Prediction of reserves using multivariate power-normal mixture distribution

Ang Siew Ling and Pooi Ah Hin

Citation: AIP Conference Proceedings **1782**, 050003 (2016); doi: 10.1063/1.4966093

View online: http://dx.doi.org/10.1063/1.4966093

View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1782?ver=pdfcov

Published by the AIP Publishing

Prediction of Reserves Using Multivariate Power-Normal Mixture Distribution

Ang Siew Ling[1, a)] and Pooi Ah Hin[2, b)]

[1, 2] *Sunway University Business School,*
Centre for Actuarial Studies, Applied Finance and Statistics,
No. 5, Jalan Universiti, Bandar Sunway,
47500, Subang Jaya, Selangor.

[a)]siewlinga@sunway.edu.my
[b)]ahhinp@sunway.edu.my

Abstract. Recently, in the area on stochastic loss reserving, there are a number of papers which analyze the individual claims data using the Position Dependent Marked Poisson Process. The present paper instead uses a different type of individual data. For the *i*-th $(1 \le i \le n)$ customer, these individual data include the sum insured s_i together with the amount paid y_{ij} and the amount a_{ij} reported but not yet paid in the *j*-th $(1 \le j \le 6)$ development year. A technique based on multivariate power-normal mixture distribution is already available for predicting the future value ($y_{i\,j+1}$, $a_{i\,j+1}$) using the present year value (y_{ij}, a_{ij}) and the sum insured s_i . Presently the above technique is improved by the transformation of distribution which is defined on the whole real line to one which is non-negative and having approximately the same first four moments as the original distribution. It is found that, for the dataset considered in this paper, the improved method gives a better estimate for the reserve when compared with the chain ladder reserve estimate. Furthermore, the method is expected to provide a fairly reliable value for the Provision of Risk Margin for Adverse Deviation (PRAD).

INTRODUCTION

Outstanding claims reserve in non-life insurance is established to provide for the future liability for claims which are incurred but not yet reported (IBNR) or which have been reported and not yet settled (IBNS). The setting and monitoring of outstanding claims reserve is a vital task of an actuary. The actuary makes use of a variety of available methods to calculate and set the reserve needed for an insurer. In the current practice, most of the actuaries in the non-life insurance use the chain-ladder technique to obtain the best estimate of the reserve.

The distribution-free chain ladder model of Mack is a frequently used model for stochastic claims reserving [1]. Renshaw and Verrall have tried to link chain ladder with different stochastic models [2]. England and Verrall further discussed the stochastic methods based on the framework of generalised linear models [3]. Instead of using the data aggregated in run-off triangles, some authors used the individual claims data to study loss reserving. The work on individual claims estimation using Position Dependent Marked Poisson Processes can be found in [4-10].

In this paper we use the data for the sum insured together with the amount paid and the amount reported but not yet paid in the *j*-th development year for $1 \le j \le 6$. This data is essentially a summarized version of the individual claims data. A model based on Multivariate Power-Normal Mixture (MPNM) distribution [11-12] is proposed for calculating the aggregate claims liabilities. For the dataset considered, the prediction interval based on the distribution for the aggregate claim liabilities is found to have good ability of covering the observed aggregate claim liabilities. The estimate for the aggregate claim liabilities based on the mean of the distribution is found to have a smaller mean absolute percentage error when compared with the chain ladder reserve estimate.

The 4th International Conference on Quantitative Sciences and Its Applications (ICOQSIA 2016)
AIP Conf. Proc. 1782, 050003-1–050003-7; doi: 10.1063/1.4966093
Published by AIP Publishing. 978-0-7354-1444-0/$30.00

The layout of the paper is as follows. In Section 2, we give a short introduction to the MPNM distribution. In Section 3, we give an updated version of the method in [13] for estimating reserves. Section 4 gives some numerical results and section 5 concludes the paper.

