On probability-based inference under data missing by design

Olli Saarela

Department of Chronic Disease Prevention
National Institute for Health and Welfare, Helsinki, Finland

and

Department of Mathematics and Statistics
University of Helsinki, Finland

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in auditorium CK112, Exactum (Gustaf Hällströmin katu 2B), on 24th of September 2010, at 12 o'clock noon

Helsinki 2010

Supervised by:

Professor Elja Arjas
Department of Mathematics and Statistics
University of Helsinki

Docent Sangita Kulathinal
University of Helsinki
University of Tampere
Indic Society for Education and Development, Nashik, India

Reviewed by:

Professor Sven Ove Samuelsen
Department of Mathematics
University of Oslo

Docent Aki Vehtari
Department of Biomedical Engineering and Computational Science (BECS)
Aalto University

Opponent:

Professor Ørnulf Borgan
Department of Mathematics
University of Oslo

ISBN 978-952-92-7801-5 (Paperback)
ISBN 978-952-10-6419-7 (PDF, `http://ethesis.helsinki.fi`)
Helsinki University Print
Helsinki 2010

Abstract

Whether a statistician wants to complement a probability model for observed data with a prior distribution and carry out fully probabilistic inference, or base the inference only on the likelihood function, may be a fundamental question in theory, but in practice it may well be of less importance if the likelihood contains much more information than the prior. Maximum likelihood inference can be justified as a Gaussian approximation at the posterior mode, using flat priors. However, in situations where parametric assumptions in standard statistical models would be too rigid, more flexible model formulation, combined with fully probabilistic inference, can be achieved using hierarchical Bayesian parametrization. This work includes five articles, all of which apply probability modeling under various problems involving incomplete observation. Three of the papers apply maximum likelihood estimation and two of them hierarchical Bayesian modeling.

Because maximum likelihood may be presented as a special case of Bayesian inference, but not the other way round, in the introductory part of this work we present a framework for probability-based inference using only Bayesian concepts. We also re-derive some results presented in the original articles using the toolbox equipped herein, to show that they are also justifiable under this more general framework. Here the assumption of exchangeability and de Finetti's representation theorem are applied repeatedly for justifying the use of standard parametric probability models with conditionally independent likelihood contributions. It is argued that this same reasoning can be applied also under sampling from a finite population.

The main emphasis here is in probability-based inference under incomplete observation due to study design. This is illustrated using a generic two-phase cohort sampling design as an example. The alternative approaches presented for analysis of such a design are full likelihood, which utilizes all observed information, and conditional likelihood, which is restricted to a completely observed set, conditioning on the rule that generated that set. Conditional likelihood inference is also applied for a joint analysis of prevalence and incidence data, a situation subject to both left censoring and left truncation. Other topics covered are model uncertainty and causal inference using posterior predictive distributions. We formulate a non-parametric monotonic regression model for one or more covariates and a Bayesian estimation procedure, and apply the model in the context of optimal sequential treatment regimes, demonstrating that inference based on posterior predictive distributions is feasible also in this case.

Keywords: Bayesian nonparametric regression, case-cohort design, causal inference, conditional likelihood, full likelihood, incidence, model selection, monotonic regression, nested case-control design, prevalence, probability-based inference, two-phase study design

Contents

List of original publications

The thesis consists of the introductory part and the following five articles, referred to in the text by Roman numerals (I-V).

I. Saarela, O. and Kulathinal, S. (2007). Conditional likelihood inference in a case-cohort design: an application to haplotype analysis. *The International Journal of Biostatistics*, 3. Available from `http://www.bepress.com/ijb/vol3/iss1/1`.

II. Saarela, O., Kulathinal, S., Arjas, E., and Läärä, E. (2008). Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Statistics in Medicine*, 27:5991–6008.

III. Saarela, O., Kulathinal, S., and Karvanen, J. (2009). Joint analysis of prevalence and incidence data using conditional likelihood. *Biostatistics*, 10:575–587.

IV. Saarela, O. and Arjas, E. (2010). A method for Bayesian monotonic multiple regression. Accepted for publication in *Scandinavian Journal of Statistics*.

V. Arjas, E. and Saarela, O. (2010). Optimal dynamic regimes: presenting a case for predictive inference. *The International Journal of Biostatistics*, 6. Available from `http://www.bepress.com/ijb/vol6/iss2/10`.

Authors' contributions

I. The research problem was conceived by SK and the authors were jointly responsible for formulating the methods. OS was mainly responsible for writing of the paper and he implemented the methods. SK helped in writing the paper.

II. The research problem was conceived by EL and the authors were jointly responsible for formulating the methods and writing of the paper. EA proved the theorems in Appendix A and OS was responsible for implementing the methods.

III. The research problem was conceived by SK and JK. OS was mainly responsible for formulating the methods and writing of the paper and he implemented the methods. SK and JK helped in writing the paper and in formulating the methods.

IV. The research problem and the general model construction were formulated by EA. OS was mainly responsible for writing of the paper and he refined the model construction and designed and implemented the sampling algorithm. EA helped in writing the paper.

V. The research problem and the model construction were formulated by EA. The authors were jointly responsible for writing of the paper. OS was responsible for designing and implementing the sampling algorithm.

1 Introduction

Two objectives guide this presentation. First, statistics without a theoretical framework would be but a collection of unrelated tricks (Cox, 2006). Accepting without reservations the need for a framework, the second goal here is to present it in the most minimalist way possible. According to Dawid (1984), "the only concept needed to express uncertainty is probability". A corollary of this is that, inasmuch statistical inference is about making informed statements on unobserved quantities, we can present a theoretical framework for statistical inference using probability as the only concept.

Although the articles included in this work involve a variety of statistical inference problems, what is common to all five papers is that they all apply probability models as the tool for solving the problems. Whether the actual estimation of the parameters of the probability models is carried out using maximum likelihood methods or Bayesian computation is of less importance in some of these problems. However, there are compelling reasons for choosing the Bayesian approach as the theoretical framework for the presentation herein. First, maximum likelihood inference may be presented as a special case or approximation of fully probabilistic Bayesian inference (see Section 3.2), while the opposite is not possible. For instance, many problems involving hierarchical parametrization involve adoption of the Bayesian approach. Even though under certain conditions (see exchangeability below) frequency-based reasoning will, in the limit, produce results similar to the Bayesian approach, generally the frequency-based concept of probability is too limited to cover all statistical inference problems. (As a side note, the properties of Bayesian computation are evaluated using frequency-based reasoning, see Bayarri and Berger, 2004, p. 64 and Section 3.1.) Thus, due to the stated striving for a minimalist presentation, we will avoid presenting two theoretical frameworks side by side by choosing to present only the more general one. Moreover, the theoretical toolbox needed for the Bayesian approach is considerably lighter compared to the alternatives. Since the uncertainty on all kinds of unobserved quantities may be expressed using conditional (posterior) probability distributions given the observed quantities (data), we only need to know the Bayes' theorem (Section 2.2) to be able to express the posterior distribution as a function of the probability model for the data (the likelihood) and our prior information on the unobserved quantities. Applying the Bayes' theorem is essentially updating prior knowledge based on new observed information. Due to the adoption of the Bayesian approach in this presentation, we will also re-derive some results presented in the original articles using the toolbox equipped herein.

According to Lindley and Novick (1981, p. 45), "inference is a process

whereby one passes from data on a set of units to statements about a further unit". Though not all statistical inference problems involve the concept of a population consisting of units or individuals, the epidemiological applications considered in the five articles herein do. Thus in addition to conditional probability, we need to know how to utilize observations made on different individuals to make inference on quantities interpreted to represent something that is a property of a population of individuals (or a generic individual, that is, the further unit in the above quote). For this purpose, we utilize the concept of exchangeability and the related result known as de Finetti's theorem (Section 2.2). The exchangeability postulate means that the units or individuals can be exchanged in such a way that the joint information learned when observing some characteristic on a finite set of such units does not depend on which observation was made on which unit. This is usually expanded by assuming always a further unit (and the resulting infinite sequence of further units) onto which the same property applies. Now the similarity of the units implied by their exchangeability can be put to use in statistical inference by applying de Finetti's representation theorem which states that the joint distribution of observations made on a finite set of such units can be represented in terms of a prior distribution for parameters and a parametric probability model where the individual contributions are conditionally independent given the parameters. This justifies the use of conventional i.i.d. models, even though in terms of information learned, observations are not independent (Rubin, 1987, p. 40). It should be noted that without the extension of exchangeability onto infinite sequences, the representation theorem holds only approximately (Diaconis, 1977; Diaconis and Freedman, 1980). In the following we apply the representation theorem only for introducing parameters which we can interpret to be properties of a generic unit (or any population of such units; what this means in the case of sampling from a finite population is discussed in Section 4.2). We do not apply it for quantities which are properties of study designs such as inclusion indicators in finite population sampling or treatment assignments in an experimental design, which cannot be naturally extended outside some finite context. This does not imply that fully probabilistic Bayesian inference would be invalid in situations where the exchangeability does not hold; only that the resulting probability expressions will likely be more complicated.

The plan is as follows. Section 2 reviews the basic concepts of axiomatic (measure theoretical) system of probability and briefly discusses the information based interpretation of probability. These concepts are then applied in the context of statistical inference, with a review of conditional probability, conditional expectation, Bayes' theorem, exchangeability and de Finetti's theorem. Predictive inference, a natural extension of the exchangeability

postulate, is also discussed. As examples of probability models, we review marked point processes, and a special case of these, survival models. Section 3 discusses Bayesian computation, and as an alternative, maximum likelihood estimation. Section 4 introduces the topics covered in the five articles, presenting these under the umbrella term of incomplete observation, using this term in a slightly wider meaning than, for instance, Andersen et al. (1993, Chapter III). We argue that from the Bayesian point of view, we do not need to make a conceptual distinction between unobserved observables and unobservables, since the inference on both kinds of quantities proceeds in exactly the same way. The main topic covered here is the missing by design situation, using a generic two-phase study design as an example, and the alternative approaches of full likelihood, which utilizes all observed information, and conditional likelihood, which is restricted to completely observed set, conditioning on the rule that generated that set. In addition, we introduce the topics of model uncertainty and causal inference using posterior predictive distributions. We summarize the conclusions in Section 5.

2 Probability-based statistical inference

2.1 Probability

A. N. Kolmogorov's (1933) mathematical (axiomatic) definition of probability is straightforward, although it says little on the philosophical essence of probability and randomness. A collection of subsets of the sample space Ω, closed with respect to countable set operations, is known as a σ-algebra. We denote this as $\mathcal{F}$. The pair $(\Omega, \mathcal{F})$ is a measurable space, where events $A \in \mathcal{F}$ are said to be measurable sets. A probability measure is a mapping $P : \mathcal{F} \to [0, 1]$, for which $P(\Omega) = 1$ and $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ for disjoint sets $A_i \in \mathcal{F}$. The triple $(\Omega, \mathcal{F}, P)$ is called a probability space. Commonly in applications the events $A \in \mathcal{F}$ can not be observed directly and the probability space represents merely an abstraction of the random phenomenon of interest. A random variable maps the outcomes $\omega \in \Omega$ onto outcomes on some observable space. A real valued ($\mathcal{F}$-measurable) random variable is defined as $X : \Omega \to \mathbb{R}$, for which $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ-algebra of $\mathbb{R}$. (Some examples of non-real valued random variables are introduced in Sections 2.4 and 4.6.) The combined mapping $P_X \equiv P \circ X^{-1}$ is called the distribution of X. It is a probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$. Random variable X is said to be continuous if its distribution is absolutely continuous with respect to the Lebesgue measure l. Then the distribution can be written as $P_X(B) = \int_B f_X(x)\, l(\mathrm{d}x)$

for all $B \in \mathcal{B}$, where $f_X : \mathbb{R} \to \mathbb{R}_+$ is the Radon-Nikodym derivative of P_X with respect to l, and is known as the density function of X. The cumulative distribution function $F_X(x) \equiv P_X((-\infty, x])$ uniquely determines the distribution of X. A random vector $X = (X_1, \ldots, X_p) : \Omega \to \mathbb{R}^p$ can be defined analogously to above. Then the (joint) distribution P_X is a probability measure on $(\mathbb{R}^p, \mathcal{B}_p)$, where $\mathcal{B}_p$ is the Borel σ-algebra of $\mathbb{R}^p$.

With the above mathematical concepts reiterated, we are left with the question of what are probability and randomness. Here we take the view that uncertainty is essentially lack of information on physically existing quantities and deterministic (causal) events occurring in our single material universe. This corresponds most closely to the "support" (or evidence) interpretation of probability (Shafer, 1992), and differs from the "belief" (or subjective) interpretation mainly in that in the former case the degree of belief is what a (generic) rational individual would hold given specific degree of evidence, while in the latter case the degree of belief may be different even between individuals holding the same information. What evidence, or information, means in terms of probabilities comes clearer in Section 2.2 with the discussion on conditional probabilities. Shafer (1992) calls the support interpretation of probability as "rational degree of belief" while Cox (2006) uses the term "impersonal degree of belief". The support interpretation could be brought closer to subjective interpretation by considering also individuals' other properties as "information" which affects their decisions. Thus there does not need to be a contradiction between objective reality and subjective probability. Consistent with the support interpretation would be to think that in the case of complete information there is no randomness. Naturally, this brings us to the controversies of quantum mechanics. However, with a reference to Jaynes (2003, p. 327-330), we shall proceed with the assumption that the events of interest are sufficiently macro-level for such questions to be less relevant. To sum up, we assume the existence of a single reality, mechanistic in nature; our objective is to collect more information on that reality, and to attempt to quantify how much is still unknown given this information.

2.2 Bayes' and de Finetti's theorems

As noted before, the objective in statistical inference is to make informed statements on unobserved quantities based on observed data. A very general tool for this purpose discussed here is the probability model. A parametric probability model is a probability distribution specified in terms of parameters, which may be broadly interpreted to represent some underlying properties of a mechanism which has produced the observations, hopefully capturing some systematic components of interest (e.g. Cox and Hinkley,

1974, p. 5). A more concrete interpretation of parameters, which we will adopt for this presentation, is discussed below. It is usually not contended that such a model is an accurate representation of reality, nor is such accuracy necessary. What is important is that the model is able to capture some essential characteristics of reality and simplify them into a manageable form (cf. Cox and Hinkley, 1974, p. 5-6).

Let now $Y = (Y_1, \ldots, Y_n) : \Omega \to \mathbb{R}^n$ and $\Theta = (\Theta_1, \ldots, \Theta_p) : \Omega \to \mathbb{R}^p$ both be ($\mathcal{F}$-measurable) random vectors. Here Y represents observations while Θ represents parameters. That the parameters are taken to be random variables on the same σ-algebra as the random variables representing observable quantities, already implies that the present approach is Bayesian. Despite its name, this particular school of thought is attributable to Bruno de Finetti (1906-1985) rather than Thomas Bayes (c. 1702-1761) (Jaynes, 2003, p. 655; see also de Finetti, 1974, and Stigler, 1982). The Bayesian approach will be chosen throughout, as it considerably simplifies the theoretical framework needed for making statistical inference (Cox and Hinkley, 1974, p. 364; Efron, 1978, p. 236). In fact, the only results needed in addition to the basic probability theory are the two named in the title of the present section. Since the objective was to make inference about parameters Θ based on observations, a natural starting point is the conditional expectation $E(g(\Theta) \mid Y) \equiv E(g(\Theta) \mid \sigma(Y))$, where $\sigma(Y) \subset \mathcal{F}$ is the sub-σ-algebra induced by the random vector Y (that is, the smallest σ-algebra with respect to which Y is measurable) and $g : \mathbb{R}^p \to \mathbb{R}$ is some Borel-measurable function. In decision theoretic framework g would be a loss function, but we do not consider formal decision making here; a decision theoretic approach to statistical inference is presented by e.g. Young and Smith (2005). The latter representation of the conditional expectation is central to its interpretation, since σ-algebras can be interpreted as information; the larger $\sigma(Y)$ is, the more information it can potentially convey on Θ. On the other hand, if Y involves no information, that is, $\sigma(Y) = \{\Omega, \emptyset\}$, then $E(g(\Theta) \mid \{\Omega, \emptyset\}) = E(g(\Theta))$, the unconditional expectation. It is also worth recalling here the definition of independence; random vectors Θ and Y are independent (denoted as $\Theta \perp\!\!\!\perp Y$) if $P(A_1 \cap A_2) = P(A_1)P(A_2)$ for all $A_1 \in \sigma(\Theta)$ and $A_2 \in \sigma(Y)$, or equivalently, in terms of conditional probabilities, $P(A_1 \mid A_2) = P(A_1)$. The relationship between conditional probability and conditional expectation is $P(A_1 \mid A_2) = E(\mathbf{1}_{A_1} \mid A_2)$, where $\mathbf{1}_{A_1}$ denotes the indicator function of event A_1. The interpretation of independence between the random variables Θ and Y is that Y involves no information relevant to learning on Θ (Dawid, 1979, p. 3). Independence can be defined equivalently in terms of probability distributions: $\Theta \perp\!\!\!\perp Y$ if the joint distribution of the concatenated random vector (Θ, Y) has the product form

$P_{\Theta,Y}(\Theta \in B_1, Y \in B_2) = P_\Theta(\Theta \in B_1)P_Y(Y \in B_2)$ for all $B_1 \in \mathcal{B}_p$ and $B_2 \in \mathcal{B}_n$.

