Linköping Studies in Science and Technology. Dissertations
No. 814

Computer Based Statistical Treatment in Models with Incidental Parameters

Inspired by Car Crash Data

Anna Vadeby

Mathematical Statistics
Department of Mathematics
Linköping University, SE-581 83 Linköping, Sweden
ISBN 91-7373-625-2 ISSN 0345-7524

ISBN 91-7373-625-2
ISSN 0345-7524
Printed in Sweden by Uni Tryck, Linköping 2003

Abstract

Bootstrap and Markov chain Monte Carlo methods have received much attention in recent years. We study computer intensive methods that can be used in complex situations where it is not possible to express the likelihood estimates or the posterior analytically. The work is inspired by a set of car crash data from real traffic.

We formulate and develop a model for car crash data that aims to estimate and compare the relative collision safety among different car models. This model works sufficiently well, although complications arise due to a growing vector of incidental parameters. The bootstrap is shown to be a useful tool for studying uncertainties of the estimates of the structural parameters. This model is further extended to include driver characteristics. In a Poisson model with similar, but simpler structure, estimates of the structural parameter in the presence of incidental parameters are studied. The profile likelihood, bootstrap and the delta method are compared for deterministic and random incidental parameters. The same asymptotic properties, up to first order, are seen for deterministic as well as random incidental parameters.

The search for suitable methods that work in complex model structures leads us to consider Markov chain Monte Carlo (MCMC) methods. In the area of MCMC, we consider particularly the question of how and when to claim convergence of the MCMC run in situations where it is only possible to analyse the output values of the run and also how to compare different MCMC modellings. In Metropolis-Hastings algorithm, different proposal functions lead to different realisations. We develop a new convergence diagnostic, based on the Kullback-Leibler distance, which is shown to be particularly useful when comparing different runs. Comparisons with established methods turn out favourably for the KL.

In both models, a Bayesian analysis is made where the posterior distribution is obtained by MCMC methods. The credible intervals are compared to the corresponding confidence intervals from the bootstrap analysis and are shown to give the same qualitative conclusions.

Key words: Bootstrap, MCMC, incidental parameters, maximum likelihood, convergence diagnostics, Kullback-Leibler, relative collision safety.

List of Included Papers

This thesis is based on the following papers, referred to in the text by the letters A-E.

A: Modelling and Inference of Relative Collision Safety in Cars.
Anna Vadeby, Linköping Studies in Science and Technology. Theses No. 685.

B: Including Driver Characteristics in a Model of Relative Collision Safety.
Anna Vadeby, Technical Report, Linköping University, LiTH-MAT-R-2000-19.
Submitted to *Accident Analysis and Prevention*.

C: Estimation in a Model with Incidental Parameters.
Anna Vadeby, Technical Report, Linköping University, LiTH-MAT-R-2002-02.

D: The Empirical KL-Measure of MCMC Convergence.
Urban Hjorth and Anna Vadeby
Submitted to *Statistics and Computing*.

E: On Gibbs Sampler and Metropolis-Hastings Applied to Pairwise Poisson and Car Crash Data.
Anna Vadeby.

Acknowledgements

My foremost gratitude is directed towards my supervisor Associate Professor Urban Hjorth. You have always been an inspiring supervisor and there has always been time for questions and fruitful discussions. Thank you for all your help and encouragement.

I am grateful to Vinnova and former KFB for financial support of this work.

I want to thank Associate Professor Eva Enqvist for introducing me to the area of mathematical statistics. Your excellent teaching and enthusiasm inspired me to begin my graduate studies in mathematical statistics. During my time as a graduate student, you have always been very supportive no matter the issue.

I also want to thank my present and former friends and colleagues at the Department of Mathematics. Thanks for your support and interesting discussions.

Furthermore, I thank my parents Inger and Lars, for always believing in me and for all your love and enthusiastic help with the children.

Last, and above all, I am grateful to my children Emma and David for making the sun shine even the cloudiest day and to my husband Per for sharing my life.

Linköping, April 2003

Anna Vadeby

Contents

Part I

1 Introduction

Modelling, parameter estimation and uncertainty evaluation are three corner stones in statistics. This thesis represents all three of them in a non-standard situation with indirect observations that can be related to a quality parameter, here mostly the collision safety of the driver in a car.

Depending on the modelling, this may give a growing set of parameters or a corresponding set of unobserved random variables that we want to eliminate or balance out in order to concentrate on the interesting parameters. Although parameter estimates are a necessary part of all solutions, the main focus in this thesis is on methods for uncertainty evaluation by computer intensive technique. This technique includes bootstrap analysis in a classical frequentist modelling and Markov chain Monte Carlo analysis in a Bayesian setting.

This thesis consists of this overview and five papers A, B, C, D and E, listed at the end and referred to as Paper A, Paper B etc.

2 A New Model for Car Crash Data

A new model for car crash data which aims to estimate and compare the relative collision safety among different car models is formulated in Paper A. This model uses data with all the variation in speed, direction and hitting points that occur in real traffic, in contrast to designed crash tests. Analysis with this model is relevant primarily for the driver safety. With a different classification, other goals such as minimising accident costs can also be analysed by similar modelling.

2.1 Earlier Models

The problem of estimating relative collision safety from real data has been studied earlier. In Sweden, the insurance company Folksam (Hägg et al. [48]), has developed methods enabling different car models to be ranked due to their respective safety level. There exists several other methods with similar purpose. Jeremy Broughton and the British Department of Transport [16] have worked out a method comparable to the method of Folksam. They use logistic regression methods and calculate two so called DOT indices by considering factors that are likely to influence the likelihood of being injured. Logistic regression models are also used by Tapio and Ernvall [91] and Cameron et al. [17] and [67], to obtain ranking lists of cars due to collision safety. Tapio et al. [90] have also developed another model with the same purpose, but with a different structure, as in Tapio and Ernvall [91].

2.2 Our Model

The aim with our model in Paper A is to estimate the relative collision safety among different car models. The main goals in the development are that the model should be theoretically well defined, the mechanisms behind the model should be easy to understand and the model should offer possibilities to estimate the parameter uncertainties without depending on dubious approximations to normality.

We introduce a parameter α_k for the relative risk after correction for the weight of a car. Each α_k will be connected to one specific car model. Our information comes from police reported collisions in combination with hospital information about injury classes. See Paper A for details. The driver injuries are classified into four classes: 0,1,2 and 3, where 0 represents that the driver is unhurt and with an increasing scale up till 3, which represents that the driver is dead. Further information from the data base are car model and weight of the car.

In a head on crash, both cars are exposed to the same force, but the violence to the drivers are different. This difference might have many explanations, but one of the most intuitive factors that seems to influence the injury outcome is the almost instantaneous change of speed that the driver or the passengers are exposed to. Such information is, however, not given in the data base, but we shall introduce parameters closely related to the change of speed in each crash. Other important factors are the type of car being driven and the masses of the two cars involved in the crash. It has also been shown that a person's age and sex influence the injury risk in accidents that are otherwise similar.

We let t be a measure of how much violence the driver is exposed to in the crash, i.e. t is a function of the car model being driven, car mass and change of speed. We use the following criteria in the search of a suitable model: in the case of a gentle crash, the violence to the driver should be small and the probability that the driver is unhurt should be close to one. On the other hand, in the case of a heavy crash, the probability of a minor injury should be close to zero and the probability that the driver is badly hurt or dead ought to be close to one. When a person ends up in, for example, injury class 2, he or she has in some sense passed through the states 0 and 1. That is, given that one has reached a certain state, there is a probability to move to a higher state. This kind of model structure, with an ordinal model and a probability to move from one state to another, is the same structure as in a pure birth process. See for example Ross [83]. We use the structure given from that type of process, with an expression of the violence instead of a time parameter. This model is well defined and Kolmogorov's equations enable us to express the likelihood for the whole data set. We use maximum likelihood estimation of the parameters. Though possible to express, it is not possible to maximise the likelihood analytically to obtain expressions for the parameter estimates. We have therefore used computer based techniques to obtain the parameter estimates. The parameters in our model are of three different types: the relative risk parameters $\boldsymbol{\alpha}$, the birth rates $\boldsymbol{\lambda}$ in the birth process and the parameters introduced to replace the information of change of speed, $\boldsymbol{\theta}$. Therefore, a three step procedure is iterated in the estimation, one step for each parameter type. Our parameter of interest is $\boldsymbol{\alpha}$, while $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are nuisance parameters.

In Paper B, the model for car crash data is extended to include parameters related to the driver's age and sex. These factors are known to affect the risk of death (Evans [30]), and probably also the probability for the other injury classes in the same direction. If some vehicle types have very different driver populations in these respects, the relative safety parameters can be misleading if the model does not include parameters to allow for this. A fourth parameter type is included to take care of the driver characteristics and this introduces a fourth step in the iteration procedure. Only minor effects on the estimated safety parameters are observed in this set of data, but the main contribution of Paper B is that the possibility to allow for such factors is investigated.