MULTIVARIATE POWER-NORMAL MIXTURE DISTRIBUTION

The random variable $\tilde{\varepsilon}$ defined by the transformation [14]

$$\tilde{\varepsilon}=\psi(\lambda^{+},\lambda^{-},z)=\begin{cases}[(z+1)^{\lambda^{+}}-1]\big/\lambda^{+} & ,z\geq 0,\lambda^{+}\neq 0\\ \log(z+1) & ,z\geq 0,\lambda^{+}=0\\ -[(-z+1)^{\lambda^{-}}-1]\big/\lambda^{-} & ,z<0,\lambda^{-}\neq 0\\ -\log(-z+1) & ,z<0,\lambda^{-}=0\end{cases} \tag{1}$$

of the standard normal random variable z is said to have a power-normal distribution with parameters λ^{+} and λ^{-}. The vector **y** is said to have a *k*-dimensional power-normal distribution with parameters $\boldsymbol{\mu},\mathbf{H},\lambda_i^{+},\lambda_i^{-},\sigma_i,1\leq i\leq k$ if

$$\mathbf{y}=\boldsymbol{\mu}+\mathbf{H}\boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\mu}=E(\mathbf{y})$, $\mathbf{H}$ is an orthogonal matrix and $\varepsilon_1,\varepsilon_2,...,\varepsilon_k$ are uncorrelated,

$$\varepsilon_i=\frac{\sigma_i\left[\tilde{\varepsilon}_i-E\left(\tilde{\varepsilon}_i\right)\right]}{\left\{\operatorname{var}\left(\tilde{\varepsilon}_i\right)\right\}^{1/2}} \tag{3}$$

$\sigma_i>0$ is a constant, and $\tilde{\varepsilon}_i$ has a power-normal distribution with parameters λ_i^{+} and λ_i^{-} [11].

For $0\leq p_i\leq 1$, i =1, 2, define

$$\tilde{\varepsilon}_i^{*}=\begin{cases}\bar{\tilde{\varepsilon}}_i^{(1)}-\mu_{i1} \text{ with probability } p_i\\ \bar{\tilde{\varepsilon}}_i^{(2)}+\mu_{i2} \text{ with probability } q_i=1-p_i\end{cases}$$

where μ_{i1} and μ_{i2} are constants,

$$\bar{\tilde{\varepsilon}}_i^{(j)}=\frac{\tilde{\varepsilon}_i^{(j)}-E\left(\tilde{\varepsilon}_i^{(j)}\right)}{\left\{\operatorname{var}\left(\tilde{\varepsilon}_i^{(j)}\right)\right\}^{1/2}},\ \ j=1,2$$

$\tilde{\varepsilon}_i^{(j)}$ has a power-normal distribution with parameters $\lambda_i^{(j)+}$ and $\lambda_i^{(j)-}$ and

$$\mathrm{E}\ (\tilde{\varepsilon}_i^{*})=0 \tag{4}$$

In order to achieve the condition given by Equation (4), the constant μ_{i2} should be given by

$$\mu_{i2}=\frac{p_i\mu_{i1}}{q_i}$$

the variance of $\tilde{\varepsilon}_i^{*}$ is then given by

$$\operatorname{var}\left(\tilde{\varepsilon}_i^{*}\right)=1+\left(p_i+p_i^2/q_i\right)\mu_{i1}^2$$

The random variable $\tilde{\varepsilon}_i^{*}$ is said to have a power-normal mixture distribution with parameters $p_i,\mu_{i1},\lambda_i^{(j)+}$ and $\lambda_i^{(j)-},j=1,2$ [12].

