ATLANTIS PRESS
Journal of Statistical Theory and Applications, Vol. 17, No. 2 (June 2018) 307–323

A Tutorial on Levels of Granularity: From Histograms to Clusters to Predictive Distributions

STANLEY L. SCLOVE
Department of Information & Decision Sciences
University of Illinois at Chicago
601 S. Morgan St.
Chicago, IL 60607-7124
slsclove@uic.edu

Received 13 March 2017
Accepted 30 May 2017

Abstract

Consider the problem of modeling datasets such as numbers of accidents in a population of insured persons, or incidences of an illness in a population. Various levels of detail or granularity may be considered in describing the parent population. The levels used in fitting data and hence in describing the population may vary from a single distribution, possibly with extreme values, to a bimodal distribution, to a mixture of two or more distributions via the Finite Mixture Model, to modeling the population at the *individual* level via a compound model, which may be viewed as an infinite mixture model. Given a dataset, it is shown how to evaluate the fits of the various models by information criteria. Two datasets are considered in detail, one discrete, the other, continuous.

Keywords: Cluster Analysis, Finite Mixture Model, Bayesian models; Compound models; prior distribution, infinite mixture

AMS subject classification: 62-07, 62F15

1 Introduction

1.1 Background and Summary

Consider the problem of modeling a dataset of numbers accidents in a population of insured persons, or the incidence of an illness in a population. One can consider a spectrum of *levels of granularity* in describing the data and the corresponding population, from histograms, to a single distribution from a parametric family , to a bimodal distribution, to a mixture of two or more distributions, to modeling the population at the *individual* level via compound (predictive) distributions. Examples included are a dataset of employee days ill (a discrete variable) and a dataset of family expenditures on groceries (a continuous variable). Fits to the data obtained by various levels of granularity are compared using the Bayesian Information Criterion.

1.2 Levels of "Granularity"

In discussing the level at which to analyze a dataset, it will be shown how to go from individuals to histograms and modes to clusters or mixture components, back to individuals. The various levels proceed from histograms to clusters to predictive distributions

That is, one can go from **modes** to **sub-populations** back to **individuals.** These choices might be called choosing the level of "granularity" at which to analyze the dataset. Methods for various levels of granularity include those indicated in the table.

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Table 1: Methods for various levels of granularity

Method	parameters	Model
Histograms	different bin widths	Multinomial
Cluster Analysis	parameter values for clusters of individuals	Finite Mixture Model
Predictive distribution	parameter value for each individual	Bayesian prior distribution

There will be two extended analyses of datasets. These datasets are neither new nor large nor high-dimensional but hopefully will still be found to be interesting, particularly from the point of view of this paper. These two datasets are:

- Expenditure in a week on fruits and vegetables for 60 English families (Connor & Morrell, *Statistics in Theory & Practice*)
- Days ill in a year for 50 miners (hypothetical data from Kenkel, *Statistics for Management & Economics*)

2 Days Ill dataset

The first, with discrete data, concerns a (hypothetical) dataset of days ill in a year of n = 50 miners (Kenkel 1984). The days ill are of course integer values. They range from 0 to 18 days in the year.

Table 2: Frequencies of days ill

days	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
freq	2	3	5	5	2	5	5	4	6	3	0	1	4	1	2	0	0	1	1

The histogram (shown here with bins 0, 1-2, 3-4, . . . ,17-18) suggests bimodality, with modes at about 7 days and 12 days. (Strictly speaking, it can be argued that there is an error in this figure; namely, the rectangular bar for the value 0 has the same width as that for 1-2, 3-4, etc., even though these span two values instead of one. As will be discussed below, the width should be chosen so that relative frequency is proportional to area, and area = width $\times$ height, so that the height should be doubled if the bin width is halved. This will be discussed further below.)

The sample mean is about $\bar{x} = 6.6$ days and the sample variance is $s^2 = 19.07,$ that is, the sample standard deviation is $s = 4.37$ days. A single Poisson would not provide a good fit — for a Poisson distribution, the mean and variance are equal, but here the variance is much larger than the mean.

2.1 Extreme values

But, if one does fit a Poisson to the dataset, omitting, say, the upper two values 17 and 18 days, what conclusions might be drawn? Would the 17 and 18 be considered to be particularly unusual? That is to say, later, having fit a distribution, we can assess the probability of such extreme observations.

For now, we note that the mean of the other 50-2 = 48 observations is 5.1 days; the variance, 17.47, still very different values, so a single Poisson would not provide a good fit even after omitting the two largest, possibly outlying, observations. Note that also, besides the gap at 15 and 16

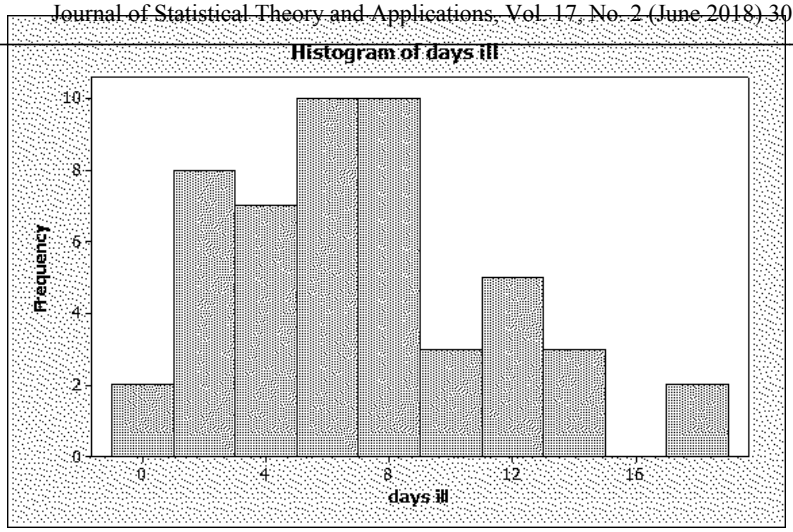

Figure 1: Histogram of days ill

days, there is a gap at 10 days, nobody having been ill that number of days. Perhaps a mixture of distributions would make more sense.

2.2 Poisson mixture model

Consequently, a mixture of two Poissons was fit. The mixture model has p.m.f. (probability mass function) $p(x) = \pi_1\, p_1(x) + \pi_2\, p_2(x)$, where $p_1(\cdot)$ is the p.m.f. of a Poisson distribution with parameter λ_1 and $p_2(\cdot)$ is the p.m.f. of a Poisson distribution with parameter λ_2.

