Learning from multiple analogies: an Information Theoretic framework for predicting criminal recidivism

Bhati, Avinash

2007

Learning from Multiple Analogies: An Information Theoretic Framework for Predicting Criminal Recidivism

Avinash Singh Bhati*

Paper to be presented at
the 2007 ASC Annual Meeting, Atlanta GA

Abstract

If recidivism is defined as rearrest within a finite period following release from prison, then the kinds of outcomes typically available to researchers include: (i) whether or not the individual was rearrested within the follow-up period; (ii) how many times the individual was rearrested; and (iii) what was the duration from release to first (or subsequent) rearrest. Since these outcomes are all different manifestations of the same underlying stochastic process, they provide multiple analogies from which to recover information about it. This paper develops a semi-parametric approach for utilizing information in these, and several other related outcomes, to predict criminal recidivism and presents preliminary findings.

1. OVERVIEW

The statistical analysis of criminal recidivism performs two critical functions within the field of criminal justice research. First, it is typically a key measure

*Senior Research Associate, Justice Policy Center, The Urban Institute, 2100 M Street, NW, Washington, DC 20037.

used for evaluating the efficacy of crime control policy and practice (Lipton, Martinson and Wilks 1975; Sherman et al. 1997). Second, it is used almost exclusively as the measure with which to develop and validate actuarial risk assessment instruments for offender populations (Glaser 1955; Gottfredson and Gottfredson 1986). The contribution that the analysis of criminal recidivism can make to our understanding of criminal justice policy and practice, therefore, cannot be overstated.

Criminal recidivism can best be thought of as a process that results in one or more events. Defined nominally as "the reversion of an individual to criminal behavior after he or she has been convicted of a prior offense, sentenced and (presumably) corrected" (Maltz 1984, pg 1), the empirical study of recidivism usually requires operationalizing, measuring, and modeling this concept.

The first of these components—operationalization—is informed largely by the purpose of the analysis. What do we wish to study when we analyze criminal recidivism? The answer to this is theoretically motivated and typically leaves little room for an empirical analyst's subjective interpretation; either the operationalization is appropriate or is not. The second component—measurement—is a function of the availability of accurate measures of an appropriate operationalization. Again, there is little room for analysts' subjective assessment here; either measures are accurate or they are not. It is at the modeling stage—the last component—where an analyst confronts a crucial choice: If a process yields several manifestations, which one(s) to analyze? This seemingly innocuous choice can play a surprisingly influential role in the ensuing results.

This paper seeks to demonstrate the utility of an information theoretic framework for constructing and estimating semi-parametric models of criminal recidivism which build on the fundamental recognition that multiple manifestations of a process provide multiple analogies about it. As such, models consistent with several related analogies ought to yield keener insights into the process.

Although the framework builds on contributions from two fairly voluminous literatures in the social sciences—*Information Theory* and *Event-History*

Analysis—it offers more than a marriage of these two sub-fields. It makes novel contributions to each of these sub-fields independently, while advancing the state-of-the-art in predicting criminal recidivism. Preliminary findings, using a real world data set, suggest that the analytical strategy works well in within-sample, out-of-sample, as well as off-the-support prediction problems.

The rest of this paper is organized as follows. The next section motivates the work and reviews the relevant literature. Following that, the information theoretic framework is described. To demostrate the potential of the analytical strategy, the paper then discusses an empirical application. As the work is ongoing, only preliminary findings are presented and discussed. The paper concludes with a brief discussion of the findings and enumerates promising directions for future work. Technical details are provided in a mathematical appendix.

2. MOTIVATION AND BACKGROUND

Suppose re-arrest within a finite period following some punishment is an appropriate operationalization of recidivism and suppose an analyst is able to obtain accurate dated re-arrest information from a police department's electronic data system. How should (s)he proceed? The statistics literature offers several seemingly distinct approaches for analyzing the evidence. For example, the analyst could dichotomize the recidivism measured within the follow-up period and analyze that as a binary choice. Similarly, the analyst could count up the number of times the individual recidivates within the follow-up period and analyze that as a count measure. The analyst could also compute the duration from release to first (or subsequent) re-arrest and analyze that using survival analysis techniques. Despite the general recognition that, mathematically, these outcomes (and, it can be shown, several others) are all related to one another, current practice (and most software) require analysts to choose among them. Scholars not comfortable making that choice sometimes analyze several of the outcomes independently and, not surprisingly, can reach conflicting conclusions (Mitchell and Moore 2002).

This is an unfortunate situation for the empirical analyst since these outcomes (binary, count, and duration, among others) are all different manifestations of the same underlying stochastic process thereby providing multiple analogies from which to recover information about it. The main motivation of this research effort is to develop and assess the utility of an information theoretic framework for incorporating information from several related manifestations—the multiple analogies—into a single model describing the stochastic process of criminal recidivism. Presumably, the more structure we can build into this model, the keener its insights will be regarding the process under study.

Scholars from most social science disciplines recognize well that event history data can be presented and analyzed in numerous ways (Kalbfleisch and Prentice 1980; Allison 1982, 1984; Mayer and Tuma 1990; Lancaster 1990; Yamaguchi 1991; Beck 1998; Box-Steffensmeier and Jones 2004). They also recognize that these seemingly disparate approaches are all connected through the underlying *hazard* (failure, arrival, or event) rate. Texts or other expositional materials, for example, typically begin with an explication of the hazard rate and its various transformations—the survival rate and event probability function. Monographs covering criminal recidivism are no exception (Maltz 1984; Schmidt and Witte 1988).

Recently Alt, King, and Signorino (2001) have explicitly pointed out the connection between the choice of outcome analyzed and the ensuing inferences. Their work builds on an extensive literature dealing with aggregation bias and the ecological inference problem (King 1977; King, Rosen and Tanner 2004). Their main concern, however, is that "they [we] do not want the form in which the data is collected to influence the substantive idea they [we] can explore" (Alt, King, and Signorino 2001, pg 22). As a solution, the authors define models that—under some special conditions—allow them to recover the same underlying parameters regardless of the manifestation analyzed. But what if analysts have access to multiple manifestations? Surely, any attempt to use only one or two of them in isolation would be a tremendous waste of available evidence.

To be sure, methods for including information in more than one manifestation of a process do exist. For example, researchers oftentimes estimate zero-inflated or hurdle models of event-counts that analyze count and binary manifestations of the events simultaneously (Cameron and Trivedi 1998; Greene 2000). Similarly, an impressive array of models dealing with inter-event duration dependence in event-count models have been invented (King 1989; Winkelmann 1995); as have methods for combining duration information in recurring failure events (Lin, Wei and Ying 1998; Ezell, Land and Cohen 2003).