We next define

$$\varepsilon_i^{*}=\begin{cases}\sigma_i^{*}\ (\bar{\tilde{\varepsilon}}_i^{(1)}-\mu_{i1})\Big/\left\{\operatorname{var}\left(\tilde{\varepsilon}_i^{*}\right)\right\}^{1/2} \text{ with probability } p_i\\ \sigma_i^{*}\ (\bar{\tilde{\varepsilon}}_i^{(2)}+\mu_{i2})\Big/\left\{\operatorname{var}\left(\tilde{\varepsilon}_i^{*}\right)\right\}^{1/2} \text{ with probability } q_i\end{cases} \tag{5}$$

and consider a vector $\mathbf{y}^{*}$ consisting of *k* correlated random variables. The vector $\mathbf{y}^{*}$ is said to have a *k*-dimensional power-normal mixture distribution with parameters $\boldsymbol{\mu}^{*},\mathbf{H}^{*}$, p_i, μ_{i1}, σ_i^{*}, $\lambda_i^{(j)+}$, $\lambda_i^{(j)-}$, $1\leq i\leq k$, $1\leq j\leq 2$ if

$$\mathbf{y}^* = \boldsymbol{\mu}^* + \mathbf{H}^*\boldsymbol{\varepsilon}^* \tag{6}$$

where $\mu^* = E[\mathrm{y}^*]$; $\mathbf{H}^*$ is an orthogonal matrix, $\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_k^*$ are uncorrelated and ε_i^* is given by Equation (5).

METHOD BASED ON MULTIVARIATE POWER-NORMAL MIXTURE DISTRIBUTION FOR ESTIMATING RESERVES

As in [13], we consider the following data for the i –th customer among a group of N customers from n customers: (i) the sum insured S_i, (ii) the amount paid Y_{ij} and (iii) the amount A_{ij} reported but not yet paid in the j-th development year for $1 \le j \le 6$. Table 1 shows an example of the data for N customers. The i-th entry in first column of the Table 1 is S_i. For the subsequent columns in Table 1, the (i, j) entry is (Y_{ij}, A_{ij}).

TABLE 1: Original Claims Data

Sum Insured	Development year (j)					
	1	2	3	4	5	6
S_1	(Y_{11}, A_{11})	(Y_{12}, A_{12})	(__ , __)	(__ , __)	(__ , __)	(__ , __)
S_2	(Y_{21}, A_{21})	(Y_{22}, A_{22})	(Y_{23}, A_{23})	(__ , __)	(__ , __)	(__ , __)
S_3	(Y_{31}, A_{31})	(__ , __)	(__ , __)	(__ , __)	(__ , __)	(__ , __)
S_4	(Y_{41}, A_{41})	(Y_{42}, A_{42})	(Y_{43}, A_{43})	(Y_{44}, A_{44})	(Y_{45}, A_{45})	(__ , __)
S_5	(Y_{51}, A_{51})	(Y_{52}, A_{52})	(Y_{53}, A_{53})	(Y_{54}, A_{54})	(Y_{55}, A_{55})	(Y_{56}, A_{56})
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
S_N	(Y_{N1}, A_{N1})	(Y_{N2}, A_{N2})	(Y_{N3}, A_{N3})	(Y_{N4}, A_{N4})	(__ , __)	(__ , __)

Note: The symbol "__" in the table denotes the value which is still unknown.

The unknown values of the i-th row in TABLE 1 may be denoted as (Y_{ij}, A_{ij}), $j = j_i, j_i + 1, \ldots, 6$, and the estimated value of the unknown (Y_{ij}, A_{ij}) may be denoted as $(\hat{Y}_{ij}, \hat{A}_{ij})$. The aggregate claim liabilities

$$R = \sum_{i=1}^{N}\left[\left(\sum_{j=j_i}^{6} Y_{ij}\right) + A_{i6}\right]$$

may then be estimated by

$$\hat{R} = \sum_{i=1}^{N}\left[\left(\sum_{j=j_i}^{6} \hat{Y}_{ij}\right) + \widehat{A}_{i6}\right]. \tag{7}$$

To estimate (Y_{ij}, A_{ij}), let us first assumed that we have a fairly large amount of historical data given by $(s_i\ , y_{i1}\ , a_{i1}, \ldots, y_{i6}, a_{i6})$, $i = 1, 2, \ldots, n$, and use the method described below to fit multivariate power-normal mixture distributions to the historical data.

