

A Bayesian Approach to Cross-Validation in Pedestrian Accident Reconstruction

2011-01-0290

Published
04/12/2011Gary A. Davis
University of Minnesota

Copyright © 2011 SAE International

doi:10.4271/2011-01-0290

ABSTRACT

In statistical modeling, cross-validation refers to the practice of fitting a model with part of the available data, and then using predictions of the unused data to test and improve the fitted model. In accident reconstruction, cross-validation is possible when two different measurements can be used to estimate the same accident feature, such as when measured skidmark length and pedestrian throw distance each provide an estimate of impact speed. In this case a Bayesian cross-validation can be carried out by (1) using one measurement and Bayes theorem to compute a posterior distribution for the impact speed, (2) using this posterior distribution to compute a predictive distribution for the second measurement, and then (3) comparing the actual second measurement to this predictive distribution. An actual measurement falling in an extreme tail of the predictive distribution suggests a weakness in the assumptions governing the reconstruction. This paper describes an implementation of these ideas using the Bayesian freeware WinBUGS. A statistical version of a throw distance model was fitted to data from 55 staged collisions between adult pedestrian dummies and sedan-type vehicles, while a standard braking-to-stop model was used to relate skidmark length to vehicle speed. Cross-validation is then illustrated using an additional 15 staged collisions. In 13 of the 15 cases the skidding and throw models yielded consistent results, while in two cases the resulting impact speed estimates were substantially different. In both of the discrepant cases it was possible to identify the sources of the discrepancies, and produce consistency once these were accounted for.

INTRODUCTION

Forensic problems can involve using less-than-perfect evidence to evaluate case-related hypotheses. When either the

quality of the evidence or its relation to these hypotheses is uncertain, probability theory has been offered as a logic for reasoning about these uncertainties [1,2,3]. Given a prior assessment of probabilities for competing hypotheses, and a likelihood function which gives the probability distribution of the evidence given these hypotheses, Bayes theorem can be used to update the probabilities assigned to the hypotheses in light of the available evidence. A Bayesian approach to accident reconstruction was introduced implicitly by Wood and O'Riordain [4], in a reconstruction of a two-vehicle collision. Here, Monte Carlo simulation was used to estimate probability distributions for a set output variables, including initial and impact speeds, given probability distributions for the input variables and deterministic equations linking the outputs to the inputs. What gave this exercise its Bayesian flavor was that acceptable ranges were specified for three of the output variables, and those input combinations producing values falling outside these ranges were rejected. This procedure is essentially an application of rejection sampling to produce a Monte Carlo sample from a truncated distribution, or equivalently, from a Bayesian posterior distribution where the likelihood function for the constraining output variables is a product of uniform probability densities. A similar approach was used by Rose et al.[5] to reconstruct a two-vehicle collision, where constraints were placed on the resulting normal and tangential restitution coefficients, while Kimbrough [6] has used rejection sampling to assess the relative plausibility of different collision scenarios.

Before proceeding it may be helpful to give an example of Bayesian inference, where a vehicle's initial speed is to be estimated from a measurement of skidmark length [7, pp. 6-7]. In [7], the measured skidding distance was $s = 33$ meters, and the authors gave this an uncertainty range of ± 1 meter, while the measured value of the frictional drag coefficient was $f = 0.7$ with an uncertainty range of ± 0.10 .

After applying the bounding method for describing uncertainty the authors reported an estimated speed of 21.3 meters/second with an uncertainty of ± 1.8 meters/second. To use Bayes theorem on this problem we first need to specify a likelihood function $g(s|v, f)$, giving the conditional probability density for s given the initial speed v and the drag coefficient f and a probability density describing the prior uncertainty for v and f , $h(v, f)$. The posterior probability density for v and f , given the measurement s , is then found by applying Bayes theorem for probability densities

$$h(v, f | s) = \frac{g(s | v, f)h(v, f)}{\int g(s | v, f)h(v, f)dvdf} \quad (1)$$