This ability to incorporate information in multiple manifestations comes at a price, however. Some restrictive assumptions must be tolerated. Unfortunately, the estimation and inferential implications of these assumptions are not benign (Dean and Balshaw 1997), especially as they relate to the form of unobserved heterogeneity (Heckman and Singer 1984a,b). In order to develop and estimate models that are robust to some of these limitations—while, at the same time, able to incorporate information from multiple analogies—it would be helpful to rely less on unverifiable assumptions and more on a full utilization of all available evidence. Such a strategy would approach crucial aspects like proportionality and duration dependence in an agnostic fashion and be more concerned about profitably utilizing all available manifestations of the process. One such approach is explained next.

3. ANALYTICAL STRATEGY

Consider, as a point of departure, the following scenario. A cohort of individuals is released from prison and followed for a period of T years. We wish to study their failure process to better understand and predict this behavior. Suppose we define re-arrest as "failure" and suppose we have available re-arrest dates for each individual through the follow-up period. Let us ignore, for the moment, the problems associated with re-incarceration (when individuals are taken

off the street and are therefore not at risk of failing again).[1] Let us also assume explicitly, as is typically done implicitly, that the stochastic process we are interested in studying does not change during the follow-up period. If the individual is re-arrested within the follow-up period T, then let b denote a binary outcome coded 1; let c denote the number of times the individual is re-arrested; and let d denote the duration to the first re-arrest event. If the individual is not re-arrested during this period, let each of these manifestations be set to 0. Now, define

$$y(t) = 1[t = d] \qquad \text{and} \qquad f(t) = 1[t \leq \min(d, T)] \qquad \forall t \in \mathscr{T}$$

where $\mathscr{T} = \mathbb{R}_+$ and $1[\cdot]$ is an indicator function returning 1 if the condition inside $[\cdot]$ is satisfied, else 0. Consequently, $y(t)$ is simply a function flagging when the event actually occurs and $f(t)$ is a function flagging when the event is at risk of occurring.[2]

Suppose, next, that we define $r(t)$ as the unknown hazard that reflects the stochastic process resulting in the event flagged by $y(t)$. Since an individual cannot fail if (s)he is not at risk of failing, we can use both $y(t)$ and $f(t)$ to derive conditional links between the hazard and the event as:

$$f(t)y(t) \approx f(t)r(t) \qquad \forall t \in \mathscr{T}. \tag{1}$$

Besides yielding the familiar non-parametric hazard rate estimates,[3] this approximation allows us to derive analogies between the hazard rate and several of its manifestations.

[1] This is for expositional purposes only. The framework may be readily adapted to account for such *not at risk* spells, should they exist.

[2] By altering the definition of $y(t)$ and $f(t)$ we can characterize multiple events and by redefining $f(t)$ appropriately we can characterize spells when an individual is not at risk of experiencing the event. For ease of exposition, these nuances are omitted here.

[3] Assuming that the hazard is fixed across individuals, taking the unconditional expectation of (1) and re-arranging terms yields $\hat{r}(t) = \mathbb{E}[f(t)y(t)]/\mathbb{E}[f(t)]\ \forall t \in \mathscr{T}$. This is a non-parametric estimate of the hazard rate—the number of people expected to fail at t divided by the number of people expected to be at risk of failing at t.

3.1. Identifying Suitable Analogies

First, suppose that we integrate both sides of (1) over the domain $\mathscr{T}$ and assume that this procedure converts the approximation into an equality. Since $y(t) = 1$ if and only if $f(t) = 1$, clearly, this integration will yield the binary outcome b on the left hand side. Hence, this procedure yields our first analogy linking the hazard to a manifestation.

$$b = \int_{\mathscr{T}} f(t) r(t) \, dt$$

Next, consider, pre-multiplying both sides of (1) by t and then taking the integral. The left hand side of this equation would now yield d—the duration to first re-arrest (and 0 if the observation is censored)—since the only time when $f(t) = y(t) = 1$ is when $t = d$. Consequently, we would have identified another analogy.

$$d = \int_{\mathscr{T}} t \, f(t) r(t) \, dt$$

Of course, there may be reason to believe that the hazard of recidivism is independently affected by a stochastic process that progresses with age (e.g., the age-crime curve). Hence, the age at first rearrest event, and not duration to first rearrest event, may be the more appropriate quantity to model. Denoting age at release as g, we define $a = g + d$. Multiplying both sides of (1) by $a + t$ and integrating yields another analogy linking the hazard to age at first rearrest.

$$a = \int_{\mathscr{T}} [g + t] f(t) r(t) \, dt$$

Quadratic or cibic transformation of an outcome like age at first rearrest can also be introduced in a parallel fashion.

Next, let us consider the number of rearrests within a follow-up period. If the duration to an event is η years,[4] what knowledge does that provide about the

[4]In this paragraph, I resort to using the notation η to denote duration to first event instead

number of events likely to be accumulated through some follow-up period T?

To motivate this link, note that an increase in the duration to first re-arrest event can be expected to reduce the total number of events (rearrests) that an individual should accumulate over any fixed follow-up period. Moreover, this decrease should be proportional to the annual offending rate. Why? Because if duration to the first event increases by one year, then the total number of events that can be accumulated should be reduced by the re-arrest rate of that one year. A crude proxy for the annual offending rate, on the other hand, can be obtained by inverting the duration to first event. In other words, we can postulate the following first order differential equation connecting the total number of rearrests to the duration to first rearrest:

$$\frac{dc}{d\eta} = -\frac{1}{\eta}$$

Solving this equation by integration yields

$$\begin{aligned} c &= -\int \frac{1}{\eta} d\eta \\ &= -\log(\eta) + \varkappa \end{aligned}$$

where $\varkappa$ is the constant of integration. Noting the terminal condition that $c = 1$ if $\eta = T$, we can solve for the constant of integration to get $\varkappa = 1 + \log(T)$. Reverting back to the original notation d to denote duration to first event, we can now write

$$c = 1 + \log \frac{T}{d}.$$

This link between c and d allows us to derive another analogy. Suppose we multiply both sides of (1) by $1 + \log(T/t)$ and then integrate over the relevant domain. The left hand side would yield c and we would have derived another

of d. Otherwise, the differential of the duration variable ends up being denoted dd.

analogy between the hazard and a manifestation.

