Group-Based criminal trajectory analysis using cross-validation criteria

J.D. Nielsen[1], J.S. Rosenthal[2], Y. Sun[3],
D.M. Day[4], I. Bevc[5], and T. Duchesne[6]

(Last revised July 30, 2012.)

[1] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

[2] Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada, M5S 3G3.

[3] Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario, Canada, M5G 1X5.

[4] Department of Psychology, Ryerson University, 350 Victoria Street, Toronto, Ontario, Canada, M5B 2K3.

[5] The Hincks-Dellcrest Centre, 1645 Sheppard Avenue West, Toronto, Ontario, Canada, M3M 2X4.

[6] Département de mathématiques et de statistique, Pavillon Alexandre-Vachon, Université Laval, 1045, av. de la Medecine, Quebec City, Quebec, Canada, G1V 0A6.

Running Head: Trajectory Analysis using Cross-Validation

Acknowledgements. We are very grateful to Lianne Rossman for collecting and coding the Toronto offender data and thank Xinyue Liao for running the *Mplus* analysis. We thank the anonymous referee for many helpful comments.

Author Notes: This research was supported by grants from The Hincks-Dellcrest Centre, Ryerson University, the Samuel Rogers Memorial Trust, the Natural Science and Engineering Research Council of Canada, the National Crime Prevention Centre, and the Ministry of Children and Youth Services, Youth Justice Services. Correspondence regarding this article should be addressed to Jeffrey S. Rosenthal, Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada, M5S 3G3, jeff@math.toronto.edu.

Group-based Criminal Trajectory Analysis using Cross-Validation Criteria

Abstract

In this paper, we discuss the challenge of determining the number of classes in a family of finite mixture models with the intent of improving the specification of latent class models for criminal trajectories. We argue that the traditional method of using either the *Proc Traj* or *Mplus* package to compute and maximize the Bayesian Information Criterion (BIC) is problematic: *Proc Traj* and *Mplus* do not always compute the MLE (and hence the BIC) accurately, and furthermore BIC on its own does not always indicate a reasonable-seeming number of groups even when computed correctly. As an alternative, we propose the new freely available software package, *crimCV*, written in the R-programming language, and the methodology of *cross-validation error* (CVE) to determine the number of classes in a fair and reasonable way. In the present paper, we apply the new methodology to two samples of $N = 378$ and $N = 386$ male juvenile offenders whose criminal behavior was tracked from late childhood/early adolescence into adulthood. We show how using CVE, as implemented with *crimCV*, can provide valuable insight for determining the number of latent classes in these cases. These results suggest that cross-validation may represent a promising alternative to AIC or BIC for determining an optimal number of classes in finite mixture models, and in particular for setting the number of latent classes in group-based trajectory analysis.

Key Words and Phrases: Group-based trajectory analysis; juvenile offenders; zero-inflated-poisson (ZIP); cross-validation; Bayesian information criterion; crimCV.

1 Introduction

Group-based trajectory models are a valuable method of modeling the relationship between age and criminal behavior in an effort to uncover the underlying or latent heterogeneity of the sample. It is widely known that criminal offenders are a diverse and varied population. Finite mixture models allow us to approximate this heterogeneity by clustering individuals into small numbers of groups that show statistically similar trajectories in terms of rate or severity of offending over time. One challenge to these methods is to determine the number of groups to use in such a model. The most common approach is to optimize the Bayesian Information Criterion (BIC) (e.g. Brame, Nagin & Wasserman, 2006; D'Unger, Land, McCall, & Nagin, 1998; Kass & Raftery, 1995; Raftery, 1995; Bartolucci et al., 2007), by computing the corresponding maximum likelihood estimator (MLE) using the software *Proc Traj* (Jones, 2001; Jones, Nagin, & Roeder, 2001; Jones, & Nagin, 2007). Although versions of this procedure have had much success, it is known that such optimization can be problematic (e.g. Nagin, 2005; Kreuter & Muthen, 2008).

In this paper we argue that this approach is fundamentally flawed in that *Proc Traj* and *Mplus* often fail to accurately compute the MLE and, hence, the BIC and, furthermore, BIC on its own does not always indicate a reasonable-seeming number of groups even when computed correctly. As an alternative, we propose the new software package, *crimCV*, together with the methodology of *cross-validation error* (CVE), to compute the MLE more accurately and provide an alternative determination of the number of latent classes to be used.

1.1 Previously-Known Problems with BIC

The BIC (Schwartz, 1978) is commonly used in analyses such as those described in Section 2.1 of this paper. However, it is known to be problematic and often suggests far too many latent classes (Ward, Day, Bevc, Sun, Rosenthal, & Duchesne, 2010). For example, the book by

Nagin (2005, pp. 74–75), *Group-based modeling of development*, contains an entire section entitled "When BIC Is Not Useful in Identifying the Best Model," which notes that:

> BIC does not always cleanly identify a preferred number of groups. Instead, in some applications the BIC score continues to increase as more groups are added. In such instances, more subjective criteria based on domain knowledge and the objectives of the analysis must be used to select the number of groups to include in the model. ... [In a particular study] BIC continued to improve for this measurement series as more groups were added.

Similarly, Loughran and Nagin (2006, p. 259; see also Blokland, Nagin, & Nieuwbeerta, 2005) conducted an application and observed that:

> In this application, the BIC was not helpful in identifying a preferred model because over the range of models explored, BIC monotonically increased with the number of groups. We settled on the four-group model ... for several [unrelated] reasons.

In D'Unger, Land, and McCall (2002), the BIC again increases monotonically with the number of classes and the authors are forced to cut off this number when they can no longer invert a certain Hessian matrix as required to assess their standard errors. Similarly, in Piquero, Blumstein, Brame, Haapanen, Mulvey, and Nagin (2001), BIC again calls for more and more groups, until with $K = 7$ their models "failed to converge on a solution," forcing them to "settle on six-class models," Finally, Yessine and Bonta (2009, pp. 446–447), using *MPlus* rather than the *Proc Traj*, also found that BIC is monotonically increasing with group number, but settled on using two groups after determining that "the mixture models failed to converge to a trustworthy solution when more than two groups were specified." Such decisions are *ad hoc*, forced by computational issues rather than genuinely indicating optimality of choice of number of groups, and are thus quite unsatisfactory since the number of groups

is selected by software limitations rather than by optimally modeling the data. Indeed, finite mixture models can be viewed as an approximation to the marginal distribution of a response variable in a heterogeneous population, and hence the number of elements should be chosen so as to ensure that the approximating model will yield adequate inference and/or prediction. Thus, despite the widespread use of BIC for determining the number of latent classes, this approach is well-known to be problematic, posing challenges for researchers in this field. As noted by Nylund, Asparouhov and Muthén (2007, p. 537):

> To date, there is not common acceptance of the best criteria for determining the number of classes in mixture modeling, despite various suggestions. This is a critical issue in the application of these models, because classes are used for interpreting results and making inferences.

Likewise, Eggleston, Laub and Sampson (2004; see also Nagin, 1999) write:

> Although the Bayesian Information Criterion has been emphasized as the primary criterion to assess the optimal number of groups, the model selection process is often more complex and, thus, group selection remains somewhat subjective.

Of course, many authors who recognize the limitations of BIC on its own argue that BIC should be combined with subject-specific *judgement* to decide the number of latent classes (see e.g., Blokland, Nagin, & Nieuwbeerta, 2005; Eggleston et al., 2004). Such judgement may indeed lead to appropriate numbers of latent classes in many cases. However, it is necessarily subjective in nature, and in any event it does not alleviate – rather, it reinforces – the fact that BIC alone is not a completely satisfactory approach in this context.

