CHAPTER 7

CONCLUSION AND FURTHER WORK

The $\text{GIT}_{3,1}$ distribution has a stochastic origin as a first-passage time distribution of a modified random walk on the half plane and is seen to be a model that is able to cater for under-, equi- and over-dispersion. Compared to the other models (GPD, COM-Poisson, double Poisson and so on) currently available with such capability, the $\text{GIT}_{3,1}$ distribution has an interesting feature in that it reduces to a non-Poisson distribution for the particular case of equi-dispersion. However it does have the Poisson distribution as a limiting distribution. Furthermore the $\text{GIT}_{3,1}$ distribution, unlike the GPD, does not encounter problem with range of the parameters for it to be a legitimate distribution.

Although the COM-Poisson distribution has a simple pmf, computation of the normalizing constant could be difficult for extreme values of the parameters. In Chapter 5, the accuracy of the infinite sum, $Z(\lambda, v)$ is studied. It is found that the formula obtained through a direct differentiation of $Z(\lambda, v)$ is favoured if higher accuracy is required.

Even though the $\text{GIT}_{3,1}$ distribution has a complicated pmf in terms of the Gauss hypergeometric function, its pmf has a simple three-term recurrence formula to facilitate computation. Furthermore the $\text{GIT}_{3,1}$ distribution has a simple probability generating function which allows parameter estimation by an alternative procedure, the pgf-based minimum Hellinger-type distance estimation (Sim and Ong, 2010), which is much simpler than maximum likelihood estimation. For the four count frequency data sets, the pgf-based estimation method is seen to be consistently as good as or better than MLE or MHD.

Apart from the considerations above, the good fits shown by $\mathrm{GIT}_{3,1}$ distribution relative to the well-known GPD and COM-Poisson distribution justify its inclusion by data analysts as a viable and flexible model for over-, equi- and under-dispersion. For future work more statistical inference can be conducted on the COM-Poisson and $\mathrm{GIT}_{3,1}$ distributions with extensions to zero-inflated and regression models and their applications.

It is of interest to consider the bivariate extensions of the $\mathrm{GIT}_{3,1}$ distribution. Two new bivariate $\mathrm{GIT}_{3,1}$ which appear as alternative choices of bivariate distributions for data analysis, are defined and named as the Type I and Type II BGITD. Figures 4.1 to 4.10 in Chapter 4 clearly showed the attractive characteristics of the proposed bivariate distributions where the shape of the distribution changes as the positive-integer parameters are varied. It is observed in Chapter 4 that we have two positive-integer parameters under Type I BGITD and three positive-integer parameters under Type II BGITD. The Type I BGITD has four parameters when the two positive-integer parameters are fixed. While for the Type II BGITD, we have six parameters by fixing the three positive-integer parameters. To improve the utility of the distributions, the extension of these positive-integer parameters to real numbers may be studied.

In addition, the correlation derived under the Type I BGITD allows a very flexible structure where it provides a full range of the correlation coefficient from -1 to 1. On the other hand, the correlation derived under the Type II BGITD only permits positive correlation among the two random variables. By having a less restrictive correlation structure, the Type I BGITD is more flexible for empirical modelling compared with the Type II BGITD. For limitations of bivariate distributions based on trivariate reduction; see Mitchell and Paulson (1981) and Lee (1999).

The properties and the statistical inference of the Type I and Type II BGITD can be explored further. Besides that, more interesting and flexible bivariate $\mathrm{GIT}_{3,1}$ distributions may also be constructed.

Frequently, the complicated form of the joint probability mass function of the bivariate distribution is a major stumbling block in applications. The computation time for parameter estimation based on the probability mass function is clearly affected by its complexity. Hence, when the distribution possesses a simple form for the probability generating function, the pgf-based minimum Hellinger-type distance estimation is proposed. As illustrated by the real life data sets (Table 6.14 to 6.17), the pgf-based minimum Hellinger-type distance estimation does not only works as well as the MLE and it also shortens the computation time.

In Chapter 6, we have examined the robustness and accuracy of the pgf-based minimum Hellinger-type distance estimation under the univariate discrete distribution through an intensive simulation study. Six estimators $T_1, T_2, T_3, T_4, T_5, T_6$ with and without weighting factors, have been examined. To illustrate the method, we considered the orthogonal-parameter negative binomial distribution. In the simulation study, data with and without contamination have been generated; for data with contamination, we considered the mixtures of two different distributions.

The simulation results suggest that for small sample sizes, the proposed estimators may be preferred over MLE and MHDE for data with outliers. For large sample sizes the estimators T_1, T_2, T_5 and T_6 may be considered. T_1 and T_2 are estimators of choice because they are simpler (without weighting factors) and execute faster. Obviously, our timings show that T_1 and T_2 compute quickly compared to the other methods. In situations where the formula for the pmf is complicated relative to the pgf, T_1 and T_2 are to be preferred over MLE or MHDE.

In methods T_3 and T_4, it is found that the estimated parameter b falls at the upper bound of our setting, indicating the parameter b intends to choose a value which is as large as possible in the range (0-4). The same result is achieved by fixing the value b at 4.0. This situation is mentioned by Gürtler and Henze (2000) where large value of b means more weight is put near to the endpoint of $t = 1$. Estimators T_5 and T_6 are not favoured as they take longer computation time than the other estimators.

Since we have only considered the existence of outliers in Chapter 6, further work can be done by examining the performance of the pgf-based minimum Hellinger-type distance estimation under the existence of inliers. An inlier is a faulty data that lies in the interior of a statistical model and it is usually difficult to find and correct. Further generalizations of the pgf-based method of estimation and consideration of their robustness are also of interest.