$$c = \int_{\mathscr{T}} [1 + \log(T/t)] f(t) r(t) \, dt$$

Just as age at first rearrest event could be derived from the duration variable, one can derive another analogy relating the hazard to the total number of events, since career initiation, accumulated by the end of the follow-up period. Denoting the number of prior (pre-release) arrests by h, the total number of arrests by follow-up as e, and multiplying bother sides of (1) by $h + 1 + \log(T/t)$, we obtain

$$e = \int_{\mathscr{T}} [h + 1 + \log(T/t)] f(t) r(t) \, dt.$$

In general, of course, we can derive a host of other analogies. Given that these outcomes are different manifestations, at least potentially, of the same underlying hazard process, let us generically denote the set of analogous claims, say J of them, as:

$$\mu_j = \int_{\mathscr{T}} \phi_j(t) f(t) r(t) \, dt \qquad \forall j \in J \tag{2}$$

where $\phi_j(t)$ are appropriate transformation of t and μ_j are the corresponding manifestations. Provided that the analogies satisfy the basic identifying restriction—that none of them are exactly implied by, or imply, another—each of them provides information about a different piece of the model that we are attempting to construct.

The analogies derived above merely provide restrictions on the shape and values that the hazard function can take. We still need some agnostic way to recover information from them (i.e., learn from them without making too many assumptions). Fortunately, information theory provides a foundation from which to approach this problem.

3.2. Learning from Multiple Analogies

Information theory builds on the pioneering work of Shannon (1948). He derived a measure of uncertainty—which he called *Information Entropy*—for quantifying a channel's capacity to communicate information. Faced with the problem of inferring individual features from aggregate properties, Edwin Jaynes, another pioneer in this field, proposed to use Shannon's Information Entropy as an agnostic criterion to maximize (since it measures uncertainty) in order to be very conservative in what we can (or cannot) infer from these aggregate properties (Jaynes 1957a,b). Viewing an experiment (or a sample) as a communication device, the Maximum Entropy procedure—as it has come to be known—is therefore a very general and powerful procedure for learning from statistical evidence (e.g., the type of analogies we have derived above).

The links between Information Theory and statistics has been very thoroughly explored (Diamond 1959; Kullback 1959; Jaynes 1979, 1986, 1988; Justice 1986; Levine and Tribus 1979; Mathai 1975; Skilling 1989; Zellner 1988; Soofi 1994, 2000). Since Shannon's measure of uncertainty was probabilistic, naturally, much of this literature develops and uses measures of information based on proper probabilities. However, if we are to learn from analogies of the type defined in (2), what we need is a measure of information that is based on the hazard rate.[5]

There is a growing statistical literature utilizing information theoretic concepts in reliability analysis (Ebrahimi, Habibullah and Soofi 1992; Soofi, Ebrahimi and Habibullah 1995; Ebrahimi and Kirmani 1996; Ebrahimi and Soofi 2003; Asadi et al. 2005). These scholars derive hazard models by utilizing the links between the hazard rates and probability functions (or survival rates) thereby converting the information-recovery problem about the hazard into one about proper probabilities. Unfortunately, this strategy is less than helpful in our cur-

[5]Some measures of information relying on positive quantities (that do not integrate to 1) have been informally proposed in the literature. They are used, for example, in image reconstruction problems (Gull and Daniell 1978; Gull 1989; Donoho et al. 1992) or for recovering regression functions (Ryu 1993).

rent situation since any transformation of the derived analogies would result in intractable transformation of the manifestations themselves (μ_j). We need a criterion that measures information in the hazard rate directly.

Denoting $\bar{r}(t)$ as a prior (pre-sample or pre-experiment) belief about the hazard rate, and using a simple set of plausibility assumptions, one can derive such a measure (see the mathematical appendix). Other than a constant scaling factor, the *net information acquired by the analyst* in terms of the hazard rate itself can be computed as:

$$I = \int_{\mathcal{T}} f(t)\left[r(t)\log\frac{r(t)}{\bar{r}(t)} - r(t) + \bar{r}(t)\right] dt. \tag{3}$$

The *inferential* task of learning from multiple analogies can now be converted into the *mathematical* problem of minimizing (3), subject to the constraints (2). This is a standard variational problem that can be solved by the method of lagrange. The primal objective function is set up as

$$\begin{aligned}\mathcal{L} \;=\; & \int_{\mathcal{T}} f(t)\left[r(t)\log\frac{r(t)}{\bar{r}(t)} - r(t) + \bar{r}(t)\right] dt \\ & + \sum_j \beta_j \left[\mu_j - \int_{\mathcal{T}} \phi_j(t) f(t) r(t)\, dt \right]\end{aligned}$$

where β_j are the lagrange multipliers associated with each of the J constraints. Solving the first order conditions provides the solution

$$r(t) = \bar{r}(t)\exp\big(\sum_j \phi_j(t)\beta_j\big) \qquad \forall t \in \mathcal{T} \tag{4}$$

and setting $\bar{r}(t) = 1 \forall t \in \mathcal{T}$ removes the possibility of analyst-induced subjectivity by making the priors completely uninformative. This solution can be used to derive a dual representation—an *un*constrained optimization problem in β_j—that can be solved using standard software (e.g., SAS or GAUSS). The dual

(unconstrainted) optimization problem is

$$\mathscr{F} = \sum_j \beta_j \mu_j - \int_{\mathscr{T}} f(t) r(t)\, dt + \int_{\mathscr{T}} f(t) \bar{r}(t)\, dt \tag{5}$$

where $r(t)$ is as derived in (4). Note also that since $\bar{r}(t)$ is not a function of any of the β_j, the last component of the objective function is really irrelevant in the optimization problem.

Individual attributes may be introduced into the strategy in a straightforward manner by replacing the μ_j with the products of individual manifestations and attributes (e.g., $\mu_{jn} x_{kn}$); by introducing subscripts of n (e.g., $r_n(t)$ and $f_n(t)$); and by summing the dual over all individuals. The dual objective with individual attributes included is defined as

$$\mathscr{F} = \sum_n \left\{ \sum_{kj} \beta_{kj} \mu_{jn} x_{kn} - \int_{\mathscr{T}} f_n(t) r_n(t)\, dt + \int_{\mathscr{T}} f_n(t) \bar{r}_n(t)\, dt \right\} \tag{6}$$

where each individual's hazard solution (path) is now defined as

$$r_n(t) = \bar{r}_n(t) \exp\left(\sum_j \phi_{jn}(t) \sum_k x_{kn} \beta_{kj} \right) \qquad \forall t \in \mathscr{T}. \tag{7}$$

The unconstrained maximization problem derived above falls under the general class of extremum estimators, $\hat{\beta} = \arg\max_\beta \mathscr{F}(\beta, \mu, \mathbf{X})$. The consistency and asymptotic normality of these estimators can be established under fairly general regularity conditions (Mittelhammer, Judge, and Miller, 2000:132–139).

