Network anomaly detection

Patrick Rubin-Delanchy
patrick.rubin-delanchy@bristol.ac.uk

December 10, 2018

Outline

- We're going to cover the main ideas of Heard, Nicholas A., et al. "Bayesian anomaly detection methods for social networks." *The Annals of Applied Statistics 4.2* (2010): 645-662.
- This is not required reading and in fact the paper relies on several concepts that could be unfamiliar and are unnecessary for this module.
- Why this paper? It has a formal mathematical grounding, with proven effectiveness in practice. For example, it's been taken up by many organisations around the world, and motivated a lot of further cyber-security research.

Let $X_1, X_2, \ldots$ denote independent replicates of a continuous random variable X with density $p(x \mid \theta)$, where θ denotes unknown parameters.

Assume a prior density $p(\theta)$ on θ. We will denote the associated distribution by G.

Given observations $x_1, \ldots, x_n$:

1. the posterior distribution of θ given $x_1, \ldots, x_n$ is:

$$p(\theta \mid x_1, \ldots, x_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta).$$

Given observations $x_1, \ldots, x_n$:

1. the posterior distribution of θ given $x_1, \ldots, x_n$ is:

$$p(\theta \mid x_1, \ldots, x_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta).$$

2. the marginal likelihood of $x_1, \ldots, x_n$ is:

$$p(x_1, \ldots, x_n) = \int p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta) \mathrm{d}\theta.$$

Given observations $x_1, \ldots, x_n$:

1. the posterior distribution of θ given $x_1, \ldots, x_n$ is:

$$p(\theta \mid x_1, \ldots, x_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta).$$

2. the marginal likelihood of $x_1, \ldots, x_n$ is:

$$p(x_1, \ldots, x_n) = \int p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta) \mathrm{d}\theta.$$

3. the predictive density of X_{n+1} is:

$$\begin{aligned} p(x_{n+1} \mid x_1, \ldots, x_n) &= \int p(x_{n+1} \mid \theta) p(\theta \mid x_1, \ldots, x_n) \mathrm{d}\theta \\ &= \frac{p(x_{n+1}, \ldots, x_n)}{p(x_1, \ldots, x_n)}. \end{aligned}$$

Example

Let $X_1, X_2, \ldots$ denote a sequence of independent normal(μ, σ^2) variables, where σ^2 is known.

Assume *a priori* that $\mu \sim$ normal(μ_0, σ_0^2).

Example continued

The posterior distribution for μ is

$$p(\mu \mid x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$
$$= \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2+\frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right\}\right].$$

Example continued

The posterior distribution for μ is

$$p(\mu \mid x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$
$$= \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2+\frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right\}\right].$$

A generally useful trick to remember is that for $x_i \in \mathbb{R}^d, c \in \mathbb{R}^d$,

$$\sum_{i=1}^{n} w_i\|x_i-c\|^2 = \sum_{i=1}^{n} w_i\|x_i-\tilde{x}\|^2 + \left(\sum_{i=1}^{n} w_i\right)\|\tilde{x}-c\|^2,$$

where

$$\tilde{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Example continued

Therefore,

$$p(\mu \mid x_1, \ldots, x_n) \propto \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\sigma_0^2}(\mu - \mu_0)^2\right\}\right]$$

Example continued

Therefore,

$$p(\mu \mid x_1, \ldots, x_n) \propto \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right\}\right]$$

$$\propto \exp\left[-\frac{1}{2}\left\{\frac{n}{\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma_0^2}(\mu_0-\mu)^2\right\}\right]$$

Example continued

Therefore,

$$
\begin{aligned}
p(\mu \mid x_1, \ldots, x_n) &\propto \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\frac{n}{\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma_0^2}(\mu_0-\mu)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\frac{\frac{n}{\sigma^2}\bar{x}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}-\mu\right)^2\right\}\right],
\end{aligned}
$$

Example continued

Therefore,

$$
\begin{aligned}
p(\mu \mid x_1, \ldots, x_n) &\propto \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\frac{n}{\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma_0^2}(\mu_0-\mu)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\frac{\frac{n}{\sigma^2}\bar{x}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}-\mu\right)^2\right\}\right],
\end{aligned}
$$

so that $\mu \mid x_1, \ldots, x_n \sim \text{normal}(\mu_n, \sigma_n^2)$, where

$$
\mu_n = \frac{\frac{n}{\sigma^2}\bar{x}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}; \quad \sigma_n^2 = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}.
$$

Example continued

We could calculate the predictive density of X_{n+1} analytically, but a simple trick is available here.