1.2 Goal of this Paper

The primary motivations and objectives of the use of finite mixture models in empirical statistical analyses arise from the fact that many empirical frequency distributions do not

conform to one of the standard frequency distributions in statistics. Because of this, the empirical frequency distributions may contain hidden heterogeneity that must be taken into consideration in the modeling process. But by their very nature, finite mixture models are approximations to a mix of conventional frequency distributions that models the hidden heterogeneity. In light of this, in this paper, we take the aforementioned concerns about BIC two steps further. First, we argue that the standard software packages, *Proc Traj* and *Mplus* do not always compute the MLE (and hence the BIC) accurately, potentially leading to incorrect conclusions. This leads us to develop our own software package, *crimCV*, which computes the MLE more reliably. Second, we argue that even when computed correctly, BIC is a flawed criterion that often fails to propose an approximation to the population distribution that is reasonable for inference or interpretation purposes because it tends to suggest finite mixture models with too many latent classes. This leads us to propose the use of CVE as an alternative and practically valuable criterion.

2 The Models and Data

Though the CVE could be used to help in finding an appropriate number of latent classes in virtually any latent class model, in this paper we will illustrate its use by considering a particular family of zero-inflation Poisson (ZIP) models for latent class trajectories of criminal careers. These are specific instances of more general finite mixture models in statistics, which have been applied to many other areas in addition to criminology; see for instance Heckman & Singer (1984) or McLachlan & Peel (2000) for general references on finite mixture models. This provides evidence that such models are a useful way to analyze a wide variety of count data which can be thought of as falling into latent classes of some kind. In this section we provide background information about the model, data and approach that we will consider.

2.1 The ZIP Latent Class Model

We first summarize the standard ZIP statistical model (Lambert, 2002) that we shall use to model criminal trajectories (Jones et al., 2001; Nagin, 2005).

Let Y_{ij} be the number of offenses by individual i at age j (in years), for $1 \leq i \leq N$ and $L \leq j \leq U$. (Here L is the lower bound on offense ages, while U is the upper bound; e.g. for our first data set $L = 8$ and $U = 38$, while for our second data set $L = 9$ and $U = 38$.) To specify the statistical model, we need to specify the probability that $Y_{ij} = y_{ij}$ for all i and j. For shorthand, let $Y_i = (Y_{i,L}, Y_{i,L+1}, \cdots Y_{iU})'$, and $y_i = (y_{i,L}, y_{i,L+1}, \cdots y_{iU})'$.

Though it is quite possible that the distribution of Y_i may be different for each individual, in finite mixture modeling we assume that each individual i can be in one of K different (homogeneous) latent classes, so that

$$P(Y_i = y_i) \;=\; \sum_{k=1}^{K} p_k \, P(Y_i = y_i \mid \text{individual } i \text{ is in class } k)\,, \tag{1}$$

for some unknown probabilities $\{p_k\}$ with $0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$. We further make the standard assumption (which surely is not strictly true, but which greatly simplifies the mathematics) that, conditional on individual i being in class k, the number of offenses Y_{ij} are independent for different ages j so that equation (1) becomes

$$P(Y_i = y_i) \;=\; \sum_{k=1}^{K} p_k \prod_{j=L}^{U} P(Y_{ij} = y_{ij} \mid \text{individual } i \text{ is in class } k)\,. \tag{2}$$

Equation (2) requires us to specify the probabilities for the Y_{ij} conditional on being in class k. To do this, we follow the zero-inflation Poisson (ZIP) model, which assumes that if individual i belongs to class k, then the distribution of Y_{ij} is a mixture of a Poisson distribution together with excess probability of zero offenses (corresponding to being in a "non-criminal" state). If we wish, we can also include year-by-year exposure times (i.e., times-at-risk), t_{ij}, which indicate the fraction of the year j in which individual i was not in

secure custody. (Thus, the t_{ij} are all between 0 and 1, and can be taken to be 1 if time-at-risk adjustments are not available or required. That is, $0 \le t_{ij} \le 1$, with $t_{ij} = 1$ indicating that individual i was at large for the entire year j.) Hence, the conditional probability distribution for Y_{ij} conditional on individual i being in class k becomes:

$$(Y_{ij} \mid \text{individual } i \text{ is in class } k) \;\sim\; (1 - q_j^k)\,\text{Poisson}(t_{ij}\lambda_j^k) + q_j^k \delta_0 \tag{3}$$

for some (unknown) mixture probabilities q_j^k and intensities λ_j^k, where δ_0 is a point-mass at zero. That is,

$$P(Y_{ij} = y_{ij} \mid \text{individual } i \text{ is in class } k) \;=\; (1 - q_j^k) e^{-t_{ij}\lambda_j^k} \frac{(t_{ij}\lambda_j^k)^{y_{ij}}}{y_{ij}!} + q_j^k I(y_{ij} = 0) \tag{4}$$

$$= \begin{cases} (1 - q_j^k) e^{-t_{ij}\lambda_j^k} \frac{(t_{ij}\lambda_j^k)^{y_{ij}}}{y_{ij}!} & \text{if } y_{ij} > 0 \\ (1 - q_j^k) e^{-t_{ij}\lambda_j^k} + q_j^k & \text{if } y_{ij} = 0 \end{cases}$$

In particular, this equation provides the estimated *mean* or *expected value*, μ_j^k, of the offense count Y_{ij} at age j of an individual i who is known to be in class k:

$$\mu_j^k \;=\; \mathbf{E}(Y_{ij} \mid \text{individual } i \text{ is in class } k) \;=\; (1 - q_j^k)(t_{ij}\lambda_j^k)\,. \tag{5}$$

2.2 The Predictor Functions

It remains to model the unknown parameters λ_j^k and q_j^k. The λ_j^k are modeled by either the quadratic predictor functions:

$$\log(\lambda_j^k) \;=\; \beta_{0k} + \beta_{1k} j + \beta_{2k} j^2 \tag{6a}$$

or the cubic predictor functions:

$$\log(\lambda_j^k) \;=\; \beta_{0k} + \beta_{1k} j + \beta_{2k} j^2 + \beta_{3k} j^3, \tag{6b}$$

where for each latent class k, the unknown values β_{ik} need to be estimated.

The q_j^k model the amount of zero inflation, that is, the excess probability of individual j being in a "non-criminal" state. They are modeled by predictor functions given as either logit-linear:

$$\text{logit}(q_j^k) \;\equiv\; \log\left(\frac{q_j^k}{1-q_j^k}\right) \;=\; \alpha_{0k} + \alpha_{1k} j \tag{7a}$$

or logit-quadratic:

$$\text{logit}(q_j^k) \;\equiv\; \log\left(\frac{q_j^k}{1-q_j^k}\right) \;=\; \alpha_{0k} + \alpha_{1k} j + \alpha_{2k} j^2 \tag{7b}$$

or proportional to $\log(\lambda_j^k)$ as in the ZIP(τ) model of Lambert (2002):

$$\text{logit}(q_j^k) \;\equiv\; \log\left(\frac{q_j^k}{1-q_j^k}\right) \;=\; -\tau_k \log(\lambda_j^k)\,, \tag{7c}$$

where the α_{ik} or τ_k are again unknown values to be estimated.

Combining these equations, the overall likelihood function is given by

$$L_K(\theta) \;=\; \prod_{i=1}^{N} \sum_{k=1}^{K} p_k \prod_{j=L}^{U} \left[(1-q_j^k) e^{-t_{ij}\lambda_j^k} \frac{(t_{ij}\lambda_j^k)^{y_{ij}}}{y_{ij}!} + q_j^k I(y_{ij}=0) \right], \tag{8}$$

where θ is a vector consisting of all of the unknown parameters (i.e., of all of the $\{p_k\}$, $\{\beta_{ik}\}$, and $\{\alpha_{ik}\}$ or $\{\tau_k\}$ as appropriate depending on which versions of the predictor functions (6a)–(6b) and (7a)–(7c) are used). Thus, the length f_K of the vector θ is equal to $7K-1$ for the quadratic-quadratic pure-ZIP model (i.e., using equations (6a) and (7b)), or $6K-1$ for the cubic-ZIP(τ) model (i.e., using equations (6b) and (7c)), etc. (The "-1" comes because

we omit p_K from the vector θ, since we must have $p_K = 1 - p_1 - p_2 - \ldots - p_{K-1}$.)