Assuming that standard regularity conditions are met, one way to conduct hypothesis tests is to construct and use the Entropy Ratio Statistic ($\mathscr{E}$). Since the value of the objective function measures the amount of uncertainty implied by the hazards, we can assess the *uncertainty reducing contribution* of each (or groups) of the associated parameters by comparing the values of the objective function from restricted and unrestricted models. Like the Likelihood Ratio statistic, the Entropy Ratio statistic has a limiting χ^2 distribution with R de-

grees of freedom (Jaynes, 1979:67). R being the number of parameters that have values fixed, either to 0 or to some other value. Denoting $\mathscr{F}_*$ and $\mathscr{F}$ as the values of the dual objective function for the restricted and unrestricted models respectively, we can compute $2 \times [\mathscr{F}_* - \mathscr{F}] = \mathscr{E} \sim \chi^2_R$ to test whether or not specific parameter(s) contribute significantly in reducing uncertainty about the structure in the data.

In a similar manner, one can obtain an estimate of the asymptotic covariance matrix of the Lagrange Multipliers by computing the negative inverted Hessian of the dual objective function. This covariance matrix can then be used to assess the stability of each of the Lagrange Multipliers without needing to estimate restricted and unrestricted versions of the models.

4. EMPIRICAL APPLICATION

The model derived in the last section was estimated and assessed using the 1994 BJS Recidivism Study (ICPSR # 3355), which provides dated criminal activities of roughly 38,000 prisoners released from 15 state prisons in 1994 (Langan and Levin 2002; BJS 2002). The data set records up to 99 dated arrest events for each of the released prisoners—including pre-incarceration as well as post-release arrests (for a follow-up period of at least three years). This allows for the computation of several manifestations of the stochastic process under study (e.g., re-arrested within the follow-up period, number of times re-arrested, duration to first re-arrest, age at first re-arrest, number of arrests accumulated from birth through the follow-up period, criminal career length, among others).

For the preliminary findings reported here, only a few explanatory variables were explored. These include age at first arrest, age at prison release, and number of prior arrests at prison release. Note that the first of these is the only truly explanatory variable used. Age at release and number of priors are part of the manifestations utilized in the analysis. More detailed analysis, using a variety of predictors, is currently lunder way.[6]

[6]This includes models that include detailed information for each releasee on the type of

Table 1: Descriptive Statistics of the California and Florida Samples Used in the Analysis.

	California		Florida	
	Mean	Median	Mean	Median
Full Sample (N)	5,773	...	2,134	...
Age at release (yr)	36	35	35	33
Age at 1st Arrest (yr)	26	22	24	21
Number of Prior Arrests	7	4	8	6
2-yr Recidivism Rate (%)	42	...	55	...
3-Yr Recidivism Rate (%)	49	...	63	...
3-Year Recidivists Sample (N)	2,904	...	1,353	...
Age at 1st Re-arrest (yr)	34	34	33	32
Number of Rearrests	2	2	3	2
Duration to 1st Rearrest (yr)	0.99	0.76	0.95	0.75

To make the estimation problem feasible (and realistic), data from only the state of California were used for estimating the model. The models, once estimated, were validated using data from the state of Florida. These two states were selected from the 15 states included in the underlying data for two reasons. First, they were the largest states in this dataset (in terms of unweighted sample sizes). Second, the unweighted three-year recidivism rate (one of the chief criterion variables) was roughly 50% for the California sample but roughly 65% for the Florida sampla. This offers some insights into the out-of-sample predictive performance of the strategy on a validation sample somewhat more criminogenic than the estimation one. Other than that, the two samples seems very similar (at least in terms of mean characteristics). Table 1 provides a prief summary of the underlying data used for the two states. With the exception of the number of prior arrests before release, for which Florida seems a bit higher than California, the other predictors and manifestations seem very similar across the two states.

prison admission, type of release, time served in prison, individual demographic attributes, and details pertaining to the offense for which the releasee was incarcerated.

A serious limitation of the BJS Recidivism data is the lack of information on the amount of time that the offender may have served in prison in several incarceration episodes prior to the current one or post-release. Despite the availability of adjudication information for each of the arrest cycles, the data do not contain clear information on the amount of time persons may have served in prison if they were incarcerated. This problem is not peculiar to this study alone. Accounting for street-time is a difficult matter in all retrospective or prospective longitudinal designs. The BJS data do, however, contain adjudication outcomes of each of the cycles and the type of adjudication at an arrest event. This information is currently being used to derive additional relevant analogies for inclusion in the model as well as to include/exclude individuals from the risk set (i.e. to define $f(t)$ appropriately). The work is ongoing and this measure is not included in the analysis reported here.

4.1. Model Estimates

Table 2 provides estimates of the lagrange multipliers from the various analogous constraints imposed to derive the model. Using the negative inverted hessian, a standard error was computed and used to derive the Wald χ^2 statistic. The statistic indicates that almost all of the constraints have informational content. That is, they provide *statistically significant* information about the process under study. The single explnatory variables used in the modeling exercise—age at first arrest—seems to indicate that offenders who initiate their careers later in life, compared to those who start earlier, have parmanently lower hazards (negative coefficient under b); have an upward pressure on the evolution of the hazard with age (positive coefficient under a), but at a decreasing rate (negative coefficient under a^2); seem to have higher numbers of crimes accumulated by the end of the follow-up period, both since release as well as since career initiation (positive coefficients under c and e); and seem to experience an upward pressure on the evolution of the hazard with time since release. Note that all of these effects enter the hazard model simultaneously, along with the intercept terms. Hence,

Table 2: Multiple Analogy Models of Criminal Recidivism in California, Parameter Estimates and Statistical Significance for Two-year and Three-year Follow-up Models.