Conditional on $X_1, \ldots, X_n$, the random variable $X_{n+1} \sim \text{normal}(\mu, \sigma^2)$, where $\mu \sim \text{normal}(\mu_n, \sigma_n^2)$.

Example continued

We could calculate the predictive density of X_{n+1} analytically, but a simple trick is available here.

Conditional on $X_1, \ldots, X_n$, the random variable $X_{n+1} \sim \text{normal}(\mu, \sigma^2)$, where $\mu \sim \text{normal}(\mu_n, \sigma_n^2)$. Therefore X_{n+1} is distributed as the sum of a $\text{normal}(0, \sigma^2)$ and a $\text{normal}(\mu_n, \sigma_n^2)$, so that:

$$X_{n+1} \mid X_1, \ldots, X_n \sim \text{normal}(\mu_n, \sigma_n^2 + \sigma^2).$$

Partial predictive p-value

Let $t = g(x_{n+1})$ be a test statistic dependent only on the $(n+1)$th observation. The partial predictive p-value for t is defined as:

$$p = \mathbb{P}(T \geq t \mid X_1 = x_1, \ldots, X_n = x_1)$$

Partial predictive p-value

Let $t = g(x_{n+1})$ be a test statistic dependent only on the $(n+1)$th observation. The partial predictive p-value for t is defined as:

$$\begin{aligned} p &= \mathbb{P}(T \geq t \mid X_1 = x_1, \ldots, X_n = x_1) \\ &= \int_t^\infty p(t \mid x_1, \ldots, x_n)\, \mathrm{d}t \end{aligned}$$

Partial predictive p-value

Let $t = g(x_{n+1})$ be a test statistic dependent only on the $(n+1)$th observation. The partial predictive p-value for t is defined as:

$$\begin{aligned} p &= \mathbb{P}(T \geq t \mid X_1 = x_1, \ldots, X_n = x_1) \\ &= \int_t^\infty p(t \mid x_1, \ldots, x_n)\, \mathrm{d}t \\ &= \int_t^\infty \int p(t \mid \theta) p(\theta \mid x_1, \ldots, x_n)\, \mathrm{d}\theta\, \mathrm{d}t. \end{aligned}$$

Partial predictive p-value

Let $t = g(x_{n+1})$ be a test statistic dependent only on the $(n+1)$th observation. The partial predictive p-value for t is defined as:

$$\begin{aligned} p &= \mathbb{P}(T \geq t \mid X_1 = x_1, \ldots, X_n = x_1) \\ &= \int_t^\infty p(t \mid x_1, \ldots, x_n)\, \mathrm{d}t \\ &= \int_t^\infty \int p(t \mid \theta) p(\theta \mid x_1, \ldots, x_n)\, \mathrm{d}\theta\, \mathrm{d}t. \end{aligned}$$

In words: the partial predictive p-value for $t = g(x_{n+1})$ is the probability of observing such a large t, given what we know about θ after n observations.

Example

In many problems we might simply be interested in large outliers, in which case we might use $t = x_{n+1}$. In our normal sequence example, the p-value is:

$$\begin{aligned} p &= \mathbb{P}(X_{n+1} \geq x_{n+1} \mid X_1 = x_1, \ldots, X_n = x_1) \\ &= \mathbb{P}(\text{normal}(\mu_n, \sigma_n^2 + \sigma^2) \geq x_{n+1}) \\ &= \mathbb{P}\left(\text{normal}(0,1) \geq \frac{x_{n+1} - \mu_n}{\sqrt{\sigma_n^2 + \sigma^2}}\right) \\ &= 1 - \Phi\left(\frac{x_{n+1} - \mu_n}{\sqrt{\sigma_n^2 + \sigma^2}}\right). \end{aligned}$$

Partial predictive p-values are uniform

We will state the following without proof.