For this example, the marginal posterior density of v is then found by integrating with respect to the drag coefficient f

$$h(v | s) = \int h(v, f | s)df \quad (2)$$

The likelihood function can be based on a standard braking distance formula relating s , v , and f

$$s = \frac{v^2}{2fg}e \quad (3)$$

where e is a random variable accounting for measurement error. If we then use lognormal distributions to describe e , and the prior uncertainties for v and f , then the fact that logarithms of lognormal random variables yield normal random variables can be exploited to evaluate the integrals appearing in equations (1) and (2). For example, if the prior uncertainty regarding v is modeled as a lognormal random variable with mean of 15 meters/second and a coefficient of variation equal to 0.5, if the prior uncertainty regarding f is modeled as a lognormal random variable with a mean of 0.7 and a coefficient of variation equal to 0.14, and if the measurement error e is lognormal with a coefficient of variation equal to 0.1, then a measurement of $s=33$ meters leads to a posterior density for v as lognormal with a mean equal to 21 meters/second and a coefficient of variation equal to 0.084, which corresponds to a posterior standard deviation of about 1.8 meters/second. Figure 1 shows this example's prior and posterior probability densities for v .

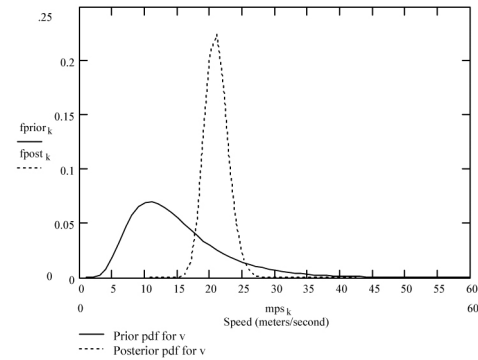


Figure-1. Prior (—) and posterior(---) probability densities for the initial speed, for illustrative example.

For the hypothetical example above we were free to choose probability distributions that made the inference problem tractable, but historically, practical application of the Bayesian approach has been hindered by the fact that the integrals appearing in equations (1) and (2) often fail to have closed forms. During the past 20 years, however, the development of effective numerical procedures for approximating these integrals, especially Markov chain Monte Carlo (MCMC) methods, has led to practical Bayesian solutions to a number of traditional statistical problems [8]. Like standard Monte Carlo methods, MCMC is a method for numerical evaluation of integrals, but instead of generating a pseudo-random sample from the distribution of interest, MCMC methods simulate realizations of Markov chains whose stationary distributions correspond to the desired Bayesian posteriors. Roughly contemporary with these advances has been the application of Bayesian thinking to the construction of expert systems [9,10,11]. Various called Bayesian belief networks or Bayesian expert systems, the idea is to model expert knowledge about a phenomenon as a set of inter-related random variables, whose joint distribution can be summarized by a directed acyclic graph (DAG). An example of DAG model is shown in Figure 3. The nodes of the graph represent the random variables entering in the model, such as initial speed, braking rate, or reaction time, while the directed links of the graph encode conditional dependencies in the variables' joint probability distribution. Early work on probabilistic graphical models for artificial intelligence, e.g. [12], identified correspondences between the connectivity of the DAG representation and decompositions of the model's joint probability distribution, leading, in some cases, to efficient methods for computing posterior distributions for the graph's variables. Early applications of Bayesian expert systems in the health sciences [13] were followed by applications in forensic science [14-15]. Regarding accident reconstruction, Davis [16] illustrated how Bayesian networks could be used to reconstruct vehicle/pedestrian accidents by approximating the desired posterior distributions using MCMC, while another exploratory application of Bayesian networks to accident reconstruction

has been given by Huygen [17]. A more formal treatment, based on Pearl's [18] theory of probabilistic causal models, is given in [19].