	2-Year Follow-up			3-Year Follow-up		
Manifestations Predictors	Lagrange Multiplier	Wald χ^2 Statistic	p-value	Lagrange Multiplier	Wald χ^2 Statistic	p value
a: Age at 1st rearrest						
Intercept	-0.0453972	328.16	0.00	-0.0571753	627.09	0.00
Age1st	0.0049772	2757.64	0.00	0.0059822	4846.11	0.00
a^2: Age at 1st rearrest squared						
Intercept	-0.0009862	1164.06	0.00	-0.0008453	1009.52	0.00
Age1st	-0.0000243	636.53	0.00	-0.0000344	1492.18	0.00
b: Rearrested at all within follow-up						
Intercept	-2.0000779	35.07	0.00	-4.1362998	144.41	0.00
Age1st	-0.2274956	241.95	0.00	-0.2298919	237.84	0.00
c: Times Rearrested within followup						
Intercept	0.1826634	4.31	0.04	0.7984504	118.71	0.00
Age1st	0.0068398	3.49	0.06	0.0002495	0.01	0.94
d: Duration to 1st rearrest						
Intercept	-0.7115693	17.22	0.00	0.7847820	34.15	0.00
Age1st	0.0177117	6.33	0.01	0.0020134	0.13	0.72
e: Times Arrested since Career Initiation						
Intercept	-0.0327985	14.08	0.00	-0.0281003	11.12	0.00
Age1st	0.0039441	104.80	0.00	0.0036626	96.30	0.00

the hazard path is an aggregation over all of these decompositions.

With few exceptions, the findings are identical across the two follow-up periods. Notably, the effects of Age1st are statistically insignificant for the last three manifestations in the larger follow-up model.

4.2. Model Predictions

The multple analogy models estimated with the California sample were next used to make predictions. The chief criterion that models' predictive efficacy is assessed on is whether they are able to predict failure at the end of the follow-up period (the binary outcome).In each case, the estimated hazard paths were integrated over the relevant domain and used to compute the probability of recidivism using the definition:

$$\Pr(b_n = 1) = 1 - \exp(-\int_{\mathscr{T}} \hat{r}_n(t)\, dt)$$

Moreover, in each case, the mean recidivism rate of the criterion of interest was used to set the cut-off point to convert this predicted probability into a binary classification.

The first set of assessments are for in-sample predictions. That is, models are assessed on their ability to predict the outcomes using the data they were estimated with. These are typically the least challenging of prediction problems. Table 3 provides a cross-tabulation of the models' predictions ($\hat{b}$) versus what was actually observed in the sample (b). The models perform fairly well—both within the 2-year ar the 3-year follow-up periods.

In the 3-year model, for example, of the 5,773 sample members, roughly half (2,853) were rearrested within three years of release. The model predicted roughly 53% (3,093) to be rearrested. Among those predicted to recidivate, two-thirds (66%) were accurate (actually did recidivate) and only 33% were erroneous redictions. Similarly, the models were able to correctly identify 72% of those that were rearrested within the follow-up period, missing about 27% of them.

Table 3: Within Sample Predictive Efficacy of Multiple Analogy Models, California Sample.

	2-year Follow-up				3-year Follow-up		
	$b=0$	$b=1$			$b=0$	$b=1$	
$\hat{b}=0$	2,418	897	3,315	$\hat{b}=0$	1,883	797	2,680
$\hat{b}=1$	953	1,505	2,458	$\hat{b}=1$	1,037	2,056	3,093
	3,371	2,402	5,773		2,920	2,853	5,773

The model in-sample predictive efficacy was somewhat lower at the 2-year window, although it was still good. Of those predicted to recidivate within the 2-year follow up period, nearly 61% did actually recidivate, and the remaining 38% were erroneous (false positives). However, of the 2-year recidivists, the models accurately identified 62% of them, but missed 38% of them.

Considering that the models used a minimal set of predictors—age at release, prior criminal history, and age at first arrest—and, of these, the first two were used to create new *dependent* variables, the predictive accuracy of the models is quite surprising. It can be expected that as additional individual level attributes and, perhaps, demographic attributes are included in the models, they will perform better yet.

Although in-sample accuracy is interesting, the more challenging prediction problems are predicting out-of-sample and off-the-support. Table 4 presents a cross tabulation for assessing the out-of-sample predictive efficacy of the models by using the California model estimates (the Lagrange Multipliers) and generating predictions—both at the 2-year and 3-year follow-up period—for the state of Florida. As one would expect, the models perform worse when estimating out-of-sample. Note the out-of-sample predictions being assessed here are for a different state. This is a different problem—a more realistic one—that taking a random subset of the estimation sample for validation purposes.

Here we find that the models—both at the 2-year and 3-year follow-up periods—under predicted the extent of recidivism. The models predicted that only 438

Table 4: Out-of-Sample Predictive Efficacy of Multiple Analogy Models, California Models Assessed on Florida Sample.

	2-yr Follow up				3-yr Follow up		
	$b=0$	$b=1$			$b=0$	$b=1$	
$\hat{b}=0$	876	820	1,696	$\hat{b}=0$	634	678	1,312
$\hat{b}=1$	87	351	438	$\hat{b}=1$	149	673	822
	963	1,171	2,134		783	1,351	2,134

and 822 individual would fail within two and three years of release, respectively, whereas 1,171 nd 1,351 individuals actually recidivated within these respective follow-up periods. That is, the models predicted recidivism rates nearly half of what was actually observed. Despite that, the news was not all bad. At least at the three year window, despite very low aggregate predictions, the models were fairly good at finding the recidivists. Of those that actually recidivated within three years of release, the models correctly identified about half of them (49%). Similarly, of those few predicted to recidivate, nearly 81% were actually accurate with a false positive rate of only 19%. Therefore, despite predicting a recidivism rate only about half of what was actually observed, the models were fairly accurate in terms of those few that they did identify as recidivists.

The performance of the model was somewhat less encouraging at the two year follow-up period, though. Once again, the models only predicted a recidivism rate about half of the actual rate. However, as in the three year case, of those that the model identified as recidivists, nearly 80% were accurate, with a 20% false positive rate. However, at the two year window, the model missed a large portion of actual recidivists. Nearly 70% of those that did fail were not accurately identifed by the models. In some sense, then, the California models were providing conservative predictions in the Florida sample. The models did not classify nearly enough sample members as recidivists. When they did, however, they were surprisingly accurate.

The most challenging prediction problems are those dealing with off-the-

Table 5: Off-the-Support Predictive Efficacy of Multiple Analogy Models, California 2 year Follow up Models Assessed on 3 Year Recidivism.

Recidivism within 3 years				Recidivism During 3rd year			
	$b=0$	$b=1$			$b=0$	$b=1$	
$\hat{b}=0$	2,080	996	3,076	$\hat{b}=0$	2,080	217	2,297
$\hat{b}=1$	840	1,857	2,697	$\hat{b}=1$	840	234	1,074
	2,920	2,853	5,773		2,920	451	3,371

support predictions. Here, models estimated on a 2-year support were used to make projections for a 3-year window. Table 5 provides cross-tabulation for assessing the predictive efficacy of the California 2-year models for projecting 3-year recidivism in California. The first cross-tabulation uses predictions of the *three-year* recidivism rate were as the second cross-tabulation uses predictions of the *third-year* recidivism rate.