In practice we are interested in the conditional expectation given a single observed realization $Y = y$. This can be calculated using the formula

$$E(g(\Theta) \mid Y = y) = \int_{\theta \in \mathbb{R}^p} g(\theta) P_{\Theta|Y}(\Theta \in \mathrm{d}\theta \mid Y = y). \tag{1}$$

Here $P_{\Theta|Y}$ is the (regular) conditional distribution of Θ given Y. This is more commonly known as the posterior distribution and is the basis of Bayesian inference. It is the representation of the uncertainty on the unknown parameters Θ after learning the information in the observed data y. The functional form of this distribution is usually not known directly, but rather is given in terms of other probability distributions by the Bayes' theorem

$$\begin{aligned} P_{\Theta|Y}(\Theta \in \mathrm{d}\theta \mid Y = y) &= \frac{P_{\Theta,Y}(\Theta \in \mathrm{d}\theta, Y \in \mathrm{d}y)}{P_Y(Y \in \mathrm{d}y)} \\ &= \frac{P_{Y|\Theta}(Y \in \mathrm{d}y \mid \Theta = \theta) P_\Theta(\Theta \in \mathrm{d}\theta)}{\int_{\theta \in \mathbb{R}^p} P_{Y|\Theta}(Y \in \mathrm{d}y \mid \Theta = \theta) P_\Theta(\Theta \in \mathrm{d}\theta)}. \end{aligned} \tag{2}$$

This formula is attributed to Bayes (1763), though Stigler's law of eponymy has been raised here by Stigler (1983). Here the distributions $P_{Y|\Theta}$ and P_Θ are known as the likelihood and the prior, respectively. Together they define the probability model for the phenomenon of interest.

Often the observations are made on several units or subjects judged to be in some sense "similar", and the interest lies in the common traits of such units. Model formulation in such cases can be justified by introducing the concept of exchangeability, which defines the required similarity, and its consequence, de Finetti's theorem, which introduces the common traits (parameters). The random vector $Y = (Y_1, \ldots, Y_n)$ is said to be exchangeable if the joint distribution P_Y is the same for every permutation of the indices $\{1, \ldots, n\}$ (Kingman, 1978). This can be interpreted to mean for example that the order in which the observations were collected is not informative (Bernardo, 1996) or in an epidemiological context, if exposure states of individuals would be exchanged, the joint distribution of outcomes would be unchanged (Greenland and Robins, 1986). If exchangeability holds for infinite sequences of such random variables, the representation theorem of de Finetti (1937) states that there exists a random vector Θ with a distribution P_Θ so

that

$$P_Y(Y \in \mathrm{d}y) = \int_{\theta \in \mathbb{R}^p} P_{Y|\Theta}(Y \in \mathrm{d}y \mid \Theta = \theta) P_\Theta(\Theta \in \mathrm{d}\theta)$$
$$= \int_{\theta \in \mathbb{R}^p} \left[\prod_{i=1}^{n} P_{Y_i|\Theta}(Y_i \in \mathrm{d}y_i \mid \Theta = \theta) \right] P_\Theta(\Theta \in \mathrm{d}\theta). \qquad (3)$$

For a proof, see Kingman (1978). It should first be noted that this result makes possible a purely functional definition for parameters, that is, parameters are that random vector for which (3) holds true. However, the theorem merely states the existence of the distributions $P_{Y_i|\Theta}$ and P_Θ, rather than specifying them (Bernardo, 1996). Thus further assumptions are needed in the actual model specification. Even if the result may not hold exactly when standard statistical models are substituted for these distributions, they may still serve as approximations. Thus the use of i.i.d. models in Bayesian inference can be justified by (3), with the parameters interpreted accordingly (Diaconis, 1977, p. 271, Rubin, 1987, p. 40). In the following we will introduce and interpret the model parameters according to the representation theorem.

The somewhat tricky concept of exchangeability is best illustrated with an example. It is said that an elephant is difficult to define but you know one when you see it. Now suppose that an observer has never seen an elephant before. However, after seeing one elephant, the observer should have a reasonably good idea how the next one will look like. In terms of random variables this means that the two observations clearly are not independent. However, it seems reasonable to assume that they are exchangeable, so that the joint information learned on elephants based on two observations does not depend on which one of the two elephants was seen first. Further, suppose that the observer identifies traits that are common to all of the observed elephants such as that they are large, gray and have trunks and tusks. With enough such traits identified, it is likely that the observer can no longer identify further ones. The observer can predict that the next elephant will have these features, but it will also have some unique features that are different from the previous observations. This means that given the identified traits (characteristic features, i.e., parameters of elephant), the observations are independent. This conditional independence property corresponds to the product form of the likelihood distribution in (3). It applies when the model is accurate in the sense that the parameters adequately describe the "form" or "idea" of elephant, or in less abstract terms, the features of any population of similar (exchangeable) elephants.

It is instructive to consider also situations where exchangeability does

not necessarily hold by elaborating the example. Now suppose the sequence of observations consists of elephants and mammoths. Considering these as fully exchangeable does no longer seem reasonable; an observed characteristic might have a different meaning depending on whether it was observed for an elephant or a mammoth (say, observing one hairy mammoth and one hairless elephant does not give the same information as observing a hairy elephant and a hairless mammoth). When trying to understand this in terms of probability expressions it is useful to note that permuting the indices of the random variables is equivalent to permuting their realized values. Suppose that now $n = 2$ and $Y_i \in \{0, 1\}$, corresponding to absence/presence of some trait common to mammoths but rare in elephants. Now if the first observation $i = 1$ happens to be an elephant and $i = 2$ a mammoth, it seems obvious that, for instance, $P(Y_1 = 0, Y_2 = 1) \neq P(Y_1 = 1, Y_2 = 0)$, and we do not have exchangeability. Let the random vector $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ indicate the subpopulation membership of observations $\{1, \ldots, n\}$ and $Y = (Y_1, \ldots, Y_n)$ the measured characteristic. In the present example with $n = 2$ and $X_i \in \{0, 1\}$ (elephant/mammoth), it easy to see that the exchangeability does not hold because the information on the subpopulation is implicitly included in the indices of the observations. The lesson to be learned from here is that the exchangeability postulate may be questionable when the random variables are doubly stochastic so that the indices of the observations are also random variables, the realized values of which contain information relevant to the problem. This issue is re-encountered in Section 4.3.

The previous discussion suggests that all observed information must be explicitly stated in the probability expression. It is now indeed more reasonable to assume that the joint distribution $P_{X,Y}$ is exchangeable in unit indices $i \in \{1, \ldots, n\}$, in which case the likelihood factors into a product form over the individual contributions $P_{X_i,Y_i|\Theta}$ (Rubin, 1987, p. 40). In the ongoing example we would then have

$$P(X_1 = 0, Y_1 = 0, X_2 = 1, Y_2 = 1) = P(X_1 = 1, Y_1 = 1, X_2 = 0, Y_2 = 0),$$

that is, the indices are no longer informative, and relabeling the observations does not change the joint probability. However, now it follows that

$$\begin{aligned}
&P(X_1 = 0, Y_1 = 0, X_2 = 0, Y_2 = 1) + P(X_1 = 0, Y_1 = 0, X_2 = 1, Y_2 = 1) \\
&+ P(X_1 = 1, Y_1 = 0, X_2 = 0, Y_2 = 1) + P(X_1 = 1, Y_1 = 0, X_2 = 1, Y_2 = 1) \\
&= P(X_1 = 0, Y_1 = 1, X_2 = 0, Y_2 = 0) + P(X_1 = 1, Y_1 = 1, X_2 = 0, Y_2 = 0) \\
&\quad + P(X_1 = 0, Y_1 = 1, X_2 = 1, Y_2 = 0) + P(X_1 = 1, Y_1 = 1, X_2 = 1, Y_2 = 0),
\end{aligned}$$

that is, $P(Y_1 = 0, Y_2 = 1) = P(Y_1 = 1, Y_2 = 0) = P(Y_2 = 0, Y_1 = 1)$, meaning that exchangeability in the marginal distribution follows from exchangeability in the joint distribution. While this may appear to be in conflict with

the earlier notion of no exchangeability in the marginal distribution of the observed traits, the marginalization here has to be interpreted as losing the information on the subpopulation membership. In contrast, if the observer possesses this information, it has to be included in the joint probability statement to achieve exchangeability.

Alternative approach would be to assume that the exchangeability applies within sequences of observations and between the indices of the sequences. Continuing the example, the model would be then parameterized in terms of traits which are common to both elephants and mammoths and traits which are specific to the two species. This idea corresponds to hierarchical Bayesian parametrization; by introducing parameter vectors $\Theta_k : \Omega \to \mathbb{R}^p$, $k = 1, \ldots, m$, corresponding to m subpopulations and a vector of hyperparameters $\Phi : \Omega \to \mathbb{R}^q$, de Finetti's theorem is then applied both within subpopulations and between subpopulation indices as (Bernardo, 1996)

$$P_Y(Y \in \mathrm{d}y) = \int_{\theta \in \mathbb{R}^{mp}} \left[\prod_{k=1}^{m} \prod_{i=1}^{n_k} P_{Y_{ki}|\Theta_k}(Y_{ki} \in \mathrm{d}y_{ki} \mid \Theta_k = \theta_k) \right] P_\Theta(\Theta \in \mathrm{d}\theta),$$

where n_k is the number of observations from the subpopulation k, $Y = (Y_1, \ldots, Y_m) : \Omega \to \mathbb{R}^{n_1 + \ldots + n_m}$, $\Theta = (\Theta_1, \ldots, \Theta_m) : \Omega \to \mathbb{R}^{mp}$ and

$$P_\Theta(\Theta \in \mathrm{d}\theta) = \int_{\phi \in \mathbb{R}^q} \left[\prod_{k=1}^{m} P_{\Theta_k|\Phi}(\Theta_k \in \mathrm{d}\theta_k \mid \Phi = \phi) \right] P_\Phi(\Phi \in \mathrm{d}\phi).$$

2.3 Predictive inference

In the previous section the main interest was in the model parameters, but an alternative approach would be to base the inference entirely upon observable quantities, by considering probability distributions of future events given past observations (Dawid, 1984). When more observations are being accumulated, the predictions become progressively more accurate. Predictive inference has direct applications in, for example, clinical decision making, even if the causal relationships between the factors involved would not be fully understood. Now suppose we have observed a realization of the random vector $Y = (Y_1, \ldots, Y_n)$ and want to predict the next observation, represented by the random variable $Y_{n+1} : \Omega \to \mathbb{R}$. The relevant probability distribution for this problem is the (posterior) predictive distribution

$$\begin{aligned} &P_{Y_{n+1}|Y}(Y_{n+1} \in \mathrm{d}y_{n+1} \mid Y = y) \\ &= \int_{\theta \in \mathbb{R}^p} P_{Y_{n+1}|\Theta}(Y_{n+1} \in \mathrm{d}y_{n+1} \mid \Theta = \theta) P_{\Theta|Y}(\Theta \in \mathrm{d}\theta \mid Y = y). \end{aligned} \tag{4}$$

The parametric probability model corresponding to the posterior distribution $P_{\Theta|Y}$ is reintroduced here for the purpose of translating the information in the past observations into information on the future observation. Following the previously introduced exchangeability reasoning, there exists a parametric probability distribution so that, given the parameters, the future observation Y_{n+1} is conditionally independent of Y, with the parameters involving all information relevant to the prediction.

It is at this point worthwhile to consider the difference between explanation (i.e., theory building or verification) and prediction tasks. According to the widely (though not universally) accepted principle of Occam's razor, the explanation should be as parsimonious as possible, that is, involving as little hypothetical quantities as possible (but no fewer than that). This means that if two theories are able to explain the same observations, the more parsimonious theory should be preferred. In statistical modeling the requirement for parsimony is self-evident, as the model fit can be progressively increased by adding more parameters. Now in terms of the marginal distribution of the data, $P_Y(Y \in \mathrm{d}y)$, the probability of observing any given realization y depends on two opposite effects of model complexity. A model with more parameters allows better fit to data, but on the other hand can accommodate a wider range of observations. This in turn means that the more complex model is more difficult to falsify with new observations (Jefferys and Berger, 1992). Further, the predictive accuracy of overly complex model suffers due to the added noise. Bayesian model selection is further discussed in Section 4.6; it is based on maximizing the marginal probability of the data and thus involves an inbuilt penalty for model complexity, functioning as an "automatic Occam's razor" (Smith and Spiegelhalter, 1980). Since the requirement for parsimony is present in both explanation and prediction tasks, the main difference between these is that the parameters of a prediction model are integrated out of the predictive distribution (4) and thus do not play a part in the actual inference. In contrast, parameters in an explanatory model would usually have some hypothesized real life counterparts, which would be the main target of the inference. This distinction has consequences to the selection of the best model; in the prediction task this is in principle straightforward: "the best model is the one which best predicts the fate of a future subject" (Clayton and Hills, 1993, p. 271). Prediction model may be validated by comparing the predictions to true outcomes. In contrast, choosing the best explanation model is a more ambiguous task; marginal probability of the data is only one of the various criteria suggested for this.

2.4 Survival models and marked point processes

The Articles I-III deal with modeling of censored time-to-event data, so as an example of probability models we recall here the basic concepts of (parametric) survival modeling. In the following, we compress the notation introduced in Sections 2.1 and 2.2 by denoting all probability distributions by P, with the argument indicating which distribution is in question. Also, we do not distinguish between random variables and their realized values if this is clear from the context. We are now concerned with pairs of random variables (T_i, E_i), $i = 1, \ldots, n$, where each $T_i \geq 0$ represents an event time and $E_i \in \{0, 1, \ldots, J\}$ indicates the type of the event at T_i. Here $E_i = 0$ indicates censoring, that is, end of the observation of subject i for reasons other than occurrence of any of the events of interest. Since the observations are accumulated progressively over time, it is useful to consider the situation in terms of stochastic processes, that is, sets of random variables indexed with respect to time. The events are identified by counting processes $N_{ij}(t) = \mathbf{1}_{\{T_i \leq t, E_i = j\}}$, $j = 0, 1, \ldots, J$. In addition, $Z_i(t)$ denotes a covariate process for subject i. Consistently with the previously discussed interpretation of σ-algebras as information, we can define the history $\mathcal{F}_{t-} = \sigma(\{N_{ij}(u), Z_i(u) : i = 1, \ldots, n; j = 0, 1, \ldots, J; 0 \leq u < t\})$, which represents the observed information up to but not including time point t. More information is accumulated as time passes, meaning that the sequence of histories is increasing in time, that is, $\mathcal{F}_{u-} \subseteq \mathcal{F}_{t-}$ for $u \leq t$. Using these quantities, and assuming T_i to be continuous, we can characterize cause-specific hazard (or intensity) functions $\lambda_{ij}(t)$ as

$$\begin{aligned} P(T_i \in \mathrm{d}t, E_i = j \mid \mathcal{F}_{t-}) &= P(N_{ij}(\mathrm{d}t) = 1 \mid \mathcal{F}_{t-}) \\ &= E(N_{ij}(\mathrm{d}t) \mid \mathcal{F}_{t-}) = \lambda_{ij}(t)\,\mathrm{d}t. \end{aligned}$$

The above corresponds to the problem of predicting whether an event of type j is going to occur for subject i at time $T_i \in \mathrm{d}t$, based on everything known until just before t. Consequently, the hazard function for any type of event occurring is

$$\lambda_i(t)\,\mathrm{d}t = P(T_i \in \mathrm{d}t \mid \mathcal{F}_{t-}) = P(N_i(\mathrm{d}t) = 1 \mid \mathcal{F}_{t-}) = E(N_i(\mathrm{d}t) \mid \mathcal{F}_{t-}),$$

where $\lambda_i(t) = \sum_{j=0}^{J} \lambda_{ij}(t)$ and $N_i(t) = \sum_{j=0}^{J} N_{ij}(t)$ (Arjas, 1989, p. 184-185). It should be noted that the additivity of hazards applies always when the event definitions are mutually exclusive and does not imply independence between the different event types.

