Proyecto Final de Fundamentos de Estadística

A. Ivonne Pineda (141194) y Dante Ruiz (183340)

12/11/2018

Contents

1 Introducción

En las compañías aseguradoras uno de los problemas a los que se enfrentan es calcular el monto total de reclamo de las pólizas en su portafolio. En este proyecto se utiliza un enfoque bayesiano para hacer inferencia sobre el número de siniestros y el monto de reclamos en la aseguradora AllState. Se utiliza información de tres años 2005, 2006 y 2007 para hacer inferencia sobre el 2008.

Se implementa un enfoque de estimación bayesiano similar al propuesto en el trabajo "Bayesian Analysis of aggregate loss models".[1] En dicho trabajo se utilizó la distribución Coxian y se implementaron métodos de cadenas de Markov (MCMC), para modelar el número y montos reclamados de una empresa de seguros. En este trabajo se utiliza el mismo enfoque con base en los modelos Poisson y Normal para el número de siniestros y montos de reclamos respectivamente.

Por un lado, la Poisson es una distribución de probabilidad discreta que expresa a partir de una frecuencia media la probabilidad de que ocurra un determinado número de eventos durante cierto periodo de tiempo; en el caso de este proyecto se utilizará para expresar la ocurrencia de siniestros considerando el número de pólizas anuales. Por otro lado, la distribución normal permite modelar gran parte de fenómenos naturales y desconocidos por la enorme cantidad de variables incontrolables que en ellos intervienen, particularmente en el ámbito de la economía y finanzas; el uso del modelo normal para modelar los montos reclamados se justificará asumiendo que cada observación se obtiene como la suma de unas pocas causas independientes.

El enfoque de este trabajo es identificar si la información de los tres años de la aseguradora AllState es relevante para el cálculo de la prima de riesgo para el año 2008. El documento se encuentra dividido en seis secciones. En primer lugar, se presentan los resultados del análisis exploratorio de los datos que nos permite entender el comportamiento de los dos fenómenos de interés. Posteriormente, se realiza la modelación de manera individual para frecuencia de siniestros y severidades, de modo que con los resultados se pueda generar una distribución compuesta. Finalmente, se presenta un análisis de los resultados de manera general para posteriormente derivar en el cálculo del Value at Risk (VaR) para el portafolio de la compañía.

2. Objetivo

El objetivo de este trabajo es definir un modelo para calcular la prima de riesgo individual en seguro de autos para la empresa AllState en el año 2008.

Particularmente:

1) Modelar el número de reclamos.
2) Modelar el monto de reclamos por póliza siniestrada.
3) Obtener la distribución conjunta del monto agregado de siniestros.
4) Calcular el VaR para el monto agregado.

3. Análisis exploratorio de los datos

Tenemos una base de datos para la empresa aseguradora AllState que contiene un portafolio de 1,048,575 de pólizas (observaciones) organizadas con las siguientes ocho variables:

- `Household_ID`: Identificador numérico de la póliza (una por hogar).
- `Vehicle`: identificador del vehículo asegurado por póliza (dentro de cada hogar).
- `Calendar_Year`: Año calendario en el que el vehículo fue asegurado.
- `Model_Year`:: Año calendario de fabricación/venta del vehículo.

[1] (Ausín, et al, 2010)

- `Blind_Make`: Fabricante del vehículo (discrecional).
- `Claim_Amount`: Monto de reclamo asociado con el vehículo (montos en USD).

Ya que el objetivo de este proyecto es el cálculo de la prima de riesgo individual a continuación se definen los siguientes conceptos:

$$Tasa\ de\ siniestros = \frac{número\ siniestros}{número\ de\ asegurados}$$

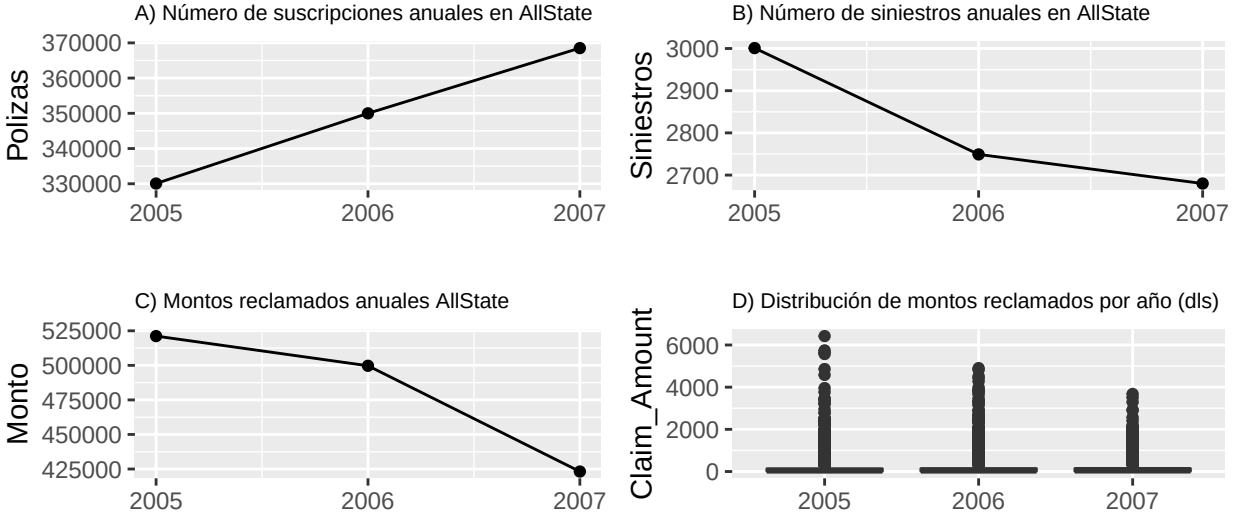
$$Costo\ medio = \frac{costo\ total\ de\ siniestros}{número\ de\ siniestros}$$

Una vez definidos estos conceptos se calculan las siguientes estadísticas del portafolio de AllState:

Table 1: Resumen de los siniestros y montos reclamados por año

Año	Siniestros	Asegurados	Monto_Reclamos	Tasa	Costo_medio
2005	3001	330065	521181.0	0.0090921	173.6691
2006	2749	349991	499706.8	0.0078545	181.7777
2007	2680	368519	423182.4	0.0072724	157.9039

En la tabla y en la gráfica B) se observa que el número de siniestros es decreciente y se encuentra entre 3001 y 2680 eventos. Sin embargo, el número de asegurados, gráfica A) ha crecido pasando de 330,065 a 368,519. Asimismo, el monto de reclamos, gráfica C), a disminuido pasando de 521,181 a 423,182 dólares. Se observa que la tasa de siniestros también ha disminuido pasando de 0.0090 a 0.0072 en el periodo. El costo medio del portafolio ha disminuido de 173.6 a 157.9 dólares. En la gráfica D) se muestra que la distribución de los montos reclamados no ha cambiado en los tres años y se encuentra segada a la derecha. Asimismo, se muestra que existen muchos valores extremos lo cual afecta el cálculo de la prima del portafolio.



Para poder calcular la prima de riesgo individual en 2008 se tienen dos elementos desconocidos, uno es el número de casos siniestrados y dos las severidades individuales.

4. Modelación

En esta sección se desarrolla el modelo predictivo del valor de la prima de riesgo individual. El modelo que se propone es un modelo conjugado que depende de dos variables aleatorias (v.a.), la frecuencia de siniestros (Número de reclamos) y las severidades (Montos de reclamo).

La v.a. de frecuencia de siniestros se va a modelar con una distribución Poisson, porque la variable en cuestión representa conteos y por lo tanto solo toma valores en los números reales enteros y positivos.

Por otro lado, la v.a. de severidades se va a modelar con una distribución Normal con los datos transformados en escala logarítmica.

Modelación de la frecuencia de siniestros (Número de reclamos)

La v.a. de frecuencia de siniestros se modelará con una distribución poisson ya que la variable siniestros corresponde a un conteo.