Theorem

Assume the test statistic T has a continuous distribution. Under a generative model where $\theta \sim G$ (i.e. the "prior holds"), the partial predictive p-values form a sequence of independent uniform random variables on $[0, 1]$. If the test is discrete, the sequence of p-values is i.i.d. from a distribution that is stochastically larger than uniform.

Therefore, a lot of the methodology we've covered so far transfers to this Bayesian learning framework.

Conjugate priors

Definition

The prior and posterior distributions are said to be conjugate for a given model (or likelihood) if they are from the same distributional family.

In this module, you will be expected to remember that the conjugate prior distributions for:

1. The normal likelihood with known variance is a normal distribution (as we saw)
2. The Bernoulli likelihood is a Beta distribution (as we will see)
3. The multinomial likelihood is a Dirichlet distribution
4. The Poisson likelihood is a Gamma distribution (as we will see)

Network modelling

An important application of anomaly detection is network monitoring, which is of course a critical cyber-security problem, but is just as important in (intensive) healthcare monitoring, fraud detection, fault detection, and more.

We will now consider a (simplified) dynamic network modelling scenario where we observe connections occur over time between different entities on a network.

Network modelling

Specifically, assume that for a pair of entities (i, j) there is a *discrete-time* counting process $N_{ij}(t) \in \{0, 1, \ldots\}$ observed at discrete time points $\{1, \ldots, T\}$.

$N_{ij}(t)$ is the total number of connections between i and j over $\{1, \ldots, t\}$.

We will denote by $\mathrm{d}N_{ij}(t)$ the number of events occurring at time t, for $t \in \{1, \ldots, T\}$, so that:

$$N_{ij}(t) = \sum_{\tau=1}^{t} \mathrm{d}N_{ij}(\tau).$$

The framework advocated by Heard (2010) is to use "simple, conjugate Bayesian models for discrete time counting processes to track the pairwise links of all nodes in the graph to assess normality of behaviour" in order to pick out a greatly reduced set of anomalous nodes for more careful analysis.

To cope with the "bursty" nature of network communications, we will consider the hurdle model (i, j index temporarily dropped):

$$\mathrm{d}N(t) = \mathrm{d}A(t)\{\mathrm{d}B(t) + 1\},$$

where $A(t)$ and $B(t)$ are independent discrete time counting processes with $\mathrm{d}A(t) \in \{0, 1\}$, $\mathrm{d}B(t) \geq 0$.

To cope with the "bursty" nature of network communications, we will consider the hurdle model (i, j index temporarily dropped):

$$\mathrm{d}N(t) = \mathrm{d}A(t)\{\mathrm{d}B(t) + 1\},$$

where $A(t)$ and $B(t)$ are independent discrete time counting processes with $\mathrm{d}A(t) \in \{0, 1\}$, $\mathrm{d}B(t) \geq 0$.

We will now analyse the case where:

1. $\mathrm{d}A(t)$ is a sequence of independent and identically distributed Bernoulli variables with success probability μ.
2. $\mathrm{d}B(t)$ is a sequence of independent and identically distributed Poisson variables with mean λ.

Bayesian inference

Recall that the Beta(α, β) distribution has density:

$$\frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $\alpha, \beta > 0$ and

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Let $N(\cdot)$ (resp. $A(\cdot)$) denote $N(1), \ldots, N(T)$ (resp. $A(1), \ldots, A(T)$).

Assuming a Beta(α, β) prior for μ we have

$$p\{\mu \mid N(\cdot)\} = p\{\mu \mid A(\cdot)\}$$

Let $N(\cdot)$ (resp. $A(\cdot)$) denote $N(1), \ldots, N(T)$ (resp. $A(1), \ldots, A(T)$).