For j =1, 2,…, 5, we form sub-tables labelled as (j, r, 0, 0, d), (j, r, r, 0, d), (j, r, 0, r, d), (j, r, r, r, d), (j, r, 0, 0, 0, d), (j, r, 0, 0, r, d), (j, r, 0, r, 0, d), (j, r, r, 0, 0, d), (j, r, r, 0, r, d), (j, r, 0, r, r, d), (j, r, r, r, 0, d) and (j, r, r, r, r, d) [14]. For a given sub-table, let y_i denote the value of the i-th non-zero column, $1 \le i \le k$. After finding a MPNM distribution for $\boldsymbol{y}$, we find a conditional distribution for the k-th variable when the values of the initial k-1 variables are given. The resulting conditional distribution may not be non-negative. However, we may transform the distribution of the k –th variable y_k to that of the variable y_k^+ which is non-negative and having approximately the same first four moments as the original k-th variable. The transformation may be chosen to be

$$y_k^+ = \begin{cases} 0, & \text{with probability } q \\ y_k^* \big/ P\left(y_k^* \geq 0\right), & \text{with probability } p = 1 - q \end{cases} \tag{8}$$

where $p > 0$ is a constant,

$$y_k^* = \mu^* + \sigma^* \frac{\tilde{\varepsilon} - E(\tilde{\varepsilon})}{\left\{Var(\tilde{\varepsilon})\right\}^{1/2}} ,$$

μ^*, σ^* are constants and $\tilde{\varepsilon}$ is given by Equation (1).

We may use the following procedures to examine the goodness of fit provided by the multivariate power-normal mixture distribution.

Suppose the given table with k non-zero columns has N_w rows. For the i-th $\left(1 \leq i \leq N_w\right)$ set of values of the first $k - 1$ variables given in the table, we generate a value for the k-th variable. The N_w observed and generated values for the k-th variable are sorted individually in a non-descending order, and plotted against i. The plot then essentially compares the quantiles of the marginal distributions of the observed and fitted values for the last variable y_k .

There may be cases in which the goodness of fit of the procedure based on the multivariate power-normal mixture distribution is not satisfactory. The lack of fit may typically be due to (A) some extreme observed values of the last variable y_k or (B) the following fairly deterministic relationship between the last two variables y_{k-1} and y_k :

$$y_k = \begin{cases} 0 & \text{with probability } P_Z \\ y_{k-1} & \text{with probability } P_{\bar{Z}} = 1 - P_Z \end{cases}$$

For case (A), suppose that there are N_E extreme values and $N_{\bar{E}}$ non-extreme values of y_k . We may initially fit a multivariate power-normal mixture distribution to the $N_{\bar{E}}$ values of $\boldsymbol{y}$ when the values of y_k (denoted as $y_k^{(\bar{E})}$) are non-extreme. When the initial (k -1) values are given, the value for y_k may now be assumed to take the form

$$y_k = \begin{cases} y_k^{(\bar{E})} & \text{with probability } N_{\bar{E}} \big/ \left(N_E + N_{\bar{E}}\right) \\ \text{(a value selected at random} & \\ \text{from the } N_E \text{ extreme values)} & \text{with probability } N_E \big/ \left(N_E + N_{\bar{E}}\right) \end{cases}$$

For case (B), suppose there are N_Z cases in which y_k is zero and $N_{\bar{Z}}$ cases in which $y_k = y_{k-1}$. When the initial (k -1) values are given, the value for y_k may now be assumed to be given by

$$y_k = \begin{cases} 0 & \text{with probability } N_Z \big/ \left(N_Z + N_{\bar{Z}}\right) \\ y_{k-1} & \text{with probability } N_{\bar{Z}} \big/ \left(N_Z + N_{\bar{Z}}\right) \end{cases}$$

The goodness of fit of the distribution may be examined further by using the following procedure given in [14]:

(I) Select at random N rows from the n rows of historical data, and for the i-th selected row, assign a value selected at random from $\{1, 2, \ldots, 5\}$ to j_i -1 and treat $\{(Y_{ij}, A_{ij}),\ \ j = j_i,\, j_i + 1, \ldots, 6\}$ as the set of unknown values for the i-th customer.