Parameter estimates for the Poisson mixture model. Starting values for iterative estimation were obtained by a Student's t method of clustering into two groups (Sclove 2016). In this method, each cut-point is tried.

In more detail: Obtain the *order statistic*, resulting from ordering the the observations $\{x_1, x_2, \ldots, x_n\}$ as

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x(n).$$

Candidate cut-off points $c_1, c_2, \ldots, c_{n-1}$ are chosen in between the order statistics:

$$x_{(1))} < c_1 < x_{(2)} < c_2 < \cdots < x_{(n-1)} < c_{n-1} < x_{(n)}.$$

For example, one can take c_j to be the midpoint $c_j = (x_{(j)} + x_(j+1))/2$. One then computes two-sample t for each clustering, that is, for each cut-point. The best clustering into two clusters is the one which gives the largest value of $|t|$, or, equivalently, of t^2. This is because t^2 compares the within-groups sum of squares and the between-groups sum of squares.

Pooled t assumes equal variances in the two clusters. Unpooled t avoids this assumption. Not assuming equal variances, one can consider the standardized difference between means as an objective function, that is, one can use an unpooled two-sample t as the criterion, maximizing it over cut points. The unpooled t statistic is given by $t^2 = (\bar{x}_1 - \bar{x}_2)^2 \;/\; (s_1^2/n_1 + s_2^2/n_2)$, that is, $|t| = |\bar{x}_1 - \bar{x}_2| \,/\, \sqrt{s_1^2/n_1 + s_2^2/n_2}$.

The value of the two-sample, unequal variance Student's t is computed for each cut-point, and the cut-point giving the largest t^2 is chosen.

The means of the resulting clusters are taken as tentative estimates of the distributional means; the cluster relative frequencies, as tentative estimates of the mixing probabilities. These estimates were 2.1 days, 8.9 days, with cluster frequencies of 17 out of 50 and 33 out of 50, that is, mixing probabilities of .34, .66. By doing a grid search in the vicinity of these initial estimates to maximize the mixture-model likelihood, the following estimates were obtained: $\hat{\lambda_1} = 2.77$ days, $\hat{\pi_1} = .40$, $\hat{\lambda_2} = 9.12$ days, $\hat{\pi_2} = .60$. Also, taking these as starting points for the EM (Expectation-Maximization) algorithm, only slightly different estimates were obtained; these were $\hat{\lambda_1} = 2.84$ days, $\hat{\lambda_2} = 9.20$ days, $\hat{\pi_1} = .41$, $\hat{\pi_2} = .59$.

Extreme value assessment with the fitted Poisson mixture. With the fitted Poisson mixture, the estimate of $\Pr\{X > 16\}$ is .007, fairly small. So perhaps results of 17 or more days could and should be considered outliers.

2.3 Comparison of models by model-selection criteria

The two fits, by histogram and by Poisson mixture, were compared by means of *model-selection criteria.* Given K alternative models, indexed by $k = 1, 2, \ldots, K$, penalized-likelihood model-selection criteria are smaller-is-better criteria that can be written in the form

$$\text{MSC}_k \;=\; -2\,\text{LL}_k \;+\; a(n)\,m_k,$$

where m_k is the number of free parameters used in fitting Model k, LL_k is the log maximum likelihood of Model k, and $a(n) = \ln n$ for BIC (Bayesian Information Criterion; Schwarz 1978) and $a(n) = 2$ for all n for AIC (Akaike's Information Criterion; Akaike 1973, 1974; Kashyap 1982; Sakamoto 1992). That is, for $k = 1, 2, \ldots, K$ alternative models,

$$\text{AIC}_k \;=\; -2\,\text{LL}_k \;+\; 2\,m_k,$$

and

$$\text{BIC}_k \;=\; -2\,\text{LL}_k \;+\; (\ln n)\,m_k.$$

The number of parameters for the Poisson mixture is 2 means plus 2 mixing probabilities, less 1 because the probabilities must add to 1. That is 3 free parameters for the Poisson mixture. The number of parameters for the histogram, scored by the multinomial distribution with 17 categories (0 through 18, but 15 and 16 are missing), less 1 because the multinomial probabilities must add to 1, leaving 16 free parameters.

The results are in the next table. The histogram wins by a bit according to AIC, but the Poisson mixture wins by a wide margin according to BIC. To see this, note that BIC is derived (Schwarz 1978) as the first terms in the Taylor series expansion of (-2 times) the posterior probability of Model k, $\Pr(\text{Model } k \,|\, \text{data}) = \text{pp}_k$, say. That is,

$$-2 \ln \text{pp}_k \;\approx\; \text{Const.} + \text{BIC}_k, \text{ or } \text{BIC}_k \approx C\,\exp(-\text{BIC}_k/2).$$

Table 3: Comparison of two models

Model, k	- 2 LL_k	m_k	AIC_k	BIC_k	pp_k
$k = 1$: histogram	261.642	16	293.642	324.234	5.0×10^{-7}
$k = 2$: Poisson mixture	283.473	3	289.473	295.209	≈ 1

Table 4: Calculation of posterior probabilties of alternative models

Model, k	BIC_k	same - 295	$\exp(-\text{same}/2)$	pp_k
1	324.234	29.234	4.5×10^{-7}	4.98×10^{-7}
2	295.209	0.209	0.90085	1.000
		sum =	0.90085	

To calculate the posterior probabilities, one subtracts a large constant from each, divides by 2, exponeniates the negative of this, and sums these, dividing by the sum to normalize.

Different bin widths. If one makes several histograms, with different bin widths, how should the likelihood for histograms be computed? Given data points $x_1, x_2, \ldots, x_n$, the likelihood for a given p.m.f. $p(\cdot)$ is

$$L \;=\; \Pi_{i=1}^{n}\, p(x_i).$$

Here $p(x_i)$ is the p.m.f. at the data point x_i. (For continuous data, we would write the p.d.f., $f(x_i)$.) But in the context of histograms what we can take $p(x_i)$ to be?

Denote the number of bins by J. Let the bin width be denoted by h . This is an increment along the x-axis.