2.3 Fitting the Model

To fit the model, we require a maximum likelihood estimator (MLE), that is, a vector $\hat{\theta}$ of estimated values of all the unknown parameters that maximizes the likelihood function $L_K(\theta)$ in equation (8). Such $\hat{\theta}$ consists of estimated values $\{\hat{p}_k\}$, $\{\hat{\beta}_{ik}\}$, and $\{\hat{\alpha}_{ik}\}$ or $\{\hat{\tau}_k\}$ as appropriate, which in turn implies estimated values $\{\hat{\lambda}_j^k\}$ and $\{\hat{q}_k\}$, and thus allows us to estimate all relevant quantities and probabilities associated with the model with K latent classes.

After obtaining the estimator $\hat{\theta}$, we can then use (8) to estimate the *a posteriori* probability π_i^k that individual i belongs to group k, by

$$\hat{\pi}_i^k \;=\; \frac{\hat{p}_k \prod_{j=L}^{U} \left[(1-\hat{q}_j^k) e^{-\hat{\lambda}_j^k} \frac{(\hat{\lambda}_j^k)^{y_{ij}}}{y_{ij}!} + \hat{q}_j^k I(y_{ij}=0) \right]}{\sum_{m=1}^{K} \hat{p}_m \prod_{j=L}^{U} \left[(1-\hat{q}_j^m) e^{-\hat{\lambda}_j^m} \frac{(\hat{\lambda}_j^m)^{y_{ij}}}{y_{ij}!} + \hat{q}_j^m I(y_{ij}=0) \right]} \,. \tag{9}$$

If we wish, we can also identify the most probable latent class of each individual i, as

$$\text{most probable latent class of individual } i \;=\; \arg\max_k \hat{\pi}_i^k \,. \tag{10}$$

We can then classify the individuals by most probable latent class and investigate the criminal characteristics of each such group.

Now, the length f_K of the vector θ grows fairly quickly with K. So, for even moderately large K, computing the MLE is computationally challenging. Such issues are considered further below.

2.4 Prediction for New Subjects

This model also provides a method for *predicting* offense trajectories of new individuals. Indeed, if an $(N+1)^{\text{st}}$ individual comes along, with offense data counts $y_{N+1,L}, y_{N+1,L+1}, \ldots, y_{N+1,U}$, we could estimate the class membership probabilities $\hat{\pi}^k_{N+1}$ from (9). Then, combining the values $\hat{\pi}^k_{N+1}$ with the conditional means given by (5), we conclude that

$$\hat{y}_{N+1,j} \;\equiv\; \text{estimated mean of } Y_{N+1,j} \;=\; \sum_{k=1}^{K} \hat{\pi}^k_{N+1}\,(1-q^k_j)(\lambda^k_j)\,. \tag{11}$$

If $\hat{y}_{N+1,j}$ is close to $y_{N+1,j}$, this indicates that the model does a good job of predicting future individuals' offense patterns. Such issues will be important in Section 3.2 below.

3 Number of Latent Classes

The above discussion provides a complete and unambiguous recipe for estimating the entire multidimensional vector of parameters θ, as well as the *a posteriori* group membership probabilities π^k_i in (9), at least once we have specified the number K of latent classes. However, the choice of number of classes K must first be made. As we have already argued, this choice should be made so as to obtain an "optimal" approximation to the distribution of Y_i in a heterogenous population. In practice, this is usually done by considering several potential values of K and choosing the one which yields the best value of a given criterion. In the remainder of this section, we present such criteria and argue that the CVE criterion tends to propose values of K that correspond to models that are more sensible in practice.

3.1 BIC and AIC

A conventional method for selecting the number of classes, or more generally the number of parameters to use in a model, is to minimize the Bayesian Information Criterion (BIC),

given by

$$\text{BIC} \;=\; -2\;\log(\text{likelihood}) + (\text{number of free parameters})\,\log(\text{number of observations})\,. \tag{12}$$

For the model described in Section 2.1, the likelihood function is $L_K(\theta)$, and the number of free parameters is equal to the length f_K of the vector θ and hence is equal to $7K-1$ for the quadratic-quadratic model, or $5K-1$ for the cubic-tau model, and so on. Also, the number of observations is taken to be $N(U-L+1)$, the total number of y_{ij} data values. (Note that some implementations by default use N instead of $N(U-L+1)$. For consistency we stick with $N(U-L+1)$ here, but in any case such difference will not greatly affect the results.) Hence, the formula becomes

$$\text{BIC} \;=\; -2\;\log(L_K(\hat{\theta})) + f_K\;\log\big(N\,(U-L+1)\big)\,. \tag{13}$$

The BIC criterion would then have us choose the value of K that minimizes (12).

Closely related to BIC is the Akaike information criterion (AIC), proposed by Akaike (1974), defined by

$$\text{AIC} \;=\; -2\;\log(\text{likelihood}) + 2\,(\text{number of free parameters})\,. \tag{14}$$

which in our case becomes

$$\text{AIC} \;=\; -2\;\log(L_K(\hat{\theta})) + 2f_K\,. \tag{15}$$

The AIC criterion would then have us choose the value of K that minimizes (15). A comparison of (14) and (12) (or of (15) and (13)) shows that AIC is similar in spirit to BIC, and indeed we shall see below that it appears to have similar difficulties and limitations to BIC. In a related direction, Brame et al. (2006) compare AIC and BIC in simulations, concluding

that "AIC generally outperformed BIC in identifying the correct number of groups. However, this is not uniformly the case and both criteria performed well with large samples", but they do not compare AIC and BIC to other criteria such as CVE.

We note that BIC is sometimes defined as $-1/2$ times the value described in (12) and (13) (and similarly for AIC). In this case, the criterion would instead have us choose the value of K that *maximizes* the BIC. So, to compare our results with other results using the alternative definition, simply multiply all of our BIC (and AIC) values by $-1/2$.

Below, we shall consider both BIC and AIC, and will find (as previously observed) that neither provides a satisfactory resolution for determining the number of latent classes K. Because of these difficulties, we will consider an alternative method, cross-validation error (CVE), for determining the number of latent classes K.

3.2 The Cross-Validation Approach

The previous section discussed the challenge of accurately computing the BIC. However, we shall argue below that even when computed correctly (e.g. with *crimCV*), BIC on its own still does not provide a very useful criterion for determining the number K of latent classes. Thus, we propose an alternative criterion for this purpose. Specifically, we propose to make use of *cross-validation* (Hélie, 2006; Stone, 1974). This is a method of using data to test the predictive power of a model. If the model predicts well, then this suggests that it is a good and appropriate model which provides a good "fit" of the data and should be used.

In the present context, we proceed as follows. For each possible choice of number of latent classes K, we compute a cross-validation error (CVE) indicating the extent to which the model fails to perfectly model the data. The final choice of K is then the one that minimizes this CVE value.

More specifically, we use *leave-one-out* cross-validation. This method measures the accuracy of the fit for individual i by using estimates of the model parameters θ based on data for all the *other* individuals but *not* individual i. Specifically, for a given choice of the

number K of latent classes, we proceed as follows. For each individual i, we *omit* the data for individual i, and then determine an MLE $\hat{\theta}^{[-i]}$ for θ based on data for the *other* $N-1$ individuals only. Then, using the parameter values $\hat{\theta}^{[-i]}$, we then use (9) and (11) to obtain estimates $\hat{y}_{ij}^{[-i]}$ for the various y_{ij} values ($L \leq j \leq U$). That is, we obtain the best possible estimates for the offense counts for individual i, using a model whose parameters have been fit *without* using the offense data of individual i. This provides a fair measure of the accuracy of the model, without encouraging over-fitting (see e.g. Day et al., 2007).