(II) For $j = j_i,\, j_i + 1, \ldots, 6$, use the fitted multivariate power-normal mixture distributions to iteratively

(i) generate a value $\hat{Y}_{ij}$ for Y_{ij} , given the values of $S_i, \hat{Y}_{ij-1}$ and $\hat{A}_{ij-1}$;

(ii) generate a value $\hat{A}_{ij}$ for A_{ij} , given the values of $S_i, \hat{Y}_{ij-1}$, $\hat{A}_{ij-1}$ and $\hat{Y}_{ij}$;

where $\hat{Y}_{ij_i-1} = Y_{ij_i-1}$ and $\hat{A}_{ij_i-1} = A_{ij_i-1}$.

(III) Estimate the aggregate claim liabilities by using Equation (7).

(IV) Repeat Steps (I) to (III) N_g number of times.

The average value $\bar{R}$ of the N_g generated values of $\hat{R}$ may then treated as the *best* estimate based on multivariate power-normal mixture distribution for R. Denoting [v] as the largest integer which is smaller than or equal to v, we sort the N_g generated values of $\hat{R}$ in an ascending order, and use the $[N_g(\alpha/2)]$ -th and $[N_g(1-\alpha/2)]$ -th sorted values to estimate respectively the $100\,(\alpha/2)\,\%$ and $100\,(1-\alpha/2)\,\%$ points of the distribution of $\hat{R}$. Let the estimated percentage points be denoted as L and U respectively. Then $[\,L\,,U\,]$ may be taken to be a nominally $100\,(1-\alpha)\,\%$ prediction interval for the unknown aggregate claim liabilities *R*.

Steps (I) to (IV) are repeated N_r times. For the i_r–th time of repetition, let $\left[L_{i_r}, U_{i_r}\right]$ be the corresponding prediction interval for *R*, $\bar{R}_{i_r}$ the corresponding best estimate for *R* and $R_{i_r}^{(0)}$ the corresponding observed value of *R*. The mean absolute percentage error given by

$$\text{MAPE} = \frac{1}{N_r}\sum_{i_r=1}^{N_r}\frac{\left|\bar{R}_{i_r} - R_{i_r}^{(0)}\right|}{R_{i_r}^{(0)}}\times 100\%$$

is then a measure of the prediction accuracy. Furthermore, the estimated coverage probability of the prediction interval for *R* may be used to judge the goodness of fit of the multivariate power-normal mixture distribution. A value of the estimated coverage probability which is close to the target value 1 – α is an indication that the fit given by the multivariate power-normal mixture distribution is adequate.

NUMERICAL RESULTS

The data for $(s_i\,, y_{i1}\,, a_{i1}, \dots, y_{i6}, a_{i6}),\, i = 1, 2, \dots, 1000$, used in this study are obtained by coding the data from an insurance company in Malaysia. The goodness of fit of the fitted multivariate power-normal mixture distributions is investigated by using the procedures in the previous section.

If a sub-table of *k* non-zero columns has less than 10 rows, or the values in the last column are all equal to zero, then the value of y_k is taken to be zero irrespective of the values $y_1, y_2, \dots, y_{k-1}$ of the initial *k* – 1 variables.

When the number of rows in a sub-table is at least 10, then a multivariate power-normal mixture distribution is fitted to the data. If the fit is not satisfactory, then the numbers N_E and $N_{\bar{E}}$ (or N_Z and $N_{\bar{Z}}$) are noted. Examples of the values of N_E , $N_{\bar{E}}$ and N_Z , $N_{\bar{Z}}$ are given in Tables 2 and 3.