Let the bins be indexed by $j,\ j = 1, 2, \ldots, J$. The class limits are $x_0, x_0 + h, x_0 + 2h, \ldots, x_0 + Jh$. The class intervals (bins) are $[x_0, x_0 + h), [x_0 + h, x_0 + 2h), \ldots, [x_0 + (J-1)h, x_0 + Jh)$. In the present application, $x_0 = 0$. Now, let $j(x_i)$ denote the bin containing x_i and $n_{j(x_i)}$ be the frequency in that bin. To approximate $f(x_i)$, motivated by $f(x_1) \approx [F(x_2) - F(x_1)]\,/\,(x_2 - x_1) = [F(x_1 + h) - F(x_1)]/\,[(x_i + h) - x_1] \;=\; [F(x_1 + h) - F(x_1)]/\,h$, write

$$f(x_I) \approx [F(x) - F(x_i)]\,/\,h$$

and

$$\begin{aligned} f(x_i) \;&=\; \text{probability density at } x_i \\ &\approx\; \text{probability in bin containing } x_i \,/\, \text{width of bin} \\ &=\; [n_{j(x_i)}\,/n]/h \\ &=\; n_{j(x_i)}/nh. \end{aligned}$$

That is, the concept is that probability density is probability per unit length along the x axis. Thus the likelihood is

$$L \;=\; \Pi_{i=1)}^{n} p(x_i) \;=\; \Pi_{i=1}^{n}\,(n_{j(x_i)}/nh) \;=\; (1/h^n)\,\Pi_{i=1}^{n}\,(n_{j(x_i)}/n).$$

Note that $p(x_1, x_2, \ldots, x_n) \;=\; \Pi_{i=1}^{n}\, p_{j(x_i)} \;=\; \Pi_{j=1}^{J}\, p_j^{n_j}$ is a **multinomial** p.m.f. with probabilities p_j and frequencies n_j for the J categories. The maximized likelihood L is this multinomial (with p_j estimated as n_j/n), divided by h^n, which may be viewed as an adjustment to the likelihood due to the bin width h. In computing the likelihood, the probability *density* is to be used, where "density" is probability / bin width. Note that with a continuous variable we would compute probability density as $f(x_i)/h$, that is, $f(x_i)/$(Lebesgue measure of the bin interval), whereas with a discrete variable we are really computing probability density as $p(x_i)/h$, where now h is the *counting measure* of the bin interval.

g bin widths. In the case of non-constant bin widths, with a bin-width of h_j for the j-th interval, take the probability density at x_i to be $n_{j(i)}\,/h_{j(i)})$, where $h_{j(i)}$ is the width of the interval in which x_i falls and $n_{j(i)}$ (short for $n_{j(x_i)}$) is the frequency (count) in that interval. The likelihood is

$$L \;=\; \Pi_{i=1}^{n}\, p(x_i) \;=\; \Pi_{i=1}^{n}\, (n_{j(i)}\,/\,n)\,/\,h_{j(i)} \;=\; (1/n^n)\,\Pi_{i=1}^{n}\, n_j/\,h_{j(i)}.$$

Table 5: Sample distribution with a bin width of 2

days	0 -1	2-3	4-5	6-7	8- 9	10- 11	12- 13	14- 15	16-17	18-19
freq	5	10	7	9	9	1	5	2	1	1

Table 6: Sample distribution with varying bin widths: bins 0, 1, 2-3,4-5.6-7, . . . , 16-17, 18

days	0	1	2-3	4-5	6-7	8- 9	10- 11	12- 13	14- 15	16-17	18
bin width	1	1	2	2	2	2	2	2	2	2	1
freq	2	3	10	7	9	9	1	5	2	1	1

Table 7: Comparison of models

Model, k	- 2 LL_k	m_k	AIC_k	BIC_k	pp_k
histogram, bin width h=1	261.6	16	293.6	324.2	.000
histogram, bin width h=2	273.2	9	291.2	308.4	.001
histogram, varying bin widths	267.8	9	285.8	303.0	.020
Poisson mixture	283.5	3	289.5	295.2	.978

According to AIC, the histogram with varying bin widths wins, the Poisson mixture coming in second. According to BIC (and, equivalentlly, posterior probability), the Poisson mixture scores the best, by far. But the point is not just which model wins, but that such a comparison, comparing histograms on the one hand with fitted distributions on the other, can be made.

Levels of granularity, cont'd. Perhaps another level of granularity is approached by *predictive distributions,* which may be viewed as getting to the individual level of granularity. Predictive distributions may be viewed in the light of compound distributions resulting from a prior distribution on the parameter at the individual level. From the viewpoint of modern statistics, a predictive distribution is merely the marginal distribution of the observable random variable, having integrated out the prior on the parameter. (Details to follow.)

The Yule-Greenwood model approaches modeling at the individual level, stating that each individual may have his or her own accident rate λ and so is an example of a *compound model.* In terms of granularity, the Yule-Greenwood model is a classical example at the level of the individual in that it employs a Poisson model for each individual's accident rate λ and then puts a (Gamma) distribution over the population of values of λ. The model is the Gamma-Poisson model (sometimes called the Poisson-Gamma model) and is a prime example of a **compound model.** The Gamma is a **conjugate** prior distribution for the Poisson, meaning that the posterior distribution of λ is also a member of the Gamma family. We discuss this further below; first, however, we fit histograms and mixtures to a dataset with continuous data.

Example with continuous data

The variable in the next example is expenditure in a week (£) of $n = 60$ English families on fruits and vegetables (Connor and Morrell 1977, data from the British Institute of Cost and Management Accountants). The data are reported to two decimals. This sort of measurement, treated as continuous, contrasts with the integer-valued variable considered in the example above.

Here are the data, sorted from smallest to largest:

0.21, 0.33, 0.36, 0.38, 0.41, 0.46, 0.48, 0.48, 0.51, 0.51, 0.51, 0.58, 0.64, 0.66, 0.69, 0.69, 0.71, 0.74, 0.74, 0.78, 0.78, 0.79, 0.84, 0.87, 0.87, 0.88, 0.89, 0.91, 0.91, 0.93, 0.98, 0.98, 1.03, 1.03, 1.05, 1.08, 1.12, 1.16, 1.17, 1.19, 1.24, 1.25, 1.26, 1.26, 1.28, 1.33, 1.38, 1.44, 1.48, 1.51, 1.53, 1.58, 1.61, 1.62, 1.76, 1.78, 1.79, 1.83, 1.96, 2.13