Once we have estimates $\{\hat{y}_{ij}^{[-i]}\}_{j=L}^{U}$, then the cross-validation error for individual i, CVE(i), can be measured in terms of the average absolute difference between the true values $(y_{i,L}, y_{i,L+1}, \cdots, y_{i,U})$ and the predicted values $(\hat{y}_{i,L}^{[-i]}, \hat{y}_{i,L+1}^{[-i]}, \cdots, \hat{y}_{i,U}^{[-i]})$, i.e.

$$\text{CVE}(i) \;=\; \frac{1}{U-L+1} \sum_{j=L}^{U} \left| y_{ij} - \hat{y}_{ij}^{[-i]} \right| .$$

By repeating this process for each of the N individuals i, and then averaging the cross-validation errors CVE(i) over all N individuals, we get our final cross-validation error (CVE):

$$\text{CVE} \;=\; \frac{1}{N} \sum_{i=1}^{N} \text{CVE}(i) \, .$$

The value CVE thus provides a fair measure of how appropriate the approximation given by the finite mixture model with the chosen group number K is for the given data in terms of how accurately a model with that number of groups is able to predict the offender data of new individuals. If CVE is large, this indicates that the model with K groups is not a good statistical model for this data. By contrast, if CVE is small, then the model with K groups is doing a good job of predicting offender data of new individuals.

The cross-validation criterion for number of groups then involves simply choosing the value of K that minimizes CVE. The software package, *crimCV*, computes CVE for any criminal offense data and any number K of latent groups and is thus appropriate for applying

this cross-validation criterion. (Another possible method is the bootstrap likelihood ratio test [BLRT], not considered here; see Nylund, Asparouhov & Muthén, 2007; Kreuter & Muthén, 2008.) We shall see in Section 6 that the cross-validation criterion usually provides a sensible, stable recommendation for the number of latent classes K, and thus represents a valuable alternative to other criteria such as BIC and AIC.

4 Toronto Juvenile Offender Samples (TO1 and TO2)

To study this model, as well as the *Proc Traj* and *crimCV* software, and the BIC and CVE criteria, we consider two Toronto juvenile offender data sets, referred to as TO1 ($N = 378$) and TO2 ($N = 386$), previously studied by Day, Bevc, Duchesne, Rosenthal, Rossman, and Theodor (2007) (see also Day, Bevc, Rosenthal, Duchesne, Rossman, & Theodor, 2003; Day, Bevc, Duchesne, Rosenthal, Sun, & Theodor, 2008; Day, Nielsen, Ward, Sun, Rosenthal, Duchesne, Bevc, & Rossman, 2011; Ward et al., 2010). These data sets are publicly available for inspection in completely de-identified form, by contacting I. Bevc or by installing the *crimCV* software package discussed below.

All the juvenile offenders in this study had served a sentence between January 1, 1986 and December 30, 1997 at one of two open custody facilities (i.e., group homes) in Toronto, Canada, operated by a children's mental health centre. During this period, a total of 764 male offenders served a sentence at one of the two sites; therefore, our research involves the entire population of youth from these facilities during this period. Information about the two samples is presented in Table 1.

[INSERT TABLE 1 ABOUT HERE.]

The first sample, "TO1", comprised a randomly selected sample of 378 youth, from the population of youth at these facilities during this period. The sample was, on average, 17.6

years at the time of admission into the facility and the average sentence length was 124.6 days (Median = 92 days). Their criminal activity was tracked for an average of 18.7 years (range = 12.3 – 29.3 years), from their first recorded involvement with the justice system to September 26, 2007. The average age at the end of the follow-up was 34.1 years (range = 28.7 – 40.5 years).

The remaining 386 offenders from this population constituted the second sample, "TO2". This group was, on average, 17.7 years at the time of admission into the facility and the average sentence length was 122.6 days (Median = 93 days). Their criminal activity was tracked for an average of 16.4 years (range = 9.8 – 28.7 years), from their first recorded involvement with the justice system up to and including September 26, 2007. Their average age at first court contact was 15.6 years (range = 9.6 – 19.4 years) and the average age at the end of the follow-up was 32.0 years (range = 26.3 – 40.2 years). A strength of these samples is that the follow-up periods extended from late childhood (for offenses committed under the Juvenile Delinquents Act [JDA] in Canada, for which the minimum age of criminal liability was 7 years, unlike the Young Offenders Act [YOA], to which most of the criminal data apply, and the current Youth Criminal Justice Act [YCJA], for which the minimum age is 12 years) and early adolescence into adulthood, well beyond the challenging period of emerging adulthood of the early 20's (Arnett, 2000, 2007).

4.1 Criminal Data

The criminal data for TO1 were received, initially ,on March 17, 2001 and updated on September 26, 2007, from four sources: (1) the (Ontario) Ministry of Community and Social Services (MCSS); (2) the (Ontario) Ministry of Community Safety and Correctional Services (MCSCS); (3) the Canadian Police Information Centre (CPIC); and (4) the Predisposition Reports (PDRs) maintained by the children's mental health centre. The criminal data for TO2 were received from three sources: (1) the (Ontario) Ministry of Community Safety and Correctional Services (MCSCS); (2) the Canadian Police Information Centre (CPIC); and

(3) the predisposition reports (PDRs) maintained by the children's mental health centre. These data sources were used to ensure a high degree of completeness and accuracy of the information, which is essential for research that requires an accurate temporal sequencing of criminal activity (Smith, Smith, & Norma, 1984).

The reason only three data sources were used for TO2 (as well as for the second follow-up for TO1 as noted below) was that, at some point between March 17, 2001 and September 26, 2007, the MCSS system was integrated into the MCSCS and thus two sources became one main source. Due to the nature of the data integration process and the importance of continued monitoring of active offenders, historical data for juvenile offenders between the ages of ages of 12- 15 years (referred to as Phase I offenders) were available only for those offenders who were still actively serving youth sentences at the time of our inquiry. Thus, due to the limitations of accessible data at the time, the majority of these Phase 1 data for TO2 were only retrieved through the PDRs.

As well, to aid the accuracy of our time-at-risk adjustments (see below), supplemental movement data containing (institutional) location start and exit dates were provided by the MCSCS for our data requested on September 26, 2007. Due to the implementation of a new criminal record monitoring system between the first and second follow-up periods, TO2 data were received in electronic format from the Ministry and contained more information than that received on the paper profiles (i.e. "rap sheets") for the initial follow-up of TO1. Not only could the charges and dates of convictions be discerned as before, but now the movement data associated with each charge was provided and allowed for a more accurate account of time-served per date of conviction. As a result of these differences in data received between the two follow-up periods, we decided not to combine the data for TO1 and TO2 into one large data set for analyses. The criminal data for these samples gave us year-by-year offense counts $\{y_{ij}\}$; for TO1 we have $1 \leq i \leq N = 378$ and $L = 8 \leq j \leq U = 38$, while for TO2 we have $1 \leq i \leq N = 386$ and $L = 9 \leq j \leq U = 38$.

4.2 Time-at-Risk Adjustments

As described in Day et al. (2007), the time-at-risk adjustment contained additional challenges since the data included dates of court contact but not dates of offense. This necessitated computing *estimates* t_{ij} of the fraction of the year at age j when subject i was at large. These t_{ij} values can then be used as offsets as in the previous section (as in Piquero et al., 2001, p 59). Unfortunately, *Proc Traj* cannot easily handle time offsets. Alternatively, we can use the "divide and round" (DAR) approximation of Ward et al. (2010) (see also Eggleston et al., 2004, p. 6 bottom), consisting of adjusting each count y_{ij} by dividing it by the corresponding exposure time t_{ij} and rounding the result to the nearest integer (truncated at a maximum of 25). We feel that time offsets are a far preferable method, since DAR does not distinguish between, for example, "5 offenses in 3 months at large" and "20 offenses this year" even though the effect of these two different realities on our conclusions should be far from identical. However, DAR does provide a clear and unbiased data set with which to compare *Proc Traj* to other software, as we do in the next section.