Assuming a Beta(α, β) prior for μ we have

$$\begin{aligned} p\{\mu \mid N(\cdot)\} &= p\{\mu \mid A(\cdot)\} \\ &\propto p\{A(\cdot) \mid \mu\} p(\mu) \end{aligned}$$

Let $N(\cdot)$ (resp. $A(\cdot)$) denote $N(1), \ldots, N(T)$ (resp. $A(1), \ldots, A(T)$).

Assuming a Beta(α, β) prior for μ we have

$$\begin{aligned} p\{\mu \mid N(\cdot)\} &= p\{\mu \mid A(\cdot)\} \\ &\propto p\{A(\cdot) \mid \mu\} p(\mu) \\ &\propto \mu^{A(T)} (1-\mu)^{T-A(T)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \end{aligned}$$

Let $N(\cdot)$ (resp. $A(\cdot)$) denote $N(1), \ldots, N(T)$ (resp. $A(1), \ldots, A(T)$).

Assuming a Beta(α, β) prior for μ we have

$$\begin{aligned} p\{\mu \mid N(\cdot)\} &= p\{\mu \mid A(\cdot)\} \\ &\propto p\{A(\cdot) \mid \mu\} p(\mu) \\ &\propto \mu^{A(T)}(1-\mu)^{T-A(T)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &\propto \mu^{\{A(T)+\alpha\}-1}(1-\mu)^{\{T-A(T)+\beta\}-1}, \end{aligned}$$

Let $N(\cdot)$ (resp. $A(\cdot)$) denote $N(1), \ldots, N(T)$ (resp. $A(1), \ldots, A(T)$).

Assuming a Beta(α, β) prior for μ we have

$$\begin{aligned} p\{\mu \mid N(\cdot)\} &= p\{\mu \mid A(\cdot)\} \\ &\propto p\{A(\cdot) \mid \mu\} p(\mu) \\ &\propto \mu^{A(T)}(1-\mu)^{T-A(T)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &\propto \mu^{\{A(T)+\alpha\}-1}(1-\mu)^{\{T-A(T)+\beta\}-1}, \end{aligned}$$

from which we recognise the functional form of a

$$\text{Beta}\{\alpha + A(T), \beta + T - A(T)\}$$

distribution.

Let $N(\cdot)$ (resp. $A(\cdot)$) denote $N(1), \ldots, N(T)$ (resp. $A(1), \ldots, A(T)$).

Assuming a Beta(α, β) prior for μ we have

$$\begin{aligned} p\{\mu \mid N(\cdot)\} &= p\{\mu \mid A(\cdot)\} \\ &\propto p\{A(\cdot) \mid \mu\} p(\mu) \\ &\propto \mu^{A(T)}(1-\mu)^{T-A(T)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &\propto \mu^{\{A(T)+\alpha\}-1}(1-\mu)^{\{T-A(T)+\beta\}-1}, \end{aligned}$$

from which we recognise the functional form of a

$$\text{Beta}\{\alpha + A(T), \beta + T - A(T)\}$$

distribution.

By a similar argument,

$$\mu \mid N(1), \ldots, N(t) \sim \text{Beta}\{\alpha + A(t), \beta + t - A(t)\}.$$

Predictive distribution

To obtain the predictive density for $\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)$ we again use a trick.

We have: E()=p(=1)1+p(=0)0=p(=1)

$$\mathbb{P}\{\mathrm{d}A(t+1) = 1 \mid N(1), \ldots, N(t)\} = \mathbb{E}\left[\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)\right]$$

Predictive distribution

To obtain the predictive density for $\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)$ we again use a trick.

We have:

$$\begin{aligned}\mathbb{P}\{\mathrm{d}A(t+1) = 1 \mid N(1), \ldots, N(t)\} &= \mathbb{E}\left[\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left\{\mathrm{d}A(t+1) \mid \mu, N(1), \ldots, N(t)\right\} \mid N(1), \ldots, N(t)\right]\end{aligned}$$

Predictive distribution

To obtain the predictive density for $\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)$ we again use a trick.