TABLE 2: Number of Extreme Observations in the sub-table

Sub-tables	N_E	$N_{\bar{E}}$
(1, r, 0, 0, d)	2	411
(1, r, 0, r, d)	1	311
(1, r, r, r, d)	1	48
(1, r, 0, r, r, d)	4	196
(2, r, 0, r, r, d)	1	67

TABLE 3: Number of Zero Observations in the sub-table

Sub-tables	N_Z	$N_{\bar{Z}}$
(1, r, 0, r, 0, d)	15	97
(1, r, r, r, 0, d)	20	18
(2, r, 0, r, 0, d)	16	138
(2, r, r, r, 0, d)	10	26
(3, r, 0, r, 0, d)	5	148
(3, r, 0, r, r, d)	39	3
(3, r, r, r, 0, d)	1	12
(4, r, 0, r, 0, d)	17	121
(4, r, 0, r, r, d)	29	3
(4, r, r, r, 0, d)	1	8
(5, r, 0, r, 0, d)	1	121
(5, r, 0, r, r, d)	5	0
(5, r, r, r, 0, d)	0	3

For the data of which the values of the last column are non-extreme values, we fit them with a multivariate power-normal mixture distribution and find a conditional distribution for the *k*-th variable when the values of the initial *k*-1 variables are given. The resulting conditional distribution is transformed by using the transformation given by Equation (8) and later used to generate a value for the last variable y_k .

The plots of the sorted versions of the observed and generated values for the last variable y_k are shown in Figure 1 and Figure 2.

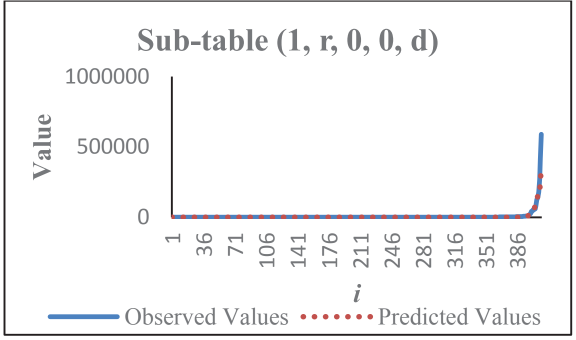

FIGURE 1. The plot of the observed and generated values of the last variable y_k after sorting in the case when the data in the sub-table (1, r, 0, 0, d) are used.

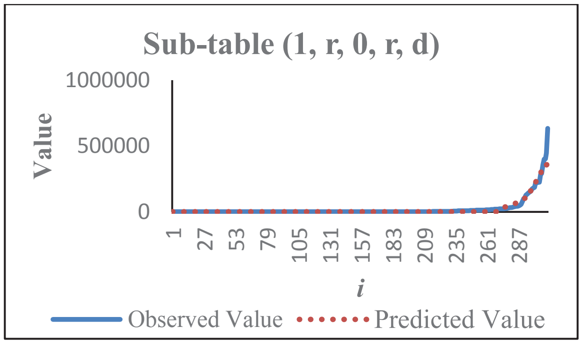

FIGURE 2. The plot of the observed and generated values of the last variable y_k after sorting in the case when the data in the sub-table (1, r, 0, r, d) are used.

Figures 1 - 2 show that for the data in the sub-tables (1, r, 0, 0, d) and (1, r, 0, r, d), the observed and generated values of the last variable y_k have about the same empirical distributions. Similar goodness of fit results can also be obtained for the data in the remaining sub-tables.

When N_r is chosen to be 100, the value of MAPE for the estimate based on multivariate power-normal mixture distribution is found to be 54.5% while that for the estimate based on chain ladder is 60.7%. Thus for the dataset considered, the reserve estimate based on multivariate power-normal mixture distribution performs better than that of the chain ladder procedure.