Supongamos que en el periodo 2005-2007 hay una tasa única de siniestros que sigue la siguiente distribución.

$$y_i \sim Poisson(e_i\lambda)$$

donde y_i es el número de siniestros observados en cada año i, e_i es el número de asegurados, y λ es la tasa de siniestros medida como número siniestros entre número de asegurados por cada año.

Debido a que no contamos con información inicial acerca de la tasa de siniestros asignamos a λ una distribución inicial no informativa de la forma:

$$g(\lambda) \propto \frac{1}{\lambda^{\frac{1}{2}}}$$

Sea $n = (y_1, y_2, y_3)$, usamos el Teorema de Bayes para calcular la densidad posterior de λ.

$$g(\lambda|y) \propto g(\lambda)f(y|\lambda) = g(\lambda)\prod_{i=1}^{n} f(y_i|\lambda)$$

$$= \frac{1}{\lambda}\prod_{i=1}^{n}\left(\frac{exp(-e_i\lambda)(e_i\lambda)^{y_i}}{y_i!}\right)$$

$$\propto \lambda^{(\sum_{i=1}^{n} y_i - 1)} exp(-\sum_{i=1}^{n} e_i\lambda)$$

identificamos esta densidad posterior como una $Gamma(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} e_i)$ donde expresamos la función de densidad de una distribución $Gamma(\alpha, \beta)$ usando el parámetro de forma α y el inverso del parámetro de escala β de manera que la función de densidad es,

$$f(x|\alpha, \beta) = \beta^{\alpha}\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}I_{0,\infty}(x)$$

$$\text{para } x \geq 0 \text{ y } \alpha, \beta > 0.$$

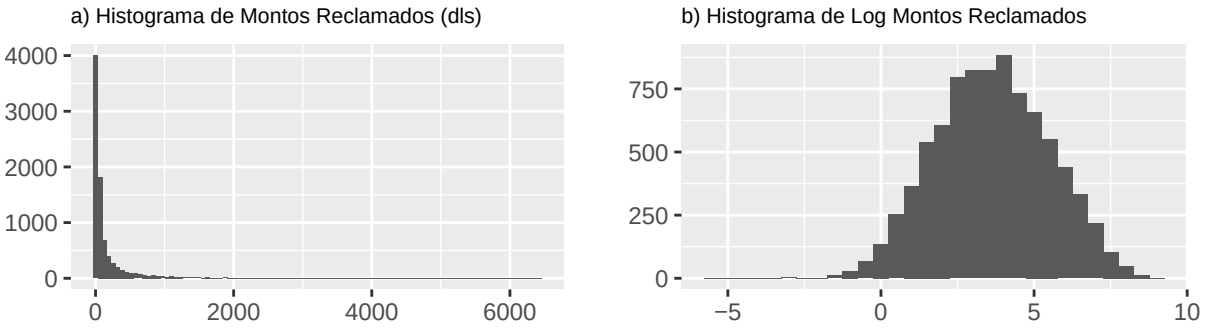
Denotaremos el número de siniestros por año y_i* con póliza e_i (exposiciones) en un año futuro condicional a la λ, y_i* se distribuye $Poisson(e_i\lambda)$, no conocemos el verdadero valor de λ, sin embargo nuestro conocimiento actual está contenido en la densidad posterior $g(\lambda|y)$. Por lo tanto, la distribución predictiva posterior de y_i* está dada por:

$$f(y_i * |e_i, y) = \int f(y_i * |e_i\lambda)g(\lambda|y)d\lambda$$

donde $f(y_i * |e_i, y)$ es una densidad Poisson con media λ. La densidad predictiva posterior representa la verosimilitud de observaciones futuras basadas en el modelo ajustado. Para este caso, la densidad $f(y_i * |e_i, y)$ representa el número de siniestros que se predecirán para un año con exposición e_i.

Modelación de las severidades (Montos de reclamo)

El primer paso para plantear un modelo sobre las severidades es conocer como se encuentran distribuidos durante el periodo.



En el panel a) se muestra el histograma de la variable monto reclamado. Se observa que la distribución de los datos se encuentra sesgada a la derecha y es unimodal. En el panel b) se muestra el histograma de los datos después de aplicar una transformación logarítmica. Se puede ver que la distribución es unimodal y simétrica.