We have:

$$\begin{aligned}
\mathbb{P}\{\mathrm{d}A(t+1) = 1 \mid N(1), \ldots, N(t)\} &= \mathbb{E}\left[\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{\mathrm{d}A(t+1) \mid \mu, N(1), \ldots, N(t)\right\} \mid N(1), \ldots, N(t)\right] \\
&= \mathbb{E}\{\mu \mid N(1), \ldots, N(t)\}.
\end{aligned}$$

Predictive distribution

To obtain the predictive density for $\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)$ we again use a trick.

We have:

$$\begin{aligned}\mathbb{P}\{\mathrm{d}A(t+1) = 1 \mid N(1), \ldots, N(t)\} &= \mathbb{E}\left[\mathrm{d}A(t+1) \mid N(1), \ldots, N(t)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left\{\mathrm{d}A(t+1) \mid \mu, N(1), \ldots, N(t)\right\} \mid N(1), \ldots, N(t)\right] \\ &= \mathbb{E}\{\mu \mid N(1), \ldots, N(t)\}.\end{aligned}$$

But,

$$\mu \mid N(1), \ldots, N(t) \sim \mathrm{Beta}\{\alpha + A(t), \beta + t - A(t)\},$$

so that

$$\begin{aligned}\mathbb{E}\{\mu \mid N(1), \ldots, N(t)\} &= \frac{\alpha + A(t)}{\alpha + A(t) + \beta + t - A(t)} \\ &= \frac{\alpha + A(t)}{\alpha + \beta + t}.\end{aligned}$$

Partial predictive p-value

It follows that the partial predictive p-value for the test statistic $Y = \min\{\mathrm{d}N(t+1), 1\} = \mathrm{d}A(t+1)$ is 要么1要么0

$$p = \begin{cases} 1 & \text{if } y = 0, \\ \frac{\alpha + A(t)}{\alpha + \beta + t} & \text{if } y \geq 1. \end{cases}$$

(Here we are using Y, y for test statistics instead of the usual T, t because the latter are currently being used to indicate time).

Recall that the Gamma(a, b) distribution has a density:

$$\frac{\beta^a}{\Gamma(a)} x^{a-1} \, \mathrm{e}^{-bx},$$

and the Poisson(λ) distribution has a probability mass function:

$$\frac{\lambda^k \, \mathrm{e}^{-\lambda}}{k!}.$$

Inference for λ

Assuming a Gamma(a, b) prior for λ we have:

$$p\{\lambda \mid N(\cdot)\} = p\{\lambda \mid (\mathrm{d}B(t) : \mathrm{d}A(t) = 1)\}$$

通过这个N()看poisson的，就要通过dB来看，
但dB只能在dA=1的时候看到

Inference for λ

Assuming a Gamma(a, b) prior for λ we have:

$$\begin{aligned} p\{\lambda \mid N(\cdot)\} &= p\{\lambda \mid (\mathrm{d}B(t) : \mathrm{d}A(t) = 1)\} \\ &\propto p\{(\mathrm{d}B(t) : \mathrm{d}A(t) = 1) \mid \lambda\} p(\lambda) \end{aligned}$$

Inference for λ

Assuming a Gamma(a, b) prior for λ we have:

$$\begin{aligned} p\{\lambda \mid N(\cdot)\} &= p\{\lambda \mid (\mathrm{d}B(t) : \mathrm{d}A(t) = 1)\} \\ &\propto p\{(\mathrm{d}B(t) : \mathrm{d}A(t) = 1) \mid \lambda\} p(\lambda) \\ &\propto \lambda^{N(T) - A(T)} \, \mathrm{e}^{-A(T)\lambda} \, p(\lambda) \end{aligned}$$