Parametric survival models can now be defined in terms of hazard functions. For simplicity we take the covariates to be constant over time, with

the realized value $Z_i = z_i$ treated as fixed, and assume the current state of each individual to be conditionally independent of other individuals' histories, given the individual's own history and a vector of parameters $\Theta = \theta$. It should be noted that this assumption would not be valid, for example, in the context of contagious diseases (Kalbfleisch and Prentice, 2002, p. 152). The cause-specific hazard function now simplifies into

$$\lambda_{ij}(t)\,\mathrm{d}t = P(T_i \in \mathrm{d}t, E_i = j \mid T_i \geq t, z_i, \theta). \tag{5}$$

The (sub)distribution of the events of type j is given by (Kalbfleisch and Prentice, 2002, p. 251-252)

$$\begin{aligned} P(T_i \in \mathrm{d}t, E_i = j \mid z_i, \theta) &= P(T_i \in \mathrm{d}t, E_i = j \mid T_i \geq t, z_i, \theta) P(T_i \geq t \mid z_i, \theta) \\ &= \lambda_{ij}(t)\,\mathrm{d}t \exp\left\{ -\int_0^t \sum_{k=0}^{J} \lambda_{ik}(u)\,\mathrm{d}u \right\}. \end{aligned}$$

The above holds true without any further assumptions on the parametrization of the model. However, suppose that the parameter vector is partitioned as $\Theta = (\Theta_0, \Theta_1, \ldots, \Theta_J)$, where Θ_j, $j \in \{0, 1, \ldots, J\}$ are parameters describing specifically events of type j. Typically some of the Θ_j are not of interest (are nuisance parameters); most commonly this is the case for Θ_0, the parameters describing the censoring events. It would be desirable to avoid specification of the model components which are not of interest. This is possible if in (5) we could assume conditional independence $(T_i \in \mathrm{d}t, E_i = j) \perp\!\!\!\perp \Theta_{-j} \mid T_i \geq t, z_i, \theta_j$, where $\Theta_{-j} \equiv \{\Theta_0, \Theta_1, \ldots, \Theta_p\} \setminus \{\Theta_j\}$. Now (5) may be parameterized in terms of θ_j only and the subdistribution becomes

$$P(T_i \in \mathrm{d}t, E_i = j \mid z_i, \theta) \overset{\theta_j}{\propto} \lambda_{ij}(t) \exp\left\{ -\int_0^t \lambda_{ij}(u)\,\mathrm{d}u \right\}.$$

However, the required conditional independence assumption stated above is untestable in practice (Arjas, 1989, p. 204-205). What actually is assumed here is best illustrated with an example. Suppose an experiment where two (different kinds of) components are connected in series in an electrical circuit. Now $J = 1$, with parameters Θ_0 and Θ_1 describing the properties of the two component types, including their expected lifetime. Covariates z_i might include the properties of the experimental set-up other than those of the two components, such as the current running through the circuit. If the interest is in making inference on, say, the expected lifetime of the components of type 1, failure of the component of type 0 censors the observation on the other component. However, given that z_i involves all relevant attributes

of the experiment, it may be reasonable to assume such censoring to be noninformative (or non-innovative, Arjas, 1989, p. 204), that is, that it does not depend on the properties of the components of type 1, characterized by the parameters Θ_1. For observations from n repeats of such an experiment, we can now write a likelihood expression

$$\prod_{i=1}^{n} P(T_i \in \mathrm{d}t_i, E_i = e_i \mid z_i, \theta) \overset{\theta_1}{\propto} \prod_{i=1}^{n} \left[\lambda_{i1}(t_i)^{e_i} \exp\left\{ - \int_0^{t_i} \lambda_{i1}(u)\, \mathrm{d}u \right\} \right].$$

In both of the parameter estimation methods to be discussed in Sections 3.1 and 3.2 it is sufficient to define the likelihood function only up to a constant. With an additional requirement that the random vectors Θ_0 and Θ_1 are a priori independent, this means that if the parameters Θ_0 are not of interest, they need not be estimated (Rubin, 1976). Technically, a priori independence of parameter vectors is defined as in Section 2.2. However, since the parameters are unobservable quantities, the practical interpretation of this assumption may be difficult to grasp at first. In the previous example this could mean that if the two components originated from different manufacturers and factories, even if we had some prior information on the expected lifetime or other properties of such components in general, we do not have a reason to believe that the manufacturing processes of the two factories would be similar in such a way that the properties of the produced components would be more similar than indicated by the marginal prior distributions. A priori independence of parameters can be best understood as conditional independence given the prior information, even if such conditioning is not always explicitly written.

In Articles IV and V the underlying model structure is defined in terms of marked point processes, which, in addition to being models for spatial phenomena, are a flexible tool for constructing probability models with less rigid parametric assumptions (see e.g. Arjas and Heikkinen, 1997). In one dimension, survival models can be presented as a special case of the marked point process framework (Arjas, 1989). The following definition is from Møller and Waagepetersen (2004, p. 241-242), and is taken up here as an example of non-real valued random variables. Let $S \subseteq \mathbb{R}^p$. Further, let $B \in \mathcal{B}_0 : B \subseteq S$, where $\mathcal{B}_0$ is the class of bounded Borel sets. Point configurations $x \subseteq S$ are locally finite if $n(x_B) < \infty$, where $x_B = x \cap B$. The space of all locally finite point configurations is defined as $N_{\mathrm{lf}} = \{x \subseteq S : n(x_B) < \infty \,\forall\, B \in \mathcal{B}_0\}$. The σ-algebra induced by such sets is $\mathcal{N}_{\mathrm{lf}} = \sigma(\{x \in N_{\mathrm{lf}} : n(x_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0)$. A point process is an ($\mathcal{F}$-measurable) random variable $X : \Omega \to N_{\mathrm{lf}}$, for which $X^{-1}(F) \in \mathcal{F}$ for all $F \in \mathcal{N}_{\mathrm{lf}}$. Its distribution P_X is a probability measure on $(N_{\mathrm{lf}}, \mathcal{N}_{\mathrm{lf}})$ and its realizations are locally finite point configurations $x = (x_1, \ldots, x_{n(x)})$, where $x_j \in S$, $j = 1, \ldots, n(x)$. A

marked point process Y (with mark space $A \subseteq \mathbb{R}$) is obtained by attaching a random variable (mark) $E_j : \Omega \to A$ to each point $T_j \in S$ of a point process X: $Y = \{(T_j, E_j) : T_j \in X\} \subset S \times A$. The connection to the earlier time-to-event data situation is obvious from this notation; if $S = [0, \infty)$, the point locations can be interpreted as event times, and marks indicate what happens at each event time, that is, type of the event or censoring.

A building block for more complicated point processes is often the Poisson point process. Let $\rho : S \to [0, \infty)$ be an intensity function and $\mu(B) = \int_B \rho(x_j)\,\mathrm{d}x_j$ an intensity measure. The distribution for a Poisson point process on S with intensity function ρ can be written for any $B \subseteq S : \mu(B) < \infty$ and $F \subseteq N_{\mathrm{lf}}$ as (Møller and Waagepetersen, 2004, p. 15)

$$
\begin{aligned}
P(X_B \in F) &= \sum_{n=0}^{\infty} P(X_B \in F \mid n) P(N(B) = n) \\
&= \sum_{n=0}^{\infty} \int_{x_1 \in B} \cdots \int_{x_n \in B} \mathbf{1}_{\{(x_1,\ldots,x_n) \in F\}} \left[\prod_{j=1}^{n} \frac{\rho(x_j)}{\mu(B)}\,\mathrm{d}x_j \right] \frac{\mu(B)^n}{n!} \exp\{-\mu(B)\} \\
&= \sum_{n=0}^{\infty} \frac{\exp\{-\mu(B)\}}{n!} \int_{x_1 \in B} \cdots \int_{x_n \in B} \mathbf{1}_{\{(x_1,\ldots,x_n) \in F\}} \left[\prod_{j=1}^{n} \rho(x_j)\,\mathrm{d}x_j \right]. \qquad (6)
\end{aligned}
$$

Here the (random) number of points $N(B)$ is Poisson distributed with mean $\mu(B)$ and the point configuration given the number of points consists of independent and identically distributed points with density $f(x_j) = \rho(x_j)/\mu(B)$ (this is known as a binomial point process). The above distribution is fully defined by the intensity function. If ρ is constant, the process is called a homogeneous Poisson process.

3 Estimation

3.1 Markov chain Monte Carlo

Formula (1) already suggested how Bayesian inference on parameters $\Theta : \Omega \to \mathbb{R}^p$ might be carried out. A closed form for the posterior distribution can be obtained only in special cases where the likelihood and the prior are conjugate so that application of Bayes' formula gives a posterior distribution of similar form as the prior, with "updated" parameter values. In the general case the inference utilizes simulation. The term "Monte Carlo method" was coined by Metropolis and Ulam (1949) and refers to a family of computational methods where simulations based on computer generated random numbers are used to find approximate solutions to mathematical

problems. In Monte Carlo integration the integral of the type (1) is approximated by $\overline{g}_m = \frac{1}{m}\sum_{k=1}^{m} g\left(\theta^{(k)}\right)$, where $\left(\theta^{(1)}, \ldots, \theta^{(m)}\right)$ is an independent random sample from the posterior distribution $P_{\Theta|Y}$. By the strong law of large numbers, now $\overline{g}_m \overset{\text{a.s.}}{\to} E(g(\Theta) \mid Y = y)$ when $m \to \infty$ (Robert and Casella, 2004, p. 83). If the functional form of $P_{\Theta|Y}$ is unknown, drawing independent random samples from it is not straightforward either. Fortunately, the above convergence still holds true if the sample is obtained from a Markov chain with stationary distribution $P_{\Theta|Y}$. A Markov chain is a sequence of random variables $\Theta^{(0)}, \Theta^{(1)}, \Theta^{(2)}, \ldots$ which, for any $k \in \mathbb{N}$, has the conditional independence property $\Theta^{(k+1)} \perp\!\!\!\perp \left(\Theta^{(0)}, \Theta^{(1)}, \ldots, \Theta^{(k-1)}\right) \mid \Theta^{(k)}$. The probability of a state transition in a Markov chain is often denoted as $K(\theta, A) = \int_{\theta' \in A} K(\theta, \mathrm{d}\theta') = P_{\Theta^{(k+1)}|\Theta^{(k)}}\left(\Theta^{(k+1)} \in A \mid \Theta^{(k)} = \theta\right)$, where K is known as a transition kernel. The transition kernel and the resulting chain are here taken to be time-homogeneous (stationary), meaning that the superscript can be dropped from the notation. A sufficient condition for the chain to have the target stationary distribution $P_{\Theta|Y}$ is that the transition kernel satisfies

$$P_{\Theta|Y}(\Theta \in A \mid Y = y) = \int_{\theta \in \mathbb{R}^p} K(\theta, A) P_{\Theta|Y}(\Theta \in \mathrm{d}\theta \mid Y = y) \tag{7}$$

for all $A \in \mathcal{B}_p$. A well behaving Markov chain should also be irreducible, recurrent and aperiodic. Omitting formal definitions, these three properties mean that the chain moves (communicates) between all states $A \in \mathcal{B}_p$, that each state is visited often enough, and that the moves are not deterministic in nature.

There are a variety of methods for constructing a Markov chain fulfilling (7); we will review here one which is widely used in applications, known as "Metropolized Gibbs sampler" or "Metropolis-within-Gibbs" (Robert and Casella, 2004, p. 392-394). First, the transition kernel is chosen as

$$\begin{aligned} K(\theta, \mathrm{d}\theta') = {} & P_{\Theta_1|\Theta_{-1},Y}(\Theta_1 \in \mathrm{d}\theta_1' \mid \theta_2, \ldots, \theta_p, y) \\ & \times P_{\Theta_2|\Theta_{-2},Y}(\Theta_2 \in \mathrm{d}\theta_2' \mid \theta_1', \theta_3, \ldots, \theta_p, y) \\ & \times \ldots \times P_{\Theta_p|\Theta_{-p},Y}(\Theta_p \in \mathrm{d}\theta_p' \mid \theta_1', \ldots, \theta_{p-1}', y), \end{aligned}$$

where $\Theta_{-j} \equiv \{\Theta_1, \ldots, \Theta_p\} \setminus \{\Theta_j\}$. This corresponds to updating each of the components Θ_j, $j = 1, \ldots, p$, in turn from their univariate full conditional distributions. This algorithm is known as the Gibbs sampler. The name was introduced by Geman and Geman (1984), who devised the algorithm for sampling from Gibbs distributions, a very general distribution family. Despite being nondescriptive, the name has stuck (Banerjee et al., 2004, p. 113). A result known as the Hammersley-Clifford theorem states that the

joint distribution can always be expressed as a function of the set of full conditional distributions, which thus contain sufficient information for sampling from the joint distribution (e.g. Robert and Casella, 2004, p. 377). As a side note, the converse is not true, that is, a set of full conditional distributions does not necessarily define any joint distribution. Now, each of the subchains $\Theta_j^{(0)}, \Theta_j^{(1)}, \Theta_j^{(2)}, \ldots$ is also a Markov chain. Using shorthand notations $\theta'_{1:(j-1)} \equiv (\theta'_1, \theta'_2, \ldots, \theta'_{j-1})$ and $\theta_{j:p} \equiv (\theta_j, \theta_{j+1}, \ldots, \theta_p)$, the transition kernel for updating a subchain j is chosen as

$$\begin{aligned} K(\theta_j, \mathrm{d}\theta'_j) &= P(\Theta_j \in \mathrm{d}\theta'_j \mid \theta'_{1:(j-1)}, \theta_{j:p}, y) \\ &= P(U < \alpha \mid \theta'_{1:(j-1)}, \theta_{j:p}, \theta'_j, y) P_{\widetilde{\Theta}_j \mid \Theta, Y}(\widetilde{\Theta}_j \in \mathrm{d}\theta'_j \mid \theta'_{1:(j-1)}, \theta_{j:p}, y) \\ &\quad + (1 - a)\mathbf{1}_{\{\theta_j \in \mathrm{d}\theta'_j\}}, \end{aligned}$$

where

$$a = \int_{\theta'_j \in \mathbb{R}} P(U < \alpha, \mid \theta'_{1:(j-1)}, \theta_{j:p}, \theta'_j, y) P_{\widetilde{\Theta}_j \mid \Theta, Y}(\widetilde{\Theta}_j \in \mathrm{d}\theta'_j \mid \theta'_{1:(j-1)}, \theta_{j:p}, y).$$