The usual approach to obtain uncertainty estimates in large sample theory is to study the inverse of the information matrix. The information matrix is defined as: $I(\xi) = \|I_{jk}(\xi)\|$, where $I_{jk}(\xi) = -E\left[\frac{\partial^2}{\partial\xi_j\partial\xi_k} log(f(X;\xi))\right]$ $j, k = 1, ..., s$ and ξ denotes a parameter vector of dimension s. See for example Lehmann [53]. This is not a straightforward solution in the car crash model, since the dimension of the parameter $\boldsymbol{\theta}$ equals the number of crashes and is growing with the number of observations. The dimension of the information matrix is therefore also growing with the same rate.

A naive and unrealistic solution to this dimension problem is to replace the nuisance parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ with their estimated values and study the information matrix for $\boldsymbol{\alpha}$ alone. This is a solution which underestimates the variance of the estimated parameter $\hat{\boldsymbol{\alpha}}$, since the extra uncertainty introduced by the unknown nuisance parameters is not taken into account. Compare Lehmann [53], page 438, who states that the asymptotic variance of an efficient estimator when some of the parameters are unknown never falls below the value when they are known.

In the following section we will discuss different approaches and solutions to problems with similar parameter structure as the car crash model.

3 Review of some Incidental Parameter Problems

3.1 Classical Examples

In many studies one formulates a statistical model where only some of the parameters are of scientific interest. The other parameters are called "nuisance parameters" and are required to complete the probability mechanism but are not of essential value in themselves.

The elimination of nuisance parameters is a central but difficult problem in statistical inference. It is well known that the presence of nuisance parameters can make inference more complicated and that classical asymptotic results might not be valid. To describe the situation of interest, we can use the concept of structural and incidental parameters introduced by Neymann and Scott [68] in an early article written in 1948. Consider a sequence of random variables $X_1, X_2, X_3, \ldots$. The distribution of X_i depends on the parameters τ and θ_i, where the value of τ is independent of i, while the value of θ_i changes with i. The parameter τ is called a structural parameter and θ an incidental parameter. Our main issue is to estimate the structural parameter τ.

Neymann and Scott [68] show that in some cases when the number of parameters increases as the number of observations, the maximum likelihood estimator of the structural parameters need not be consistent; much data is no guarantee that estimates lie close to the true parameter values. Even if the maximum likelihood estimator of the structural parameter is consistent, the maximum likelihood estimator need not possess the property of asymptotic efficiency. To show this they considered two examples: "The many means example" which illustrates the inconsistency and "The many variances example" which illustrates the lack of asymptotic efficiency. These two examples can shortly be described as follows:

The many means example: Suppose that $\{X_{ij}\}$ are independently distributed according to a normal distribution with density

$$p_{ij}(x_{ij}|\alpha_i, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_{ij}-\alpha_i)^2}{2\sigma^2}}, \quad i = 1, ..., s,\ j = 1, .., n,\ s \to \infty.$$

The α_i are the incidental parameters and σ^2 is the structural parameter and also the parameter of interest. The maximum likelihood estimates of α_i and

σ are $\hat{\alpha}_i = \bar{x}_i$ and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}{sn} \sim \frac{\sigma^2}{sn}\chi^2(s(n-1))$$

with expectation $\sigma^2(n-1)/n$ for every s. Thus the maximum likelihood estimate $\hat{\sigma}^2$ is not consistent when $s \to \infty$.

The many variances example: Suppose that $\{X_{ij}\}$ are independently distributed according to a normal distribution with density

$$p_{ij}(x_{ij}|\alpha, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x_{ij}-\alpha)^2}{2\sigma_i^2}}, \quad i = 1, ..., s,\ j = 1, .., n,\ s \to \infty.$$

Consider the mean parameter α to be the structural parameter of interest and the many variances σ_i^2 to be the incidental parameters. Neymann and Scott study the maximum likelihood estimates and show that the estimate $\hat{\alpha}$ tends to α in probability and $\hat{\alpha}$ is consequently a consistent estimate of α. Furthermore, they show that though consistent, the maximum likelihood estimate need not be the most efficient estimator of α.

Many of the articles concerning the incidental parameter problem refer to the examples and results in the Neymann and Scott [68] article. A review of the incidental parameter problem in statistics and economics is written by Lancaster [52] to mark the 50th anniversary of the Neymann and Scott [68] article.

There are different ways to handle the problem with incidental or nuisance parameters and one can see two main approaches. The first approach is to treat the nuisance parameters as unknown constants and the second is to treat them as independent and identically distributed variables. The attitude towards the treatment of nuisance parameters seems to vary with the author and sometimes even within the same author. We see no reason to insist on only one formulation, but prefer to have an open mind and try different alternatives.

3.2 Preliminaries

In this section we establish some notation. Suppose that we have a sample of size n, $Y = (Y_1, ..., Y_n)$ and that each Y_i is a vector of length d. The vectors Y_i are independent and identically distributed with density $f(y, \theta)$. The parameter θ can be partitioned as $\theta = (\psi, \lambda)$, where $\psi = (\psi_1, ..., \psi_r)$ is the parameter of interest and $\lambda = (\lambda_1, ..., \lambda_p)$ is the nuisance parameter.

Let $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$ be the overall maximum likelihood estimate and denote by $\hat{\psi}_\lambda$ the maximum likelihood estimate of ψ for fixed λ and similarly $\hat{\lambda}_\psi$. The likelihood function is denoted by $L(\theta)$ and the log-likelihood function is $l(\theta)$.

A statistic S is said to be sufficient for Y if the conditional distribution for Y given $S = s$ is independent of θ for all s. A sufficient statistic is said to be minimal if there is no other sufficient statistic that provides greater reduction of data.

A statistic $A(Y)$ is said to be ancillary for the parameter θ if its distribution does not depend on θ. This is also termed distribution constant. By an ancillary statistic, one often means a distribution constant statistic which together with the maximum likelihood estimator constitutes a sufficient statistic. By an asymptotic ancillary statistic we mean a statistic A that has a distribution not depending on θ up to an appropriate order of approximation. See Barndorff-Nielsen and Cox [6] for a further description.

3.3 Deterministic Nuisance Parameters

If the attitude towards the nuisance or incidental parameters is that they should be regarded as unknown constants, there are several different proposals of how to proceed. One of the most common likelihood approaches is to replace the nuisance parameters with their conditional maximum likelihood estimates $\hat{\lambda}_\psi$, leading to the profile likelihood. The profile log-likelihood for ψ is defined by:

$$l_p(\psi) = l(\psi, \hat{\lambda}_\psi) = \max_\lambda l(\psi, \lambda),$$

see for example Hinkley et al. [46]. The maximum being over all λ that are consistent with the given value of ψ. The profile log-likelihood is then used as an ordinary log-likelihood and estimates of the structural parameters are achieved by maximisation. The profile likelihood is, however, not a real likelihood in the sense that it is proportional to the sampling distribution of an observable quantity. In Paper C, the estimates obtained by the profile likelihood in a pairwise Poisson example are compared to the results from a bootstrap analysis and first order approximations given by the delta method. In this example, all three methods show convergence to the same normal distribution with the same asymptotic variance.

Conditions under which the profile log-likelihood has a maximum $\hat{\psi}$ that converges in probability to ψ are studied in Mak [61]. He also gives a proper formula for the asymptotic covariance matrix. The conditions given in Mak [61] are not easy to verify. To solve the problem that the profile likelihood sometimes gives inconsistent or inefficient estimates of the structural parameters, especially if there are incidental or a large number of nuisance parameters involved, different adjustments of the profile likelihood that aim to approximate the likelihood more closely are suggested. One of them is the modified profile log-likelihood due to Barndorff-Nielsen [2] in 1983, defined by:

$$l_{mp}(\psi) = l_p(\psi) - \frac{1}{2}log(det(j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi))) + log(det\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}) \qquad (3.1)$$

where $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is the observed information matrix evaluated for fixed ψ at the corresponding maximum likelihood estimate $\hat{\lambda}_\psi$ and $\hat{\lambda}$ is considered as a function of $(\hat{\psi}, \hat{\lambda}_\psi, a)$, a being an asymptotic ancillary. Asymptotically, $l_{mp}(\psi)$ and $l_p(\psi)$ are equivalent to first order. Cox and Reid [26] study the difference between the profile and the modified profile likelihood and show that the modification of the profile likelihood is small if the parameter of interest is a mean parameter but is of more importance if the parameter of interest is a canonical parameter. Whether there is a difference of importance depends primarily on the expected value of a certain third order derivative of the log likelihood. The modified profile likelihood requires in general explicit specification of an ancillary statistic, so in situations where there is no obvious such ancillary statistic Barndorff-Nielsen suggests other adjusted versions of the profile likelihood, see Barndorff-Nielsen [4], [5] and Barndorff-Nielsen and Cox [6].