Del análisis gráfico se justifica que se puede modelar la v.a. de frecuencia de severidades ya sea con una distribución Gamma o Normal con datos transformados en escala logarítmica. En este trabajo se decidió que x_i sigue una distribución Normal:

De modo que al aplicar el logaritmo a los montos reclamados se utiliza la siguiente ecuación $x_i = log(Monto\ Reclamado)\ \epsilon\ \mathbb{R}$ y los datos se distribuyen del siguiente modo:

$$x_i \sim N(x|\mu, \sigma^2)$$

Es decir, que los montos reclamados siguen una distribución Log-Normal que se define:

$$monto_i \sim logN(monto|\mu, \sigma^2)$$

Para el resto del desarrollo $monto_i$ se denominará z_i.

Debido a que no contamos con información inicial acerca de las severidades (montos reclamados) asignamos a los parámetros una distribución inicial no informativa de la forma:

$$g(\mu, \sigma^2) \propto \frac{1}{\sigma}$$

Así, es posible encontrar la distribución posterior. El desarrollo completo se encuentra en el **Anexo 1**.

$$g(\mu, \sigma^2|x) \propto g(\mu, \sigma^2) f(x|\mu, \sigma^2) = g(\mu, \sigma) \prod_{i=1}^{n} f(x_i|\mu, \sigma)$$

$$\pi(\mu, \sigma^2|z_1, ..., z_n) \propto (\frac{1}{\sigma}) \prod_{i=1}^{n} N(z_i|\mu, \sigma^2)$$

La ecuación anterior tiene el kernel de una distribución Gamma como se expresa a continuación.

$$\propto GaIn\left(\sigma^2|\frac{n+1}{2}, \frac{1}{2}\sum_{i=1}^{n}(z_i - \bar{z})^2\right) \cdot N\left(\mu|\bar{z}, \sigma^2 + \sum_{i=1}^{n} z_i^2\right)$$

Donde para la función de distribución de Gamma Inversa sus parámetros de forma (α_1) y escala (β_1) están dados por:

$$\alpha_1 = \frac{n+1}{2}, \quad \beta_1 = \frac{1}{2}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

Asimismo, para la función de distribución Normal los parámetros m_1 y s_1 son:

$$m_1 = \bar{z}, \quad s_1 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

Modelación de la distribución compuesta y predicción

Una vez calculadas las distribuciones posteriores para los tres parámetros de interés λ, σ^2 y μ es posible generar una distribución compuesta. Esta consiste en tres modelos diferentes que al combinarlos no es posible encontrar una expresión analítica cerrada conocida. Por esta razón se simularán los valores tanto para los siniestros como para los montos reclamados 2008, a través de un código en R que se adjunta en el entregable.

La distribución compuesta es la siguiente:

$$\mathbb{P}(\sum_{j=1}^{J_{2008}} X_{2008,j}|datos) = \sum_{n=0}^{\infty}\mathbb{P}(N_{2008} = n|datos)\cdot\mathbb{P}(\sum_{j=1}^{n} X_{2008,j}|datos)$$

$$\pi(\lambda, \mu, \sigma^2) = Ga(\lambda|\alpha_1, \beta_1)\cdot GI(\sigma^2|\alpha_1, \beta_1)\cdot N(\mu|m_1, \sigma^2 s_1)$$