As an algorithm this works by first drawing an auxiliary random variate $\widetilde{\Theta}_j$ from an instrumental (proposal) distribution. The proposed value θ'_j is accepted with probability

$$\begin{aligned} \alpha &= \min \left\{ 1, \frac{P_{Y \mid \Theta}(Y \in \mathrm{d}y \mid \theta'_{1:j}, \theta_{(j+1):p}) P_\Theta(\Theta_{1:j} \in \mathrm{d}\theta'_{1:j}, \Theta_{(j+1):p} \in \mathrm{d}\theta_{(j+1):p})}{P_{Y \mid \Theta}(Y \in \mathrm{d}y \mid \theta'_{1:(j-1)}, \theta_{j:p}) P_\Theta(\Theta_{1:(j-1)} \in \mathrm{d}\theta'_{1:(j-1)}, \Theta_{j:p} \in \mathrm{d}\theta_{j:p})} \right. \\ &\qquad \left. \times \frac{P_{\widetilde{\Theta}_j \mid \Theta, Y}(\widetilde{\Theta}_j \in \mathrm{d}\theta_j \mid \theta'_{1:j}, \theta_{(j+1):p}, y)}{P_{\widetilde{\Theta}_j \mid \Theta, Y}(\widetilde{\Theta}_j \in \mathrm{d}\theta'_j \mid \theta'_{1:(j-1)}, \theta_{j:p}, y)} \right\} \qquad (8) \\ &= \min \left\{ 1, \frac{P_{\Theta, Y, \widetilde{\Theta}_j}(\Theta_{1:j} \in \mathrm{d}\theta'_{1:j}, \Theta_{(j+1):p} \in \mathrm{d}\theta_{(j+1):p}, Y \in \mathrm{d}y, \widetilde{\Theta}_j \in \mathrm{d}\theta_j)}{P_{\Theta, Y, \widetilde{\Theta}_j}(\Theta_{1:(j-1)} \in \mathrm{d}\theta'_{1:(j-1)}, \Theta_{j:p} \in \mathrm{d}\theta_{j:p}, Y \in \mathrm{d}y, \widetilde{\Theta}_j \in \mathrm{d}\theta'_j)} \right\} \end{aligned}$$

by drawing an auxiliary random variate $U \sim \mathrm{Uniform}[0, 1]$. Otherwise, the chain stays at the current state θ_j. Depending on the choice of the proposal distribution, this method is known as the Metropolis or Metropolis-Hastings algorithm (by Metropolis et al., 1953; Hastings, 1970). Two things should be noted here. First, the target distribution here is the posterior full conditional distribution $P_{\Theta_j \mid \Theta_{-j}, Y}$, but this has been replaced in (8) with the joint posterior $P_{\Theta \mid Y}$. This can be done because the two are proportional with respect to Θ_j. Second, the Bayes' formula (2) has been applied here to present the posterior as a function of the likelihood $P_{Y \mid \Theta}$ and the prior P_Θ, with the denominators P_Y canceling out. Also, any multiplicative terms

in the likelihood not depending on the parameters cancel out and thus it is sufficient to specify the likelihood only up to a constant. The functional form of what is left is known and can be numerically evaluated. One commonly used choice for a proposal distribution, known as random walk Metropolis, is $\widetilde{\Theta}_j \mid \theta_j \sim N(\theta_j, s^2)$, where the (fixed) standard deviation s may be adjusted during an initial "burn-in" period to achieve an acceptance rate which produces a well mixing chain. This proposal is symmetric and thus the proposal ratio cancels out of (8).

In the Metropolis-Hastings ratio (8) the dimension p of the parameter vector was fixed, with the same joint measure (the product of target and proposal distributions) $P_{\Theta,Y,\widetilde{\Theta}_j}$ appearing in both the numerator and denominator. Use of this standard form is limited to moves within a fixed parameter subspace, since in moves between parameter subspaces of different dimension the measures in the numerator and denominator would typically not match. An important extension to Markov chain Monte Carlo was made by Green (1995) who showed that such moves are possible by requiring that the joint measure is absolutely continuous with respect to some appropriate symmetric measure. By the Radon-Nikodym theorem, the joint measure then has a density with respect to the symmetric measure. The Metropolis-Hastings ratio can then be written as a ratio of these densities. The symmetry ensures reversibility of the moves, hence the name "reversible jump MCMC". It should be noted that also standard Metropolis-Hastings moves are reversible (satisfy the detailed balance condition, e.g. Robert and Casella, 2004, p. 230), so again the name may not be the most descriptive, unless it is taken to mean the generalized Metropolis-Hastings procedure as a whole, and not just the Green's extension. A variable dimension algorithm would be needed, for instance, in simulating realizations from a point process of the type (6), due to the random number of points in a realization (see Møller and Waagepetersen, 2004, p. 112-115, for an example of such an algorithm).

In addition to the methods outlined above, Bayesian inference requires specification of a prior distribution. Sometimes no strong prior information exists or its use is not seen as appropriate. Thereby, much attention in Bayesian literature has been paid to the development of noninformative (or reference) priors (see e.g. Berger and Bernardo, 1992, for a review). The noninformativeness is then defined with respect to some specific criterion; the best known example is the Jeffreys' prior (Jeffreys, 1946) which is proportional to the square root of the determinant of the Fisher information. It is invariant to one-to-one transformation of parametrization. Although Fisher information does not depend on the data other than through the number of observations, Jeffreys' prior ties the prior specification to the likelihood specification, and thus has been seen as violating the likelihood principle (see

Berger and Wolpert, 1988, p. 21). Nevertheless, in many cases it produces an intuitively plausible result; for example, in the case of Gaussian observations $Y_i \mid \mu, \sigma \sim N(\mu, \sigma^2)$, the Jeffreys' priors for the mean (location) and standard deviation (scale) parameters are $f(\mu) \propto 1$ (flat over the real line) and $f(\sigma) \propto 1/\sigma$, respectively. The latter prior assigns the same probability to any interval (a, ca) with fixed c and thus is noninformative on the "scale" of the observations. These priors are improper (nonintegrable) but can be used as long as the posterior remains proper (is integrable, and thus is a probability measure).

The noninformativeness may well be understood without any information criterion other than the probability measures involved. If we wanted to be a priori noninformative about the parameters, we would like the inference to be driven only by the data and not the prior. It is easy to see from (8) that this can be achieved by specifying the prior to be flat (uniform) over its support. The posterior is then proportional to the likelihood, with the Metropolis-Hastings moves with a symmetric proposal based only on the likelihood ratio. If the support interval is infinite, such a prior is improper, in which case it is required that the likelihood is integrable with respect to the parameters. Since this may be difficult to check in practice (e.g. Robert and Casella, 2004, p. 406), it may be preferable to favor proper priors where possible. An important application of improper priors are intrinsic autoregressive models, (see e.g. Rue and Held, 2005), which are defined through full conditional distributions, but do not have a joint distribution. Such distributions may be used as a part of hierarchical parametrization, but not as a model for observed data.

Non-invariance of flat priors to transformation of parametrization is sometimes seen as a critical issue in the Bayesian approach (e.g. Cox, 2006, p. 73-75). However, if the prior is specified as noninformative for the current parametrization of interest, it is difficult to see why non-informativeness with respect to some other parametrization would be relevant. Since no universally accepted solution for "objective" prior specification exists, the Bayesian approach has been rejected by many, albeit sometimes acknowledging its superiority in some other aspects, on the basis of striving for scientific objectivity (e.g. Efron, 1986; Cox, 2006, p. 199). Bayesians, for their part, value more highly the internal coherency and simplicity of their framework. Moreover, the fundamental difference between the Bayesian approach and its competitors lies in the interpretation of the concept of probability, rather than the prior specification (cf. Jaynes, 2003, p. 499-500). If noninformativeness in prior specification is not the aim, then it is enough to require that the prior assumptions made are explicitly stated and that the analysis is repeatable with alternative priors.

3.2 Maximum likelihood

Should only the likelihood function be used for inference on the unknown parameters, the Bayes' formula can be better interpreted in the form

$$\frac{P_{\Theta|Y}(\Theta \in \mathrm{d}\theta \mid Y = y)}{P_{\Theta}(\Theta \in \mathrm{d}\theta)} \overset{\theta}{\propto} P_{Y|\Theta}(Y \in \mathrm{d}y \mid \Theta = \theta),$$

where the proportion on the left hand side reflects the information lost when the prior is ignored. In the previous section it was noted that flat priors over their support are noninformative in the sense that then the inferences are based on the likelihood only. While formula (1) might have implied that posterior mean would have a special status as the "estimator" to be used in the inference, any other descriptive statistic of the posterior would be equally allowed. Assuming that the relevant densities exist, now if $f_{\Theta}(\theta) \propto 1$, it so happens that

$$\arg\max_{\theta} f_{\Theta|Y}(\theta \mid y) = \arg\max_{\theta} f_{Y|\Theta}(y \mid \theta). \tag{9}$$

With flat priors over infinite intervals, the posterior may or may not be a proper distribution. However, the posterior mode may be a sensible statistic regardless of whether the posterior distribution is proper, which would not be the case for, say, the posterior mean. Hence, with such flat priors, the right hand side of (9), known as the maximum likelihood estimator, may be used for making inference on the parameters without the requirement for a proper posterior distribution. We limit the discussion here to unimodal likelihoods; multimodal likelihoods are usually pathological cases. In addition to a point estimate, some kind of estimate for its accuracy, based on the observed data, is needed as well. As noted previously, the posterior distribution itself is a (probability) measure for this accuracy (cf. Jaynes, 2003, p. 501), which can then be expressed with any appropriate descriptive statistic, such as standard deviation or credible interval, calculated from a random sample drawn from the posterior using the methods outlined in Section 3.1. An alternative approach would be to use deterministic Gaussian (Laplace) approximation for the posterior. Although this method in direct posterior approximation has seen limited use since the breakthrough of MCMC methods, it is useful in constructing multivariate proposal distributions (e.g. Rue and Held, 2005, p. 167-171). Recently, Laplace approximations in Bayesian inference have been resurrected by Rue et al. (2009). Use of Gaussian approximation is better justified in the above situation with flat priors, since the posterior is then proportional to the likelihood and by (3), the likelihood is often a product of conditionally independent contributions, making the approximation more

accurate as n increases (Clayton and Hills, 1993, p. 78). Introducing the notations $\log f_{Y|\Theta}(y \mid \theta) \equiv l(\theta)$ and $\arg\max_\theta f_{Y|\Theta}(y \mid \theta) = \arg\max_\theta l(\theta) \equiv \hat{\theta}$, the approximation is based on a second order Taylor expansion for the log-likelihood at its mode:

$$\begin{aligned} l(\theta) &= l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2 + O((\theta - \hat{\theta})^3) \\ &\overset{l'(\hat{\theta})=0}{=} l(\hat{\theta}) + \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2 + O((\theta - \hat{\theta})^3) \\ \Rightarrow l(\theta) - l(\hat{\theta}) &\approx \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2. \end{aligned}$$

The above suggests that, with flat priors, the posterior around its mode could be approximated by the Gaussian distribution $N(\hat{\theta}, -l''(\hat{\theta})^{-1})$ (with the mode $\hat{\theta}$ considered as fixed), and that the approximation becomes progressively more accurate when the number of observations increases. The asymptotic result holds true also with proper priors with some added regularity conditions (Walker, 1969; Cox and Hinkley, 1974, p. 399-400). Standard asymptotic theory for the maximum likelihood estimator (e.g. Cox, 2006, p. 100-102) obtains a very similar looking result; approximate convergence in distribution $\hat{\theta} \overset{\sim}{\to} N(\theta, E[-l''(\theta)]^{-1})$ when $n \to \infty$. Here $E[-l''(\theta)]$, the Fisher information, is unknown, and is in practice replaced with the observed information $-l''(\hat{\theta})$ at the maximum likelihood point. The roles of the parameter values and estimator have here been reversed, with the mode $\hat{\theta}$ now considered as random and θ fixed. In practice, this reversal is of no consequence (though in principle the consequences are drastic), since the inference is in both cases based on using $\hat{\theta}$ as a point estimator and $-l''(\hat{\theta})^{-1}$ as an estimator for its accuracy. Thus maximum likelihood inference, whether understood in terms of its dedicated theory, or as approximate Bayesian inference, as in the presentation herein, can be carried out by first maximizing the log-likelihood function, calculating its second derivative (or Hessian matrix) at the maximum likelihood point, and inverting the latter. Standard numerical tools are available for these tasks.

4 Inference and incomplete observation

4.1 A framework for incomplete observation problems

D. B. Rubin (1976) in his paper "Inference and missing data" laid out some general conditions for valid statistical inference in the presence of missing data, to be reviewed below. The term "missing data" traditionally refers to elements in a data matrix which for one reason or another have not been

observed (e.g. Little and Rubin, 1987, p. 3). This meaning is narrow in the sense that it refers only to observable quantities which would have been observed, say, had the data collection methods been better, had there been more resources or had the study subjects been more cooperative. An example of a situation involving such missing data is nonresponse in sample surveys (e.g. Rubin, 1987; Särndal and Lundström, 2005), where the data items to be collected are decided in advance, but all of them are not obtained for reasons mostly outside the survey makers' control. Instead of missing data we prefer here to talk about incomplete observation or even more generally, incomplete information. As discussed in Section 2.2, statistical inference essentially is quantifying our incomplete information on the phenomena of interest, given actually observed information, using probability measures as instruments. In practice incomplete observation is ever-present in statistical inference and what unifies different situations involving missing data or other kinds of incomplete observation is that all of them can be handled using exactly the same statistical methods. This is because for a Bayesian any unobserved quantity, be it a parameter, missing data item, latent variable, random effect, or an underlying true value with an imprecise measurement, is the same kind of mathematical object, that is, a random variable without an observed realization. Random variables with an observed realization can be conditioned upon, and given the observations, the uncertainty that is left can be expressed by the joint posterior distribution of all unobserved quantities. One may on philosophical grounds argue for separation of quantities into "ontological" and "epistemological" (e.g. Skrondal and Rabe-Hesketh, 2004, p. 4); not wanting to wade into such discussion here, it is sufficient to note that in probability-based statistical inference such a distinction does not exist, and thus all inference problems involving incomplete observation can be presented under a single unified framework, that is, probability.

4.2 Two-phase study design

Rubin (1987) presented a probabilistic approach to sample surveys subject to nonresponse. Here we apply similar principles to two-phase study designs where the first phase is, for instance, a sample survey of some population of interest and the second phase involves a subsampling within the first phase sample, utilizing the information collected at the first phase (e.g. Särndal et al., 1992, p. 344). As a side note, in epidemiological literature the term "two-stage design" is sometimes used interchangeably with "two-phase" (e.g. Breslow and Cain, 1988; Zhao and Lipsitz, 1992; Langholz, 2007); however, in light of sampling literature, the latter term is correct, with the former referring to clustered designs where subsampling is carried out within clusters

selected in the first stage (see Särndal et al., 1992, p. 134). The use of two-phase data collection is usually motivated by cost-effectiveness concerns, so that information which is easier or less expensive to obtain is collected on a larger group in the first phase, and this information in turn can be utilized for constructing a more efficient second phase sampling design to collect further information on a smaller subgroup (see e.g. Karvanen et al., 2009). A related concept here is that of sampling frame; according to Särndal et al. (1992, p. 9), this is defined as any materials or devices which identify and allow access to the elements of the (finite) target population of interest. The sampling frame includes also any auxiliary information which may be utilized in the sampling; using the auxiliary information, the sampling design can be made more efficient or tailor-made for specific study purposes. Population registers are commonly used as sampling frames in the Nordic countries. "Study base" is a term sometimes used in the same meaning as sampling frame (Langholz and Goldstein, 2001, for example).