The model projected considerably well when considering the three-year recidivism measure. The two year model projected a slightly lower overall 3-year recidivism rate (47%) than was observed in the sample (49%). Moreover, of the 2,697 individuals predicted to recidivate within three years of release, 68% actually did. About 31% of these projections were erroneous. On the other hand, of the 2,852 individuals who did recidivate within three years of release, the models correctly identified 65% of them. The models only missed about a third of them.

These forecasts, although made for a 3-year window using a 2-year model, are really a blend of in-sample and off-the-support prediction problems. A stricter criterion for assessing these projections is to compare the predictive efficacy of the models, conditional on the individual having survived (not recidivated) by the end of the 2-year follow-up period. The second cross-tabulation in Table 5 provides information for that comparison. The findings are mixed. The models clearly over-predicted the third year recidivism rate. They predicted that

1,074 individuals would recidivate within the third year of release where as only about 451 actually did. With this over-projections also came a high false positve rate. Of the 1,074 individuals predicted to recidivate during the third year, only about 22% actually did. The false positve rate was very high (nearly 81%). On the other hand, the model had a surprisingly decent hit rate. Of the 451 individuals that did recidivate within the third year of release, the models correctly identified 52% of them.

5. CONCLUSION

The goal of this paper was to develop and apply a semi-parametric, information theoretic approach for utilizing knowledge in multiple analogies for studying and predicting criminal recidivism. It was expected that, despite relying on a minimal set of predictors, models consistent with several analogies simultaneously should perform well. Although the relative performance of multiple analogy models—both relative to other types of modeling strategies or to models using fewer/more analogies—is yet to be gauged, preliminary findings presented in this paper suggest that the strategy holds promise.

5.1. Discussion of Findings

Limited analysis using data from the states of California (for estimation and validation) and Florida (used only for validation) were conducted. A minimal set of predictors—the age at release, criminal history, and the age at first arrest—were used in the modeling strategy to simultansously models several outcomes of interest—including age at first rearrest, age at first rearrest squared, a binary indicator of failure within a finite follow-up period, a count of the number of rearrests during the follow-up period, duration to the first rearrest event, and the number of arrests accumulated from the initiation of the career to the end of the follow-up period. Tests of statistical significance suggest that each of the multiple analogies included in the models did infact reduce the analysts uncertainty

about the recidivism process "significantly."

The estimated models were next used to make predictions. Not surprisingly, the in-sample predictive performance of the models were very good. The false positive and false negative rates of the three year model were 34% and 30% respectively and those of the two year model were 38% and 29% repectively. Similarly, the hit rates—proportion of actual recidivists correctly identified—for the three and two year models were, respectively, 72% and 63%.

The estimated California models were also used to predict recidivism among the Florida sample. These out-of-sample predictions were less enocouraging. Despite having acceptible false positive and negative rates (respectively, 18% and 51% for the three year predictions and 20% and 48% for the two year predictions), the hit rate was usually low, particularly for the 2 year predictions (50% and 30% for the three and two year predictions, respectively). The models were conservative, though. They predicted very few recidivists but, among those predicted, the error rate was very low.

Finally, the models were gauged on their ability to predict off-the-support. Here the findings were mixed. The two year models predicted failure within three years of release fairly well. However, they were not that accurate at predicting failure within the third year of release. The two-year models predicted nearly 3 times as many third-year recidivists as there really were in the sample. An unacceptibly large proportion of these were false positives (78%). However, even these predictions had a decent hit rate. The predictions were able to accurately identify nearly half of those that actually did recidivate during the third year.

The analysis conducted here was very limited. However, the preliminary findings are encouraging and offer several insights.

First, the multiple analogy models performed better when estimated using three-year samples than when using the two-year samples. This is not suprising given that each of the analogies have more information when the analyst is able to follow the cohort thorough a longer priod.

Second, the out-of-sample predictions were fairly low. This findings is also not suprising, on hind sight. Table 1 shows that the sample of Florida was very similar to the sample in California, as least with respect to the attributes used. However, the failure rate in the Florida sample was considerably higher than that in the California sample. This suggests, perhaps, that the main differences between California and Florida samples was not the individual attributes of the offenders, but the punitiveness of the system or data definition or collecting peculiarities. Since the modeling strategy does not takes these differences into account, it should be expected that model estimates from California would underpredict recidivism when used on the Florida sample. The systematic differences between states can be accounted for by appropriately defining the priors (in this analysis the $\bar{r}_n(t)$). Such advances are currently being developed. However, it is encouraging that despite the overall underprediction of the phenomenon in Florida, the models were conservative (low false positive rates).

Third, the analysis did not take into account various cut-off criteria to make predictions. Judicious selections of an optimal cut-off point for each model may further improve the predictive performance of the models.

5.2. Ongoing and Future Research

The preliminary set of models developed in this paper were designed to gain some insights into the working of the approach and to gauge its predictive performance. Given the minimal set of predictors included, the models seem to perform surprisingly well. There are several aspects of the strategy that need further exploration.

First, models that include more individual attributes are currently being tested. It should be expected that such knowledge may improve the predictive efficacy of the models as it allows for more flexibility. However, compromising model parsimony is not without cost. Ultimately, the extent to which additional predictors will increase the out-of-sample or off-the-support predictive efficacy of the multiple analogy models is an empirical issue.

Second, other relevant analogies need to be developed that can allow analysts to incorporate knowledge about crime-type-specific recidivism manifestations. For example, a hazard models of general recidivism could have two sub-components—one relating to violent and another to non-violent crimes. Each of these sub-components should yield nuanced insights into the process although the ultimate interest may be in predicting general recidivism. This work is currently under way.

Third, the models, so far, have completely ignored the issues relating to re-incarceration. This needs to be addressed. Although the data source provides less than perfect informationn about this aspect of the model, analogies are currently being developed that will allow the model to learn from such outcomes as whether or not the individual was reincarcerated within the follow-up period, how many times, and the duration to first reincarceration event in addition to outcomes relating to the binary, count, and duration manifestations of rearrest.

Finally, a more elaborate assessment of the predictive efficacy of this modelling strategy needs to be undertaken. This includes comparing the multiple analogy models with other state-of-the-art survival models as well as using tools like the receiver operating characteristic curves to compare models across a range of cut-off points. This work is currently underway.