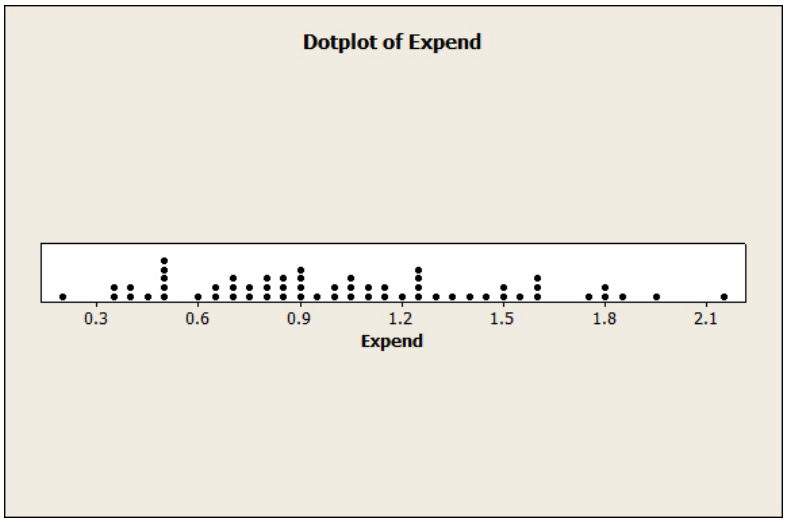

Figure 2: Dotplot of Expenditure

The minimum is 0.21 £; the maximum, 2.13 £. The sample mean is $\bar{x} =$ 1.022 £, the sample standard deviation, $s = 0.4562$ £(sample variance $s^2 = 0.2081$). The frequency distribution (see below in Figure 2 and Table 7) suggests possible bimodality.

Table 8: Frequency distribution of weekly expenditure (£)

lower limit	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91	1.01	1.11
upper limit	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20
Frequency	1	3	4	4	4	6	5	5	4	4
lower limit	1.21	1.31	1.41	1.51	1.61	1.71	1.81	1.91	2.01	2.11
upper limit	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00	2.10	2.20
Frequency	5	2	2	3	2	3	1	1	0	1

The distribution as tabulated here has a bin width h of 0.10. We consider below also the results for $h = 0.2$, for fitting a single Gamma and also for fitting a mixture of two (Gaussian) distributions. Below we compare these four fits by means of AIC and BIC.

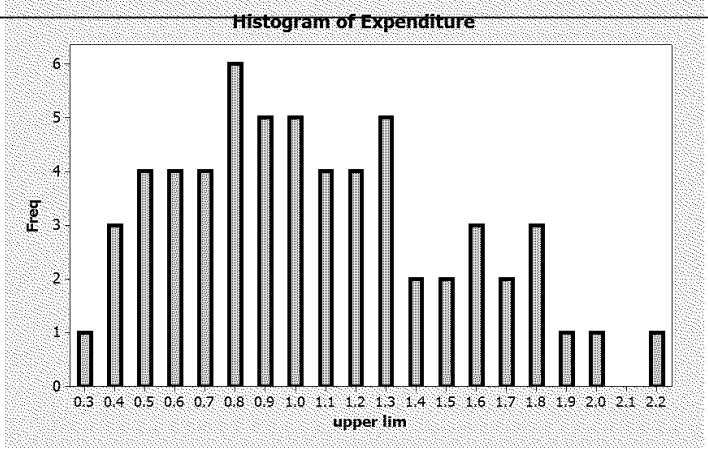

Figure 3: Histogram of Expenditure

3.1 Fitting a Gamma distribution

The two-parameter Gamma p.d.f. is

$$f(x) \;=\; \lambda^{m-1}e^{-x/\beta}/\left[\Gamma(m)\,\beta^m\right],\; m>0, \lambda>0, \beta>0, x>0.$$

The mean is $m\beta$. The variance is $m\beta^2$. Method-of-moments estimates are, for the scale parameter $\beta = \sigma^2/\mu,\; \hat{\beta} = \; s^2/\bar{x} = 0.2081/1.022 = 0.2035.$ and for the shape parameter $m = \mu/\beta,$ so $\hat{m} = \bar{x}/\hat{\beta} = \; 1.022/0.2035 = 5.0246.$

3.2 Fitting mixture models

3.2.1 Gaussian mixture

The mixture model has p.d.f. $f(x) \;=\; \pi_1\, f_1(x) \;+\; \pi_2\, f_2(x)$, where $f_1(\cdot)$ is the p.d.f. of a Gaussian distribution with mean μ_1 and variance σ_1^2 and $f_2(\cdot)$ is the p.d.f. of a Gaussian with mean μ_2 and variance σ_2^2. The estimates are $\hat{\mu_1} \;=\; 0.72£,\; \hat{\mu_2} \;=\; 1.46£,\; \hat{\sigma_1} \;=\; 0.23£,\; \hat{\sigma_2} \;=\; 0.27£,\; \hat{\pi_1} \;=\; .62,\; \hat{\pi_2} = .38.$ The results were obtained by approximate maximization of the likelihood doing an EM (Expectation-Maximization) iteration. Starting values were obtained by the Student's t method (Sclove 2016). As mentioned above, in this method, each cut-point is tried. Two-sample, unequal variance Student's t is computed for each cut-point, and the cut-point giving the largest t is chosen. This gives starting values obtained from the means and variances of the two resulting clusters. Then, given starting values $\mu_1(0), \sigma_1^{(0)},\; \mu_2^{(0)}, \sigma_2^{(0)}$, one then computes $f_j(x_i; \mu_j^{(0)}, \sigma_j^{2\,(0)}),\; j = 1, 2,\; i = 1, 2, \ldots, n,$ and posterior probabilities of group membership, $pp(j \mid x_i),$ the posterior probability that x_i arose from population j; at step s, for populations $j = 1, 2,$

$$pp^{(s)}(j \mid x_i) = \pi_j^{(s)}\, f_j(x_i; \mu_j^{(s)}, \sigma_j^{2\,(s)}) / \;\left[\pi_1^{(s)} f_1(x_i; \mu_j^{(s)}, \sigma_j^{2\,(s)}) \;+\; \pi_2^{(s)} f_2(x_i; \mu_j^{(s)}, \sigma_j^{2\,(s)})\right].$$

Then the estimates are updated. (See McLachlan and Peel (2000), p. 82.) For $j = 1, 2,$

$$\mu_j^{(s+1)} \;=\; \Sigma_{i=1}^n\, pp^{(s)}(j|x_i)\, x_i \,/\, \Sigma_{i=1}^n\, pp^{(s)}(j|x_i).$$

As a check, note that if for one group j, it happened that $pp^{(s)}(j|x_i) = 1$ for all cases $i,$ then the new estimate of the mean for that j is simply $\Sigma_{i=1}^n\, x_i/n = \bar{x}$. The updated estimates of the second moments are, for $j = 1, 2,$

$$\mu_{2j}^{(s+1)} \;=\; \Sigma_{i=1}^n\, pp^{(s)}(j|x_i)\, x_i^2 \,/\, \Sigma_{i=1}^n\, pp^{(s)}(j|x_i).$$

ance equals raw second moment minus square of mean, or

$$\sigma_j^{2\,(s+1)} \;=\; \mu_{2j}^{(s+1)} \;-[\mu_j^{(s+1)}]^2.$$

Note that the numerical estimates of σ_1 and σ_2 are somewhat different; the ratio of variances is $(0.27/0.23)^2 = 0.075/0.051 = 1.46$. Given this, it does not seem particularly worthwhile in this case to trying fitting two Gaussians with equal variances.