Another adjustment of the profile likelihood is suggested by Cox and Reid [25]. For a scalar parameter ψ of interest, Cox and Reid define the conditional profile likelihood by:

$$l_{cp}(\psi) = l_p(\psi) - \frac{1}{2}det(n(j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi))) \qquad (3.2)$$

This definition requires that the parameters λ and ψ are orthogonal in the sense that $E[\partial^2 l/\partial\psi\partial\lambda_j] = 0, j = 1, ..., p$. The effect of the second term is to penalise values of ψ for which the information about λ is relatively large. The difference between the modified profile likelihood (3.1) and the conditional profile likelihood (3.2) is that the use of orthogonal parameters in (3.2) allows us to ignore the term $log(det(\partial\hat{\lambda}_\psi/\partial\hat{\lambda}))$ in (3.1). Barndorff-Nielsen and McCullagh [3] show that in a number of instances in which the parameters are not orthogonal, the conditional profile likelihood (3.2) agrees with the modified profile likelihood (3.1). By agreement, they mean that the likelihoods only differ by terms of order $O(n^{-1})$.

Another concept, the directed likelihood, is defined as follows:

$$r(\psi) = sgn(\hat{\psi} - \psi)(2\{l_p(\hat{\psi}) - l_p(\psi)\})^{1/2}. \tag{3.3}$$

Different versions of directed likelihoods are studied by Barndorff-Nielsen [4], [5] and Sartori et al. [84]. Sartori et al. show that even though the directed modified profile likelihood has a standard normal distribution only up to first order, it often performs much better than expected. Sartori et al. [84] study exponential families and compare the directed modified profile likelihood with modified directed likelihood, which is a higher order approximation, and show that these two methods provide essentially the same inferences.

Confidence intervals associated with adjusted likelihoods in models with one parameter of interest and one nuisance parameter are studied by Mukerjee and Reid [66]. Pierce and Peters [71] use saddle point approximations to study confidence regions in the presence of nuisance parameters.

Since these methods require either specification of an ancillary or that the parameters are orthogonal, they seem to be restricted to rather special frameworks such as exponential families or transformation groups. McCullagh and Tibshirani [62] try a different approach to the adjustment of the profile likelihood based on properties of the score function. The basic idea behind the adjustment is that since the mean of a regular maximum likelihood score function is zero and the variance is the negative of the expected second derivative matrix computed in the true parameter point, it would be reasonable to require the same properties to hold for the profile likelihood score function in the point (ψ, λ_ψ). The adjusted profile log-likelihood is given by the integral of the adjusted score function. McCullagh and Tibshirani [62] give no formal proof, but indicate that one justification for the adjustment can be found in the theory of optimal estimation equations. The method seems to give satisfactory performance and gives result similar to the modified and conditional profile log-likelihoods. Rather than adjusting the profile likelihood for the effect of nuisance parameter estimation, Stafford [85] considers the use of the McCullagh and Tibshirani [62] adjustment for robust purposes.

Another score based adjustment of the profile likelihood is proposed by Stern [86]. The adjustment aims to reduce the score and information bias and is particularly applicable when the parameter of interest is vector valued.

Several other methods to construct likelihoods suited for situations with many nuisance parameters have been suggested. These are often called "marginal likelihoods" or "conditional likelihoods". They arise when components of the sufficient statistics have marginal or conditional distributions that depend on the structural parameter, but not on the nuisance parameter.

Andersen [1] suggests that instead of maximising the likelihood directly, one should eliminate the incidental parameters by considering the conditional distribution given a minimal sufficient statistic for the λ:s. The value of ψ that maximises this conditional distribution is called the conditional maximum likelihood estimator. Andersen [1] proves that under certain regularity conditions, the conditional maximum likelihood estimator is consistent and asymptotically normal, but that the asymptotical variance of the estimator is not in general equal to the Cramer-Rao bound. If we instead condition on a statistic that is ancillary for λ and study the estimate obtained from that conditional distribution, Andersen [1] shows that the Cramer-Rao lower bound of variance is attained asymptotically. The problem of obtaining efficiency bounds when the nuisance parameters are non random is also treated in Strasser [87] and [88] and Pfanzagl [69] and [70]. Well known information bounds that are valid when the nuisance parameters are treated as random are also shown to remain valid for nonrandom nuisance parameters under certain regularity conditions. The relation between the number of parameters p (not necessarily nuisance parameters) and the sample size, n is studied by Portnoy [72]. In the exponential family, he shows normal approximation results if $p/n \to 0$ or $p^2/n \to 0$, depending on the model considered.

Various conditional likelihoods are also suggested in Kalbfleisch and Sprott [50]. They suggest the use of maximum relative likelihoods which can be described as the profile likelihood standardised with respect to its maximum over ψ. Similar to the profile likelihood, the maximum relative likelihood can be misleading in both precision and location. If the distribution of interest can be factorised into two distribution functions where one factor contains no information concerning the nuisance parameter λ in the presence of the parameter of interest ψ, then Kalbfleisch and Sprott [50] suggest the use of the marginal likelihood (in principle the first factor) when making inference for λ. The use of sufficient and ancillary statistics when constructing conditional distributions for inference is reviewed in Reid [75].

Methods that sometimes require less information than likelihood methods are estimating functions, having zero mean in the true parameter point. The use of estimating functions in the presence of nuisance parameters are studied by, for example: Godambe [43], Ferguson et al. [31], Liang and Zeger [55] and Yuan and Jennrich [95].

3.4 Random Nuisance Parameters

From a Bayesian point of view the treatment of nuisance parameters is clear. Simply integrate them out from the likelihood with respect to a prior distribution. Consequently nuisance parameters are not something that require a separate study.

There are, however, other ways than the strictly Bayesian way of looking at the nuisance parameters as random. Berger et al. [7] study integrated likelihood methods for the elimination of nuisance parameters. They mean that even if one is not willing to carry through a subjective Bayesian analysis, the use of integrated likelihood methods should be encouraged. Berger et al. [7] focus on elimination of the nuisance parameter λ by simple integration resulting in a uniform integrated likelihood:

$$L^u(\theta) = \int L(\theta, \lambda) d\lambda.$$

They claim that there are several advantages using the integrated likelihood compared to for example using the profile likelihood or other maximisation methods. One advantage is that the integrating methods automatically account for some of the nuisance parameter uncertainty. Another advantage is that one can easily perform a sensitivity analysis by simply trying other prior distributions for the nuisance parameter λ and see how $L^\star(\theta)$ varies. One disadvantage pointed out by Lancaster [52], is that this naive construction of a uniform integrated likelihood does not, in general, solve the issue of consistency.

Kalbfleisch and Sprott [50] suggest that if a prior density for λ in the form $p(\lambda; \psi)$ is known, this prior information can be combined with the likelihood function, and integration gives a function of ψ only, to be used as a likelihood. The difficulty with this method pointed out by Kalbfleisch and Sprott [50] is that that precise type of prior information is not often available. This approach is investigated in Paper C, where the incidental parameters are treated as random variables, which are integrated out. The remaining likelihood is then used to obtain maximum likelihood estimates of the structural parameters.

The consistency is considered in an article written by Kiefer and Wolfowitz [51]. They study the incidental parameter problem in a slightly different way by postulating that the incidental parameters are independently distributed change variables having an unknown distribution G, where G does not belong to a certain parametric class. They showed that under certain regularity conditions, the maximum likelihood estimate of a structural parameter is strongly consistent and as a by-product, G can also be estimated by the maximum likelihood method.

Lindley [57] adopts a stricter Bayesian attitude towards the incidental parameters and advocates the use of a prior distribution. He sees no justification for marginal, conditional or other quasi-likelihoods that are not derived by integration with respect to a prior distribution.

The use of a non-informative prior called the reference prior is recommended by Liseo [59]. The reference prior was proposed by Bernado [9] in 1979. See also Berger and Bernado [8]. Liseo compares Bayesian techniques with likelihood methods by comparing credible sets derived from the reference prior approach with those computed with likelihood methods. The Bayesian methods are here shown to be better.

Levine and Casella [54] consider the use of a certain type of priors, called matching priors. These priors lead to posterior confidence regions which have approximate frequentist validity. Matching priors are constructed as solutions to a partial differential equation and it might be a complex task to obtain the solutions analytically. In Levine and Casella [54], a numerical Monte Carlo approach is suggested to overcome the computation difficulties.

For certain complex problems, where no analytical solutions are available, there is need for more computer based solution techniques, and we therefore move on the the area of computer intensive methods.

4 Bootstrap

Many of the previous methods are designed for analytically tractable problems. In certain complicated models, such as our model for car crash data, bootstrap might be a useful tool for inference about the parameters. The method dates back to Efron [28], 1979, and is now an important standard tool in many application areas. See e.g. monographs by Efron and Tibshirani [29], Hjorth [47] and Davison and Hinkley [27]. For completeness some of the basic ideas will be given.