5 Software Considerations

As noted above, the computation of the MLE for these models is very challenging, especially for larger numbers of latent classes. To compute the MLE, we originally used the SAS procedure *Proc Traj* provided by Jones (2001; see also Jones et al., 2001; Jones, & Nagin, 2007). However, we found that *Proc Traj* sometimes failed to converge consistently, and sometimes gave different values when used with different versions of the program or different initialisations, and sometimes failed to find the true MLE (see Section 5.2).

Another commonly-used software package is *Mplus* (www.statmodel.com; Muthén and Muthén, 2001), which provides some improvements over *Proc Traj*, but still sometimes fails to find the true MLE (see Section 5.2). In addition, both *Proc Traj* and *Mplus* are proprietary packages that must be purchased and do not allow for inspection and modification of the

source code. For all these reasons, we eventually decided to write our own software package, *crimCV*, described next.

5.1 A New Software Package: crimCV

Our software package incorporates the above ZIP models for criminal offender data, computes the MLE (and BIC and AIC) more reliably than *Proc Traj*, and also computes the quantity CVE. Furthermore, it is available as a free, publicly available package to augment the statistical software R at no cost to the user and with full access to the source code. We note that while our R software package is called "crimCV" and was developed specifically for analyzing criminal offense trajectories, the analysis it performs could be applied more generally to many other types of data involving finite mixture models as well; we hope to develop these ideas further in subsequent work.

To run *crimCV*, one should proceed as follows:

1) Obtain the R programming environment by following the instructions at:

`http://www.r-project.org/`

2) Run R, by double-clicking the R icon or typing 'R' at the terminal.

3) At the R prompt '`>`', type '`chooseCRANmirror()`' and select the closest repository.

4) At the R prompt '`>`', type '`install.packages("crimCV")`', which will install the *crimCV* package. (Note: you need to have administrative privileges on your computer to install packages. Alternatively, *crimCV* can be obtained directly from `www.probability.ca/crimCV`.)

5) At the R prompt '`>`', type '`help("crimCV")`' for a description of the software, options and an example of usage.

After following the above steps, *crimCV* can be freely run, both on the TO1 and TO2 data to check our results herein, and on any other criminal offense data as desired.

5.2 Comparison of the Different Software Packages

To compare these various software packages, we had each of *Proc Traj*, *Mplus* and *crimCV* maximize the log-likelihood and compute the corresponding BIC and AIC values for the TO1 and TO2 data sets. Since *Proc Traj* cannot easily handle time offsets, nor was it designed for the ZIP(τ) model, we therefore focused on the quadratic-quadratic pure-ZIP model together with the DAR method of correcting for time-at-risk.

The computed results are presented in Table 2.

[INSERT TABLE 2 ABOUT HERE.]

These results show that for small numbers of latent groups, *Proc Traj*, *Mplus* and *crimCV* obtain the same values for the maximum log-likelihood (and hence for BIC and AIC, as well), as we would hope. However, for larger numbers of latent groups, discrepencies start to arise. In such cases, the *crimCV* values for the log-likelihood are consistently *larger* (i.e., less negative) than those computed by the other software packages. This shows that it is *Proc Traj* and *Mplus* that sometimes fail to maximize the log-likelihood and thus obtain too small an estimate of the MLE and thus a correspondingly too small estimate of the values of BIC and AIC.

We note that this run used the *Proc Traj* default starting values. If *Proc Traj* is instead initialised with excellent starting values, for example, as obtained from *crimCV* itself, it will have a better chance of computing the correct answers. However, such starting values would not normally be available to the user and it does not seem likely that a small amount of "trial and error" will be sufficient to find starting values which allow *Proc Traj* to correctly compute the MLE.

Similarly, *Mplus* was run with the command "start = 100 10;." This means that in the initial stage, 100 random sets of starting values were generated. An optimization is then carried out for 10 iterations using each of the random sets of starting values. The ending

values from the optimizations with the 10 highest loglikelihoods are then used as the starting values in the final stage optimizations (EM algorithm in our case). Note that the default starting values setting in *Mplus* for mixture models is 10 2 (10 random sets in the first stage and 2 in the final stage), but we increase it to 100 10 (which is suggested by the program for $K > 2$) to improve the *Mplus* results. It is possible that increasing these values still larger would improve *Mplus*'s performance somewhat, but at the expense of even greater computation time.

We conclude that *Proc Traj* and *Mplus* sometimes fail to accurately maximize the likelihood and hence to compute correct values for BIC and AIC. They are thus sub-optimal as software for optimizing the number of latent classes. (This is not intended as a criticism of *Proc Traj* and *Mplus*, since such high-dimensional maximization is notoriously difficult, but it does illustrate the limitations of the existing software.)

By contrast, we have demonstrated that *crimCV* more accurately maximizes the likelihood and computes the BIC (and AIC) values. In addition, *crimCV* can directly handle time-at-risk offset values t_{ij}, as well as both pure-ZIP and ZIP(τ) models. Furthermore, it can also compute the CVE, as discussed further below. In addition, it is freely available including its source code which can be inspected and modified as desired.

6 Comparison of AIC, BIC, and CVE

We have already noted in the introduction that BIC has been found to be problematic in many applications. Of course, some of these difficulties may have been due to the computational limitations of *Proc Traj* as noted above. However, we now argue that even when computed correctly (e.g. with *crimCV*), BIC still does not provide as useful a criterion for determining the number K of latent classes as CVE does. We illustrate this by applying BIC and CVE (and AIC) to our two offender data sets and find that CVE provides results that suggest more reasonable numbers of latent classes.

6.1 Optimal Number of Groups for the Toronto Data

We apply our *crimCV* software to the two Toronto data sets, TO1 and TO2, with what we consider to be the most appropriate model, namely the cubic-ZIP(τ) model (i.e., using equations (6b) and (7c)), treating the time-at-risk exposure times t_{ij} as offsets in our model rather than incorporating them using the questionable DAR method. Our results are presented in Table 3.

[INSERT TABLE 3 ABOUT HERE.]

Table 3 shows that BIC and AIC decrease seemingly without limit, thus always advocating more and more latent groups. This mirrors the problems reported previously in the literature wherein BIC advocates more groups until such time as the computational software (e.g. *Proc Traj*) crashes and cannot compute larger numbers of groups.

By contrast, CVE advocates a more reasonable numbers of groups since it achieves its minimum value at a practically sensible number of groups, namely 8 for TO1, and 7 for TO2. Thus, this provides some evidence that CVE is a useful criterion for determining the number of latent classes K.

Of course, as the number of groups gets large, all of the criteria values tend to stabilize. This is illustrated in Figure 1, which shows how CVE is minimized at 8 and 7 groups respectively while BIC is never minimized, but also shows that the percentage increases get close to zero as the number of groups gets large. Thus, an analyst might perhaps wish to do the maximum likelihood classification of the members of the sample for each number of classes for which the criteria are "nearly maximal", to see whether or not the substantive conclusions from the data are affected.

[INSERT FIGURE 1 ("Fig1.png") ABOUT HERE.]

6.2 Latent Group Analysis

We have thus used *crimCV* to determine the optimal number of latent groups according to the CVE criterion, namely 8 for TO1, and 7 for TO2.

As described in Day et al. (2012; see also Day, Nielsen, Ward, Rosenthal, Sun, Bevc, Duchesne, Rossman, & Samuels, 2010), these numbers of groups then lead to interesting interpretations of the criminal behavior in the resulting latent classes. For the TO1 data, the groups were heuristically labelled Low Rate Desister (LRD), comprising 28.0% of the sample, Low Rate Chronic (LRC), comprising 26.2% of the sample, Low Rate Adolescent Peaked (LRADOLP), comprising 16.4% of the sample, Moderate Rate Chronic II (MRC-II), comprising 11.9% of the sample, Moderate Rate Chronic I (MRC-I), comprising 5.3% of the sample, High Rate Adolescent Peaked (HRADOLP), comprising 4.8% of the sample, Moderate Rate Adult Peaked (MRADLP), comprising 4.5% of the sample, and Moderate Rate Escalator (MRE), comprising 2.9% of the sample. None of the groups had less than 10 individuals in them and the mean posterior probability coefficients were quite high across all four groups, exceeding .88. The trajectory groups are plotted in Figure 2.