Simulación

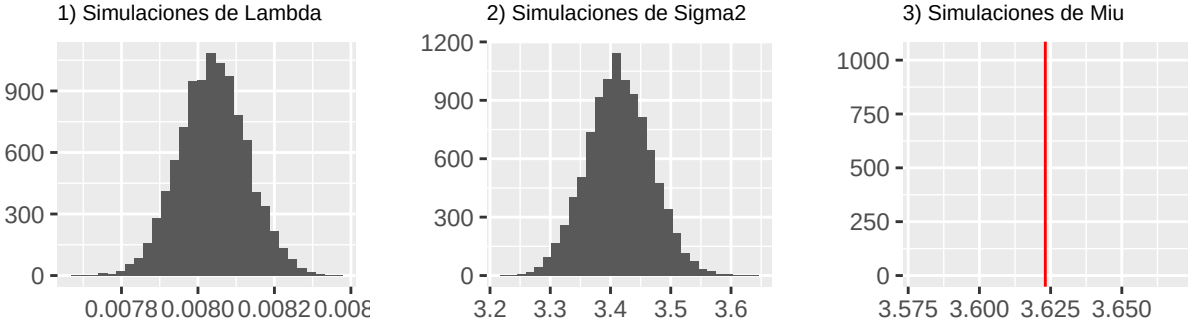
Simulamos M=10,000 veces los siguientes parámetros $\lambda^{(m)}$, $\sigma^{2(m)}$ y $\mu^{(m)}|\sigma^{2(m)}$ para obtener las predicciones de los siniestros y montos de reclamos en 2008. Todas las simulaciones que se presentan a continuación fueron generadas 10,000 veces para poder hacer inferencias válidas.

La tasa lambda se obtiene de simular M veces de una distribución posterior Gamma $\lambda^{(m)} \sim Ga(\lambda|\alpha_1, \beta_1)$ cuyos parámetros α_1 y β_1 fueron definidos en la sección anterior.

A continuación se simula la varianza de los montos reclamados para 2008 utilizando la distribución posterior Gamma Inversa $\sigma^{2(m)} \sim GaInv(\sigma^2|a_1, b_1)$ con parámetros a_1 y b_1 definidos en la sección anterior.

Para simular μ se utiliza una distribución Normal $\mu^{(m)}|\sigma^{2(m)} \sim N(\mu|m_1, \sigma^{2(m)}s_1)$ con los parámetros m_1 y s_1 que se obtienen como se indicó en la sección anterior.

A continuación se muetran las distribuciones de la 10,000 simulaciones de los parámetros λ, σ^2 y μ.



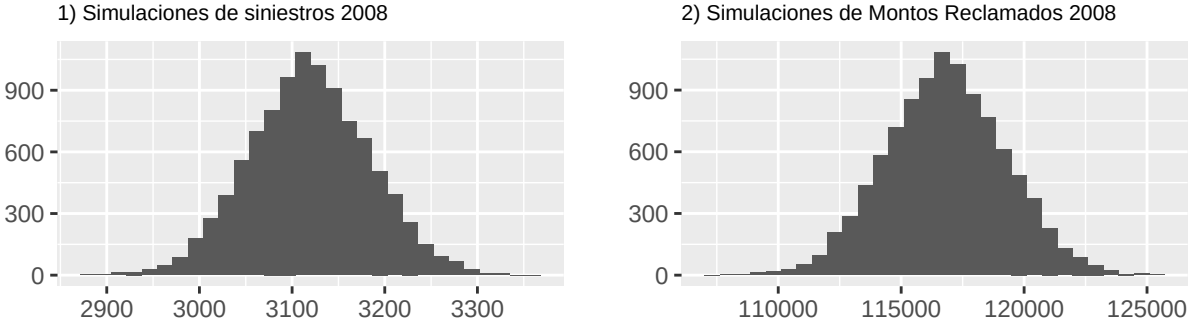
En la gráfica anterior en el panel 1) anterior muestra como se distribuyen las M simulaciones de lambda. Se puede confirmar que la magnitud de los datos coinciden con los observados en la tabla 1. Estas M simulaciones

de lambda serán utilizadas para obtener una distribución de las predicciones de los siniestros en 2008. Para este cálculo se predice el número de pólizas en 2008 utilizando una regresión lineal que es igual a 387,979 suscripciones.

En el panel 2 se observa la distribución de las σ^2 simuladas que presenta la siguientes características: unimodalidad, simetría y poca disperación. Además existen algunos datos atípicos ubicados en la cola derecha.

En el panel 3 se observa que para el parámetro μ no hay dispersión, las simulaciones se encuentran concentradas alrededor del $\mu = 3.623$. Esto se debe a que en el cálculo del parámetro s_1 la diferencia entre la suma de $z_i - \bar{z}$ es muy cercana a cero.