The table summarizes the results. According to AIC, the ranking is: Gamma, Gaussian mixture; histogram with bin width .2, histogram with bin width .1. According to BIC, the ranking is: Gaussian mixture, Gamma, histogram with bin width .2, histogram with bin width .1.

Table 9: Comparison of results

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
histogram, bin width h=0.1	61.68	18	97.68	135.38	.000
histogram, bin width h=0.2	66.25	9	84.25	103.10	.000
Gamma	71.75	2	75.75	79.94	.382
Gaussian mixture	70.79	5	80.79	78.98	.618

3.2.2 Gamma mixture

The mixture model has p.d.f. $f(x) \;=\; \pi_1\, f_1(x) \;+\; \pi_2\, f_2(x)$, where now $f_1(\cdot)$ is the p.d.f. of a Gamma distribution with shape parameter m_1 and scale parameter β_1 and $f_2(\cdot)$ is the p.d.f. of a Gamma distribution with shape parameter m_2 and scale parameter β_2. The means and variances are, for distributions $j = 1, 2,\ \mu_j \;=\; m_j\beta_j$ and $\sigma_j^2 \;=\; m_j\,/\,\beta_j^2$. The inverse expressions, for the Gamma parameters in terms of the mean and variance, are $\beta_j = \;\sigma_j^2/\mu_j$ and $m_j = \mu_j^2/\sigma_j^2$.

The resulting estimates are $\hat{m_1} \;=\; 7.459,\, \hat{\beta}_1 \;=\; 0.099,\;\; \hat{m_2} \;=\; 22.228, \hat{\beta}_2 \;=\; 0.066,\;\; \hat{\pi_1} \;=\; .61, \hat{\pi_2} = .39.$

The results were obtained by approximate maximization of the likelihood doing an EM (Expectation-Maximization) iteration. Starting values were obtained by the Student's t method (Sclove 2016). Given starting values $\mu_1(0), \sigma_1^{(0)},\, \mu_2^{(0)}, \sigma_2^{(0)}$, these were converted to starting values for $m_1, m_2, \beta_1, \beta_2$. One then computes the values of the Gamma densities,

$$f_j(x_i;\, m_j^{(0)}, \beta_j^{(0)}),\; j = 1, 2,\; i = 1, 2, \ldots, n,$$

and posterior probabilities of group membership, $pp(j \,|\, x_i),$ the posterior probability that x_i arose from population j; at step s, for populations $j = 1, 2,$

$$pp^{(s)}(j \,|\, x_i) = \pi_j^{(s)}\, f_j(x_i; m_j^{(s)}, \beta_j^{(s)})/\; [\pi_1^{(s)} f_1(x_i;\, m_j^{(s)}, \beta_j^{(s)}) \;+\; \pi_2^{(s)} f_2(x_i; m_j^{(s)}, \beta_j^{(s)})].$$

Then the estimates are updated. For $j = 1, 2,$

$$\mu_j^{(s+1)} \;=\; \Sigma_{i=1}^n\, pp^{(s)}(j|x_i)\, x_i \,/\, \Sigma_{i=1}^n\, pp^{(s)}(j|x_i).$$

As a check, note that if for one group j, it happened that $pp^{(s)}(j|x_i) = 1$ for all cases i, then the new estimate of the mean for that j is simply $\Sigma_{i=1}^n\, x_i/n = \bar{x}$. The updated estimates of the second moments are, for $j = 1, 2,$

$$\mu_{2j}^{(s+1)} \;=\; \Sigma_{i=1}^n\, pp^{(s)}(j|x_i)\, x_i^2 \,/\, \Sigma_{i=1}^n\, pp^{(s)}(j|x_i).$$

Of course, variance equals raw second moment minus square of mean, or

$$\sigma_j^{2\,(s+1)} \;=\; \mu_{2j}^{(s+1)} \;-\; [\mu_j^{(s+1)}]^2.$$

Then these are converted to updated estimates of m_j, β_j, and the iteration continues until satisfactory convergence. LL, AIC, and BIC are computed with the resulting estimates.

3.3 Comparison of models by model-selection criteria

The table summarizes the results. According to AIC, the ranking is: Gamma; Gaussian mixture; Gamma mixture; histogram with bin width .2, histogram with bin width .1. According to BIC, the ranking is: Gaussian mixture, Gamma, Gamma mixture, histogram with bin width .2, histogram with bin width .1 The Gamma mixture has only a small posterior probability.

Table 10: Comparison of results

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
histogram, bin width h=0.1	61.68	18	97.68	135.38	.000
histogram, bin width h=0.2	66.25	9	84.25	103.10	.000
Gamma	71.75	2	75.75	79.94	.382
Gaussian mixture	70.79	5	80.79	78.98	.617
Gamma mixture	71.75	5	81.75	92.22	.001

4 Compound Models in General

First, notation notation for probability functions will be reviewed.

The probability density function (p.d.f.) of a continuous random variable (r.v.) X, evaluated at x, will be denoted by $f_X(x)$. The p.d.f. of a continuous random variable Y, evaluated at y, is similarly denoted by $f_Y(y)$.

Now consider a bivariate variable $\boldsymbol{x} = (y,\, z)$. The joint p.d.f. of the r.v.s Y and Z, evaluated at (y, z) is $f_{Y,Z}(y, z)$. Example: $Y = WT, X = HT,$ the value of the joint p.d.f. at y = 80 kg and z = 170 cm is $f_{WT,HT}(80, 170)$.

Other notations include:
$f_{Y|X}(y|x)$: **conditional** probability density function of the r.v. Y, given that the value of the r.v. X is x. Example: $f_{WT|HT}(\text{wt}\,|HT = 170\text{cm})$. This represents the bell-shaped curve of weights for men of height 170 cm.

$f_{Y,Z}(y\,|\,z) \;=\; f_{Y|Z}(y|z)\, f_Z(z)$: This is the joint p.d.f. expressed as the product of the conditional of Y given Z and the marginal of Z

$f_Y(y) \;=\; \int f_{Y,Z}(y, z)\, dz \;=\; \int f_{Y|Z}(y|z)\, f_Z(z)\, dz$: marginal p.d.f. of Y

In the development that follows, $f_Z(z)$ plays the role of the prior probability function on the parameter. That is, denoting the parameter by θ, the function $f_Z(z)$ will become $f_\Theta(\theta)$.