[INSERT FIGURE 2 ("TO1.png") ABOUT HERE.]

Comparisons on offending-related variables across the seven groups indicated that the LRD group had the latest age of first court contact (M = 15.8 years), the earliest age of last court contact (M = 20.5 years), and the shortest criminal career, spanning, on average, 4.2 years. The HRADOLP group had the earliest age of first court contact (M = 13.7 years). Not surprisingly, the MRE had the latest age of last court contact (M = 34.0 years) and the longest criminal career at 19.3 years, on average. The MRC I group had the second longest criminal career at 17.8 years, on average, followed by MRC II at 15.9 years, on average, and the LRC group at 13.5 years, on average. Over the duration of their criminal trajectories, the MRE group amassed the largest number of court contacts (adjusted by time-at-risk),

with 110.1, on average, followed by the HRADOLP group with 89.5, on average, and the MRADOLP group, 56.4, on average. By contrast, the LRD group had the fewest court contacts, on average, with 4.6.

Similar to the TO1 data, analysis of the TO2 data yielded seven trajectory groups. These groups were heuristically labeled Moderate Late Pesister (MLP), comprising 3.6% of the sample; High Late (HL), comprising 3.9% of the sample; High Early (HE), comprising 4.4% of the sample; Moderate Adolescence-Peaked (MAP), comprising 11.7% of the sample; Moderate Early Persister (MEP), comprising 14.2% of the sample; Low Desister (LD), comprising 29.8% of the sample; and Low Persister (LP), comprising 32.4% of the sample. None of the groups had less than 10 individuals in them and the mean posterior probability coefficients were quite high across all four groups, exceeding .89. These trajectories are plotted in Figure 3.

[INSERT FIGURE 3 ("TO2.png") ABOUT HERE.]

Comparisons on offending-related variables across the seven groups revealed some interesting similarities with the TO1 data. The LD group had the latest age of first court contact (M = 16.4 years), the earliest age of last court contact (M = 19.5 years), and the shortest criminal career, spanning, on average, 3.1 years (see Figure 2). Once again, the HL group had the earliest age of first court contact (M = 14.3 years). The MLP group had the latest age of last court contact (M = 31.9 years) and the longest criminal career at 16.6 years, on average. The MEP group had the second longest criminal career at 14.8 years, on average. Over the duration of their criminal trajectories, the HL group amassed the largest number of court contacts (adjusted by time-at-risk), with 78.1, followed by the HE group with 62.7, on average, and the MLP group with 52.1, on average. By contrast, the LD group had the fewest court contacts, on average, with 4.8.

7 Discussion

This paper has argued that the traditional method of computing BIC using *Proc Traj* and *Mplus*, does not maximize the likelihood in a numerically accurate manner and thus cannot be completely trusted to optimize the BIC. Inspired by this, we have presented the free software package *crimCV* for accurately maximizing likelihood and computing BIC (and AIC). *crimCV* has the additional advantages that it can handle ZIP(τ) as well as pure-ZIP models and can incorporate time-at-risk t_{ij} as a true offset variable (rather than only with the more questionable method of dividing-and-rounding (DAR) the offense count data). It can also compute the cross-validation error (CVE) in addition to BIC and AIC.

This paper has further argued that BIC, the traditional method of estimating number of latent classes (groups) when analyzing criminal offense trajectories, is flawed in that it often increases monotonically with group number – thus arguing for larger and larger numbers of groups, which eventually have to be cut off due to computational or other *ad hoc* reasons or by the use of subjective judgement on the part of the researcher. By contrast, we have argued that the cross-validation error (CVE) is a theoretically sound method of evaluating model appropriateness by directly measuring predictive ability, which makes sense logically and which also provides reasonably approximating finite mixture models in practice.

We applied our *crimCV* software to two Toronto data sets of juvenile offenders. In both cases, CVE has proved to be a more useful criterion than BIC (and AIC) to propose a reasonable finite mixture model by achieving its minimum at a practical number of latent groups (8 and 7, respectively), in contrast to BIC and AIC which continued to decrease without apparent limit. We view this as providing some evidence that cross-validation is a clear, objective, and effective method of determining the number of latent classes, and should be considered as a serious alternative to BIC or AIC. As the *crimCV* package is newly developed, our findings need to be replicated with additional data sets to further verify its practical utility for group-based trajectory analysis.

REFERENCES

Akaike H., (1974), A new look at the statistical model identification. *IEEE Trans. on Autom. Control*, *19*, 716–723.

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, *55*, 469–480.

Arnett, J. J. (2007). Emerging adulthood: What is it and what is it good for? *Child Development Perspective*, *1*, 68–73.

Bartolucci, F., Pennoni, F., & Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, series A*, *170*, 115–132.

Bloklad, A. A. J., Nagin, D., & Nieuwbeerta, P. (2005). Life span offending trajectories of a Dutch conviction cohort. *Criminology*, *43*, 919–954.

Brame, R., Nagin, D.S., & Wasserman, L. (2006). Exploring Some Analytical Characteristics of Finite Mixture Models. *Journal of Quantitative Criminology*, *22*, 31–59.

Day, D. M., Bevc, I., Rosenthal, J. S., Duchesne, T., Rossman, L., & Theodor, F. (2003), *Predicting Adult offenders' criminal trajectories from their juvenile criminal trajectories*. Poster presented at the 111$^{\text{th}}$ Convention of the American Psychological Association, Toronto, Canada.

Day, D. M., Bevc, I., Duchesne, T., Rosenthal, J. S., Rossman, L., & Theodor, F. (2007), Comparison of adult offense prediction methods based on juvenile offense trajectories using cross-validation. *Advances and Applications in Statistics*, *7*, 1–46.

Day, D. M., Bevc I.,Duchesne, T., Rosenthal, J. S., Sun, Y., & Theodor, F. (2008). Criminal trajectories from adolescence to adulthood in an Ontario sample of offenders. In G. Bourgon, R.K. Hanson, J.D. Pozzulo, K.E. Morton Bourgon & C.L. Tanasichuk (Eds.), *The Proceedings of the 2007 North American Correctional & Criminal Justice Psychology Conference (User Report)* (pp. 143-148). Ottawa: Public Safety Canada.

Day, D. M., Nielsen, J. D., Ward, A. K., Rosenthal, J. S., Sun, Y., Bevc, I., Duchesne, T., Rossman, L., & Samuels, S. (2010). Criminal trajectories of two subsamples of adjudicated Ontario youths. Final research report submitted to the National Crime Prevention Centre (NCPC), Ottawa, Ontario.

Day, D. M., Nielsen, J.D., Ward, A.K., Sun, Y., Rosenthal, R.S., Duchesne, T., Bevc, I., & Rossman, L. (2012). Long-term follow-up of criminal activity with adjudicated youth in Ontario: Identifying offence trajectories and and predictors/correlates of trajectories group membership. *Canadian Journal of Criminology and Criminal Justice*.

Dunford, F.W. & Elliott, D.S. (1984). Identifying career offenders with self-reported data. *Journal of Research in Crime and Delinquency*, *21*, 57–86.

D'Unger, A. V., Land, K. C., McCall, P. L., & Nagin, D. S. (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses. *American Journal of Sociology*, *103*, 1593–1630.

D'Unger, A.V., Land, K.C. & McCall, P.L. (2002). Sex differences in age patterns of delinquent/criminal careers: Results from poisson latent class analyses of the Philadelphia cohort study. *Journal of Quantitative Criminology*, *18*, 349–375.

Eggleston, E. P., Laub, J. H., & Sampson, R. J. (2004). Methodological sensitivities to latent class analysis of long-term criminal trajectories. *Journal of Quantitative Criminology*, *20*, 1–26.