A continuación se muestran las simulaciones de los fenómenos de interés: frecuencia de siniestros y serveridades.




```
## [1] "Datos estadísticos de las simulaciones"
```

```
##      N_SIM        MONTOS_SIM
##  Min.   :2873   Min.   :107533
##  1st Qu.:3075   1st Qu.:115174
##  Median :3118   Median :116773
##  Mean   :3119   Mean   :116808
##  3rd Qu.:3162   3rd Qu.:118425
##  Max.   :3353   Max.   :125626
```

```
## [1] "Desviación estándar del número de siniestros simulado"
```

```
## [1] 64.93
```

```
## [1] "Desviación estándar del monto reclamado simulado"
```

```
## [1] 2433.37
```

En la gráfica anterior, panel 1) se presentan las simulaciones del número de siniestros (frecuencia de siniestros). Se observa una distribución unimodal y simétrica alrededor de la media=3119 con una dispersión $\sigma = 65$. Estas simulaciones son consistentes con los resultados obtenidos del número de siniestros reportados para cada año durante el análisis exploratorio.

En el panel 2) se grafican las simulaciones de las severidades (montos reclamados). Nuevamente se observa un histograma unimodal y simétrico alrededor de la media=116,808 con una $\sigma = 2433$. Estos resultados parecen subestimar aquellos montos reclamados del análisis exploratorio previo, sin embargo, las magnitudes son consistentes con el análisis.

5. Análisis de resultados

La modelación de la distribución compuesta presentada en este proyecto probó ser adecuada para los datos pertenecientes a la aseguradora AllState. Los parámetros obtenidos a través del método Bayesiano permitieron modelar consistentemente dos fenómenos de gran relevancia para la industria de seguros. Sin

embargo, para poder sacar conclusiones de manera completa, útiles para la aseguradora se presentan cálculos correspondientes al concepto financiero de Value-at-Risk (VaR) para el portafolio de pólizas y el cálculo para la prima individual.

Cálculo del VaR para el portafolio

El cálculo del VaR se define como el percentil del 99% de la distribución de los montos reclamados. Es decir, se tiene un 99% de confiabilidad de que el monto de reclamos de la aseguradora AllState en 2008 no rebasará el siguiente monto: $122,504 usd.

Cálculo de la prima del portafolio

La prima de riesgo del portafolio de seguros de automóvil en la empresa AllState es de $ 0.32 usd, que se calcula dividiendo el VaR entre el número de suscripciones estimadas para 2008.

```
## [1] "Cálculo de prima del portafolio"
```

```
## [1] 0.32
```

6. Conclusiones

El proyecto aquí desarrollado, permite utilizar el análisis Bayesiano para modelar dos métricas de gran relevancia en la industria aseguradora: Número de siniestros anuales y Montos reclamados anuales

La distribución Poisson que se utilizó para modelar el número de siniestros y la distribución Normal que se utilizó para modelar las severidades, aproximan de manera adecuada a los datos observados y pueden dar un resultado confiable para predecir las métricas en el año 2008. Debido a que no existe mucha dispersión entre los montos reclamados para los tres años, no fue necesario utilizar el último año para predecir. A raíz de esto, la prima de riesgo para 2008 sí utiliza la información de los tres años previos.

Es de llamar la atención que la simulación de las medias condicionadas a la varianza para los montos reclamados arroja valores con la mínima dispersión, sin embargo, el resultado obtenido tiene una magnitud coherente y permite dar una predicción para la prima de riesgo individual en seguro de autos para la empresa AllState. Además, como se mencionó anteriormente, los montos reclamados simulados parecen subestimar a los observados en los datos y valdría la pena expandir en trabajos futuros las causas adyacentes a este fenómeno. Finalmente, es relevante notar que la prima de riesgo resultó menor a un dólar (frente a primas promedio en la industria de alrededor de $20 usd), y una de las causas posibles es aquel efecto no considerado por la predicción lineal creciente (supuesto realizado para este proyecto) para el número de suscripciones anuales, lo cual por definición disminuye el cálculo para la prima del portafolio. Una línea de investigación futura sería considerar el efecto del aumento o dismución de suscripciones (pólizas en el portafolio) en el cálculo de primas y de este modo relajar o cambiar ciertos supuestos hechos para este trabajo.

