Modeling Loss Given Default (LGD) by Finite Mixture Model
Chao Huang, Oklahoma State University
Liang Xie, SAS User

ABSTRACT
The loss given default or claim data, which tends to be skewed, heavy-tailed, multimodal, or kurtosis-like, from the insurance and banking industry is routinely fitted by a single distribution such as beta, lognormal, gamma, Pareto, etc., while the effects by the underlying sub-distributions within the overall distribution are generally ignored. Finite mixture model (FMM) is a widely used probabilistic model to make statistical inferences on parameters from an unknown distribution in terms of mixtures of known distributions. The new FMM procedure shipped with SAS 9.3 makes the building and diagnosis of the finite mixture model much easier. In this paper, we illustrate how to create valid finite mixture models for Loss Given Default (LGD) by an example of a credit portfolio with 317 obligors.

INTRODUCTION
In statistics, a mixture model is a probabilistic model including various sub-populations within an overall population, without much information about the distributions of each sub-population. Finite mixture model provides a robust semi-parametric framework to fit data with unknown distribution [1]. A basic example of finite mixture model can be described as two univariate normal components with the same standard deviation σ but different means μ₁ and μ₂ and their corresponding proportions π₁ and π₂, such as:

\[ f(y_i) = \pi_1 \phi(y_i; \mu_1, \sigma^2) + \pi_2 \phi(y_i; \mu_2, \sigma^2) \]

There are a few challenges against the modeling of the default or claim data for the insurance and banking industry, which tends to be skewed, heavy-tailed or kurtosis-like. In addition, the data usually shows multiple peaks, which suggests multimodal or a mixture of sub-populations. To realize the functionality of finite mixture model, Matt Flynn introduced the techniques by SAS’s NL MIXED procedure, JMP or R separately, if the starting values for the component distributions were given [2].

The experimental FMM procedure in SAS 9.3 provides maximum likelihood estimation by either Gibbs sampling or Metropolis-Hastings algorithm [3]. The ability for PROC FMM to solve the parameters from an unknown distribution in terms of mixtures of known distributions is especially helpful for understanding the messy loss data. This procedure enables us to assess the probabilities of a known distribution in fitting unknown data and evaluate the performance of a variety of distributions. In this paper, using PROC FMM, we discuss the application of finite mixture model on Loss Given Default (LGD) in the real world setting.

A CREDIT PORTFOLIO TO MODEL LOSS GIVEN DEFAULT
Loss Given Default (LGD) is the percentage of loss over the total exposure when bank’s counterparty becomes default, and is an important parameter applied in the calculation of Economic Capital or Regulatory Capital under Basel II for a banking institution [4]. Since LGD is influenced by key transaction characteristics, several categories of variables, such as firm-specific variable, macroeconomic variable and industry-specific variable, are used to build a predictive model for LGD.

In the fifth chapter of their book, Gunter and Peter [5] provided a credit portfolio of 317 obligors from a commercial bank, and then illustrated the OLS modeling solution by programming Excel/VBA. In their case, the general formula can be written as:

\[ LGD = b_0 + b_1 \times LEV + b_2 \times LGD_A + b_3 \times I\_DEF + \varepsilon \]

Here LGD is the real loss given default, LEV is the leverage coefficient by firm, LGD_A is the mean default rate by year, and I_DEF is the mean default rate by industry.

Figure 1a presents the five-number summaries of the 3 dependent variables and the independent variable in this portfolio. The distribution and the kernel density of LGD are further depicted in Figure 1b. In general, the dependent variable LGD is skewed to the left. To reduce the heteroscedasticity, Gunter and Peter used the beta linkage function to transform LGD in order to satisfy the OLS regression’s assumptions. However, the distribution of LGD displays a multimodal shape, which suggests that a single beta distribution may not well fit.
Figure 1. Summary of the dependent and independent variables. (a) A boxplot for each of the four variables in the model (the unit of X axis is the percentage). (b) A histogram overlaid with kernel density for the dependent variable LGD.

MODELING LGD WITH THE HOMOGENEOUS DISTRIBUTIONS

One of the advantages of PROC FMM is the model selection feature. For example, it will search the best component number within the range specified, if KMIN or KMAX option in the MODEL statement is activated. In this example, we need to reveal the optimized number of the components for each distribution. And we are also interested in choosing the best one out of the four homogenous mixtures such as beta, lognormal, gamma and normal distributions. To simplify the codes, a macro with six parameters that wraps PROC FMM is created.