MATHEMATICAL APPENDIX

Since a key component of the procedure for learning from multiple analogies outlined in the narrative was the functional form of the information criterion (3), this appendix provides a brief derivation of this measure based on a minimal set of plausible assumptions.

Let the information acquired about a counting process *at time* t be some function of the divergence between the prior (pre-sample or pre-experiment) assessment of the hazard, $\bar{r}(t)$, and its posterior (post-analysis or post-experiment) assessment, $r(t)$. Let us denote this quantity as $I(t) = f(r(t), \bar{r}(t))$. What is reasonable to assume about this function? In other words, what are reasonable

properties for the function f to possess?

The first set of assumptions pertain to the range of values information can take. Keeping in mind that all quantities are indexed by t (i.e., we are talking about information at a particular t), let

$$f \geq 0 \qquad \forall r, \bar{r} > 0 \tag{8a}$$

$$f = 0 \qquad \forall r = \bar{r} \tag{8b}$$

Here, (8a) states that information is a non-negative quantity for all values of the prior and posterior hazard rates and (8b) states that if the posterior is exactly the same as the prior, then no information has been acquired.

The second set of assumptions deal with how information changes as the absolute value of the posterior increases. Let

$$\frac{df}{dr} > 0 \qquad \forall r > \bar{r} \tag{9a}$$

$$\frac{df}{dr} < 0 \qquad \forall r < \bar{r} \tag{9b}$$

$$\frac{df}{dr} = 0 \qquad \forall r = \bar{r} \tag{9c}$$

These assumptions simply state that the amount of information increases if the posterior moves further away from the prior—whether or not r is higher or lower than $\bar{r}$. For example, (9a) implies that if $r > \bar{r}$ then an increase in r adds to information since it takes the analyst further away from the prior. Similarly, (9b) implies that if $r < \bar{r}$ then an increase in r brings the analyst closer to the prior. (9c) implies that f is continuous in r.

The last set of assumptions deal with the notion of diminishing marginal returns. The idea is that the same increase in the posterior hazard should imply smaller informational gains if the hazard is already high, compared to if the

hazard were low. This assumptions translates to

$$\left.\frac{df}{dr}\right|_{r=r_1} > \left.\frac{df}{dr}\right|_{r=r_2} \qquad \forall r_1 < r_2 \tag{10a}$$

or, put another way, it translates to the second order differential equation

$$\frac{d^2 f}{dr^2} = \frac{\varkappa_0}{r} \qquad \forall r \tag{10b}$$

where $\varkappa_0$ is the constant of proportionality that can be set to any arbitrary constant without loss of generality. Since the ultimate goal is to derive a measure that will be optimized (maximized or minimized), a scaling constant will make no difference to the final solution of this optimization problem.

Given these assumptions, and setting the constant of proportionality to 1, we can start by integrating (10b) to get

$$\frac{df}{dr} = \int_{\mathscr{R}} \frac{1}{r}\, dr = \log(r) + \varkappa_1$$

where the constant of integration, $\varkappa_1$, can be solved using the initial condition (9c) to get $\varkappa_1 = -\log(\bar{r})$. This yields the result

$$\frac{df}{dr} = \log\frac{r}{\bar{r}}$$

which, it can be verified, satisfies each of the conditions (9a)–(9c). This solution can be further integrated to obtain

$$f = \int_{\mathscr{R}} \log\frac{r}{\bar{r}}\, dr = r\log\frac{r}{\bar{r}} - r + \varkappa_2$$

where the constant of this integration, $\varkappa_2$, can be solved using the initial condition (8b) to get $\varkappa_2 = \bar{r}$.

This procedure yields the final functional form for f, and recognizing the

conditional (on t) aspect of this measure, we can compute the net information acquired over the entire domain $\mathscr{T}$ as

$$I = \int_{\mathscr{T}} I(t)\,dt = \int_{\mathscr{T}} \left[r(t) \log \frac{r(t)}{\bar{r}(t)} - r(t) + \bar{r}(t) \right] dt \tag{11}$$

Since the analyst modeling criminal recidivism has information only on a limited support of the domain $\mathscr{T}$ (e.g., the follow-up period) the measure in (3) appropriately restricts the computation in (11) to a limited support.

Note that (11) is a more general measure of information than the Kullback-Leibler directed divergence measure commonly used in Information Theory (Kullback 1959). To see this, note that if the prior and posteriors were in fact proper probabilities (integrating to 1) then the measure in (11) could be simplified to

$$I = \int_{\mathscr{T}} r(t) \log \frac{r(t)}{\bar{r}(t)}\,dt - \int_{\mathscr{T}} r(t)\,dt + \int_{\mathscr{T}} \bar{r}(t)\,dt = \int_{\mathscr{T}} r(t) \log \frac{r(t)}{\bar{r}(t)}\,dt$$

which is the Kullback-Leibler directed divergence measure between two proper densities. Moreover, with an uninformative or constant (over the domain) prior, the minimization of information amounts to the maximization of Entropy—precisely the procedure Edwin Jaynes initially proposed (Jaynes 1957a,b).

REFERENCES

Allison, P. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology* 13:61–98.

Allison, P. (1984). *Event history analysis: Regression for Longitudinal Data*. Newbury Park: Sage.

Alt, J.E., King, G., and Signorino, C.S. (2001). Aggregation among binary, count, and duration models: Estimating the same quantity from different levels. *Political Analysis* 9(1):21–44.

Asadi, M., Ebrahimi, N., Hamedani, G. G., and Soofi, E. (2005). Minimum Dynamic Discrimination information models. *Journal of Applied Probability* 42(3):643-660.

Bhati, A.S. (in press). Estimating the number of crimes averted by incapacitation: An Information Theoretic approach." *Journal of Quantitative Criminology*.

Bhati, A.S. (2007). *Studying the effects of incarceration on offending trajectories: An Information Theoretic approach.* Washington, DC: The Urban Institute.

Bhati, A.S., and Piquero, A.R. (under review). On the effect of incarceration on subsequent individual criminal offending: Deterrent, criminogenic, or null effects.

Beck, N. (1998). Modelling space and time: The event history approach. Pg. 191–213 in E. Scarbrough and E. Tanenbaum eds. *Research strategies in the social sciences*. New York, NY: Oxford University Press.

Box-Steffensmeier, J.M., and Jones, B.S. (2004). *Event history modeling: A guide for social scientists*. Cambridge, UK: Cambridge University Press.