Referencias

Heller, G., Mikis Stasinopoulos, D., Rigby, R., & De Jong, P. (2007). Mean and dispersion modelling for policy claims costs. Scandinavian Actuarial Journal, 2007 (4), 281-292.

Jiří Valecký. (2016). Modelling Claim Frequency in Vehicle Insurance. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis, 64 (2), 683-689.

Villar et. al. (2009). Analysis of an Aggregate Loss Model. Journal of Applied Statistics, 2009 (36), 149-166. doi: 10.1080/02664760802443921 # Anexos

Anexo 1

Desarrollo de la distribución posterior para la σ^2 y la μ

$$g(\mu,\sigma^2|x) \propto g(\mu,\sigma^2)f(x|\mu,\sigma^2) = g(\mu,\sigma)\prod_{i=1}^{n} f(x_i|\mu,\sigma)$$

$$\pi(\mu,\sigma^2|z_1,...,z_n) \propto (\frac{1}{\sigma})\prod_{i=1}^{n} N(z_i|\mu,\sigma^2)$$

$$\propto (\frac{1}{\sigma})\prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}exp\left(\frac{-1}{2\sigma^2}(z_i-\mu)^2\right)$$

$$\propto (\sigma^2)^{-(n/2+1/2)}exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(z_i-\mu)^2\right)$$

$$\propto (\sigma^2)^{-(\frac{n+1}{2})}exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(z_i-\bar{z}+\bar{z}-\mu)^2\right)$$

$$\propto (\sigma^2)^{-(\frac{n+1}{2})}exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}[(z_i-\bar{z})^2-2(z_i-\bar{z})(z_i-\mu)+(\bar{z}-\mu)^2]\right)$$

$$\propto (\sigma^2)^{-(\frac{n+1}{2})}exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(z_i-\bar{z})^2\right)\cdot exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}((\bar{z}-\mu)^2-2(z_i-\bar{z})(\bar{z}-\mu)^2)\right)$$

$$\propto (\sigma^2)^{-(\frac{n+1}{2})}exp\left(-(\frac{1}{\sigma^2})(\frac{1}{2}\sum_{i=1}^{n}(z_i-\bar{z})^2)\right)\cdot exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((\bar{z}-\mu)^2-2z_i(\bar{z}-\mu)^2+2\bar{z}(\bar{z}-\mu)^2\right))$$

Se completa el cuadrado

$$\propto (\sigma^2)^{-(\frac{n+1}{2})}exp\left(\frac{-1}{\sigma^2}(\sum_{i=1}^{n}(z_i-\bar{z})^2/2)\right)\cdot exp\left(\frac{-1}{2\sigma^2}\frac{(\mu-\bar{z})^2}{\sum_{i_1}^{n} z_i^2}\right)$$

Anexo 2

Desarrollo de la distribución compuesta.

$$\mathbb{P}(\sum_{j=1}^{J_{2008}} X_{2008,j}|datos) = \sum_{n=0}^{\infty}\mathbb{P}(N_{2008}=n|datos)\cdot\mathbb{P}(\sum_{j=1}^{n} X_{2008,j}|datos)$$

$$=\sum_{j=1}^{\infty}\int \mathbb{P}(N_{2008}=n|J_{2008}.\lambda)\cdot GaI(\lambda|\alpha_1,\beta_1)d\lambda\cdot\int N(X_{2008,1},...,X_{2008,n}|N_{2008}=n,\mu,\sigma^2)\cdot Ga(\sigma^2|\alpha_1,\beta_1)\cdot N(\mu|m_1,\sigma^2S_1)$$

$$=\int \mathbb{P}(N_{2008}=m|\lambda)\cdot\prod_{j=1}^{m} N(z_{2008,j}|\mu,\sigma^2)\cdot\pi(\lambda,\mu,\sigma^2|datos)d\lambda,d\mu,d\sigma^2$$

$$\pi(\lambda,\mu,\sigma^2) = Ga(\lambda|\alpha_1,\beta_1)\cdot GI(\sigma^2|\alpha_1,\beta_1)\cdot N(\mu|m_1,\sigma^2 s_1)$$