Bootstrap methods are based on a bootstrap resample, obtained from the empirical distribution of the data. This usually means that as many artificial data as in the original sample are drawn at random from the true data (or from an estimated distribution of the data). Suppose we have an estimator $\hat{\theta}$ of some parameter θ, and that $\hat{\sigma}$ is an estimator of the standard deviation of $\hat{\theta}$. In bootstrap, $\hat{\theta}^*$ and $\hat{\sigma}^*$ are computed from the resampled bootstrap observations, in the same way as $\hat{\theta}$ and $\hat{\sigma}$ are computed from the true observations.

The distribution for the simulated estimates obtained from the bootstrap analysis is then translated to inference for the parameter of interest. This translation relies on asymptotic results which state that, under certain regularity conditions, the bootstrap distribution $P^*(.)$ converges to the true distribution $P(.)$ in probability. See, for example, van der Vaart [93] for a detailed description.

In more complex data situations, the bootstrap resampling can be done separately for different groups of data or conditional on some (approximately) ancillary information. Returning to our model for car crash data in Paper A, we obtain a bootstrap sample with similar structure as our real sample, by creating a resampled data base, where almost every car model is involved in the same number of crashes as in the original data base. This is done to avoid, for example, situations where one car model does not appear at all in a resampled data set. More generally, we want to keep roughly the same amount of information about each vehicle type, to make the bootstrap uncertainty relevant for the true data. Since we resample collisions, not results for individual cars, the number of collisions for one specific car model is only approximately the same as in the original data set. In Paper A, we use the bootstrap samples to create two different confidence intervals for the relative risks, the simple interval and the percentile interval. See, for example Hjorth [47], for a description. These intervals give similar results, but compared to the naive intervals obtained by ignoring the uncertainty introduced by the nuisance parameters, there is a large discrepancy. The naive intervals seem to underestimate seriously the uncertainty.

5 A Poisson Model

The difficulties to compare different estimates of uncertainty in the car crash model, make us study a model with a similar, but much simpler structure in Paper C. This model is based on Poisson densities and contains a structural parameter α, incidental parameters $\theta = (\theta_1, ..., \theta_n)$ and integer valued data coming in pairs. Due to the simpler structure, it is possible here to obtain analytical expressions of both the estimators and the corresponding uncertainties. First, the incidental parameters are treated as deterministic unknown constants. Following the car crash example, we study the maximum likelihood method together with bootstrap analysis and compare the naive estimator of uncertainty with the uncertainty estimates from the bootstrap analysis. A similar underestimation is shown and expressed analytically. We also study the profile likelihood and the delta method and compare the results with the bootstrap analysis. The uncertainty estimators are shown to be asymptotically equal. Later, the incidental parameters are treated as random variables from a Gamma distribution. The incidental parameters are integrated out, and replaced by the parameters in the Gamma distribution. The problem with growing dimension of the parameter vector is now eliminated and the estimation is straight forward. Maximum likelihood estimates are calculated and uncertainty estimates are obtained from the delta method and the information matrix. The different treatments of the nuisance parameters are shown to give similar results. The similarity between variance estimates for the different approaches gives support for an assumption that the same equivalence could be valid also in the collision model. The analysis can be seen as a theoretical support for the bootstrap results already discussed, as well as for the other methods.

The results in Paper C, encourage us to move on to a new research area where the incidental parameters, as well as the structural parameters, can be treated as random variables. In Paper D we study convergence of MCMC simulations and in Paper E we use MCMC methods to explore the problems in a Bayesian setting.

6 Markov Chain Monte Carlo

Markov chain Monte Carlo methods can be seen as a computer intensive tool for solving complex estimation problems where analytical solutions are not possible. The idea is to generate a sample from a specified distribution, $\pi(.)$, by creating a Markov process with this distribution as its stationary distribution, and run the simulation so long that the sample distribution is close enough to the stationary distribution. Most MCMC methods are directed towards Bayesian statistics. There are, however, some suggestions of MCMC methods in a frequentist context. In Gelfand and Carlin [34] and Geyer and Thompson [41], MCMC methods are used in missing and depending data settings, where the likelihood involves complicated integrals. Further examples are given in for example Tanner [89].

We adopt a Bayesian view, and study the posterior density $\pi(\theta|x)$, where information from the data is summarised in the likelihood, $f(x|\theta)$ and combined with the prior distribution $\pi(\theta)$ according to Bayes formula:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

As a function of θ and for a fixed observed x, this posterior is proportional to the nominator $f(x|\theta)\pi(\theta)$. When we have a multivariate probability density known up to a normalising constant, it is possible and surprisingly simple to create a Markov chain having this density as its stationary distribution and jumping around ergodically so that the time spent in different regions will after long time be proportional to the stationary probability of the region. The Bayesian posterior distribution is conveniently studied in this way. Since MCMC only uses this proportionality, it eliminates the need to integrate the denominator. This is one of the great advantages of the method. MCMC has therefore become an attractive method for problems where earlier methods have failed because of computational difficulties. MCMC methods rely on Markov chain theory and we give a brief summary of the most common Markov chain characteristics below.

A Markov chain is a discrete time stochastic process $\{X_1, X_2, ...\}$, with the following transition distribution:

$$P(X_t|X_0, X_1, ..., X_{t-1}) = P(X_t|X_{t-1}).$$

If the Markov Chain is irreducible, aperiodic and positive recurrent, as defined in e.g. Ross [83], then the distribution of X_t will converge to a unique stationary distribution π, which does not depend on t and X_0.

Let $P_{ij} = P_{ij}(t) = P(X_t = j|X_0 = i)$. A Markov chain is said to be reversible if it is positive recurrent with stationary distribution $\pi(.)$ and $\pi(i)P_{ij} = \pi(j)P_{ji}$. The chain under time reversal then has the same probability properties as the original chain. If we can find non negative numbers x_i, summing to one and satisfying $x_iP_{ij} = x_jP_{ji}$, it follows that the Markov chain is reversible and the numbers x_i are the limiting probabilities. This follows since $x_iP_{ij} = x_jP_{ji}$, for all i, j, and $\sum_i x_i = 1$, then summing over i gives

$$\sum_i x_iP_{ij} = x_j \sum_i P_{ji} = x_j, \qquad \sum_i x_i = 1.$$

Since the limiting probabilities $\pi(i)$ are the unique solution of the above, it follows that $x_i = \pi(i)$ for all i. If instead $\pi(i)$ or numbers proportional to $\pi(i)$ are given and we can define P_{ij}-probabilities satisfying the reversibility condition, then the chain will by the same computation, have π as its stationary distribution.

There exist different MCMC algorithms, but all of them are originating from the work by Metropolis et al. [65] in 1953 and Hastings [44] in 1970.

6.1 Metropolis-Hastings Algorithm

The Metropolis algorithm [65] was developed in 1953, to investigate the equilibrium properties of large systems of particles, such as electrons in an atom. Hastings [44] generalised the algorithm in 1970.

In the Metropolis-Hastings algorithm, a Markov chain with transition matrix P satisfying $\pi(i)P_{ij} = \pi(j)P_{ji}$ is created and the chain is run until it seems to have reached stationarity.

The algorithm starts with the density of interest, $\pi(x)$, also called the target density. A conditional density $q(y|x)$, called the proposal or instrumental density is then chosen. The proposal function can be any density function that creates an irreducible and aperiodic chain, and may depend on the current point X_t. At each time t, the next state X_{t+1} is chosen by first sampling a candidate point y from the proposal distribution $q(y|x_t)$. This candidate point is accepted with probability $\alpha(x_t, y) = min(1, \frac{\pi(y)q(x_t|y)}{\pi(x_t)q(y|x_t)})$ and otherwise rejected. If the candidate point y is accepted, the chain moves to $X_{t+1} = y$. If rejected, the chain stands still and $X_{t+1} = x_t$. This transition probability is consistent with the reversibility condition $\pi(i)P_{ij} = \pi(j)P_{ji}$ if i and j are replaced by x and y. For most reasonable proposal functions, also the conditions of communicating states and aperiodicity will be trivially met. The simulated observations are dependent and forming an irreducible Markov chain with the target distribution as its stationary distribution. These realisations will be used for inference. In Paper D and E, the Metropolis-Hastings algorithm is implemented to the pairwise Poisson model and the car crash model respectively.

The performance of the chain and in particular the acceptance rate, will depend on the choice of proposal function. Gelman et al. [38] suggest that an optimal proposal function has an acceptance rate about 0.44 for one dimensional problems, and about 0.23 if the dimension of the parameter vector exceeds five. Similar results are given in Roberts et al. [81]. Guidelines for the choice of proposal function are given in Gelman et al. [37]. In Paper D, several different proposals are investigated in the pairwise Poisson model, and a method how to choose the "best" proposal is studied.