Cohort sampling designs commonly utilized in epidemiological studies, such as the case-cohort and nested case-control (risk set sampling) designs (Prentice, 1986, Langholz and Goldstein, 1996; for further references, see Kulathinal et al., 2007, and Article II), are examples of two-phase designs. Here the cohort, possibly recruited using a sample survey, is often representative of some target population of interest. The cohort, together with the information collected on all cohort members, such as longitudinal follow-up data, forms a sampling frame for the second phase selection, which is typically carried out to collect additional covariate data. We introduce some notation to cover such designs, for the most part following Rubin (1987, p. 28-30). Let $F = \{1, \ldots, N\}$ be the (fixed) first phase frame population, and $X_F = \{X_i : i \in F\}$ the variables available from the frame. Let $I_F = \{I_i : i \in F\}$ be the vector of indicators for inclusion in the first phase sample, and let $Y_F = \{Y_i : i \in F\}$ represent the variables to be collected on the subjects with $I_i = 1$. The joint probability distribution of the inclusion indicators is called a sampling design (or sampling mechanism; Rubin, 1987, p. 35). If the mechanism depends only on the information available from the frame so that $I_F \perp\!\!\!\perp Y_F \mid X_F$, it is said to be unconfounded (with Y_F) (Rubin, 1987, p. 36).

It should be noted here that traditional finite population inference considers the quantities in the population to be fixed, with the uncertainty in the inference induced only by the random sampling. Further, the parameters to be estimated are taken to be some functions of the variable values in the target population (see Särndal et al., 1992, p. 39). In contrast, here we adopt a probabilistic approach where we assume exchangeability over unit indices of the random vectors (X_i, Y_i), $i \in F$, with the exchangeability extending to

further units $N+1, N+2, \ldots$ outside the set F. The latter postulate demands an explanation, since it appears to be completely contradictory with the idea of a finite population. However, since our aim in making such a postulate is to apply de Finetti's theorem for introducing a parametric probability model and its parameters, the fundamental question here is what kind of parameters we are interested in. Again, we assume that the parameters can be interpreted to be the properties of a generic further individual, outside the set F, but in some suitable sense similar to the individuals in F. This similarity is defined by exchangeability, and the parameters introduced by applying the representation theorem characterize both the individuals in F and the further generic individual.

It is possible that instead of the above, the parameters of interest really are some simple functions of the values (X_i, Y_i) specifically for $i \in F$. An example might be the rate of unemployment in Finland, in which case, taking F to be the frame population of Finland obtained from the population register, and $Y_i = (Y_{1i}, Y_{2i})$ being the unemployment status and labor force participation of individual i, respectively, the population parameter would be defined as $\theta = \sum_{i=1}^{N} Y_{1i} / \sum_{i=1}^{N} Y_{2i}$. It seems apparent that i.i.d. probability models are not an appropriate tool for making inference on such parameters. However, as soon as the parameter of interest describes, for instance, the association between two variables, we move onto the realm of the parametric probability model. This is because we then usually believe (or hope) that the parameter has some other (say, causal) interpretation beside the narrow finite population definition. There are some similarities between the finite population parameter definition and the definition through the representation theorem (3), in the sense that, in the latter the parameters are a limit of some function of an infinite exchangeable sequence of random variables (Bernardo, 1996). However, the important difference is that in the latter case the parameters describe the common properties of every subset of such random variables, while in the former case the interpretation of parameters is confined to some specific finite set.

Alternatively, in the sense of Diaconis and Freedman (1980), we could have assumed exchangeability only in the (hopefully large enough) finite set F, and applied the representation theorem in the subset $C = \{i : I_i = 1\} \subset F$ (the realized sample), hoping that it would still give a reasonable approximation. However, due to the reasons stated, we see no real need for such a restriction, and opt to proceed with the assumption that exchangeability extends to infinite sequences.

Now by applying (3) we can write a joint probability distribution

$$
\begin{aligned}
P(X_F, Y_F) &= \int_\theta \int_\phi \left[\prod_{i=1}^{N} P(Y_i, X_i \mid \theta, \phi) \right] P(\Theta \in \mathrm{d}\theta, \Phi \in \mathrm{d}\phi) \\
&\overset{\Theta \perp\!\!\!\perp \Phi}{=} \int_\theta \left[\prod_{i=1}^{N} P(Y_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta) \\
&\quad \times \int_\phi \left[\prod_{i=1}^{N} P(X_i \mid \phi) \right] P(\Phi \in \mathrm{d}\phi) \qquad (10) \\
&= P(Y_F \mid X_F) P(X_F).
\end{aligned}
$$

Again, we suppress the subscripts determining the probability distributions and do not distinguish random variables and their realized values, whenever it is clear from the context. Here we assume that only the parameters Θ are of interest. Let $C = \{i : I_i = 1\} \subseteq F$ be a set of indices representing the subjects on whom the Y_i variables are observed; unlike F, this is a random set. Writing a likelihood expression for the set F as in (10) would involve integration over the unobserved elements $Y_{F \setminus C}$. If the frame population is very large (for instance, the population of a country), such a likelihood is rarely used directly, though exceptions exist (see e.g. Pitkäniemi et al., 2009). Here we restrict the attention to situations where we can ignore the first phase sampling design. Sampling from a finite population typically means that inclusion indicators I_F are not independent. However, by applying de Finetti's theorem and assuming unconfounded sampling mechanism, we can still obtain the customary product form for the likelihood expression. Consider now the probability distribution for observed data (I_F, X_F, Y_C), which factors into $P(I_F, X_F, Y_C) = P(Y_C \mid I_F, X_F) P(I_F \mid X_F) P(X_F)$, reflecting the order of data collection. The latter two terms give no information on the parameters Θ if X_i are observed on all $i \in F$ and the probability distribution for I_F depends on X_F only, so that $I_F \perp\!\!\!\perp \Theta \mid X_F$. The remaining term can

be further written as

$$\begin{aligned}
P(Y_C \mid I_F, X_F) &= \int_{y_{F\setminus C}} P(Y_C, Y_{F\setminus C} \in \mathrm{d}y_{F\setminus C} \mid I_F, X_F) \\
&= \int_{y_{F\setminus C}} P(Y_F \in \mathrm{d}y_F \mid I_F, X_F) \\
&\overset{I_F \perp\!\!\!\perp Y_F \mid X_F}{=} \int_{y_{F\setminus C}} P(Y_F \in \mathrm{d}y_F \mid X_F) \\
&\overset{(10)}{=} \int_{y_{F\setminus C}} \int_\theta \left[\prod_{i=1}^{N} P(Y_i \in \mathrm{d}y_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta) \\
&= \int_\theta \left[\prod_{i \in C} P(Y_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta).
\end{aligned}$$

Therefore, with these assumptions, we can ignore the first phase sampling design and write the likelihood expression directly for the obtained cohort C. Now the set C determines the sampling frame for the second phase, and observed values of Y_C are available in the second phase sampling design. Analogously to above, let $R_C = \{R_i : i \in C\}$ be the inclusion indicators in the second phase selection and let $Z_C = \{Z_i : i \in C\}$ be additional information, to be collected only on the subjects with $R_i = 1$. Further, let $O = \{i : R_i = 1\} \subseteq C$ be an index set, representing the individuals on whom the Z_i are collected. The following sections review different alternatives for analysis of such study designs.

4.3 Full likelihood

If we assume the second phase sampling design to be unconfounded (with Z_C), meaning that $R_C \perp\!\!\!\perp Z_C \mid X_C, Y_C$, the probability distribution for observed data can again be factored to reflect the order of data collection as

$$P(R_C, X_C, Y_C, Z_O) = P(Z_O \mid R_C, X_C, Y_C) P(R_C \mid X_C, Y_C) P(Y_C \mid X_C) P(X_C).$$

It seems reasonable to assume the second phase sampling mechanism to be independent of Z_C if none of the values Z_i have yet been observed when the selection probabilities are fixed. However, in many sampling designs the selection probabilities are determined only implicitly. Nevertheless, if the observed Z_i values are never utilized in the sampling procedure, it fulfils the independence condition. The second term in the above now fixes the sampling design and gives no information on the model parameters.

In the previous section we saw that with the assumptions made therein, we can ignore the first phase sampling design and consider the obtained cohort C as fixed. Analogously to the previous section, we assume exchangeability over unit indices of the random vectors (X_i, Y_i, Z_i), $i \in C$, and the extension to infinite sequences, resulting in the existence of parametric models $P(Y_i, Z_i \mid X_i, \theta)$ and $P(X_i \mid \phi)$ and the corresponding prior distributions $P(\Theta \in \mathrm{d}\theta)$ and $P(\Phi \in \mathrm{d}\phi)$. It should be noted that Rubin (1987, p. 104) assumed also the response indicators to be exchangeable. We do not make the corresponding assumption on the inclusion indicators R_i, since unlike response propensity, inclusion in the sample is not a property of the individuals themselves. In any case, such an assumption is not needed for obtaining the results that follow. Also, due to the reservations expressed in Section 2.2, below we always apply de Finetti's representation theorem in the full cohort C rather than in the random subset O.

If the parameters Φ are of no interest, we can restrict the attention to the first and third terms of the above joint distribution, which can be further written as

$$
\begin{aligned}
&P(Z_O \mid R_C, X_C, Y_C) P(Y_C \mid X_C) \\
&= \int_{z_{C \setminus O}} P(Z_O, Z_{C \setminus O} \in \mathrm{d}z_{C \setminus O} \mid R_C, X_C, Y_C) P(Y_C \mid X_C) \\
&= \int_{z_{C \setminus O}} P(Z_C \in \mathrm{d}z_C \mid R_C, X_C, Y_C) P(Y_C \mid X_C) \\
&\overset{R_C \perp\!\!\!\perp Z_C \mid X_C, Y_C}{=} \int_{z_{C \setminus O}} P(Z_C \in \mathrm{d}z_C \mid X_C, Y_C) P(Y_C \mid X_C) \\
&= \int_{z_{C \setminus O}} P(Y_C, Z_C \in \mathrm{d}z_C \mid X_C) \\
&= \int_{z_{C \setminus O}} \int_\theta \left[\prod_{i \in C} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta) \\
&= \int_\theta \left[\prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta) \prod_{i \in C \setminus O} \int_{z_i} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta).
\end{aligned}
$$

Under unconfounded sampling design and the exchangeability assumption we can then ignore the selection process behind the observations and base

the inference on the posterior

$$P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_O) \tag{11}$$
$$\overset{\theta}{\propto} \left[\prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta) \prod_{i \in C \setminus O} \int_{z_i} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta).$$

This same result was obtained in Article II from a non-Bayesian point of view, without reference to de Finetti's result. The independence assumption over the unit indices (given the parameters) there corresponds to the exchangeability herein. We call the likelihood expression in (11) a full likelihood, in the sense that, being a product over the individual contributions in the full cohort C, it utilizes all observed data.

In most of the common cohort sampling designs the selection is conditionally independent of the information to be collected under the design, given the data observed on all cohort members. However, it is possible to imagine sequential sampling designs (for instance, to meet a given quota in different exposure status categories, cf. Langholz and Goldstein, 2001, p. 70), where already observed values of Z_i can affect the sampling probabilities of later observations (for instance, stop when quota are full, continue otherwise). However, it is possible to show that a product form likelihood can be obtained also in this case. The situation is then analogous to Rubin (1987, p. 51-53), in that the joint distribution of the inclusion indicators R_C may depend on all observed data but not on the unobserved values themselves. We then say that the sampling mechanism is ignorable, or equivalently, that the unobserved data are missing at random (Rubin, 1976, p. 584). In the survey nonresponse problem discussed in Rubin (1987), the actual response mechanism is usually unknown, and likelihood-based inference requires (possibly unverifiable) assumptions about it. In contrast, the mechanism producing missingness in the second phase data collection is known; it is fixed by the second phase sampling design. We say that observations $Z_{C \setminus O}$ are missing by design (e.g. Wacholder and Weinberg, 1994; Wacholder, 1996).

Same principles apply when there are other missing data in addition to those missing by design. For instance, if the covariates to be collected are genotypes, after selection of subjects to be genotyped, it may turn out that the DNA amount or concentration is too low or the genotyping is otherwise unsuccessful. Another example involving further incomplete observation are haplotypes which are only partially identifiable from unphased genotype data (see Article I). Let $M \subseteq O$ denote the set of subjects selected in the second phase but for whom the data collection was unsuccessful. If we assume that these data are missing at random, the posterior distribution can be written

as

$$P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_{O\setminus M}) \tag{12}$$
$$\overset{\theta}{\propto} \left[\prod_{i \in O\setminus M} P(Y_i, Z_i \mid X_i, \theta) \prod_{i \in C\setminus O\cup M} \int_{z_i} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta).$$

The likelihood expression here involves integration over the missing observations. If the integration can be carried out either analytically or numerically, the inference on parameters Θ can now utilize directly either the Bayesian methods outlined in Section 3.1, or the maximum likelihood methods in Section 3.2. Alternatively, the estimation could be based on the joint posterior of all unobserved quantities $P(\Theta, Z_{C\setminus O\cup M} \mid R_C, X_C, Y_C, Z_{O\setminus M})$, and a Gibbs sampler from the corresponding full conditional distributions

$$P(Z_i \in \mathrm{d}z_i \mid R_C, X_C, Y_C, Z_{C\setminus\{i\}}, \theta) \overset{z_i}{\propto} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta), \tag{13}$$

$i \in C \setminus O \cup M$ and

$$P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_C) \overset{\theta}{\propto} \left[\prod_{i \in C} P(Y_i, Z_i \mid X_i, \theta) \right] P(\Theta \in \mathrm{d}\theta) \tag{14}$$

(see Robert and Casella, 2004, p. 357-358). This approach is known as Bayesian data augmentation (see Kulathinal and Arjas, 2006, for an application). It is equivalent to (proper) multiple imputation from the posterior distribution of missing data when the imputation and analysis models are the same (Rubin, 1987, p. 113). The advantage of using data augmentation is that the (complete data) likelihood expression in (14) is in a standard form which is often computationally more convenient than the observed data likelihood in (12); for example, conjugate distributions can then be utilized in the simulation from the posterior (14).

Maximum likelihood counterpart of Bayesian data augmentation is known as the Expectation-Maximization (EM) algorithm (Dempster et al., 1977, Robert and Casella, 2004, p. 176-177, applied here in Article I). Here the expectation step involves taking the expectation of the complete data log-likelihood with respect to the posterior distribution of the missing data at the current parameter estimates. In the maximization step this expectation is maximized with respect to the parameters. These steps are iterated until the parameter estimates converge (to the maximum likelihood estimate). The potential advantage of using the EM instead of direct maximization of the observed data likelihood again lies in the computationally more convenient

form of the (expected) complete data log-likelihood, which may be easier to maximize, especially if ready-made numerical tools or closed form maximum likelihood estimators exist for the standard form likelihood expression.

In Articles II and III the simulation studies carried out for comparing the efficiencies of different likelihood and pseudolikelihood (see section 4.5) expressions utilized real data (for instance, time-to-event responses observed for a cohort), with additional covariate data with specified distributions and effect sizes simulated using the above described two step Gibbs sampling procedure. First the additional covariates were drawn from posterior distributions of the type (13) and then the unknown (non-fixed) parameters in the simulation model were drawn from conditional distributions of the type (14). A number of datasets combining the real observations and the simulated covariates were produced by iterating these two steps. This procedure corresponds to drawing multiple imputations from a Bayesian (proper) imputation model. The aim in this approach was to make the simulation settings more realistic compared to using completely synthetic data.

4.4 Conditional likelihood

Although the full likelihood expression obtained in the previous section is optimal in the sense that it utilizes all observed data, it is not entirely without problems. For instance, if the study base is large and/or the proportion of missing data high, the integration involved may be computationally demanding. In addition, with high proportion of missing data, it is not certain that the parameters in the model have the anticipated interpretation, with the missing data also acting as unknown parameters. In extreme situations this may lead to multimodal likelihoods. Furthermore, a full likelihood expression is not applicable in situations where the study base can not be enumerated (Wacholder, 1996, p. 146). In such cases, if the mechanism that has produced the observations is known, it makes sense to confine to the set of subjects with completely observed data and condition the likelihood expression on this mechanism. Here we attempt to derive and interpret conditional likelihood from the Bayesian perspective.