As O'Hagan [20] has noted, the ability to compute Bayes estimates for complicated models has tended to outpace the ability to criticize or validate those models. Model criticism is especially important in Bayesian applications since a prior distribution which assigns zero probability to a region of values will cause Bayes theorem to assign zero posterior probability to that region. That is, if an incorrect range for a variable has been assumed a priori Bayes theorem will not, of itself, alert us to this problem. For statistical estimation problems with adequately-sized samples, this difficulty can sometimes be avoided by using improper or marginally proper priors with permissive regions of support, but this clearly will not work when, as in accident reconstruction, restrictive, proper priors are needed for some variables in order to make other important variables identifiable. More generally, even in a well-behaved parameter estimation problem the selection of a particular probability distribution for the likelihood component of a model can be viewed as placing zero prior probability on other possible likelihood functions, and so computation of a posterior distribution for that model's parameters will not necessarily tell us that we started with the wrong likelihood. If two plausible candidate models have been identified a priori then Bayesian model selection or model averaging, where both models are assigned non-zero prior probability, can be employed. This, however, is not the same as a procedure to warn of a model's weaknesses in the absence of a specified competitor. Recognition of this dogmatism in Bayes theorem is of long standing [21-22] and for this reason there is a limited consensus that updating based on Bayes theorem should be supplemented with non-Bayesian procedures for model checking [23-24]. In this paper, we will illustrate how model checking can be accomplished in accident reconstruction, by reconstructing and checking several staged collisions between vehicles and pedestrian dummies.

RECONSTRUCTION MODEL

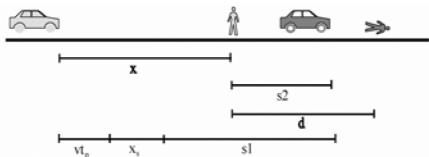


Figure-2. Diagram of vehicle/pedestrian collision.

Figure 2 illustrates a collision between a vehicle and a pedestrian, and defines several of the important variables in the reconstruction model. The essential causal story is that the vehicle was traveling at an initial speed v when at distance x from the collision point the driver noticed a pedestrian in the vehicle's path. After a reaction time of t_p the driver applied

the brakes, and after a braking transient time of t_s the vehicle's tires began making skidmarks. The vehicle struck the pedestrian with an impact speed of v_i , and as result of the impact the pedestrian traveled a distance d before coming to rest. The driver continued braking until coming to a stop, leaving a measurable skidmark of total length s_1 , with s_2 of that length being from the point of impact with the pedestrian. The objective is to use the skid mark measurements s_1 , s_2 and the measured throw distance d to draw conclusions about the vehicle's initial and impact speeds. Figure 3 is a DAG representation of the collision model, and the core of the model has two components, one describing the vehicle braking and the other describing the pedestrian throw distance. The braking model takes the form given in equation (3), where as before s is the stopping distance, v is the initial speed, and e is the braking drag coefficient which gives deceleration as a fraction of the gravitational acceleration g . e is a lognormal random variable which accounts for measurement error.

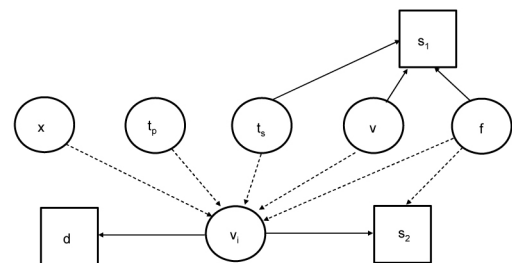


Figure-3. Directed acyclic graph (DAG) representation of vehicle-pedestrian collision model.

The model that relates the throw distance d to the impact speed v_i begins with a model proposed by Han and Brach [25]

$$d = B + \frac{v_i^2}{A^2} \quad (4)$$

where the coefficients A and B in turn depend on several physical parameters, including the masses of the vehicle and the pedestrian, the angle at which the pedestrian is launched from the vehicle, and the height of the pedestrian's center of mass above the vehicle's leading edge. Given a value for the impact speed v_i , and a probability distribution for these physical parameters, one could compute a predictive distribution for the throw distance d , as a function of the impact speed v_i . Alternatively, given a measured throw distance d and a prior distribution for the physical parameters, one could compute a posterior distribution for the impact speed. At our present state of knowledge however an appropriate probability distribution for the physical parameters is not yet available, so an expedient is to approximate the predictive relationship between impact speed