6.2 Gibbs Sampling

The Gibbs sampler was introduced by Geman and Geman [39] in 1984 and is a special case of the Metropolis-Hastings algorithm. The sampler is known from statistical physics as the "heat-bath algorithm".

In the Gibbs sampler, high dimensional sampling for a vector parameter is replaced by sampling from low dimensional blocks. The sampler is at least two-dimensional. Suppose that we can partition the parameter vector θ into r blocks $\theta = (\theta_1, ..., \theta_r)$ and that we can sample from the corresponding conditional densities $\pi_i(\theta_i|\theta_{-i})$, where θ_{-i} denotes the parameter vector without component i. Then we cycle, systematically or at random, through the parameter values θ_i, $i = 1, ..., r$, and update the parameter values by sampling from the corresponding conditional densities $\pi_i(\theta_i|\theta_{-i})$.

The densities $\pi_i(\theta_i|\theta_{-i})$ are called full conditional distributions or full conditionals and determine uniquely the joint distribution. This is an interesting feature of the distributions and is known as the Hammersley-Clifford's theorem, see for example Besag [10]. The Gibbs sampler is used to estimate the posterior density of the structural parameter in the pairwise Poisson model, Paper E.

As mentioned before, the Gibbs sampler is a special case of the Metropolis-Hastings algorithm. If we let the proposals in Metropolis-Hastings algorithm equal the full conditionals in Gibbs sampler it is easy to show that the acceptance rate equals one, (see Paper E) and that every value proposed is accepted.

6.3 Convergence Discussion

An important and sometimes complicated task in MCMC simulations is to decide how long the simulation should be and when the chain has converged to the stationary distribution. There is no general theory for predicting the required run length in advance. It is therefore necessary to perform some form of statistical analysis to assess convergence, so called convergence diagnostics. The research about convergence can be divided into two areas; methods in the first category use only the output values of the MCMC chain to assess convergence, while methods from the second category analyse the target density to determine theoretically the number of iterations that will ensure convergence in some total variation distance. Most of the methods in category two require that the target density can be expressed analytically. Although it is impossible in general to find the convergence rate (Tierney [92]), bounds can sometimes be found for special classes of Markov chains, (Rosenthal [82]). When target densities are impossible to express analytically, only output based methods can be used. The models we are most interested in are models with complicated target densities that cannot be expressed analytically. We therefore focus on methods from the first category. Here, we will discuss methods from the first category and then give a short summary of methods from the second category. Reviews on convergence diagnostics can be found in Robert [78], Brooks and Roberts [14], Cowles and Carlin [21] and Mengersen et al. [64].

6.3.1 Convergence Measures Based on the Output

There exists a variety of convergence diagnostics which are based on the output from a MCMC chain. None of the methods are able to conclude that convergence has appeared, they only indicate that it is possible that the chain may have converged. Especially, for slowly mixing Markov chains, convergence diagnostics often seem to be too optimistic and unreliable, since their conclusions are based upon output from only a small region of the state space. There are different schools concerning the question whether one should study several parallel chains, or one single chain. The advocates of parallel chains claim that the use of parallel chains with overdispersed starting values increases the possibility to detect when the chain stays in a small subset for a long time due to slow mixing, see for example Gelman and Rubin [35] and [36] and Chauveau and Diebolt [19]. Those who recommend one single chain usually argue that the parallel chain approach is computationally expensive and that a single chain may run for much longer, giving a larger chance that it reaches its stationary distribution, see Raftery and Lewis [74] and Tierney [92].

The methods that rely on the output are more or less informal. One of the most informal, and widely used not so long ago, is Gelfand et al. [33], sometimes referred to as Gelfand's thick-pen technique, where convergence is concluded if density estimates spaced far enough apart to be considered independent, differ graphically by less than the width of a thick felt-tip pen. Today, a number of more elaborate methods to determine convergence have been developed.

One popular convergence diagnostic is by Gelman and Rubin [36]. Their method is based on normal theory approximations and involves two major steps. In the first step, m parallel chains are started with overdispersed starting values. Each chain is run $2n$ iterations and the first n are discarded in order to reach stationarity. To answer the question whether these m chains are similar enough to claim approximate convergence, they compare the individual chains to the chain obtained by mixing together the mn values from all sequences. This is done by calculating a potential scale reduction factor, $\hat{R}_c$. Essentially, this means a comparison of variances within and between the chains. A large value of $\hat{R}_c$ indicates that further simulations may improve our inference about the target distribution, while a value close to 1 indicates that each of the m sets of n values is close to the target distribution. The method is generalised by Brooks and Gelman [13] who also suggest alternative scale reduction factors.

Yu and Mykland [94] propose a different convergence diagnostic, which uses a cusum path plot based on a single run and a one-dimensional summary statistic, to study the convergence of the sampler of length n, where n_0 iterations are discarded due to burn in. They calculate the observed cusum for the statistic of interest $T(\theta)$: $\hat{S}_t = \sum_{j=n_0+1}^{t}(T(\theta_j)-\hat{\mu}), \quad t = n_0+1, ..., n$, and plot $\{\hat{S}_t\}$ against t. Using a result in Lin [56], Yu and Mykland [94] argue that a smooth plot with large excursion size indicates slow mixing, while a hairy plot with small excursion size indicates a rapid mixing. The cusum plot is compared to a "benchmark" cusum path plot, $\{\hat{S}_t^b\}$ obtained from i.i.d. normal variables with mean and variance matched to the moments of the MCMC sequence. If the two plots are comparable in terms of smoothness and excursion size they conclude that the sampler is mixing well. A suggestion to make this rather subjective method more objective, is made by Brooks [12], who defines an index of "hairiness " that should lie between the bounds $0.5 \pm Z_{\frac{\alpha}{2}}\sqrt{1/4(n-n_0)}$ during $100(1-\frac{\alpha}{2})\%$ of the time if the chain is mixing well. The result relies on somewhat unrealistic assumptions of i.i.d and symmetric sequences, which however can be made approximately true by "thinning" and in some cases also by use of the empirical median instead of the mean.

A method by Raftery and Lewis [74] decides the number of burn-in iterations, the minimum sub sampling step (they store every k^{th}) and the number of further iterations needed to estimate a posterior quantile from a single run of a Markov chain. Define u to be the q^{th} quantile of θ, so that $P(\theta \leq u) = q$. For a given q, an estimate of u is easily obtained from the output. Instead of studying the original chain of interest θ^t, Raftery and Lewis [74] create a sequence $\{Z_t\} = \mathbf{1}_{\theta^t \leq u}$. Taking every kth value only, the sequence $\{Z_t^k\}$ is obtained. For sufficiently large k, this sequence approximates a two state Markov chain, which is possible to analyse explicitly. Brooks and Roberts [15] point out that, when this method is applied to problems where quantiles are not of primary interest, the method sometimes underestimates the length of burn in. The same phenomenon is mentioned in Zuur et al. [97] where the method by Raftery and Lewis [74] is too optimistic in concluding burn-in compared to the methods by Gelman and Rubin [36] and Geweke [40].

Two output based convergence diagnostics which use standard techniques from spectral analysis to gain variance estimates are Geweke [40] and Heidelberger and Welch [45].

Suppose that the intention of the analysis is to estimate the mean of some function of the simulated parameter θ and that we think that the chain has converged by time n_0. Geweke [40] suggests that one should take the first n_A and the last n_B observations out of n iterations and compute the partial expectations $\bar{\theta}_A$ and $\bar{\theta}_B$ as well as the corresponding spectral density estimates $\hat{S}_\theta^A(0)$ and $\hat{S}_\theta^B(0)$. The standardised difference between $\bar{\theta}_A$ and $\bar{\theta}_B$ is a standard normal variable. This asymptotic normality induces a convergence diagnostic which can be used to test the null hypothesis of equal location. If the hypothesis is rejected, it indicates that the chain has not converged by time n_0.

The convergence diagnostic by Heidelberger and Welch [45] uses Cramer von Mises procedures to test stationarity. The diagnostic is based on a statistic which is approximately a Brownian bridge for large samples, hence inducing a possible test.