Inference for λ

Assuming a Gamma(a, b) prior for λ we have:

$$\begin{aligned} p\{\lambda \mid N(\cdot)\} &= p\{\lambda \mid (\mathrm{d}B(t) : \mathrm{d}A(t) = 1)\} \\ &\propto p\{(\mathrm{d}B(t) : \mathrm{d}A(t) = 1) \mid \lambda\} p(\lambda) \\ &\propto \lambda^{N(T) - A(T)} \, \mathrm{e}^{-A(T)\lambda} \, p(\lambda) \\ &\propto \lambda^{a + N(T) - A(T) - 1} \, \mathrm{e}^{-\{b + A(T)\}\lambda}, \end{aligned}$$

Inference for λ

Assuming a Gamma(a, b) prior for λ we have:

$$\begin{aligned} p\{\lambda \mid N(\cdot)\} &= p\{\lambda \mid (\mathrm{d}B(t) : \mathrm{d}A(t) = 1)\} \\ &\propto p\{(\mathrm{d}B(t) : \mathrm{d}A(t) = 1) \mid \lambda\} p(\lambda) \\ &\propto \lambda^{N(T)-A(T)} \, \mathrm{e}^{-A(T)\lambda} \, p(\lambda) \\ &\propto \lambda^{a+N(T)-A(T)-1} \, \mathrm{e}^{-\{b+A(T)\}\lambda}, \end{aligned}$$

from which we recognise the functional form of a

$$\text{Gamma}\{a + N(T) - A(T), b + A(T)\}$$

distribution.

Inference for λ

Assuming a Gamma(a, b) prior for λ we have:

$$\begin{aligned} p\{\lambda \mid N(\cdot)\} &= p\{\lambda \mid (\mathrm{d}B(t) : \mathrm{d}A(t) = 1)\} \\ &\propto p\{(\mathrm{d}B(t) : \mathrm{d}A(t) = 1) \mid \lambda\} p(\lambda) \\ &\propto \lambda^{N(T)-A(T)} \, \mathrm{e}^{-A(T)\lambda} \, p(\lambda) \\ &\propto \lambda^{a+N(T)-A(T)-1} \, \mathrm{e}^{-\{b+A(T)\}\lambda}, \end{aligned}$$

from which we recognise the functional form of a

$$\text{Gamma}\{a + N(T) - A(T), b + A(T)\}$$

distribution.

By a similar argument,

$$\lambda \mid N(1), \ldots, N(t) \sim \text{Gamma}\{a + N(t) - A(t), b + A(t)\}.$$

Predictive distribution

It follows that $\mathrm{d}B(t+1) \mid N(1), \ldots, N(t)$ is distributed as

$$\mathrm{d}B(t+1) \sim \text{Poisson}(\lambda), \quad \text{where}$$
$$\lambda \sim \text{Gamma}\{a + N(t) - A(t), b + A(t)\}.$$

Predictive distribution

It follows that $\mathrm{d}B(t+1) \mid N(1), \ldots, N(t)$ is distributed as

$$\mathrm{d}B(t+1) \sim \text{Poisson}(\lambda), \quad \text{where}$$
$$\lambda \sim \text{Gamma}\{a + N(t) - A(t), b + A(t)\}.$$

Therefore (by a well-known result that is outside the scope of this module to prove),

$$\mathrm{d}B(t+1) \mid N(1), \ldots, N(t) \sim \text{negative\_binomial}(n_t, p_t),$$

where:

$$n_t = a + N(t) - A(t),$$
$$p_t = [1 + 1/\{b + A(t)\}]^{-1}.$$

Partial predictive p-value

The partial predictive p-value for the test statistic $Y = \mathrm{d}N(t+1)$ is:

$$p = \begin{cases} 1 & \text{if } y = 0, \\ \frac{a+A(t)}{a+b+t} & \text{if } y = 1, \\ \frac{a+A(t)}{a+b+t}\{1 - F_{\mathrm{NB}}(y-2)\} & \text{if } y \geq 2, \end{cases}$$

where F_{NB} denotes the cumulative distribution function of a negative binomial distribution with parameters n_t, p_t.