The elements of **compound models** are:

distribution of the observable r.v., given the parameter(s), that is, the conditional distribution of X, given the parameter(s); and the marginal distribution (predictive distribution). The marginal distribution will have the hyperparameters among its parameters.

- the prior distribution. Its parameters are called *hyperparameters.*

A generic symbol for the parameter(s) of the conditional distribution of X is the conventional $\boldsymbol{\theta}$. As a generic symbol for the hyperparameters, one could use $\boldsymbol{\alpha}$, since the prior comes first in the model when one thinks of the parameter value being given first, and then the value of the variable being observed.

For use in compound models, the probability functions include the following:

The conditional distribution of the observable r.v. X, given the value of the parameter, is $f_{X|\Theta}(x \,|\, \theta)$; p.d.f. of X for given θ.

The prior distribution on the parameter $\boldsymbol{\theta}$ with hyperparameter vector $\boldsymbol{\alpha}$, $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}; \boldsymbol{\alpha})$.

The naming of compound models takes the form, prior distribution – conditional distribution. In the Gamma-Poisson model, the conditional distribution of X given λ is Poisson(λ) and the prior distribution on λ is Gamma. In the Beta-Binomial mode, the conditional distribution of X given p is Binomial with success probability p and the prior distribution on p is Beta. (Note that some people use the form, conditional distribution – prior distribution, *e.g.,* Poisson-Gamma.)

5 The Gamma-Poisson model

5.1 Probability functions for the Gamma-Poisson model

In the Gamma-Poisson model, the distribution of X is Poisson with parameter usually called λ. The p.m.f. is

$$p(k) \;=\; e^{-\lambda}\, \lambda^k \,/\, k!,\; \lambda > 0,\; k = 0, 1, 2, \ldots \,.$$

The mean and variance are both equal to λ.

Such a distribution can be considered, say, for the number of accidents per individual per year. For the days ill dataset (days ill in a year for a sample of n = 50 miners), we have fit a single Poisson (with mean 6.58 days per year). We looked at histograms and observed bimodality. Further, the fact that the sample variance of 19.06 was considerably larger than the sample mean was a hint of inadequacy of a single Poisson. As discussed above, a mixture of Poissons was fitted, with mixing probabilities about .6 and .4 and means about 3 days and 9 days. A finer level of granularity would be obtained by saying that *each person* has his own value of λ and putting a distributon on these over the population.

5.2 Gamma family of distributions

A **Gamma** distribution could be a good choice. The choice of the Gamma family is non-restrictive in that the family can achieve a wide variety of shapes. The single-parameter gamma has a shape parameter m; the two-parameter gamma family has, in addition, a scale parameter, β. (The reciprocal of β is the **rate** parameter, so-called because it is the rate, or intensity, of the associated Poisson process.) A Gamma distribution with parameter m, has p.d.f. $\;f(\lambda) \;=\; \text{Const.}\, \lambda^{m-1}\, e^{-\lambda},\; \lambda > 0.$ The constant is $1/\Gamma(m)$. More generally, the two-parameter Gamma can be used: the p.d.f. is

$$f(\lambda) \;=\; \frac{\lambda^{m-1} e^{-\lambda/\beta}}{\Gamma(m)\, \beta^m},\; \beta > 0,\; \lambda > 0.$$

is $m\beta$. The variance is $m\beta^2$.

The Negative Exponential family of distributions. The special case of $m = 1$ in the Gamma family gives the negative exponential family of distributions. So the

$$f(\lambda) \;=\; e^{-\lambda/\beta}/\,\beta,\; \lambda > 0.$$

The mean is β. The variance is β^2.

5.3 Development of the Gamma-Poisson model

Putting a population distribution over a parameter can be a very helpful way of modeling. The resulting model is called a **compound model.** In a compound model, the random variable X is considered as the result of sampling that yields an individual and that individual's value of a parameter, and then the individual's value of X is observed, from a distribution with that value of the parameter. Note that a compound model can be viewed as an infinite mixture model.

In this discussion, focus is on a couple of particular compound models, the Gamma-Poisson, and later, the Beta-Binomial.

The Yule-Greenwood model, from a modern viewpoint, is an application of the Gamma-Poisson model to a financial, in fact, actuarial, situation. It is in terms of a model for *accident rates* in a population. Suppose that the yearly number of accidents of any given individual i in a population is distributed according to a Poisson distribution with parameter λ_i accidents per year. (This is *count data,* similar to the days ill data.) Then the probability that individual i, with parameter value λ_i, has exactly k accidents in a year, k = 0, 1, 2, . . . , is

$$e^{-\lambda_i}\,\lambda_i^k\,/k!,\;\lambda_i > 0,\; k = 0, 1, 2, \ldots\ .$$

Some individuals are more accident prone (have a higher accident rate) than others, so different individuals have different values of λ. A distribution can be put on λ to deal with this. This is the Yule-Greenwood model, dating from 1920; a precursor of the Predictive Distributions of the new Predictive Analytics, predating even Abraham Wald (1950) as a founder of modern mathematical statistics and decision theory and Jimmie Savage (1954) as a founder of modern Bayesian Statistics.

The standard choice of a prior distribution on λ is a Gamma distribution. The Gamma family is a *conjugate* family to the Poisson, meaning that the prior and posterior distributions of λ are both in the Gamma family.

The joint distribution of X and Λ. The joint probability function of X and Λ is

$$f_{X,\Lambda}(x, \lambda) = f_\Lambda(\lambda)\, p_{X|\Lambda}(x\,|\lambda),\; x = 0, 1, 2, \ldots,\; \lambda > 0.$$

The expressions for the Gamma and Poisson are put into this. That is, the weight assigned to $p_{X\,||\Lambda}(x\,|\,\lambda)$ is $f_\Lambda(\lambda)$.

The joint probability function is used to obtain

- the marginal distribution of X, by integrating out λ, and
- then the posterior distribution of Λ given x, by dividing the joint probability function by the marginal probability mass function of X.

Using the expressions for the Gamma and Poisson, it is seen that the marginal distribution of X, the number of accidents that a randomly selected individual has in a year, is of the form

$$f_X(x) \;=\; \int_0^\infty f_{X,\Lambda}(x,\lambda)\,d\lambda \;=\; \int f_{X|\Lambda}(x|\lambda)\,f_\Lambda(\lambda)\,d\lambda.$$

When the prior is Gamma and the conditional is Poisson, this marginal distribution can be shown to be *negative binomial.* Its parameters are m and $p = 1/(1+\beta)$.