A new convergence diagnostic based on the classical Kullback-Leibler (KL) measure is suggested in Paper D. The diagnostic compares empirical distributions for sub sequences or parallel runs with the full sequence of MCMC simulations. In the single run version, a long MCMC sequence consisting of T values is split into N sub sequences of length $t = T/N$. The empirical distribution for the whole sequence defines a partition of the real axis such that H classes with equal number of data is created. We let F_0 be the empirical distribution for the whole sequence, and F_i the empirical distribution for sub sequence i and compare the sub sequences with the whole sequence by calculating a Kullback-Leibler measure

$$KL_i = KL_i(\frac{F_i + \epsilon F_0}{1+\epsilon}, F_0) = \sum_{j=1}^{H} \ln \left(\frac{F_0(E_j)}{\frac{F_i(E_j)+\varepsilon F_0(E_j)}{1+\varepsilon}} \right) F_0(E_j).$$

Mixing part of F_0 with F_i solve the singularity problem when F_i has empty cells. The total measure of variability is the mean distance $KL = \sum KL_i/N$. The leading term in a series expansion leads to an interpretation in terms of cell frequencies' relative uncertainty, measured by their coefficient of variation. Simulations support the results obtained from the series expansion. A reasonable limit for convergence and good mixing sequences is suggested in Paper D, where the method also compares favourably to the cusum method and the within-between variance method.

There are other ways to interpret KL-measures. One rather close interpretation is by McCulloch [63] who compares two Bernoulli distributions $B(p)$ and $B(q)$ with probabilities $p = 0.5$ and q. Let $KL(B(0.5), B(q(k)) = k$, then $q(k) = (1 + (1 + e^{-2k})^{1/2})/2$. For KL=0.01, which usually means acceptable convergence in our setting above, we get $q(0.01) = 0.57$. This means that a KL-distance of 0.01 is comparable with describing an unobserved event as having probability 0.57 when in fact the probability is 0.5.

In our KL application, the corresponding differences are spread over many cells and individual differences are typically smaller.

6.3.2 Convergence Measures Based on the Target Density

There exists certain specialised techniques to put limits on the simulation errors. These methods use the target density to predetermine the number of iterations needed to obtain approximate convergence. Most of the methods are problem specific and can only be used to a limited number of models.

Several authors use so called coupling methods. A coupling in discrete time is a construction of two discrete time processes X_1 and X_2, marginally having the same distributions which are realisations of the Markov chain of interest, and started at different starting distributions. The joint process is supposed to be Markov. The coupling time τ is defined as $\tau = inf\{t : X_1^t = X_2^t\}$, see Lindvall [58]. The so called coupling inequality (Lindvall [58], page 12), ensures that if the probability that coupling has occurred at time t is high, then the distribution of X_2^t is a good approximation of the stationary distribution π. The coupling inequality motivates several convergence diagnostics, see for example Propp and Wilson [73], Johnson [49] and Rosenthal [82].

In Rosenthal [82], rigorous theoretical upper bounds on burn-in times are given. A suggestion of how the results by Rosenthal [82] can be applied in a more general setting is done by Cowles and Rosenthal [22]. They show a simulation based approach to obtain the bounds. Though simulation based, the method is problem specific and the amount of computation is far from negligible. This simulation approach is carried further in Cowles [23], where the original simulation approach by Cowles and Rosenthal [22] is improved.

Different weighting methods are suggested by Ritter and Tanner [76] and Zellner and Min [96]. The method by Ritter and Tanner [76] is called the "Gibbs Stopper". The basic idea behind the "Gibbs Stopper" is to study so called importance weights, defined by the ratio of a function proportional to the target density and the current approximation of the joint distribution. After each considered iteration, a histogram of the weights is produced. Convergence is indicated when the distribution of the weights convergence toward a spike distribution.

Examples of other sources where the transition kernel is used to assess convergence of the sampler are Liu, Liu and Rubin [60], Roberts [79] and [80] and Brooks et al. [11].

There are a variety of different techniques that can be used to investigate whether a MCMC simulation has converged or not. A book covering many convergence diagnostic techniques, especially methods that rely on analyses of the target distribution, is Robert [78].

7 Summary of the Papers

This thesis comprises five papers. In Paper A, a new model for estimating the relative collision safety of cars is developed. Paper B extends the model from Paper A to consider the effect of the driver's age and sex in the injury outcome of the crash. In Paper C, different estimation techniques are compared in a model with similar, but simpler, structure than the model for car crash data. Paper D focus on MCMC methods. A new convergence diagnostic based on the classical Kullback-Leibler distance is developed. In Paper E, a Bayesian view to the pairwise Poisson model and the car crash model is adopted. Gibbs sampler is used in the Poisson model and Metropolis-Hastings algorithm in the car crash model, to obtain estimates of the parameters of interest. A short summary of included papers is given below.

7.1 Paper A: Modelling and Inference of Relative Collision Safety in Cars

The question of how to extract information about the relative collision safety of different car models from collision data is considered in Paper A. A new mathematical model, based on a pure birth process, is developed to describe the probability that a driver ends up in different injury classes. This process starts in the state "unhurt", and moves at random into more serious damage. Depending on the collision forces and the safety parameter, a driver will spend a certain "time" in this process. Since the collision forces are unknown, we only have information about the ratio of these "times" for the two colliding drivers. The approach in Paper A is to define incidental parameters representing the forces, or equivalently the times spent in the damage process. If this parameter is given for one of the drivers, the other driver is given a value adjusted for the ratio of vehicle weights and the ratio of the vehicles' safety parameters. The law of conservation of linear momentum is used to express that heavy vehicles experience less change of speed than light vehicles in a collision. With this model, an explicit expression for the likelihood can be given. The model has three types of parameters: the safety parameters, the parameters of the injury class process, and the vector of collision forces. Each parameter type can be estimated relatively easy by numerical methods if the other two are known, therefore, a three-step iteration procedure is used to obtain maximum likelihood estimates of the parameters. The likelihood has no closed and explicit form, though it can be expressed and differentiated. Due to this and the growing vector of incidental parameters, standard asymptotic tools, based on likelihood derivatives and the information inequality are not useful for the variance of our maximum likelihood estimates. The approach in Paper A is to perform a bootstrap analysis. In the bootstrap analysis, a certain number of collisions is needed for each considered car model. Therefore a resampling conditional on approximately the same number of collisions as in the original data for each car model is made. The bootstrap analysis gives estimated variances and confidence intervals demonstrating that some car models are significantly better than others in their protection of the driver.

7.2 Paper B: Including Driver Characteristics in a Model of Relative Collision Safety

The driver characteristics are known to affect the risk of death, and probably also the injury risks in accidents that are otherwise similar. In Paper B, the possibility to allow for factors related to the driver's age and sex in the model from Paper A is demonstrated. If some vehicle type has very different driver populations in these respects, the relative safety parameters can be misleading without parameters related to the driver characteristics. Different models are compared and the best model, chosen by a likelihood-ratio test, includes parameters related to both age and sex. A fourth step in the estimation procedure is introduced for these new parameters and the estimates of the relative safety parameters, compensated for the driver's age and sex, are compared to relative risks from Paper A where the driver population is included. The uncertainties are studied by a bootstrap analysis, similarly to Paper A. Only minor effects on the estimated safety parameters are observed in this set of data and the magnitudes of the uncertainties in model A and model B seem to be approximately the same.
Paper B is submitted to *Accident Analysis and Prevention*.

7.3 Paper C: Estimation in a Model with Incidental Parameters

In Paper C, a simpler Poisson model of similar type as the complex model in Paper A, is studied. The model contains incidental parameters and integer valued data coming in pairs, and (Y_{i1}, Y_{i2}) are independent $Po(\theta_i)$ and $Po(\alpha\theta_i)$, given the incidental parameter θ_i, $i = 1, ..., n$. Estimates of the structural parameter α are studied for both deterministic and random θ_i. In this simpler model, different approaches can be theoretically analysed and compared. Methods that might be possible in the car crash model are of primarily interest and therefore we exclude the possibility of conditioning on the sufficient statistic $Y_{i1}+Y_{i2}$. A straightforward maximum likelihood approach leads to an estimate of α that is a ratio without mean and variance for finite samples. In the first attempt to study the asymptotical properties of this estimator $\hat{\alpha}$, the θ-values are naively replaced by their estimated values and an information matrix for α is calculated. This certainly underestimates the variance, and therefore the second attempt is to study the profile likelihood. The result obtained by the profile likelihood is also achieved asymptotically by a bootstrap analysis, despite a dependency introduced in the resampling step. The delta method, based on first order approximations, is known to be useful to study large sample behaviour of estimators without specification of a specific loss function. In the Poisson model, the delta method gives convergence of $\hat{\alpha}$ to a normal distribution with the same asymptotic variance as the profile likelihood and bootstrap. The same properties, up to first order, are seen for random as well as deterministic incidental parameters.

The similarity between variance estimates for the different approaches gives support for an assumption that the same equivalence could be valid also in the model for car crash data in Paper A. The analysis can be seen as a theoretical support for the bootstrap results as well as for the other methods.

7.4 Paper D: The Empirical KL-Measure of MCMC Convergence

A new convergence diagnostic for Markov chain Monte Carlo simulations, based on the classical Kullback-Leibler (KL) distance, is proposed in Paper D. The diagnostic is designed to compare the distribution of sub sequences in the simulation, with the result for the entire distribution, or to compare parallel simulations with the joint result. This KL-measure can also be used as a loss function when the design, including the choice of proposal function, of a Markov chain is optimised. The comparison of empirical distributions uses a Kullback-Leibler type distance over value sets defined by the output data. The singularity problem for such a measure is removed in a simple way.