%macro modselect(data = , depvar = , kmin= , kmax = , outstat = , modlist = );
/*********************************************************************/
* @param data      the name of the input dataset
* @param depvar    the name of the dependent variable
* @param kmax      the maximum number of the components
* @param outstat   the name of the output dataset
* @param modlist   the list of the distributions tested separated
*                    by space
* @return 5 fitting criterions plus the number of effective
*                    components and effective parameters
*********************************************************************/
%let modcnt = %eval(%sysfunc(count(%cmpres(&modlist),%str( )))+1);
%do i = 1 %to &modcnt;
  %let modelnow = %scan(&modlist, &i);
  ods output fitstatistics = &modelnow(rename=(value=&modelnow));
  ods select densityplot fitstatistics;
  proc fmm data = &data;
    model &depvar = / krestart kmax=&kmax dist=&modelnow;
  run;
%end;
 data &outstat;
%do i = 1 %to &modcnt;
  set %scan(&modlist, &i);
%end;
run;
%mend;
Then we run this macro to fit the mixture model with the four distributions. The model option KMAX is set as 4 to specify the maximum number of the components. The KRESTART option inhibits the results of previous models from becoming the starting values of the subsequent models. For each of the four mixture models, the density plots and the parameters are displayed in Figure 2, distinctively.

```
%modselect(data = lgddata, depvar = lgd, kmax = 4, outstat = result,
            modlist = beta lognormal gamma normal);
```

**Figure 2.** The histograms overlaid with estimated densities. (a) The density plot of LGD fitted by 3-component beta distributions. (b) The density plot by 4-component lognormal distributions. (c) The density plot of LGD fitted by 2-component gamma distributions. (d) The density plot of LGD fitted by 4-component normal distributions.

As the result, the penalized criterions, including AIC, BIC, AICC and -2 Log Likelihood, indicate that 3-component mixture model of beta distribution is significantly better than 4-component models of either Gamma distribution or lognormal distribution, while is only slightly better than 4-component model of normal distribution (Figure 3a). Furthermore, the beta distribution has higher Pearson statistic value and less numbers of parameters and components (Figure 3b). Thus, a 4-component homogeneous beta distribution is the best choice in this experiment.

Since the optimal option from the steps above is tested to be a 3-component mixture model of beta distribution, we are able to finalize the model with the three original independent variables and the LGD as dependent variables.

```
proc fmm data=lgddata;
  model lgd = lev lgd_a i_def / k = 3 dist = beta;
```
ods output parameterestimates = parmstat;
run;

As the result, PROC FMM supplies the estimation of all parameter for the three components. Each component comes with 5 parameters, including the coefficients for the mean default rates of firm, year and industry, intercept and scale (Figure 3c).

![Graph A](image1.png)
![Graph B](image2.png)
![Graph C](image3.png)

Figure 3. Comparison among the candidate models for the fitting performance. (a) Series plot of penalized criterions. The smaller values mean better. (b) Bubble plot by effective components against effective parameters (the bubble size represents the value of Pearson statistic). (c) Values of the estimated parameters for the 3-component mixture model of beta distribution

**MODELING LGD WITH THE HETEROGENEOUS DISTRIBUTIONS**

The mixture of the sub-distributions can be from different families. For example, the following code demonstrates a 2-component mixture model for LGD with one component is modeled as normally distributed while the other one is fitted by t-distribution.
**proc fmm** data=lgddata;
   model lgd = lev lgd_a i_def /dist = normal;
   model + lev lgd_a i_def /dist = t;
   output out = fmm_out pred resid(component) resid(overall);
run;

To decide which factors will contribute to classifying data points into different latent classes, the PROBMODEL statement in PROC FMM defines the model effects for the mixing probabilities and their link function. The probability model can be one of the following four options: logistic, probit, log-log, and complementary log-log. This feature empowers us to interpret latent class with insightful factors and even provide better modeling performance.

**CONCLUSION**

The FMM procedure in SAS provides us a robust platform to understand information from loss given default data by finite mixture model.

**REFERENCES**

   [http://www.ifsugproceedings.org/2012/IN/IN2.pdf](http://www.ifsugproceedings.org/2012/IN/IN2.pdf)

**CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Chao Huang
Email: hchao8@gmail.com
Blog: www.sasanalysis.com

Liang Xie
Email: xie1978@yahoo.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.