha})^n}
\]

where \(P = \prod_{i=1}^n x_i, S = \sum_{i=1}^n x_i\)
- \([\alpha|p, q] \sim \text{gamma}(p, q), [\beta|r, s] \sim \text{gamma}(r, s)\)
- \(P(\phi|D) \propto \frac{P^{\alpha-1} \cdot \exp(-\beta S)}{(\Gamma(\alpha)^r \beta^{-\alpha})^n} \cdot \alpha^{p-1} \cdot \exp(-q\alpha) \cdot \beta^{r-1} \cdot \exp(-S\beta)\)

New flavor: purchase rate and lifetime are correlated

Individuals’ purchase rates and dropout rates follow a multivariate lognormal distribution.

\[
\begin{bmatrix}
\log(\lambda) \\
\log(\mu)
\end{bmatrix} \sim \text{MVN}(\beta_0 = \begin{bmatrix} \beta_\lambda \\ \beta_\mu \end{bmatrix}, \Gamma_0 = \begin{bmatrix} \sigma^2_\lambda & \sigma_{\lambda\mu} \\ \sigma_{\mu\lambda} & \sigma^2_\mu \end{bmatrix})
\] (1)

Algorithm

Initialize the algorithm with \(\theta_i^0\) at the individual level. \(\theta_i = \begin{bmatrix} \lambda_i \\ \mu_i \end{bmatrix}\)

For each individual,

- sample \(z_i\), i.e. whether alive at \(T_i\), according to \(P(\tau_i > T_i | \lambda_i, \mu_i, D_i)\)
- If dead, sample \(\tau_i\) using a truncated exponential
- MH step ⇒ Update \(\theta_i^{t-1}\) using posterior \(P(\lambda_i, \mu_i, z_i, \tau_i | x_i, t_x, T_i)\) (likelihood · prior)
- MH step ⇒ Update \([\beta^t_0, \Gamma^t_0]\) according to standard multivariate normal regression update.
- * Draw \(\theta_i^t\) from \(\text{MVN}(\beta d_i, \Gamma_0)\) where \(d_i\) is individual level covariates
  \(\theta_i = \beta d_i + e_i\), where \(e_i \sim \text{MVN}(0, \Gamma_0)\)
References and Further Reading

- Makoto Abe: "Counting Your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model
- Lawrence Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
- Daniel Fink: A Compendium of Conjugate Priors
- Gelman et al: Bayesian Data Analysis
- Google: conjugate prior; sum of iid exponential is gamma; memoryless property of exponential...
The End

Dear Math,
Please grow up and solve your own problems,
I'm tired of solving them for you.

Cool Funny Quotes.com
Appendix A: Conjugate Priors

- Idea of conjugacy: the posterior and the prior distribution lives in the same family. In this case, the prior is called the conjugate prior for the likelihood.

- The exponential family basically covers the distribution we usually use. All distributions in exponential family when used as likelihood have conjugate priors.

- Here are a couple conjugate relationships we used in our models:
  - Bernoulli likelihood has beta as conjugate prior, and beta posterior.
  - Poisson likelihood has gamma as conjugate prior, and gamma posterior.