A series expansion shows that the leading term can be interpreted as the relative uncertainty (σ/μ) for cell frequencies in the sub sequences. A simulation study investigates the validity of the leading term in two cases with Markov dependency and selected acceptance rates. It is shown that the agreement between the leading term and the KL-measure is close, especially for long simulation times.

The KL-measure is compared to two well known methods, the variance comparison method introduced by Gelman and Rubin based on parallel runs and the method by Yu and Mykland based on cumulative sums of a single chain with a continuation due to Brooks in terms of a hairiness index for the proportion of turns in centred cumulative sums. In a Poisson example with random nuisance parameters, the performance of the measures for different proposal functions are studied, leading to a wide range of acceptance rates. The KL-measure performs very well and reacts distinctly in a systematic way on the proposals. It gives a more distinct signal than the competitors and has also a clear interpretation in terms of cell frequency uncertainty. In a second example, an analytical function with symmetry properties is defined as a posteriori density. According to recommended threshold values, the other criteria signal convergence in situations where cell probabilities are still very inaccurate according to the KL-measure. Again KL is very distinct and outperforms its competitors.
Paper D is submitted to *Statistics and Computing*.

7.5 Paper E: On Gibbs Sampler and Metropolis-Hastings Applied to Pairwise Poisson and Car Crash Data

In Paper E, a Bayesian approach to the car crash model in Paper A and the pairwise Poisson model Paper C is considered. In the Poisson model, the Gibbs sampler is used to obtain an estimated posterior distribution for the structural parameter of interest. Credible intervals for the most probable value of the structural parameter are compared to the corresponding confidence intervals calculated in Paper C. The same conclusions about the parameter are drawn from the two different analyses.

The car crash model is, due to its more complex structure, analysed in a different way. First, the original data set from Paper A, is reduced to contain only five car models instead of nineteen. Thereafter, a single component Metropolis-Hastings algorithm is implemented. The parameters are updated one by one and the different parameter types have different structure of their proposal functions. In the prior distributions, all cars are considered to be on the same safety level.

Similar to the Poisson model, the Bayesian and frequentist analyses give the same qualitative results, but the credible intervals differ slightly from the corresponding confidence intervals in centring and interval width.

References

[1] Andersen, E.B. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. J. Roy. Statist. Soc. (B) No. 32, 283-301.

[2] Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. Biometrika, Vol. 70, No. 2, 343-365.

[3] Barndorff-Nielsen, O.E., McCullagh, P. (1993). A note on the relation between modified profile likelihood and the Cox-Reid adjusted profile likelihood. Biometrika, Vol. 80, No. 2, 321-328.

[4] Barndorff-Nielsen, O.E. (1994). Adjusted Versions of Profile Likelihood and Directed Likelihood, and Extended Likelihood. Journal of the Royal Statistical Society, B, Vol. 56, No. 1, 125-140.

[5] Barndorff-Nielsen, O.E. (1995). Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. Biometrika, Vol. 82, No. 3, 489-499.

[6] Barndorff-Nielsen, O.E., Cox, D.R. (1994). Inference and Asymptotics. Chapman and Hall.

[7] Berger, J.O., Liseo, B., Wolpert, R.L. (1999). Integrated Likelihood Methods for Eliminating Nuisance Parameters. Statistical Science, Vol. 14, No. 1, 1-28.

[8] Berger, J.O., Bernado, J.M. (1992). On the Development of Reference Priors. Bayesian Statistics 4, Bernado, Berger, Dawid and Smith (eds), Oxford Science Publications, 35-60.

[9] Bernado, J.M. (1979). Reference posterior distributions for Bayesian inference. J.R. Statist. Soc. B, Vol. 41, No. 2, 113-147 (with discussion).

[10] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems" J. Roy. Statist. Soc. Ser. B, Vol. 36, 192-236.

[11] Brooks, S.P., Dellaportas, P., Roberts, G.O. (1997). An approach to Diagnosting Total Variation Convergence of MCMC algorithms. Journal of Computational and Graphical Statistics, Vol. 6, No. 3, 251-265.

[12] Brooks, S.P. (1998). Quantitative convergence assessment for Markov chain Monte Carlo via cusums. Statistics and Computing, Vol. 8, 267-274.

[13] Brooks, S.P, Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics Vol. 7, No. 4, 434-455.

[14] Brooks, S.P, Roberts, G.O. (1998). Assessing Convergence of Markov Chain Monte Carlo Algorithms. Statistics and Computing Vol. 8, 319-335.

[15] Brooks, S.P., Roberts, G.O. (1999). On quantile estimation and Markov chain Monte Carlo convergence. Biometrika, Vol. 86, No. 3, 710-717.

[16] Broughton, J. (1994). The theoretical basis for comparing the accident record of car models. Transport Research Laboratory.

[17] Cameron, M., Finch, C., Le, T. (1994). Vehicle Crashworthiness Ratings: Victoria and NSW Crashes During 1987-92, Summary Report. Monash University Accident Research Centre, Report No. 55.

[18] Cameron, M., Newstead, S., Le, T., Finch, C. (1994). Relationships Between Vehicle Crashworthiness and Year of Manufacture. Royal Automobile Club of Victoria, Report No. 94/6.

[19] Chauveau, D., Diebolt, J. (1999). An automated stopping rule for MCMC assessment. Computational Statistics, Vol. 14, 419-442.

[20] Chung, K.L. (1974). A Course in Probability Theory. Academic Press.

[21] Cowles, M.K., Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association, Vol. 91, No. 434, 883-904.

[22] Cowles, M.K., Rosenthal, J.S. (1998). A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. Statistics and Computing, Vol. 8, 115-124.

[23] Cowles, M.K. (2002). MCMC sampler convergence rates for hierarchical normal linear models: A simulation approach. Statistics and computing, Vol. 12, 377-389.

[24] Cox, D.R., Hinkley, D.V. (1974). Theoretical Statistics. Chapman and Hall.

[25] Cox, D.R., Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. With a Discussion. J.R. Statist. Soc. Ser. B, Vol. 49, No. 1, 1-39.

[26] Cox, D.R., Reid, N. (1992). A note on the difference between profile and modified profile likelihood. Biometrika, Vol. 79, No. 2, 409-411.

[27] Davison, A.C., Hinkley, D.V. (1997). Bootstrap Methods and their Applications. Cambridge University Press.

[28] Efron, B. (1979). Bootstrap methods: Another look at the jacknife. Ann. Statist., Vol. 7, 1-26.

[29] Efron, B., Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall.

[30] Evans, L. (1991). Traffic Safety and the Driver. Van Nostrand Reinhold.

[31] Ferguson, H., Reid, N., Cox, D.R. (1991). Estimating equations from modified profile likelihood. Estimating functions edited by V. P. Godambe. Oxford, Claredon Press, 273-293.

[32] Gelfand, A.E., Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association, Vol. 85, 398-409.

[33] Gelfand, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F.M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. Journal of the American Statistical Association, Vol. 85, No. 412, 972-985.

[34] Gelfand, A.E., Carlin, P. (1993). Maximum-likelihood estimation for constrained- or missing-data models. The Canadian Journal of Statistics, Vol. 21, No. 3, 303-311.

[35] Gelman, A., Rubin, D.B. (1992). A Single Series from the Gibbs Sampler Provides a False Sense of Security. Bayesian Statistics 4 Bernado, J.M., Berger, J.O., Davis, A.P. and Smith, A.F.M. (Eds.), Oxford University Press, Oxford, 625-631.

[36] Gelman, A., Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, Vol. 7, No. 4, 457-511.

[37] Gelman, A., Carlin, B., Stern, H.S., Rubin, D.B. (1995). Bayesian Data Analysis. Chapman and Hall.

[38] Gelman, A., Roberts, G.O., Gilks, W.R. (1996). Efficient Metropolis Jumping Rules. Bayesian Statistics 5, Bernado, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds), 599-607.

[39] Geman, S., Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. Pami-6, No. 6, 721-741.

[40] Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. Bayesian Statistics 4, Bernado, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds), 163-193.

[41] Geyer, C.J., Thompson, E.A. (1992) Constrained Monte Carlo Maximum Likelihood for Dependent Data. J. R. Statist. Soc. B, Vol. 54, No. 3, 657-699.

[42] Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). Markov Chain Monte Carlo in Practice. Chapman and Hall.

[43] Godambe, V. P., Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameters. The Annals of Statistics, Vol. 2, No. 3, 568-571.

[44] Hastings, W.K, (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, Vol. 57, 97-109.

[45] Heidelberger, P., Welch, P.D. (1983). Simulation run length control in the presence of an initial transient. Operations Research Vol. 31, 1109-1144.