Farrington, D. P., Coid, J. W., Harnett, L., Jolliffe, D., Soteriou, N., Turner, R., & West, D. J. (2006). *Criminal careers up to age 50 and life success up to age 48: New findings from the Cambridge Study in Delinquent Development*. Unpublished report available from the Home Office Research, Development and Statistics Directorate, 2 Marsham Street, London, UK, SW1P 4DF.

Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*, 271-320.

Hélie, S. (2006). Introduction to model selection: Tools and algorithms. *Tutorials in Quantitative Methods in Psychology*, *2*, 1–10.

Jones, B. (2001), The *Proc Traj* SAS procedure. Latest version available at: http://www.andrew.cmu.edu/user/bjones/

Jones, B.L., & Nagin, D.S. (2007). Advances in group-based trajectory modeling and a SAS procedure for estimating them. *Socioloical Methods & Research*, *35*, 542–571.

Jones, B.L., Nagin, D.S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories,*Sociological Methods & Research*, *29*, 374–393. Available at: http://www.andrew.cmu.edu/user/bjones/ref1.pdf

Kass, R.E., & Raferty, A.E. (1995). Bayes factors. *Journal of the American Statisitcal Association*, *90*, 773-795.

Kreuter, F., & Muthen, B. (2008), Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology*, *24*, 1–31.

Lambert, D. (1992), Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34* 1–14.

Loughran, T., & Nagin, D.S. (2006), Finite sample effects in group-based trajectory models. *Socioloical Methods & Research*, *35*, 250–278.

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Muthén, B., & Muthén, L. (2001), Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

Nagin, D.S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nylund, K.L. Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569.

Piquero, A. R., Blumstein, A., Brame, R., Haapanen, R., Mulvey, E. P., & Nagin, D. S. (2001). Assessing the impact of exposure time and incapacitation on longitudinal

trajectories of criminal offending. *Journal of Adolescent Research*, *16*, 54–74.

Piquero, A.R., Farrington, D.P. & Blumstein. A. (2007). *Key issues in criminal career research*. Cambridge, UK: Cambridge University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Smith, D. R. Smith, W. R., & Norma, E. (1984). Delinquent career-lines: A conceptual link between theory and juvenile offenses. *The Sociological Quarterly*, *25*, 155–172.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistcal Society*, *Series B 36*, 111–147.

Ward, A. K., Day, D. M., Bevc, I., Sun, Y., Rosenthal, J. S., & Duchesne, T. (2010). Criminal trajectories and risk factors in a Canadian sample of offenders. *Criminal Justice and Behavior*, *37*, 1278–1300.

Yessine, A.K., & Bonta, J. (2009), The offending trajectories of youthful aboriginal offenders. *Canadian Journal of Criminology and Criminal Justice*, *51*, 435–472.

Table 1. Mean (SD) characteristics of the TO1 and TO2 juvenile offender datasets.

Variable	TO1 ($N = 378$)	TO2 ($N = 386$)
End of Follow-up Period:	Sept. 26, 2007	Sept. 26, 2007
Age at admission into youth home	17.6 years (.85)	17.7 years (1.0)
Sentence length at youth home	124.6 days (109.8)	122.6 days (95.6)
Length of follow-up	18.7 years (3.0)	16.4 years (4.1)
Age at first court contact	15.5 years (1.8)	15.6 years (1.6)
Age at last court contact	26.1 years (5.5)	24.6 years (5.2)
Age at end of follow-up	34.1 years (2.6)	32.0 years (4.0)
Trajectory length	10.7 years (5.6)	9.5 years (5.6)

Table 2. Computations of the maximum log-likelihood, AIC, and BIC, using each of the software packages *Proc Traj*, *Mplus*, and *crimCV*, on the TO1 and TO2 juvenile offender datasets with the quadratic-quadratic pure-ZIP model, for various numbers of latent groups (ngr). For small ngr, the three software packages tend to agree. However, for larger numbers of groups, *crimCV* consistently finds larger (less negative) maximum log-likelihood values, and correspondingly smaller AIC and BIC values, than do *Proc Traj* and *Mplus* (and in some cases, *Proc Traj* fails to converge entirely). This illustrates that *crimCV* does a superior job of maximising the log-likelihood function.

TO1:

ngr	llike			AIC			BIC		
	Proc Traj	*Mplus*	crimCV	*Proc Traj*	*Mplus*	crimCV	*Proc Traj*	*Mplus*	crimCV
1	−13756	−13756	−13756	27524	27524	27524	27568	27568	27568
2	−13756	−12133	−12133	27524	24293	24293	27568	24389	24389
3	−11604	−11604	−11555	23249	23249	23150	23396	23396	23298
4	−11254	−11295	−11254	22561	22644	22561	22760	22843	22760
5	−11106	−11123	−11062	22281	22314	22192	22531	22565	22443
6	−11051	−10950	−10934	22184	21981	21949	22486	22347	22251
7	−10880	−10904	−10846	21857	21904	21788	22210	22126	22141
8	−11229	−10811	−10757	22568	21731	21624	22973	22137	22029
9	failed	−10770	−10711	failed	21663	21549	failed	22120	22003

TO2:

ngr	llike			AIC			BIC		
	Proc Traj	*Mplus*	crimCV	*Proc Traj*	*Mplus*	crimCV	*Proc Traj*	*Mplus*	crimCV
1	−11239	−11239	−11239	22490	22490	22490	22535	22535	22535
2	−10119	−10119	−10108	20264	20264	20241	20359	20359	20337
3	−9613	−9572	−9572	19266	19183	19183	19413	19331	19331
4	−9619	−9387	−9387	19292	18829	18829	19491	19028	19028
5	−9341	−9252	−9236	18749	18571	18540	18999	18821	18790
6	−9183	−9184	−9149	18448	18451	18381	18750	18752	18682
7	failed	−9068	−9064	failed	18232	18224	failed	18585	18577
8	failed	−9038	−8980	failed	18185	18070	failed	18590	18475
9	failed	−8951	−8928	failed	18026	17980	failed	18482	18437
10	failed	−8937	−8901	failed	18011	17940	failed	18519	18448

Table 3. Maximum log-likelihood, and the three optimality criteria (AIC, BIC, and CVE), for the datasets TO1 and TO2 with the cubic-ZIP(τ) model, for various numbers of latent groups. The smallest value for each of the three criteria is written in **boldface**. In particular, note that AIC and BIC each continue to shrink and thus advocate more and more latent groups without limit, while CVE reaches a minimum value and then increases again. (We end the runs after 9 and 8 groups, respectively, since after that the models *saturate* with one of the group probabilities becoming zero, so there are no further meaningful results to be obtained.)

TO1:

ngr	llike	AIC	BIC	CVE
1	−13967.63	27945.26	27982.26	1.0902792
2	−11929.40	23880.81	23962.22	0.9128347
3	−11424.68	22883.37	23009.18	0.9592355
4	−11191.28	22428.55	22598.77	0.9052791
5	−11016.19	22090.37	22304.99	0.8535441
6	−10886.30	21842.61	22101.63	0.8334242
7	−10805.59	21693.18	21996.60	0.8261734
8	−10732.58	21559.16	21906.99	**0.8123785**
9	−10684.54	**21475.08**	**21867.31**	0.8240060

TO2:

ngr	llike	AIC	BIC	CVE
1	−11700.095	23410.19	23447.45	0.7731111
2	−10183.745	20389.49	20471.47	0.6698342
3	−9717.829	19469.66	19596.35	0.6529144
4	−9541.180	19128.36	19299.76	0.6364674
5	−9377.273	18812.55	19028.66	0.6171476
6	−9302.665	18675.33	18936.16	0.6122683
7	−9223.541	18529.08	18834.63	**0.6108820**
8	−9165.115	**18424.23**	**18774.49**	0.6271084

Figure 1. Values (top) and percentage changes (bottom) of BIC (first and third columns) and CVE (second and fourth columns) for the TO1 (first two columns) and TO2 (last two columns) samples, illustrating that CVE is minimised at 8 and 7 groups, respectively, but that the changes are less significant as the number of groups gets large.