Conditional likelihood is not to be confused with marginal or partial likelihood. We refer to Cox (1975) for the definitions and difference between the concepts. Introduction of conditional inference is attributed to R. A. Fisher (1890-1962) by Efron (1978, p. 238). As an elementary example of this approach, it makes sense to condition the analysis on realized sample size, even if the size had originally been randomly selected. Fisher's (1956, see also Berger and Wolpert, 1988, Chapter 2) central concept here was that of relevant subset, which can be presented as a transformation of a random

vector Y into a new random vector partitioned into relevant subsets (V, W), the transformation not depending on the unknown parameter Θ. The conditional likelihood for W given the realization $V = v$ is then defined as (Cox and Hinkley, 1974, p. 16-17; Cox, 1975, p. 269)

$$P_{W|V,\Theta}(W \in \mathrm{d}w \mid V = v, \Theta = \theta).$$

It should first be noted that, although we work under this same general definition, we use conditional likelihood solely in the context of conditioning on a sampling mechanism (in the sense of e.g. Langholz and Goldstein, 2001), in contrast to the traditional maximum likelihood theory, where conditioning on an ancillary statistic (concept introduced by Fisher, 1956) is carried out to eliminate nuisance parameters (e.g. Andersen, 1970; Kalbfleisch and Sprott, 1970; Cox and Hinkley, 1974, p. 292-293). The partition into relevant subsets is specified here by the realized sample. The conditioning carried out herein does not eliminate nuisance parameters; if anything, it may increase the number of parameters needed to be estimated. Also, we will not confine ourselves to situations where the conditioning subset V (the inclusion indicators and any other information which not included in the subset W) would be an ancillary statistic. We saw in Section 4.2 (first phase sampling) that the pair of random vectors $V = (I_F, X_F)$ together indeed are an ancillary statistic, that is, their joint distribution does not depend on the parameter of interest Θ. However, we shall see that in the second phase sampling the natural conditioning subset is $V = (R_C, X_C, Y_{C\setminus O})$, which obviously depends on Θ. Such a conditioning will lose information on Θ compared to inference based on full likelihood, but will nevertheless lead to an intuitively plausible likelihood expression.

In the well known retrospective likelihood for case-control designs (see, for example, Seaman and Richardson, 2001, and Cox, 2006, p. 154-157) the conditioning corresponds to the order of the data collection. The data collection under cohort sampling designs is also retrospective in the sense that the covariates Z_O are collected after observing all of Y_C. In full likelihood inference the order of data collection is not relevant, since all observed data informative of the parameters of interest are included in the likelihood. In conditional likelihood inference the split into relevant subsets determines how much information is lost in the conditioning. In the following, to minimize the information loss, we shall proceed to write the likelihood jointly for $W = (Y_O, Z_O)$, rather than conditioning on either one. Hence, the order in which these observations were collected becomes again irrelevant to the problem.

No further justification for using the conditional approach is needed in situations where we cannot parameterize the joint distribution of (W, V) in

terms of the parameters of interest, but can do so for the conditional distribution $P_{W|V,\Theta}$ (cf. conditional exchangeability and full exchangeability, see Section 4.7). An example of a situation where an enumerable frame population is not available but the sampling mechanism known is late entry/left truncation in follow-up studies. (Target population would be the underlying unobserved birth cohort.) Because only the individuals alive at the time of a cross-sectional cohort recruitment are observed, the likelihood expression for time-to-event data will then have to be conditioned on survival until the study outset (Guo, 1993, p. 229). If the event of interest is non-fatal, individuals with a prevalent condition similar to the outcome of interest at the start of the follow-up are usually excluded from the analysis, corresponding to conditioning on healthy status at the study outset (Commenges et al., 1998, p. 1976). However, in Article III we proposed a conditional likelihood expression which utilizes the information from the individuals with a prevalent disease.

In the present context of a two-phase study design, the sampling frame is known, but for the previously mentioned reasons it may be desirable to limit the analysis to the group selected in the second phase, corresponding to conditioning the likelihood expression on the second phase sampling design. Applying Bayes' formula for the posterior distribution given all observed quantities and marginalizing over the nuisance parameters, we get

$$
\begin{aligned}
&P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_O) \\
&= \int_\phi P(\Theta \in \mathrm{d}\theta, \Phi \in \mathrm{d}\phi \mid R_C, X_C, Y_C, Z_O) \\
&\overset{\theta}{\propto} \int_\phi P(R_C, Y_C, Z_O \mid X_C, \theta, \phi) P(X_C \mid \phi) P(\Phi \in \mathrm{d}\phi) P(\Theta \in \mathrm{d}\theta) \\
&\overset{(R_C, Y_C, Z_O) \perp\!\!\!\perp \Phi \mid X_C}{=} P(R_C, Y_O, Y_{C\setminus O}, Z_O \mid X_C, \theta) P(X_C) P(\Theta \in \mathrm{d}\theta) \\
&\overset{\theta}{\propto} P(Y_O, Z_O \mid R_C, X_C, Y_{C\setminus O}, \theta) \\
&\quad \times P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta) P(\Theta \in \mathrm{d}\theta),
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
&\frac{P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_O)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta)} \\
&\overset{\theta}{\propto} P(Y_O, Z_O \mid R_C, X_C, Y_{C\setminus O}, \theta) P(\Theta \in \mathrm{d}\theta).
\end{aligned}
$$

Here the first term on the right hand side can be further developed as

$$
\begin{aligned}
&P(Y_O, Z_O \mid R_C, X_C, Y_{C\setminus O}, \theta) \\
&= \frac{P(Z_O \mid R_C, X_C, Y_C, \theta) P(R_C, Y_C \mid X_C, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta)} \\
&\overset{R_C \perp\!\!\!\perp \Theta \mid X_C, Y_C}{=} \frac{P(Z_O \mid R_C, X_C, Y_C, \theta) P(R_C \mid X_C, Y_C) P(Y_C \mid X_C, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta)} \\
&\overset{\theta}{\propto} \frac{P(Z_O \mid R_C, X_C, Y_C, \theta) P(Y_C \mid X_C, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta)} \\
&= \frac{\int_{z_{C\setminus O}} P(Z_C \in \mathrm{d}z_C \mid R_C, X_C, Y_C, \theta) P(Y_C \mid X_C, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) \int_{y_O} \int_{z_C} P(Y_C \in \mathrm{d}y_C, Z_C \in \mathrm{d}z_C \mid X_C, \theta)} \\
&\overset{R_C \perp\!\!\!\perp Z_C \mid X_C, Y_C}{=} \frac{\int_{z_{C\setminus O}} P(Y_C, Z_C \in \mathrm{d}z_C \mid X_C, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) \int_{y_O} \int_{z_C} P(Y_C \in \mathrm{d}y_C, Z_C \in \mathrm{d}z_C \mid X_C, \theta)} \\
&= \frac{\int_{z_{C\setminus O}} \prod_{i \in C} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) \int_{y_O} \int_{z_C} \prod_{i \in C} P(Y_i \in \mathrm{d}y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta)} \\
&= \frac{\prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta) \prod_{i \in C\setminus O} P(Y_i \mid X_i, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) \prod_{i \in C\setminus O} P(Y_i \mid X_i, \theta)} \\
&= \frac{\prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta)},
\end{aligned}
$$

where the products followed from the same exchangeability assumptions that were made in the previous section. We now have

$$
\frac{P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_O)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta)} \overset{\theta}{\propto} \frac{\prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta)} P(\Theta \in \mathrm{d}\theta), \tag{15}
$$

or because

$$
\begin{aligned}
&P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_{C\setminus O}) \\
&= \int_\phi P(\Theta \in \mathrm{d}\theta, \Phi \in \mathrm{d}\phi \mid R_C, X_C, Y_{C\setminus O}) \\
&\overset{\theta}{\propto} \int_\phi P(R_C, Y_{C\setminus O} \mid X_C, \theta, \phi) P(X_C \mid \phi) P(\Phi \in \mathrm{d}\phi) P(\Theta \in \mathrm{d}\theta) \\
&\overset{(R_C, Y_{C\setminus O}) \perp\!\!\!\perp \Phi \mid X_C}{=} P(R_C, Y_{C\setminus O} \mid X_C, \theta) P(X_C) P(\Theta \in \mathrm{d}\theta) \\
&\overset{\theta}{\propto} P(R_C \mid X_C, Y_{C\setminus O}, \theta) P(Y_{C\setminus O} \mid X_C, \theta) P(\Theta \in \mathrm{d}\theta),
\end{aligned}
$$

equivalently,

$$\frac{P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_C, Z_O)}{P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_{C\setminus O})} \overset{\theta}{\propto} \frac{\prod_{i\in O} P(Y_i, Z_i \mid X_i, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta)}. \tag{16}$$

Here the right hand side is the conditional likelihood. Its numerator is just the likelihood expression restricted to set O, and thereby involving no missing data. However, due to the second phase selection (which may depend on all of Y_C), the set O in itself is not necessarily representative of C (or F), and thus the numerator alone would not result in valid inferences on the parameters Θ. This is represented in the likelihood expression by its denominator, which we call a correction term (or ascertainment correction, see Clayton, 2003). It should first be noted that this term will indeed depend on the parameters since it is not conditioned on all of Y_C. Secondly, whether the correction term simplifies into a product form over the individual contributions depends on the sampling mechanism used; in the general case it will not simplify. Below we shall present two examples of sampling designs and the resulting conditional likelihood correction terms. Generally the problem in conditional likelihood correction terms is to represent them as functions of parameters estimable from the data. This may involve introducing new nuisance parameters (cf. Article III, p. 580-581).

Bayesian inference on parameters Θ is possible if the right hand side of (15) is integrable over θ. In this case the left hand side could be called a "conditional" posterior distribution (Ma et al., 2007, p. 603-604). It is obvious that this is not the same posterior distribution as (11) because the observed information on the set $C \setminus O$ is not utilized here. In fact, the form (16) shows that using (15) is equivalent to ignoring the observed information in $P(\Theta \in \mathrm{d}\theta \mid R_C, X_C, Y_{C\setminus O})$, and replacing it with just the prior $P(\Theta \in \mathrm{d}\theta)$. Maximum likelihood inference using a conditional likelihood would proceed as outlined in Section 3.2. If the data collection was unsuccessful for subjects M in the group O selected in the second phase, the likelihood expression can be written analogously to (12) as

$$\frac{\prod_{i\in O\setminus M} P(Y_i, Z_i \mid X_i, \theta) \prod_{i\in M} \int_{z_i} P(Y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta)}{P(R_C \mid X_C, Y_{C\setminus O}, \theta)}.$$

As an example, consider a frequency matching design (Langholz and Goldstein, 2001, p. 68), where all cases (individuals with an event of interest during follow-up, denoted as $Y_i = 1$) in the cohort C are selected in the second phase and m controls are selected per each case from the set of non-cases ($Y_i = 0$) in C. Let $|C| = n$ and the number of cases in the cohort

$\sum_{i \in C} Y_i = d$. Now the second phase sample size is $|O| = d + md$. Such a sampling design is characterized by the joint probability distribution

$$P(R_C \mid X_C, Y_C) = P(R_C \mid Y_C) = 1 / \binom{n-d}{md}.$$

The conditional likelihood correction term now becomes

$$\begin{aligned}
&P(R_C \mid X_C, Y_{C \setminus O}, \theta) \\
&= \sum_{y_O} \int_{z_O} P(R_C, Y_O = y_O, Z_O \in \mathrm{d}z_O \mid X_C, Y_{C \setminus O}, \theta) \\
&= \sum_{y_O} \int_{z_O} P(R_C \mid X_C, Y_C, Z_O, \theta) P(Y_O = y_O, Z_O \in \mathrm{d}z_O \mid X_C, Y_{C \setminus O}, \theta) \\
&\overset{R_C \perp\!\!\!\perp (X_C, Z_O, \Theta) \mid Y_C}{=} \sum_{y_O : \sum_{i \in O} y_i = d} \frac{1}{\binom{n-d}{md}} \int_{z_O} P(Y_O = y_O, Z_O \in \mathrm{d}z_O \mid X_C, Y_{C \setminus O}, \theta).
\end{aligned}$$

Here

$$\begin{aligned}
P(Y_O, Z_O \mid X_C, Y_{C \setminus O}, \theta) &= \frac{P(Y_C, Z_O \mid X_C, \theta)}{P(Y_{C \setminus O} \mid X_C, \theta)} \\
&= \frac{\int_{z_{C \setminus O}} \prod_{i \in C} P(Y_i = y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta)}{\sum_{y_O} \int_{z_C} \prod_{i \in C} P(Y_i = y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta)} \\
&= \frac{\prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta) \prod_{i \in C \setminus O} P(Y_i \mid X_i, \theta)}{\prod_{i \in C \setminus O} P(Y_i \mid X_i, \theta)} \\
&= \prod_{i \in O} P(Y_i, Z_i \mid X_i, \theta), \qquad (17)
\end{aligned}$$

and we get

$$P(R_C \mid X_C, Y_{C \setminus O}, \theta) = \sum_{y_O : \sum_{i \in O} y_i = d} \frac{1}{\binom{n-d}{md}} \prod_{i \in O} \int_{z_i} P(Y_i = y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta).$$

It should be noted here that the correction term depends on the parameters through the same parametric model $P(Y_i, Z_i \mid X_i, \theta)$ as the numerator of the conditional likelihood, and thus the obtained likelihood expression is indeed applicable for parameter estimation. However, the correction term in this case does not simplify into a product form (its enumeration requires going through all y_O vectors involving d cases). This reflects the interdependence of the inclusion indicators R_C in the frequency matching design where the

sample size is fixed. The baseline rate of the events is not estimable from this likelihood expression, since under this design the conditioning on R_C fixes the number of events in the cohort (Langholz and Goldstein, 2001, p. 68). This is also the case for the standard retrospective likelihood for case-control designs, which conditions directly on the event indicators (see Seaman and Richardson, 2001, p. 1076).

Bernoulli sampling is characterized by the joint probability distribution

$$P(R_C \mid X_C, Y_C) = \prod_{i \in C} P(R_i \mid X_i, Y_i) = \prod_{i \in C} \pi(X_i, Y_i)^{R_i} (1 - \pi(X_i, Y_i))^{1-R_i}.$$

Here the inclusion indicators R_i are independent while the realized sample size is random (see Särndal et al., 1992, p. 62-63). In contrast to matched designs, some case-cohort designs can be represented as Bernoulli sampling. Let us consider a simplified situation where Y_i is the case-status for subject i at the end of the follow-up period and X_i is the subject's age at the start of the follow-up. If all cases are selected, $P(R_i \mid X_i, Y_i = 1) = 1$. A subcohort of a fixed expected size is sampled independently, without regard to the case-status, with probabilities which may depend on subject's age; the selection probability for non-cases is then $P(R_i = 1 | X_i, Y_i = 0) = \pi(X_i)$. Now the expected sample size is $|O| = d + \sum_{i \in C} \mathbf{1}_{\{Y_i = 0\}} \pi(X_i)$, while the realized sample size is random. The conditional likelihood correction term under such a case-cohort design becomes

$$\begin{aligned}
&P(R_C \mid X_C, Y_{C \setminus O}, \theta) \\
&= \sum_{y_O} \int_{z_O} P(R_C, Y_O = y_O, Z_O \in \mathrm{d}z_O \mid X_C, Y_{C \setminus O}, \theta) \\
&\overset{R_C \perp\!\!\!\perp (Z_O, \Theta) | X_C, Y_C}{=} \sum_{y_O} \int_{z_O} P(R_C \mid X_C, Y_C) P(Y_O = y_O, Z_O \in \mathrm{d}z_O \mid X_C, Y_{C \setminus O}, \theta) \\
&\overset{(17)}{=} \prod_{i \in C \setminus O} P(R_i = 0 \mid X_i, Y_i = 0) \\
&\quad \times \prod_{i \in O} \int_{z_i} \sum_{y_i \in \{0,1\}} P(R_i = 1 \mid X_i, Y_i = y_i) P(Y_i = y_i, Z_i \in \mathrm{d}z_i \mid X_i, \theta) \\
&\overset{\theta}{\propto} \prod_{i \in O} \int_{z_i} \left[P(Y_i = 1, Z_i \in \mathrm{d}z_i, \mid X_i, \theta) + \pi(X_i) P(Y_i = 0, Z_i \in \mathrm{d}z_i, \mid X_i, \theta) \right].
\end{aligned}$$

Thus in this case the conditional likelihood expression assumes the product form

$$\prod_{i \in O} \frac{P(Y_i, Z_i \mid X_i, \theta)}{\int_{z_i} \left[P(Y_i = 1, Z_i \in \mathrm{d}z_i, \mid X_i, \theta) + \pi(X_i) P(Y_i = 0, Z_i \in \mathrm{d}z_i, \mid X_i, \theta) \right]}, \quad (18)$$

which corresponds to the result obtained in Article I, p. 10-11.