[46] Hinkley, D.V., Reid, N., Snell, E.J. (1991). Statistical Theory and Modelling: In honour of Sir David Cox, FRS. Chapman and Hall.

[47] Hjorth, J.S.U. (1994). Computer Intensive Statistical Methods. Chapman and Hall.

[48] Hägg, A., Kamren, B., v Koch, M., Kullgren, A., Lie, A., Malmstedt, B., Nygren, Å. , Tingwall, C. (1992). Folksam Car Model Safety Rating 1991-1992. FOLKSAM research.

[49] Johnson, V.E. (1996). Studying Convergence of Markov Chain Monte Carlo Algorithms Using Coupled Sample Paths. Journal of the American Statistical Association, Vol. 91, No. 433, 154-166.

[50] Kalbfleisch J.D., Sprott D.A. (1970). Application of Likelihood Methods to Models Involving Large Number of Nuisance Parameters. J. Roy. Statist. Soc. Ser. B, Vol. 32, 175-208.

[51] Kiefer, J., Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely many Incidental Parameters. Ann. Math. Statist. Vol. 27, 887-906.

[52] Lancaster, T. (2000). The Incidental Parameter Problem since 1948. Journal of Econometrics, Vol. 95, 391-413.

[53] Lehmann, E.L. (1983). Theory of Point Estimation. John Wiley & Sons.

[54] Levine, A., Casella, G. (2003) Implementing matching priors for frequentist inference. Biometrika, Vol. 90, No. 1, 127-137.

[55] Liang, K-Y., Zeger, S.C. (1995). Inference Based on Estimating Functions in the Presence of Nuisance Parameters. Statist. Sci. Vol. 10, 158-173.

[56] Lin, Z.Y. (1992). On the increments of partial sums of a ϕ-mixing sequence. Theory of probability and its applications, Vol. 36, 316-326.

[57] Lindley, D.V. (1971). The estimation of many parameters. Foundations of statistical inference/Ed. by V.P Godambe, D.A. Sprott, 435-455.

[58] Lindvall, T. (1992). Lectures on the Coupling Method. Wiley.

[59] Liseo, B. (1993). Elimination of nuisance parameters with reference priors. Biometrika, Vol. 80, No. 2, 295-304.

[60] Liu, C., Liu, J., Rubin, D.B. (1992). A Variational Control Variable for Assessing Convergence of the Gibbs Sampler. In Proceedings of the American Statistical Association, Statistical Computing Section, 74-78.

[61] Mak, T.K., (1982). Estimation in the presence of incidental parameters. The Canadian Journal of Statistics. Vol. 10, No. 2, 121-132.

[62] McCullagh, P., Tibshirani, R. (1990). A Simple Method for the Adjustment of Profile Likelihoods. J.R. Statist. Soc. B, Vol. 52, No. 2, 325-344.

[63] McCulloch, R.E. (1989). Local Model Influence. Journal of the American Statistical Association, Vol. 84, No. 406, 473-478.

[64] Mengersen, K.L., Robert C.P., Guihenneuc-Jouyaux, C. (1999). MCMC Convergence Diagnostics: A Reviewww. Bayesian Statistics 6, Bernado, J.M., Berger, J.O., Davis, A.P., and Smith, A.F.M. (Eds.), Oxford University Press, Oxford, 415-440.

[65] Metropolis, N., Rosenbluth, M.N., Teller, A.H., Teller, E., (1953). Equations of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, Vol. 21, No. 6, 1087-1092.

[66] Mukerjee, R., Reid, N. (1999). On confidence intervalsassociated with the usual and adjusted likelihoods. J.R. Statist. Soc B, Vol. 61, Part 4, 945-953.

[67] Newstead, S., Cameron, M., Skalova, M. (1996). Vehicle Crashworthiness Ratings: Victoria and NSW Crashes During 1987-94. Monash University Accident Research Centre, Report No. 92.

[68] Neymann, J., Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrica, Vol. 16, No. 1, 1-32.

[69] Pfanzagl, J. (1970). Consistent estimation in the presence of incidental parameters. Metrika, Vol. 15, 141-148.

[70] Pfanzagl, J. (1993). Incidental Versus Random nuisance parameters. The Annals of Statistics, Vol. 21, No. 4, 1663-1691.

[71] Pierce, D.A., Peters, D. (1992). Practical Use of Higher Order Asymptotics for Multiparameter Families. J.R. Statist. Soc. B, Vol. 54, No. 3, 701-737.

[72] Portnoy, S. (1988). Asymptotic behaviour of likelihood methods for exponential families when the number of parameters tends to infinity. The Annals of Statistics, Vol. 16, No 1, 356-366.

[73] Propp, J.G., Wilson, D.B. (1996). Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics. Random Structures and Algorithms, Vol. 9, 223-252.

[74] Raftery, A.E., Lewis, S.M. (1992). How many iterations in the Gibbs Sampler? Bayesian Statistics 4, Bernado, J.M., Berger, J.O., Davis, A.P. and Smith, A.F.M. (Eds.), Oxford University Press, Oxford, 763-773.

[75] Reid, N. (1995). The Roles of Conditioning in Inference. Statistical Science, Vol. 10, No. 2, 138-199.

[76] Ritter, C., Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. Journal of the American Statistical Association, Vol. 87, No. 419, 861-862.

[77] Robert, C.P., Casella, G. (1999). Monte Carlo Statistical Methods. Springer-Verlag, New York.

[78] Robert, C.P. (1998). Discretisation and MCMC Convergence Assessment. Lecture Notes in Statistics, Springer Verlag.

[79] Roberts, G.O.(1992). Convergence Diagnostics of the Gibbs Sampler. Bayesian Statistics 4, Bernado, J.M., Berger, J.O., Davis, A.P. and Smith, A.F.M. (Eds.), Oxford University Press, Oxford, 775-782.

[80] Roberts, G.O. (1994). Methods for Estimating L^2 Convergence of Markov Chain Monte Carlo. Bayesian Statistics and Econometrics: essays in Honour of Arnold Zellner, eds D. Berry, I. Chaloner and J. Geweke, Amsterdam, North Holland, 373-384.

[81] Roberts, G.O., Gelman, A., Gilks, W.R. (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. Annals of Applied Probability, Vol. 7, 110-120.

[82] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. Journal of the American Statistical Association, Vol. 90, No. 430, 558-566.

[83] Ross, S.M. (1993). Introduction to Probability Models. Academic Press.

[84] Sartori, N., Bellio, R., Salvan, A., Pace, L. (1999). The directed modified profile likelihood in models with many nuisance parameters. Biometrika, Vol. 86, No. 3, 735-742 .

[85] Stafford, J.E. (1996). A robust adjustment of the profile likelhood. The Annals of Statistics, Vol. 24, No. 1, 336-352.

[86] Stern, S.E. (1997). A Second-order Adjustment of the Profile Likelihood in the case of Multidimensional Parameter of Interest. J. R. Statist. Soc. B, Vol. 59, No. 3, 653-665.

[87] Strasser, H. (1996). Asymptotic efficiency of estimates for models with incidental nuisance parameters. The Annals of Statistics, Vol. 24, No. 2, 879-901.

[88] Strasser, H. (1998). Perturbation invariant estimates and incidental parameters. Mathematical methods of statistics, Allerston Press, Inc, Vol. 7, No. 1, 1-26.

[89] Tanner, M.A. (1996). Tools for Statistical Inference. Springer-Verlag New York.

[90] Tapio, J., Pirtala, P., Ernvall, T. (1995). The Accident Involvement and Injury Risk Rates of Car Models. University of Oulu, Publications of Road and Transport Laboratory, Report No. 30.

[91] Tapio, J., Ernvall, T. (1995). Logistic Regression in Comparison of the Passive Safety of Car Models. University of Oulu, Publications of Road and Transport Laboratory, Report No. 32.

[92] Tierney, L. (1994). Markov Chains for exploring posterior distributions (with discussion). Annals of Statistics, Vol. 22, 1701-1762.

[93] van der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics.

[94] Yu, B., Mykland, P. (1998). Looking at Markov Samplers through Cusum Path Plots: a Simple Diagnostics Idea. Statistics and Computing, Vol. 8, 275-286.

[95] Yuan, K-H., Jennrich, R.I. (2000). Estimating equations with nuisance parameters: theory and applications. Ann. Inst. Statist. Math., Vol. 52, No. 2, 343-350.

[96] Zellner, A., Min, C. (1995). Gibbs Sampler Convergence Criteria. Journal of the American Statistical Association, Vol. 90, 921-927.

[97] Zuur, G., Garthwaite, P.H., Fryer, R.J. (2002). Practical Use of MCMC Methods: Lessons from a Case Study. Biometrical Journal, Vol. 44, No. 4, 435-455.

Part II

Included Papers