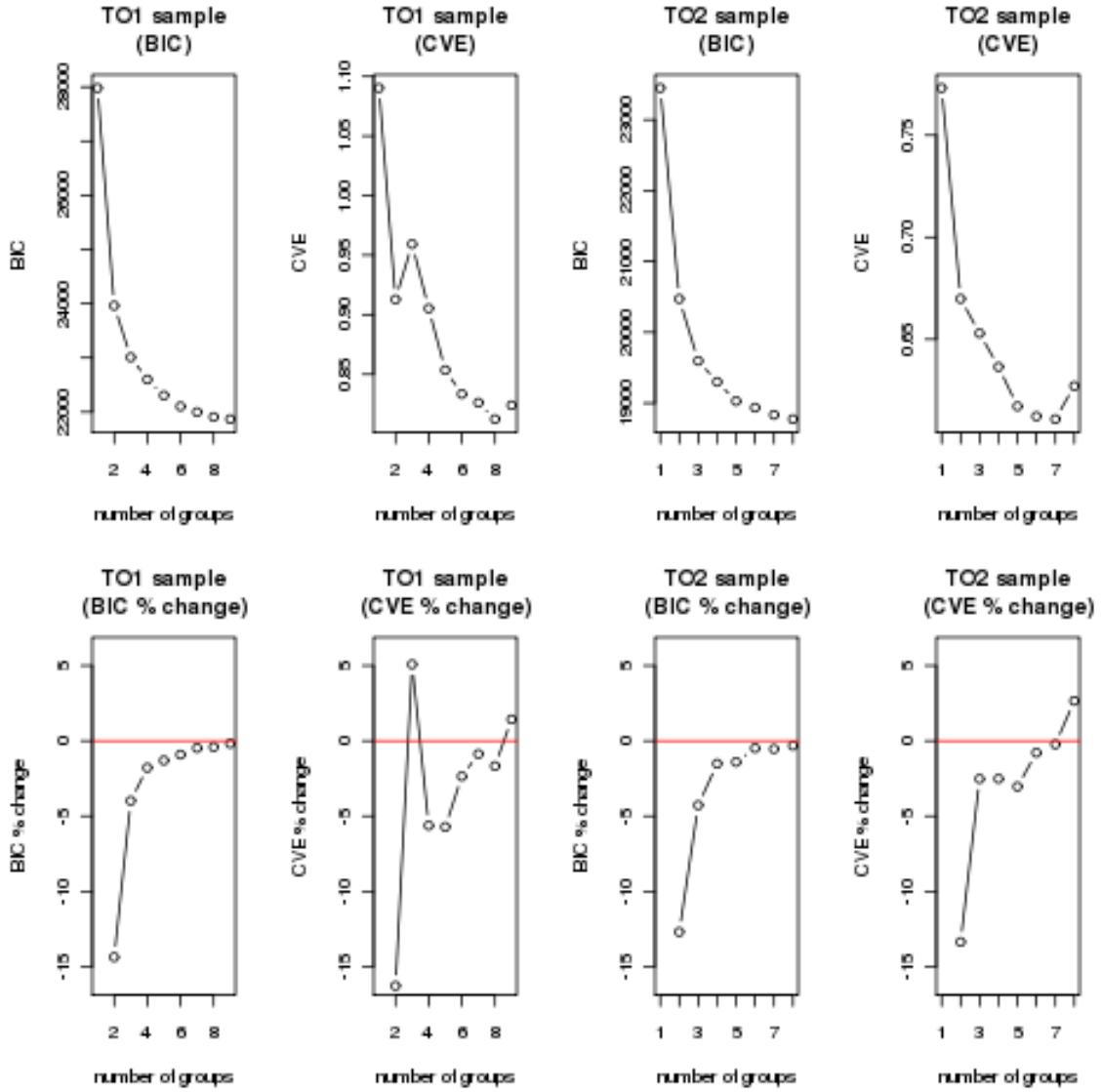

Figure 2. Estimated Criminal Trajectories for the Eight-Group Model for TO1.

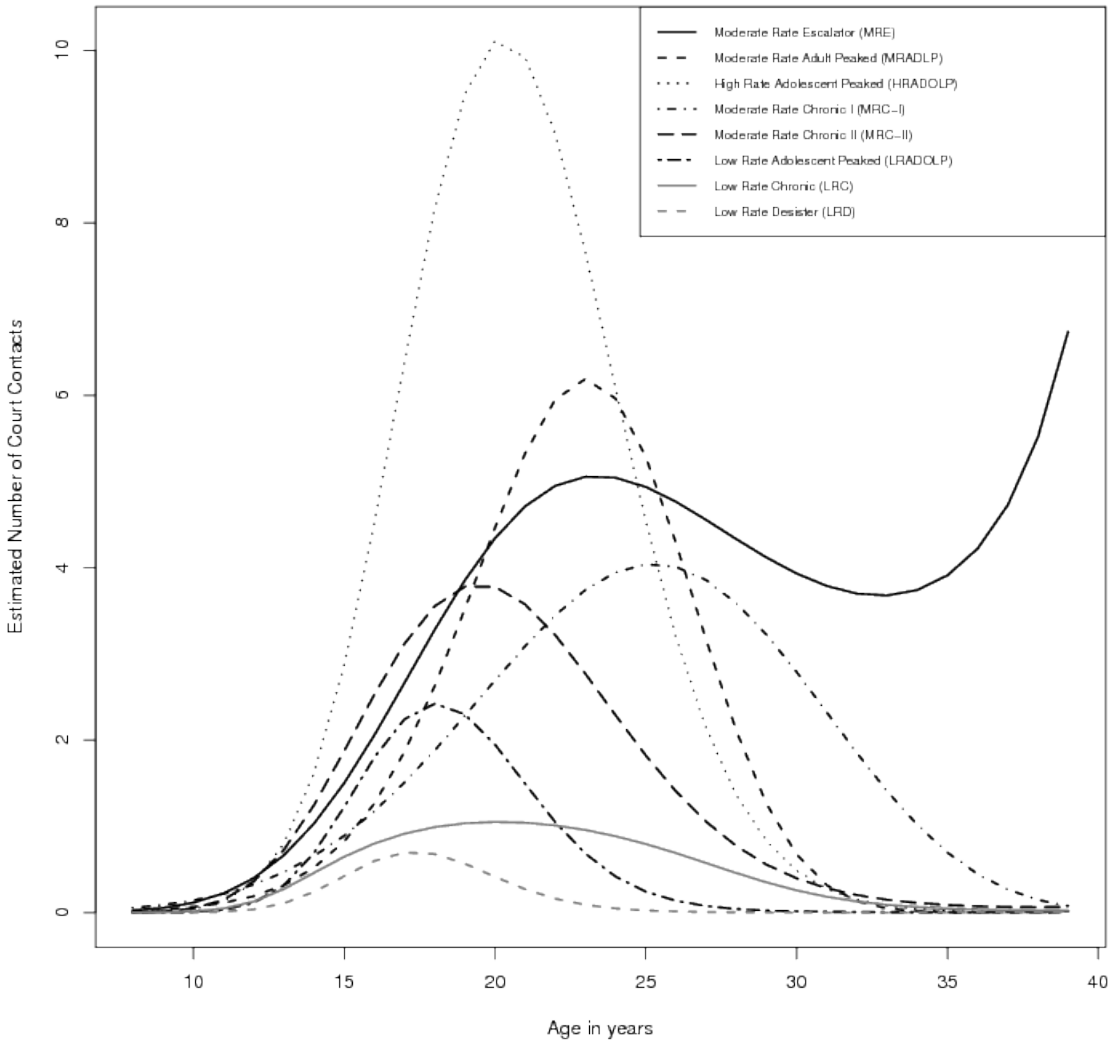

Figure 3. Estimated Criminal Trajectories for the Seven-Group Model for TO2.