4.5 Pseudolikelihood

As a comparison to the previously discussed probability-based approach to cohort sampling designs, we review here briefly an alternative approximative method discussed by, for example, Kalbfleisch and Lawless (1988), Samuelsen (1997) and Samuelsen et al. (2007). Consider now conditional probabilities $P(R_i = 1 \mid X_C, Y_C) \equiv \pi_i$, $i \in C$, known as the first-order inclusion probabilities (Särndal et al., 1992, p. 31). It should be noted that, with the exception of Bernoulli sampling, the product of such terms does not specify the sampling mechanism, since in sampling without replacement (with a fixed sample size) from a finite population, the indicators R_i are not independent (see Särndal et al., 1992, p. 30-33). Kalbfleisch and Lawless (1988, p. 153) used the first-order inclusion probabilities for constructing an approximation for the true complete data log-likelihood (in the full cohort) as

$$\sum_{i \in O} \frac{1}{\pi_i} \log P(Y_i, Z_i \mid X_i, \theta) \approx \sum_{i \in C} \log P(Y_i, Z_i \mid X_i, \theta). \tag{19}$$

They refer to such an expression as a pseudolikelihood. Cox (2006, p. 152) uses the same term for any likelihood expression that ignores some dependencies between the variables. As an "estimator" for the true log-likelihood, (19) corresponds to the Horvitz-Thompson estimator for a population total (Horvitz and Thompson, 1952; Särndal et al., 1992, p. 42-43). Since in many sampling mechanisms the first order inclusion probabilities are specified only implicitly, the application of (19) may require estimation of the quantities π_i. Samuelsen (1997, p. 382) presented an estimator for the first order inclusion probabilities in nested case-control sampling. Samuelsen et al. (2007) suggested poststratification, with the cases defining their own stratum, and the inclusion probabilities for non-cases estimated using the realized sampling fractions within strata. With a time-to-event response $Y_i = (T_i, E_i)$, where $E_i = 1$ indicates a case and $E_i = 0$ a non-case, if the only parameters of interest are regression coefficients β in the semiparametric proportional hazards model (Cox, 1972), the pseudolikelihood simplifies into (Kalbfleisch and Lawless, 1988, p. 156; Samuelsen, 1997, p. 383)

$$\sum_{i \in O} E_i \left[\beta'(X_i, Z_i) - \log \sum_{j \in \mathcal{R}_i} \frac{1}{\pi_j} \exp\{\beta'(X_j, Z_j)\} \right],$$

where $\mathcal{R}_i \subseteq O$ denotes the risk set at event time T_i. If not specified by the sampling design, the inclusion probabilities π_j are replaced with estimates $\hat{\pi}_j$. In designs where all cases are selected, $\pi_j = 1$ if $E_i = 1$, and

the cases contribute to the risk sets with unit weights, while the non-cases assume weights greater or equal to one. Parameter estimation would proceed by maximizing the pseudolikelihood expression with respect to β, and possibly using resampling-based variance estimators (see e.g. Barlow, 1994) for the resulting estimator $\hat{\beta}$, if Gaussian approximation is not thought to be appropriate.

The connection between the approximative likelihood here and the conditional likelihood (18) derived for a case-cohort design are the first-order inclusion probabilities. In Bernoulli sampling the actual inclusion probabilities are equivalent to the first-order inclusion probabilities (Särndal et al., 1992, p. 32). This suggests that if the inclusion indicators in the sampling design used are approximately independent, the case-cohort conditional likelihood might give a reasonable approximation, with the first order inclusion probabilities for cases and non-cases substituted in (18).

4.6 Model uncertainty and selection

In addition to missing data, another problem for which the Bayesian approach provides a seemingly natural solution is model uncertainty, which we understand broadly to mean the uncertainty on the adequacy of the chosen parametric probability distributions in prediction and explanation tasks. However, as noted in Section 2.3, the predictive ability of a model is much easier to evaluate than its ability to explain the unobserved parts of reality we are interested in. While the Bayes' formula (2) is commonly applied in the setting of one fixed parametric probability model, it may as well be applied in the context of a (countable) set of alternative models (possibly corresponding to competing hypotheses or theories) $M = \{h_1, h_2, \ldots\}$, as (Sisson, 2005, p. 1077)

$$P_{H|Y}(H = h_k \mid Y = y) = \frac{P_{Y|H}(Y \in \mathrm{d}y \mid H = h_k) P_H(H = h_k)}{\sum_{l=1}^{|M|} P_{Y|H}(Y \in \mathrm{d}y \mid H = h_l) P_H(H = h_l)}, \quad (20)$$

where $H : \Omega \to M$ is a discrete random variable; the associated probability statements are to be understood as the uncertainty about the "true" model (the concept of which is further discussed below). Here the left hand side is known as the posterior model probability. Because the denominator is a constant, if the alternative models are given equal prior probabilities $P_H(H = h_k) = 1/|M|$, the model with the highest posterior probability is given by $\arg\max_k P_{Y|H}(Y \in \mathrm{d}y \mid H = h_k)$, which is the model selection criterion already discussed in Section 2.3. The probability here was referred to as the marginal probability of the data because it is given by integration

over the parametrization of the model as

$$P_{Y|H}(Y \in \mathrm{d}y \mid H = h_k) = \int_{\theta_k \in \mathbb{R}^{p_k}} P_{Y|\Theta_k}(Y \mid \Theta_k = \theta_k) P_{\Theta_k}(\Theta_k \in \mathrm{d}\theta_k), \quad (21)$$

where $\Theta_k : \Omega \to \mathbb{R}^{p_k}$ is a parameter vector specific to the model h_k. A rough approximation for this quantity is the Bayesian Information Criterion (BIC, Schwarz, 1978; see also Raftery, 1995, p. 130-133). Here it is important to note that the parameter vectors associated with the alternative models may be of a different dimension. It is not necessary that the models are nested, though in practical problems this is often the case. A typical example is variable selection where the space of alternative models may be represented as $M = \{0, 1\}^p$, where p is the total number of possible covariates and each binary indicator corresponds to exclusion/inclusion of the covariate in the model. As the total number of alternative models is then $|M| = 2^p$, the model space easily gets too large (say, with $p = 30$, over one billion) for the evaluation of every model to be feasible in practice, especially as numerical evaluation of integrals of the type (21) is not straightforward. This is because direct application of Monte Carlo integration by simulating from the prior distribution P_{Θ_k} is very inefficient if the prior involves little information (Kass and Raftery, 1995, p. 779). Although alternative methods for evaluating such integrals have been proposed (see Kass and Raftery, 1995, for a review), application of Bayesian model selection became fully feasible in practice only after invention of Markov chain Monte Carlo techniques which allow simultaneous exploration of both parameter and model spaces (Sisson, 2005). Here the extension by Green (1995) to the Metropolis-Hastings algorithm (see Section 3.1) was central. Reviews of the methods in the context of variable selection are provided by Dellaportas et al. (2002) and O'Hara and Sillanpää (2009).

Denoting the concatenated random vector of all parameters as $\Theta = (\Theta_1, \ldots, \Theta_{|M|}) : \Omega \to \mathbb{R}^{p_1 + \ldots + p_{|M|}}$, the aim in simultaneous inference on the model and parameter spaces is to produce random samples from the posterior distribution $P_{\Theta, H|Y}(\Theta \in \mathrm{d}\theta, H = h_k \mid Y = y)$. With such a sample available, if the posterior probabilities for specific models are of interest, they can be easily obtained by the integration

$$P_{H|Y}(H = h_k \mid Y = y) = \int_{\theta \in \mathbb{R}^{p_1 + \ldots + p_{|M|}}} P_{\Theta, H|Y}(\Theta \in \mathrm{d}\theta, H = h_k \mid Y = y).$$

If some parameters, say Φ, are common to all models, posterior inference on these parameters may be carried out by integrating out the model specific

parameters and the model indices themselves as

$$P_{\Phi|Y}(\Phi \in \phi \mid Y = y) \tag{22}$$
$$= \int_{\theta \in \mathbb{R}^{p_1+\ldots+p_{|M|}}} \sum_{l=1}^{|M|} P_{\Phi,\Theta,H|Y}(\Phi \in \mathrm{d}\phi, \Theta \in \mathrm{d}\theta, H = h_l \mid Y = y).$$

This approach is known as model averaged inference (see e.g. Burnham and Anderson, 2004). Yet another, and possibly the most important, generalization is to consider the posterior predictive distributions

$$P_{Y_{n+1}|Y}(Y_{n+1} \in \mathrm{d}y_{n+1} \mid Y = y) \tag{23}$$
$$= \int_{\theta \in \mathbb{R}^{p_1+\ldots+p_{|M|}}} \sum_{l=1}^{|M|} P_{Y_{n+1},\Theta,H|Y}(Y_{n+1} \in \mathrm{d}y_{n+1}, \Theta \in \mathrm{d}\theta, H = h_l \mid Y = y),$$

integrated over both the parameter and model spaces (Sisson, 2005, p. 1077). We can still further generalize the present framework by noting that the model indicator H need not be a discrete random variable; instead the distribution P_H in (20) can be defined as a general mixing measure, in the sense of Teicher (1960), for example. Now the distribution P_H of the random variable $H : \Omega \to M$ is a probability measure on $(M, \mathcal{M})$ and we have

$$P_{H|Y}(H \in \mathrm{d}h \mid Y = y) = \frac{P_{Y|H}(Y \in \mathrm{d}y \mid H = h) P_H(H \in \mathrm{d}h)}{\int_{h \in M} P_{Y|H}(Y \in \mathrm{d}y \mid H = h) P_H(H \in \mathrm{d}h)},$$

where the notation $\mathrm{d}h$ is to be interpreted as $\mathrm{d}h \in \mathcal{M}$. It should be noted that in this general case the space of possible models need not be countable. One example is the Diriclet process mixture model (see Neal, 2000, for a review), while in the model devised in Article IV and applied in Article V, the model space was defined by Poisson point processes (see Section 2.4). Although in these cases the space of possible models is not countable, the realizations of H, specifying the dimension of the parameter space of each model, are finite, so (21) still applies, and a posterior predictive distribution $P_{Y_{n+1}|Y}(Y_{n+1} \in \mathrm{d}y_{n+1} \mid Y = y)$ may be obtained by integrating over all realizations of models and the related parameters encountered during a finite MCMC run. However, in the general case the notion of posterior model probability (20), attributed to a specific model realization, loses its meaning, although posterior probabilities may still be attributed to some suitable set of such realizations.

It is now in order to consider more closely the interpretation of the concept of posterior model probability. Gelman and Rubin (1995) in their commentary to Raftery (1995) questioned the need to do model selection in the

first place, criticizing especially assigning posterior probabilities to a fixed set of candidate models. Although the selection criterion based on the maximal marginal probability of the data will consistently select the true model if such a model is included in the set of candidate models, as was already noted in Section 2.2, we rarely truly believe that any of the candidate models actually is the true data generating mechanism. Instead, the truth may be interpreted as equivalent to complete information on the reality itself, a situation which is mostly unattainable in practice. In any case, had we complete information available, there might be no need to resort to probability-based inference in the first place. If the true model is not included in the model space M, but posterior model probability is nevertheless used to select the best model, the resulting selection may or may not be meaningful, depending on the intended application of the model. The resulting selection may be characterized as the model with the best prequential predictive ability (Raftery, 1995, p. 777) or, asymptotically with increasing sample size, approaching from below (in terms of model complexity) the most parsimonious of the models with minimal Kullback-Leibler information loss with respect to truth or full reality (Burnham and Anderson, 2004, p. 275-280).

To sum up, posterior probability-based inferences concerning the "true" model require a priori assumption that such a model is in the model space. Such an assumption may be more reasonable in situations corresponding to the classical hypothesis testing, with two competing hypotheses, formulated so that one or the other is supposed to be correct (see Raftery, 1995, p. 776). In the general case with multiple alternative models which all are at best approximations of the truth, one should be careful and not attach the posterior model probabilities with meanings they do not in reality possess. This, however, does not change the fact that model uncertainty is a real problem. In any case, selection of a single best model may tend to hide this problem. Especially in non-Bayesian statistics, the selection of a single, in some sense optimal, model is often followed by inferences entirely in the setting of that model, as if any other candidate models had never existed (Draper, 1995, p. 45-46). For these reasons, instead of model selection, we prefer to utilize the methods reviewed in the present section in situations where we want to relax possibly too rigid parametric assumptions, by specifying the model through random functions, instead of confining ourselves to a single parametric distribution. The inference then utilizes model averaged posterior distributions of the form (22) or posterior predictive distributions (23), integrated over the model space. In such distributions the uncertainty involved in the model specification is appropriately quantified.

4.7 Causality and potential outcomes

Previously we have discussed inference based on posterior distributions for parameters and posterior predictive distributions, and also explanatory and predictive modeling. It should be emphasized that the use of posterior predictive distributions is not limited only to prediction task, but that they are also important tools in explanatory analysis. As an example we discuss here utilization of posterior predictive distributions in causal inference. Although the role of the randomized experimental study design is often emphasized in the discussion of statistical inference on causal effects (see, for example, Rubin, 1978, and Holland, 1986), observational studies are also carried out in hope to learn something about causal relationships; such studies would not be funded to find and report mere statistical associations without the belief that there may be something real at work behind them. Moreover, in many situations an experimental study is not a feasible option due to ethical or practical reasons. Therefore, the question of causal inference is by no means limited to randomized designs. It is also too wide of an area to deal with herein; we refer to Pearl (2009) and concentrate here on one simplified example, broadly following the principles outlined by Arjas and Parner (2004).