In the Bayesian model, the parameter of the conditional distribution of X, say θ, is treated as a random variable Θ.

In the Gamma-Poisson model, θ is the Poisson parameter λ.

The conditional distribution of X given that $\Theta \;=\; \theta$ is Poisson(λ). The probability mass function is

$$p_{X|\Lambda}(x;\lambda) \;=\; e^{-\lambda}\,\lambda^x\,/\,x!,\; x = 0, 1, 2, \ldots, .$$

The joint p.d.f. of X and Λ can be written as $p_{X\,|\,|\Lambda}(x\,|\,\lambda)$ which is $f_\Lambda(\lambda)(x,\lambda) = \; f_\Lambda\, p_{X|\Lambda}(x|\lambda.$

As mentioned above, from this, the posterior distribution of Λ, that is, the distribution of Λ given x, can be computed, and the marginal distribution of X can be computed. Marginal maximum likelihood estimation can be applied to obtain estimates of the remaining parameters.

Posterior distribution of Λ. Analogous to $\Pr(B|A) \;=\; \Pr(A\cap B)/\Pr(A)$, the p.d.f. of the posterior distribution is the joint p.d.f. , divided by the marginal p.d.f. of X :

$$f_{\Lambda|X}(\lambda|x) \;=\; f_{X,\Lambda}(x,\lambda)\,/\,f_X(x).$$

This will turn out to be a Gamma distribution, that is, it is in the same family as the prior. The Gamma is a *conjugate prior* for the Poisson.

Marginal distribution of X. In **Predictive Analytics,** the marginal distribution of X is computed as a model of a future observation or observations of X.

The *marginal distribution* of X is obtained by integrating the joint distribution with respect to the parameter. Note that this computation combines information, by weighting the conditional distribution of X given λ with the prior on λ. This computation of the p.d.f. is, as stated above, $f_X(x) = \; \int_0^\infty t\, f(x|\,\lambda)\, f_\Lambda(\lambda)\,d\lambda$.

Moments. The mean of the marginal distribution of X is $mq/p \;=\; m\,\beta$. The variance of the marginal distribution of X is $mq/p^2 = m\,\beta(1+\beta)$.

5.4 Empirical Bayes estimation

Empirical Bayes estimation, at least in the present context, means estimating the parameters of the prior using observations from the marginal distribution.

The hyperparameters in terms of the moments of the marginal. The parameters of the prior are called *hyperparameters.* In this case, they are λ and β. Suppose we solve for them in terms of the first two moments of the marginal.

Estimating the prior parameters from the marginal. Estimates of the prior parameters m and β can be obtained by, for example, taking the expressions for the hyperparameters m and β in terms of the first two raw moments and plugging in estimates m_1' and m_2'. Given a sample $X_1, X_2, \ldots, X_n$, we have $m_1' = \bar{X} = \; \sum_{i=1}^N \; X_i\,/\,N$ and $m_2' \;=\; \sum_{i=1}^n \; X_i^2\,/\,n$.

Application to the days ill dataset

Returning to the days ill dataset for n = 50 miners, the days ill in a year ranged from 0 to 18; the distribution seems to be bimodal.

The p.m.f. of the Negative Binomial distribution with parameters m and p is where k is the number of trials in excess of m required to get m Heads. In the Gamma-Poisson model, the marginal distribution of X is Negative Binomial with parameters with parameters m and $p = 1/(1+\beta)$.

Given that the true mean of the marginal ("predictive distribution") Negative Binomial is $\mu = mq/p = m\beta$ and the true variance is $\sigma^2 = mq/p^2 = m\beta(1+\beta)$, and the sample mean $\bar{x} = 6.58$ and the sample variance $s^2 = 19.07,$ one can set up two equations and solve for method of moments estimates of the hyperparameters m and β in the Gamma prior for λ.
The equations are $[1] : m\,\beta = 6.58;\ [2] :\ m\,\beta(1+\beta) = 19.07.$
Putting [1] in [2] gives $6.58(1+\beta) = 19.07,\ 1+\beta = 19.07/6.58 \approx 2.898,\ \hat{beta} \approx 1.898$. Then $m \approx 6.58/\beta = 6.58/1.898 \approx 3.467$. Now, $\mu = m(1-p)/p = m/p - m,\ \mu + m = m/p,\ p = m/(\mu+m)$ or, estimating $p = 3.467/(6.58+3.467) = 3.467/10.05 = 0.345$. So now we have estimates of the hyperparameters.

To estimate the mean and variance of the Gamma prior, one can proceed as follows. The mean of the prior is $m\beta$, estimated as 6.58 days ill per year. The variance of the prior is $m\beta^2$, estimated as $6.58(1.898) \approx 12.49$. The standard deviation is thus estimated as $\sqrt{12.49} \approx 3.03$ days ill per year.

Maximum likelihood estimates are not in closed form but numerical values for them could be obtained by numerical maximization of the likelihood function. It is helpful to use the method of moments as a quick and simple method to get an idea of the values of the parameters.

The table includes the marginal negative binomial with $m = 3$ and $p = .344$ in the comparison.

Table 11: Comparison of models, cont'd

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
histogram, bin width h=1	261.6	16	293.6	324.2	.000
histogram, bin width h=2	273.2	9	291.2	308.4	.000
histogram, varying bin widths	267.8	9	285.8	303.0	.002
Poisson mixture	283.5	3	289.5	295.2	.100
marginal Negative Binomial	283.0	2	286.0	290.8	.898

According to AIC, the histogram with varying bin widths still wins, the Negative Binomial coming in second. According to BIC (and, equivalentlly, posterior probability), the Negative Binomial scores the best, by far. This Negative Binomial is unimodal with a mode of .115 at 3 days. Because it is unimodal, it perhaps does not capture the flavor of the original data, which is reflected better by the Poisson mixture.

6.1 Beta-Binomial model

Another compound model is the Beta-Binomial model.In this model, the conditional distribution of X given p is Binomial(n,p). The prior on p is Beta(α, β). The posterior distribution of p given x is Beta($\alpha + x$, $\beta + n - x$). It is as if there had been a first round of $\alpha + \beta$ trials, with α successes, followed by a second round of n trials, with x successes. Method of Moments estimates of the parameters of the prior can be relatively easilty obtained. So can the Bayes estimates.