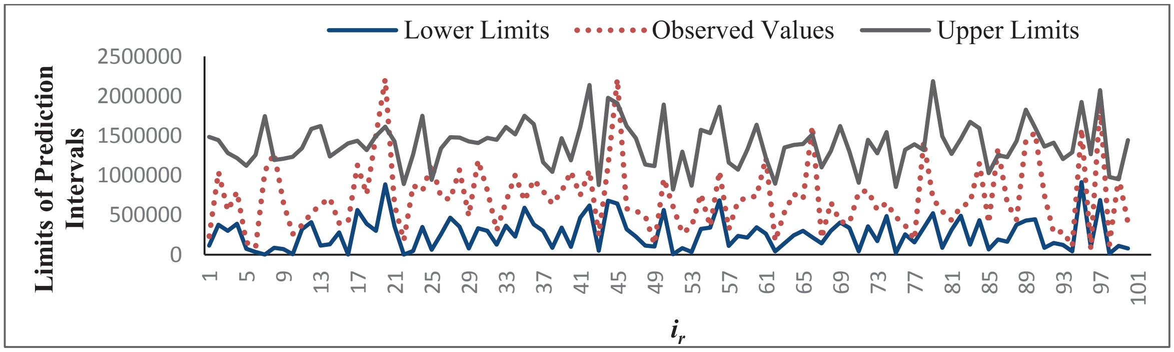

FIGURE 3. Prediction Interval for the Aggregate Claims Liabilities. [$n = 1000$, $N = 500$, $N_r = 100$, $N_g = 100$, $\alpha = 0.05$, Estimated Coverage Probability = 0.90]

Figure 3 shows the $N_r = 100$ prediction intervals for the aggregate claims liabilities. The estimated coverage probability of 0.90 is not too far from the target value 0.95. Thus the fit given by the multivariate power-normal mixture distribution is fairly satisfactory, and the difference between the 75% point of the distribution of the estimated reserve and the best estimate of the reserve is expected to provide a fairly reliable value for the Provision of Risk Margin for Adverse Deviation (PRAD).

CONCLUSION

The method based on the multivariate power-normal mixture distribution for estimating reserve is promising. The method yields a reserve estimate which could be better than the chain ladder reserve estimate. Furthermore, the method permits an estimation of the distribution of the reserve which can be used to estimate the Provision of Risk Margin for Adverse Deviation (PRAD).

It is also possible that the multivariate power-normal mixture distribution may provide an alternative way of analyzing the individual claims data which have previously been analyzed by using the Position Dependent Marked Poisson Process.

REFERENCES

1. T. Mack, ASTIN Bulletin, **23**, 213-225 (1993).
2. A. E. Renshaw and R.J. Verrall, British Actuarial Journal, **4**, pp.903 – 923 (1998).
3. P.D. England and R.J.Verrall, "Stochastic Claims Reserving in General Insurance" Presented to the Institute of Actuaries (2002).
4. E. Arjas, ASTIN Bulletin, **19**, 139 – 152 (1998).
5. R. Norberg, ASTIN Bulletin, **23**, 95-115 (1993); *ibid* **29**, 5-15 (1999).
6. S. Haastrup and E. Arjas, ASTIN Bulletin, **26**, 139 – 164 (1996).
7. C.R. Larsen, ASTIN Bulletin, **37**, 113-132 (2007).
8. X.B.Zhao, X. Zhou and J.L. Wang, Insurance: Mathematics and Economics, **45**, 1-8 (2009).
9. X.B. Zhao and X. Zhou, Insurance: Mathematics and Economics, **46**, 290-299 (2010).
10. K. Antonio and H.J. Plat. Scandinavian Actuarial Journal, **7**, 649-669 (2014).
11. A. H. Pooi, Applied Mathematical Sciences, **6**, pp. 5735 – 5748 (2012).
12. A. H. Pooi, Applied Mathematical Sciences, **8**, pp.5613-5623 (2014).
13. S. L. Ang and A. H. Pooi, "Internal Modelling under Risk-Based Capital Framework" in *Proceedings of the 2nd Innovation and Analytics Conference & Exhibition* 2015 (IACE 2015), Vol. 1691, pp. 050003.
14. I.K. Yeo and R. A. Johnson, Biometrika, **87**, 954 – 9 (2000).