Consider an experimental (but not necessarily randomized) study involving the individuals $C = \{1, \ldots, n\}$. Let $X_C = \{X_i : i \in C\}$ be a collection of pre-treatment covariates measured on the n individuals, and $A_C = \{A_i : i \in C\}$, $A_i \in \{0, 1\}$, a vector of indicators telling whether each individual received placebo or active treatment/intervention. Further, let $Y_C = \{Y_i : i \in C\}$ be response variables, measured after a sufficient time interval after application of the treatments. The joint probability distribution of the observed data at the end of the study period can be factored representing the order of the data collection as

$$P(A_C, X_C, Y_C) = P(Y_C \mid A_C, X_C)P(A_C \mid X_C)P(X_C).$$

Because the treatments, in addition to taking into account the individual covariate values X_i, may be administered according to some pre-defined quota, it does not seem reasonable to assume full exchangeability over the unit indices of the triples (A_i, X_i, Y_i). However, it may be justified to assume exchangeability (over unit indices) of the responses Y_i, given the covariates X_i and the treatment A_i. This postulate is known as conditional exchangeability (see Lindley and Novick, 1981, p. 47). It should be noted that the weaker condition of conditional exchangeability follows from full exchangeability but not vice versa. If we further assume that treatments given to other individuals do not affect the outcome of a given individual ("no interference between

treatments"; this assumption would not be realistic in vaccination trials, for instance), we get

$$P(Y_C \mid A_C, X_C) = \int_\theta \left[\prod_{i=1}^{n} P(Y_i \mid A_C, X_C, \theta)\right] P(\Theta \in \mathrm{d}\theta)$$
$$\overset{Y_i \perp\!\!\!\perp (A_{C\setminus\{i\}}, X_{C\setminus\{i\}}) \mid (A_i, X_i, \Theta)}{=} \int_\theta \left[\prod_{i=1}^{n} P(Y_i \mid A_i, X_i, \theta)\right] P(\Theta \in \mathrm{d}\theta).$$

In addition, it is assumed that that the treatment assignment is carried out according to a known rule which may depend only on the (then observed) covariates X_C, and $P(A_C \mid X_C)$ is therefore uninformative on the parameters Θ. This situation is fully analogous to the unconfounded and ignorable sampling mechanisms discussed in Section 4.3. Inferences on the parameters Θ may now be based on the posterior

$$P(\Theta \in \mathrm{d}\theta \mid A_C, X_C, Y_C) \overset{\theta}{\propto} \left[\prod_{i=1}^{n} P(Y_i \mid A_i, X_i, \theta)\right] P(\Theta \in \mathrm{d}\theta), \tag{24}$$

and causal inferences will be based on the comparison between two posterior predictive distributions concerning a generic individual indexed as $n+1$, namely the predictive distribution of the response given no treatment

$$\begin{aligned} &P(Y_{n+1} \mid A_{n+1} = 0, X_{n+1}, A_C, X_C, Y_C) \\ &= \int_\theta P(Y_{n+1} \mid A_{n+1} = 0, X_{n+1}, \theta) P(\Theta \in \mathrm{d}\theta \mid A_C, X_C, Y_C) \end{aligned}$$

and the predictive distribution of the response given the treatment

$$\begin{aligned} &P(Y_{n+1} \mid A_{n+1} = 1, X_{n+1}, A_C, X_C, Y_C) \\ &= \int_\theta P(Y_{n+1} \mid A_{n+1} = 1, X_{n+1}, \theta) P(\Theta \in \mathrm{d}\theta \mid A_C, X_C, Y_C). \end{aligned}$$

For instance, we might be concerned with the expected treatment effect

$$\begin{aligned} &E(Y_{n+1} \mid A_{n+1} = 1, X_{n+1}, A_C, X_C, Y_C) \\ &- E(Y_{n+1} \mid A_{n+1} = 0, X_{n+1}, A_C, X_C, Y_C), \end{aligned}$$

which is easily obtained if an MCMC sample is available from both predictive distributions. It is notable that the predictive approach to causal inference entirely avoids the need to introduce the somewhat controversial counterfactual notation (see Dawid, 2000), and the related random variables. Indeed,

the unobservability of counterfactuals is named as the fundamental problem of causal inference by Holland (1986, p. 947), implying that causal inference is impossible. The statistical solution noted by Holland is to consider averaged population level causal effects instead of individual level causal effects. This can be naturally formulated in terms of exchangeability; the very spirit of the exchangeability postulate is that we can make probability statements concerning a further generic unit, while the tool to make such statements is the posterior predictive distribution. The potential outcomes considered herein simply correspond to two different conditional probability distributions, which are obtained using the same exchangeability and conditional independence assumptions that were applied in deriving the likelihood expression for the observed data.

The above situation was straightforward because the treatment assignment could be chosen at will, in which case it should be easy to make sure that no confounding occurs, as long as the relevant covariates are appropriately taken into account in the analysis. The same probabilistic framework can be applied also in purely observational studies, but then further assumptions are needed as the treatment assignment will be outside the control of the researcher, with a possibly unknown mechanism. To illustrate this, we introduce a new set of variables $U_C = \{U_i : i \in C\}$, representing all the factors which are unobserved in the study but may have potentially contributed to the treatment assignment. The joint distribution may now be factored as

$$P(U_C, A_C, X_C, Y_C) = P(Y_C \mid A_C, X_C, U_C)P(A_C \mid X_C, U_C)P(X_C \mid U_C)P(U_C).$$

It is now obvious that if these unobserved factors may have an effect on both the treatment assignments A_C and the responses Y_C, inferences based on the marginal posterior distribution (24) obtained above will not tell about the true causal effects of the treatment. The inferences will be valid only if either one of the further conditional independence assumptions $Y_C \perp\!\!\!\perp U_C \mid A_C, X_C$ and $A_C \perp\!\!\!\perp U_C \mid X_C$ holds true. The latter corresponds to "no unmeasured confounders" (see Robins, 1997), and means assuming that the treatment assignments have been based only on information which is available also in the observational study. The latter assumption seems more reasonable in practice than the former, since it is at least verifiable to a degree, if more information is obtained on the assignment mechanism. If we are confident that enough covariates have been obtained to assume no unmeasured confounders, the inferences can proceed utilizing the same posterior distribution (24) and the predictive distributions as in the experimental situation (see Arjas and Parner, 2004, p. 175-176). The probabilistic framework outlined here was applied in the context of optimal sequential treatment regimes in Article V.

5 Discussion

Even complex cohort sampling designs can be easily handled by using the full likelihood approach. It is needed that the sampling design satisfies a missing at random type assumption, which seems to apply to most of the commonly used cohort sampling designs, including quota sampling and countermatching. In addition, an exchangeability postulate extending for infinite sequences is needed; the sensibility of this depends on the parameters of interest. If it seems reasonable to generalize the results from observed data to a further generic unit, the postulate may be justified. In practice full likelihood inference involves an integration over the unobserved elements of the variables collected in the second phase sampling. Such integration is conveniently handled by using Bayesian data augmentation. Alternatives are the EM-algorithm or direct maximization of the numerically integrated observed data likelihood. However, the full likelihood approach is available only when an enumerable study base exists. This is commonly not the case in retrospective case-control studies.

Conditioning on the rule which produced the realized second phase sample will also result in a plausible likelihood expression. Here the likelihood is written over the fully observed subset instead of the full cohort. The conditioning will in many cases lose information compared to a corresponding full likelihood which utilizes all observed data. However, conditioning may be necessary when the study base is very large, making the required integration impractical, or when an enumerable study base is not available. Left truncation in cross-sectional cohort recruitment is an example of the latter situation. Conditional likelihood involves a correction term which will generally depend on the parameters of interest and on the sampling mechanism used.

We have seen that the inference problems in the five articles included in this work can be covered under a general probabilistic framework, with maximum likelihood estimation interpreted as a special case or approximation. However, the fully probabilistic Bayesian approach can handle many situations for which likelihood by itself would be inadequate. Examples covered here include utilizing hierarchical parametrization in more flexible model definition and posterior predictive distributions in causal inference.

Acknowledgements

This work began in the summer of 2004 when I started working as a statistician in the MORGAM project at the International CVD Epidemiology Unit of the National Public Health Institute (KTL), which later became part of the Chronic Disease Epidemiology and Prevention Unit of the National Institute for Health and Welfare (THL). Many of the research problems addressed in this thesis originated from this project which is an international pooling of cardiovascular cohorts. This is also where one of my supervisors, Sangita Kulathinal, was working at that time. In the same summer I also registered as a postgraduate student at the Department of Mathematics and Statistics of the University of Helsinki, with Elja Arjas as my other supervisor.

First and foremost, I wish to express my gratitude to my supervisors, Elja and Sangita, without whose knowledge, enthusiasm and persistence this work would never have seen the light of day. They are also the ones who exposed me to Bayesian statistics. Secondly, I would like to thank my boss Kari Kuulasmaa for providing a stable and encouraging working environment during the course of this work. Thanks also go to Juha Karvanen, with whom I shared an office room for many years, for countless scientific discussions, to Esa Läärä for the collaboration which produced the second publication herein, and to the KTL/THL statistics reading group for the theoretical debates which offered welcome relief from the day-to-day practical work at the institute. I would like to thank the two reviewers, Sven Ove Samuelsen and Aki Vehtari, for taking time from their busy schedule to read my work, and Jukka Corander for taking care of various practical issues in organizing the public defense. Finally, I am grateful to my father, who did not live to see this, and to my mother for their support over the years.

I acknowledge the financial support granted by the Emil Aaltonen foundation in 2009 for finalizing this work, though unfortunately in the end this could not be utilized due to some unexpected circumstances at THL at the time. The development of statistical methods for the MORGAM project was partly supported through the European Union Fifth Framework Programme GenomEUtwin Project (Contract No. QLG2 CT-2002-01254), and by the Medical Research Council London (G0601463, identification No. 80983: Biomarkers in the MORGAM Populations).

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32:283–301.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Arjas, E. (1989). Survival models and martingale dynamics. *Scandinavian Journal of Statistics*, 16:177–225.

Arjas, E. and Heikkinen, J. (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Computational Statistics*, 12:385–402.

Arjas, E. and Parner, J. (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, 31:171–187.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC, Boca Raton, FL.

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, 50:1064–1072.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19:58–80.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian statistics 4*, pages 35–60. Oxford University Press, Oxford.

Berger, J. O. and Wolpert, R. L. (1988). *The likelihood principle, Second edition*. Institute of Mathematical Statistics, Hayward, CA.

Bernardo, J. M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–121.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75:11–20.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods Research*, 33:261–304.

Clayton, D. (2003). Conditional likelihood inference under complex ascertainment using data augmentation. *Biometrika*, 90:976–981.

Clayton, D. and Hills, M. (1993). *Statistical models in epidemiology*. Oxford University Press, Oxford.

Commenges, D., Letenneur, L., Joly, P., Alioum, A., and Dartigues, J.-F. (1998). Modelling age-specific risk: application to dementia. *Statistics in Medicine*, 17:1973–1988.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 74:187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press, Cambridge.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall, London.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41:1–31.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:278–292.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95:407–424.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68.

de Finetti, B. (1974). Bayesianism: its unifying role for both the foundations and applications of statistics. *International Statistical Review*, 42:117–130.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12:27–36.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36:271–281.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, 8:745–764.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 57:45–97.

Efron, B. (1978). Controversies in the foundations of statistics. *The American Mathematical Monthly*, 85:231–246.

Efron, B. (1986). Why isn't everyone a Bayesian. *The American Statistician*, 40:1–11.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver and Boyd, Edinburgh.

Gelman, A. and Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, 25:165–173.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15:413–419.

Guo, G. (1993). Event-history analysis for left-truncated data. *Sociological Methodology*, 23:217–243.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge University Press, Cambridge.

Jefferys, W. H. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80:64–72.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society, Series A*, 186:453–461.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7:149–160.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition*. Wiley, New Jersey.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B*, 32:175–208.

Karvanen, J., Kulathinal, S., and Gasbarra, D. (2009). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis*, 53:1782–1793.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kingman, J. F. C. (1978). Uses of exchangeability. *The Annals of Probability*, 6:183–197.

Kolmogorov, A. N. (1933). *Grundbegriffe der wahrscheinlichkeitsrechnung*. Julius Springer, Berlin.

Kulathinal, S. and Arjas, E. (2006). Bayesian inference from case-cohort data with multiple end-points. *Scandinavian Journal of Statistics*, 33:25–36.

Kulathinal, S., Karvanen, J., Saarela, O., and Kuulasmaa, K. (2007). Case-cohort design in practice - experiences from the MORGAM project. *Epidemiologic Perspectives & Innovations*, 4. Available from `http://www.epi-perspectives.com/content/4/1/15`.

Langholz, B. (2007). Use of cohort information in the design and analysis of case-control studies. *Scandinavian Journal of Statistics*, 34:120–136.

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiological cohort studies. *Statistical Science*, 11:35–53.

Langholz, B. and Goldstein, L. (2001). Conditional logistic analysis of case-control studies with complex sampling. *Biostatistics*, 2:63–84.

Lindley, D. V. and Novick, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics*, 9:45–58.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley, New York.

Ma, J., Amos, C. I., and Daw, E. W. (2007). Ascertainment correction for Markov chain Monte Carlo segregation and linkage analysis of a quantitative trait. *Genetic Epidemiology*, 31:594–604.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC, Boca Raton, FL.

Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4:85–118.

Pearl, J. (2009). *Causality, Second Edition*. Cambridge University Press, Cambridge.

Pitkäniemi, J., Varvio, S.-L., Corander, J., Lehti, N., Partanen, J., Tuomilehto-Wolf, E., Tuomilehto, J., Thomas, A., and Arjas, E. (2009). Full likelihood analysis of genetic risk with variable age at onset diseasecombining population-based registry data and demographic information. *PLoS One*, 6. e6836.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer, New York.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor, *Latent variable modeling and applications to causality. Lecture notes in statistics (120)*, pages 69–117. Springer-Verlag, New York.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6:34–58.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields*. Chapman & Hall/CRC, Boca Raton, FL.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84:379–394.

Samuelsen, S. O., Ånestad, H., and Skrondal, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*, 34:103–119.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley, Chichester.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag, New York.

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

Seaman, S. R. and Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika*, 88:1073–1088.

Shafer, G. (1992). What is probability? In Hoaglin, D. C. and Moore, D. S., editors, *Perspectives on contemporary statistics*, pages 93–106. Mathematical Association of America, Washington, DC.

Sisson, S. A. (2005). Transdimensional Markov chains: a decade of progress and future perspectives. *Journal of the American Statistical Association*, 100:1077–1089.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, Boca Raton, FL.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, 42:213–220.

Stigler, S. M. (1982). Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society, Series A*, 145:250–258.

Stigler, S. M. (1983). Who discovered Bayes's theorem? *The American Statistician*, 37:290–296.

Teicher, H. (1960). On the mixture of distributions. *The Annals of Mathematical Statistics*, 31:55–73.

Wacholder, S. (1996). The case-control study as data missing by design: estimating risk differences. *Epidemiology*, 7:144–150.

Wacholder, S. and Weinberg, C. R. (1994). Flexible maximum likelihood methods for assessing joint effects in case-control studies. *Biometrics*, 50:350–357.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B*, 31:80–88.

Young, G. A. and Smith, R. L. (2005). *Essentials of statistical inference*. Cambridge University Press, Cambridge.

Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, 11:769–782.

Summaries of the original publications

I. Saarela, O. and Kulathinal, S. (2007). Conditional likelihood inference in a case-cohort design: an application to haplotype analysis. *The International Journal of Biostatistics*, 3. Available from `http://www.bepress.com/ijb/vol3/iss1/1`.

Conditional likelihood inference under case-cohort design is formulated and applied using the EM-algorithm in the context of joint estimation of regression coefficients and population haplotype frequencies. The conditional and full likelihood approaches are compared in a simulation study. The conditional likelihood correction term is derived under multinomial regression model and survival model with type I censoring.

II. Saarela, O., Kulathinal, S., Arjas, E., and Läärä, E. (2008). Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Statistics in Medicine*, 27:5991–6008.

Full likelihood inference is considered under multiple nested case-control selections and compared to non-likelihood-based alternatives in a simulation study based on real data. The model is estimated by maximizing the observed data likelihood function. It is shown that under certain assumptions the full likelihood approach is applicable for general cohort sampling designs.

III. Saarela, O., Kulathinal, S., and Karvanen, J. (2009). Joint analysis of prevalence and incidence data using conditional likelihood. *Biostatistics*, 10:575–587.

A conditional likelihood expression is formulated for follow-up data subject to both left censoring and left truncation due to individuals with a prevalent disease status observed at the time of cross-sectional recruitment of a follow-up cohort. The model is estimated by maximizing the observed data likelihood function. The proposed method is compared in a simulation study based on real data to a standard approach where the prevalent cases are excluded from the analysis. Based on this, the new method results in a gain in efficiency.

IV. Saarela, O. and Arjas, E. (2010). A method for Bayesian monotonic multiple regression. Accepted for publication in *Scandinavian Journal of Statistics*.

A non-parametric monotonic regression model for one or more covariates and a Bayesian estimation procedure are formulated. These are shown to work in simulated and real data examples. The monotonic

construction is based on marked point process realizations, and allows reduction into lower dimensional submodels. The actual inference is based on model averaged results over the realizations.

V. Arjas, E. and Saarela, O. (2010). Optimal dynamic regimes: presenting a case for predictive inference. *The International Journal of Biostatistics*, 6. Available from `http://www.bepress.com/ijb/vol6/iss2/10`.

Causal inference using posterior predictive distributions is discussed in the context of optimal sequential treatment regimes. A hierarchical model formulation and a Bayesian estimation procedure are presented for an example application, utilizing the model proposed in Article IV.