Bureau of Justice Statistics. (2002). *Recidivism of prisoners released in 1994, Codebook for dataset 3355*. Downloaded from NACJD in April 2007.

Cameron, C., and Trivedi, P. (1998). *The analysis of count data*. New York, NY: Cambridge University Press.

Dean, C.B., and Balshaw, R. (1997). Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. *Journal of the American Statistical Association* 92(440):1387–1398.

Diamond, S. (1959). *Information and error: An introduction to statistics*. New York, NY: Basic Books.

Donoho, D.L., Johnstone, I. M., Joch, J. C., and Stern, A. S. (1992). Maximum entropy and nearly black objects. *Journal of the Royal Statistical Society B*

54:41-81.

Ebrahimi, N., Habibullah, M., and Soofi, E. (1992). Testing exponentiality based on Kullback-Leibler information. *Journal of the Royal Statistical Society B* 54(3):739-748.

Ebrahimi, N., and Kirmani, S. N. U. A. (1996). A characterization of the proportional hazard model through a measure of discrimination between two residual life distributions. *Biometrika* 83(1):233-235.

Ebrahimi, N., and Soofi, E. (2003). Static and dynamic information for duration analysis. Invited presentation given at *A conference in honor of Arnold Zellner: Recent developments in the theory, method, and application of information and entropy econometrics*. Washington, D.C. Accessed February, 2007 from http://www.american.edu/cas/econ/faculty/golan/golan/Papers/8_20soofi.pdf

Ezell, M.E., Land, K.C., and Cohen, L.E. (2003). Modeling multiple failure time data: A survey of variance-corrected proportional hazard models with applications to arrest data. *Sociological Methodology* 33(1):111-167.

Glaser, D. (1955). The efficacy of alternate approaches to parole prediction. *American Sociological Review* 20:283–287.

Gottfredson, S.D., and Gottfredson, D.M. (1986). Accuracy of prediction models. Pg. 212–290 in eds. Blumstein, Cohen, Roth, and Visher *Criminal careers and "career criminals" volumes II*. Washington, DC: National Academy Press.

Greene, W. (2000). Models with discrete dependent variables. Pg. 811–895 in *Econometric analysis* (4th edition). Upper Saddle River, NJ: Prentice Hall.

Gull, S.F. (1989). Developments in maximum entropy data analysis. in *Maximum entropy and Bayesian statistics* ed. Skilling, J. Boston, MA: Kulwer.

Gull, S.F., and Danielle, G.J. (1978). Image reconstruction from incomplete and noisy data. *Nature* 272:686-690.

Heckman, J., and Singer, B. (1984a). Econometric duration analysis. *Journal of Econometrics* 24:63–132.

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

Jaynes, E.T. (1957a). Information Theory and Statistical Mechanics. *Physics Review* 106:620–630.

Jaynes, E.T. (1957b). Information Theory and Statistical Mechanics II. *Physics Review* 108:171–190.

Jaynes, E.T. (1979). Where do we stand on maximum entropy? Pg. 15–118 in R.D. Levine and M. Tribus (eds.) *The maximum entropy formalism*. Cambridge, MA: MIT Press.

Jaynes, E.T. (1986). Bayesian methods: An introductory tutorial. Pg. 1–25 in J.H. Justice (ed.) *Maximum entropy and Bayesian methods in applied statistics*. Cambridge, UK: Cambridge University Press.

Jaynes, E.T. (1988). Discussion. *American Statistician*. 42:280–281.

Justice, J.H. ed. (1986). *Maximum entropy and Bayesian methods in applied statistics*. Cambridge, UK: Cambridge University Press.

Kalbfleisch, J.D., and Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

King, G. (1977). *A solution to the ecological inference problem*. Princeton, NJ: Princeton University Press.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science* 33(3):762–784.

King, G., Rosen, O., and Tanner, M.A. eds. (2004). *Ecological inference: New methodological strategies*. Cambridge, UK: Cambridge University Press.

Kullback, S. (1959). *Information theory and statistics*. New York, NY: John Wiley.

Lancaster, T. (1990). *The analysis of transition data*. New York, NY: Cambridge University Press.

Langan, P.A., and Levin, D.J. (2002). *Recidivism of prisoners released in 1994*. Special report. Washington, DC: Bureau of Justice Statistics.

Levine, R.D. and Tribus, M. eds. *The maximum entropy formalism*. Cambridge, MA: MIT Press.

Lin, D.Y., Wei, L.J., and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* 85(3):605–618.

Lipton, D., Martinson, R., and Wilks, J. (1975). *The effectiveness of correctional treatment: A survey of treatment valuation studies*. New York, NY: Praeger Press.

Maltz, M.D. (1984). *Recidivism*. Orlando, FL: Academic Press, Inc.

Mathai, A.M. (1975). *Basic concepts in Information Theory and statistics: Axiomatic foundations and applications*. New York, NY: John Wiley.

Mayer, K.U. and Tuma, N.B. ed. (1990). *Event history analysis in life course research*. Madison, WI: The University of Wisconsin Press.

Michell, S.M., and Moore, W.H. (2002). Presidential uses of force during the Cold War: Aggregation, truncation, and temporal dynamics. *American Journal of Political Science* 46(2):438–452.

Mittelhammer, R. C., Judge, G. G., and Miller, D. J. (2000). *Econometric Foundations*. Cambridge, UK: Cambridge University Press.

Ryu, H.K. (1993). Maximum entropy estimation of density and regression functions. *Journal of Econometrics* 56:397–440.

Schmidt, P., and Witte, A.D. (1988). *Predicting recidivism using survival models*. New York, NY: Springer-Verlag.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27:379-423.

Sherman, L.W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., and Bushway, S. (1997). *Preventing crime: What works, what doesn't, and what's promising?* A Report to the United States Congress. Washington, DC: National Institute of Justice. (http://www.ncjrs.gov/works)

Skilling, J. (1989). *Maximum entropy and Bayesian methods*. Dordrecht, the Netherlands: Kluwer.

Soofi, E.S. (1994). Capturing the intangible concept of information. *Journal of the American Statistical Association* 89(428):1243–1254.

Soofi, E.S. (2000). Principal Information Theoretic approaches. *Journal of the American Statistical Association* 95(452):1349–1353.

Soofi, E.S., Ebrahimi, N., and Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association* 90(430):657–668.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of business & economic statistics* 13(4):467–474.

Yamaguchi, K. (1991). *Event history analysis*. Newbury Park, CA: Sage Publications.

Zellner, A. (1988). Optimal information processing and Bayes' theorem. *American Statistician* 42:278–284.