and throw distance using a statistical model. Figure 4 shows a plot of the measured throw distances versus measured impact speeds for 55 crash tests between sedan-type vehicles and adult-sized dummies, tabulated in Chapter 14 of [26]. Since the variability in the throw distance appears to increase as the impact speed increases, the measured throw distance was assumed to be a lognormal random variable with expected value given by equation (4). Bayes estimates for the

parameters A, B, and the lognormal variance σ_d^2 were computed using the MCMC software WinBUGS [27], and Table 1 summarizes the estimation results. It is noted that these estimates are similar to those reported by Han and Brach [25, Table 2] for adult-dummy crash tests with wrap trajectories. Figure 5 shows a probability plot test for normality of the lognormal residuals, indicating that the lognormal assumption is plausible for these data.

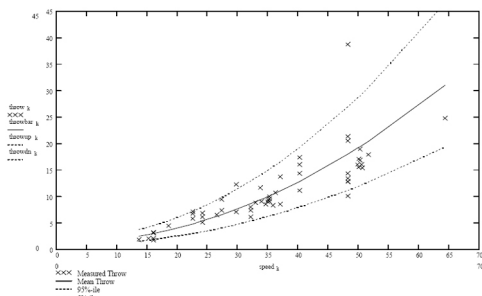


Figure 4. Measured throw distances and impact speeds from 55 vehicle/pedestrian crash tests. Speed is in km/h while throw distance is in meters. Also shown are predicted means(—) and 95% confidence intervals (---), based on equation (4) and the estimates in Table 1.

Table-1. Bayes estimates for throw model parameters: Impact speed in meters/second, throw distance in meters.

Posterior Summary				
Parameter	Mean	Standard Deviation	2.5 percentile	97.5 percentile
A	3.27	0.11	3.07	3.49
B	1.11	0.36	0.44	1.86
σ_d	0.27	0.03	0.22	0.33

To start a Bayesian accident reconstruction it is necessary to specify prior distributions for the model's exogenous variables, in this case those in Figure 3 having no arrows pointing into them. While it is possible to use relatively non-informative priors for the vehicle's initial speed and distance, the available measurements are not sufficient to identify all the collision model's exogenous variables, and informative priors, reflecting expert opinion, are needed for the braking drag coefficient (f), the duration of transient phase between the initiation of braking and the start of the measurable skidding (t_s), and the driver's reaction time (t_p). A discussion of, and justification for, these prior distributions has been given in [19], and will simply be summarized. The braking drag coefficient f was taken to be uniformly distributed between 0.5 and 0.9, corresponding to friction factors on dry asphalt or concrete pavements [28]. The braking transient time t_s was taken to be uniformly distributed between 0.1 seconds and 0.5 seconds [29-30], and the driver's reaction time was taken to be uniformly distributed between 0.5 seconds and 2.5 seconds [31]. The coefficient of variation for the skidmark measurement error was taken on to be 0.1 [32]. The pedestrian throw distance model described above differs somewhat from that used in [19], and the prior for the throw distance parameters A and B was taken to be bivariate normal, with means and standard deviations as given in Table 1, and a correlation of 0.753, reflecting the correlation observed in the estimates summarized in Table 1. The reconstruction model was implemented using the MCMC software WinBUGS, and the WinBUGS code for this model is listed in the Appendix.

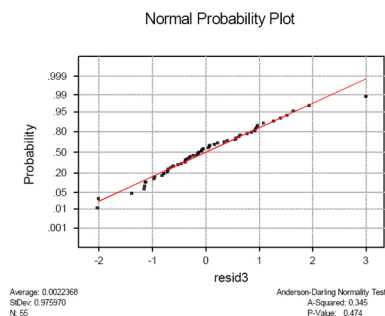


Figure-5. Normality test for lognormal residuals of throw-distance model.