6.2 Normal-Normal model

We have considered the Gamma-Poisson model and, briefly, another prominent compound model, the Beta-Binomial model, with a Beta prior on the Binomial success probability parameters. Still another compound model is the Normal-Normal model.

In the Normal-Normal model, X is distributed according to $\mathcal{N}(\mu, \sigma^2)$, the Gaussian distribution with mean μ and variance σ^2. The prior on μ can be taken to be $\mathcal{N}(\mu_0,\, \sigma_0^2)$, or perhaps a Gaussian with a different mean if there is some particular reason to do this.

The posterior distribution is again Normal. That is, the Normal family is the conjugate family for the Normal distribution. The marginal distribution is also Normal, with mean $\mathcal{E}[X] = \mathcal{E}[\mathcal{E}[X|\mu] = \mathcal{E}[\mu] = \mu_0$. The variance of the marginal distribution is the mean of the conditional variance plus the variance of the conditional mean, $\mathcal{V}[X] = \mathcal{E}[\sigma^2] + \mathcal{V}[\mathcal{E}[\mu] = \sigma^2 + \mathcal{V}[\mu] = \sigma^2 + \sigma_0^2$. These two terms are the "components of variance". The decomposition of the variance can be obtained also by doing the requisite algebra on the product of the prior and conditional. This model is similar to a Random Effects (Model II) model in ANOVA. The group effects (group mean minus overall mean) are considered as a sample from $\mathcal{N}(0,\, \sigma_0^2)$.

7 Proposed Extensions and Issues

Multivariate models. Multivariate generalizations, where the response is a vector rather than a scalar, could be interesting, both in general and in the context of the analysis of variance.

Parameter-space boundary issues. Chen and Szroeter (2016) provide a more widely applicable version of BIC (originally called the Schwarz Information Criterion, SIC). This version deals with boundary issues in the parameter space. These issues can be important when dealing with mixture models, as the parameter space changes as the number of component distributions changes..

A BIC for Maximum Marginal Likelihood. In view of the fact that, in this paper, compound models have been mentioned as a level of granularity for modeling at the individual level, it is worth mentioning that Spiegelhalter *et al.* (2002) use an information theoretic argument to derive a measure p_D for the effective number of parameters in a model as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. In general this measure p_D approximately corresponds to the trace of the product of Fisher's information and the posterior covariance, which in normal models is the trace of the 'hat' matrix projecting observations onto fitted values, that is, the matrix of estimated regression coefficients. The properties of the measure in exponential families are explored in their paper. The posterior mean deviance is suggested as a Bayesian measure of fit or adequacy, and the contributions of individual observations to the fit and complexity can give rise to a diagnostic plot of deviance residuals against leverages. Adding p_D to the posterior mean deviance gives a deviance information criterion for comparing models, which is related to other information criteria and has an approximate decision theoretic justification.

Acknowledgments

An earlier version of this paper was presented at the 2016 Annual Meeting of the Classification Society, June 1-4, 2016, University of Missouri, and also at ICOSDA 2016, the 2016 International Conference on Statistical Distributions and Applications, October 14-16, 2016, Niagara Falls, Ontario, Canada, organized by Professor Ejaz Ahmed of Brock University and Professors Felix Famoye and Carl Lee of Central Michigan University. This paper was presented in the topic-invited session "Extreme Value Distributions and Models", organized by Professor Mei Luang Huang, Brock University; sincere thanks are extended to Professor Huang for the invitation to speak at the conference.

Bibliography. The books mentioned below on predictive analytics, those of Murphy and Bishop, do not discuss the Gamma-Poisson model explicitly. Those who wish to consult these books may however refer to them to find Murphy, p. 41 on the Gamma family of dsitributions and/or Bishop, p. 688 on the Gamma family. As mentioned, an original paper, anticipating the subject of compound models, is that of Major Greenwood and G. Udny Yule (1920). To review background in Probability Theory in general, see, for example, Emanuel Parzen (1992) or Sheldon Ross (2014). See also Parzen, *Stochastic Processes* (1962) or Ross, *Applied Probability Models* (1970).

References

[1] H. Akaike (1973). "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Càski, Budapest: Akademiai Kiàdo, pp. 267–281.

[2] H. Akaike (1974). "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control,* **19** (6): 716–723.

[3] C. M. Bishop (2006). *Pattern Recognition and Machine Learning.* Springer.

[4] L.-Y. Chen and and J. Szroeter (2016). Extension of the Schwarz Information Criterion for models sharing parameter boundaries. *Journal of Statistical Planning and Inference,* **Vol. 17, No. 4,** 68–84.

[5] L. R. Connor and A. J. H. Morrell (1977). *Statistics in Theory and Practice. 7th ed.* London: Pitman.

[6] M. Greenwood and G. U. Yule (1920). "An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents". *Journal of the Royal Statistical Societ,* **83:** 255–279.

[7] R. Kashyap (1982). "Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Volume:PAMI-4, Issue: 2), 99–104.

[8] J. Kenkel (1984) *Introductory Statistics for Management and Economics. 2nd ed.* Duxbury Press, Boston, MA. Exercise 4, p. 31.

[9] G. J. McLachlan and D. Peel (2000). *Finite Mixture Models.* John Wiley and Sons, New York.

[10] K. P. Murphy (2012). *Machine Learning: a Probabilistic Perspective.* The MIT Press.

[11] E. Parzen (1960). *Modern Probability Theory and its Applications.* Wiley. (Reprinted in 1992 as a Wiley Classics Edition.)

[12] E. Parzen (1962). *Stochastic Processes.* Wiley, New York. Dover Publications (Reprint of the original, published by Wiley.)

[13] S. M. Ross (1970). *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco. Reprinted, Dover, New York, 1992.

[14] S. M. Ross (2014). *Introduction to Probability Models. 11th ed.* Elsevier, Amsterdam, The Netherlands; Waltham, MA; San Diego, CA.

[15] Y. Sakamoto (1992). *Categorical Data Analysis by AIC.* Springer, New York.

[16] L. J. Savage (1954). *The Foundations of Statistics.* John Wiley and Sons, New York.

[17] G. Schwarz (1978). "Estimating the Dimension of a Model". *Annals of Statistics,* **6,** 461-464.

[18] S. L. Sclove (2016). "t for Two (Clusters)". Presented June 2, 2016, at the Classification Society Annual Meeting, University of Missouri, Columbia. Submitted for consideration for publication.

[19] D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B,* **64** , 583–616.

[20] A. Wald (1950). *Statistical Decision Functions*. John Wiley and Sons, New York; Chapman and Hall, London.