Coarsening at random: characterizations, conjectures, counter-examples

Richard D. Gill
Mark J. van der Laan
James M. Robins

ABSTRACT The notion of *coarsening at random* (CAR) was introduced by Heitjan and Rubin (1991) to describe the most general form of randomly grouped, censored, or missing data, for which the coarsening mechanism can be ignored when making likelihood-based inference about the parameters of the distribution of the variable of interest. The CAR assumption is popular, and applications abound. However the full implications of the assumption have not been realized. Moreover a satisfactory theory of CAR for continuously distributed data—which is needed in many applications, particularly in survival analysis—hardly exists as yet. This paper gives a detailed study of CAR. We show that grouped data from a finite sample space *always* fit a CAR model: a nonparametric model for the variable of interest together with the assumption of an arbitrary CAR mechanism puts no restriction at all on the distribution of the observed data. In a slogan, `CAR is everything`. We describe what would seem to be the most general way CAR data could occur in practice, a sequential procedure called *randomized monotone coarsening*. We show that CAR mechanisms exist which are not of this type. Such a coarsening mechanism uses information about the underlying data which is not revealed to the observer, without this affecting the observer's conclusions. In a second slogan, `CAR is more than it seems`. This implies that if the analyst can argue from subject-matter considerations that coarsened data is CAR, he or she has knowledge about the structure of the coarsening mechanism which can be put to good use in non-likelihood-based inference procedures. We argue that this is a valuable option in multivariate survival analysis. We give a new definition of CAR in general sample spaces, criticising earlier proposals, and we establish parallel results to the discrete case. The new definition focusses on the distribution rather than the density of the data. It allows us to generalise the theory of CAR to the important situation where coarsening variables (e.g., censoring times) are partially observed as well as the variables of interest.

0. Overview

The phenomena of missing data in multivariate analysis (some components of a multivariate vector not observed), censoring in survival analysis, and grouped data in general, have in common that rather than observing a random variable (or vector) X of interest, one is only able to observe that X

takes a value in some possibly randomly determined set of values; by 'randomly determined' we mean that the set not only depends on X itself but also possibly on auxiliary random variables. The notion of '*coarsening at random*' was introduced by Heitjan and Rubin (1991) to single out exactly those situations in which the coarsening mechanism can be ignored when making inference on the distribution of X.

The same notion, but restricted in application to missing observations in a multivariate vector, goes back to Rubin (1976) and Little and Rubin (1987); in this context it is called 'missing at random'. These papers, and those of Heitjan (1993, 1994) have studied the statistical consequences of MAR and CAR in parametric models. Heitjan (1993) gives significant biomedical examples of CAR. Robins and Rotnitzky (1992), van der Laan (1993, 1995), Robins, Rotnitzky and Zhao (1994), and Robins (1996a) study statistical consequences in non- and semi-parametric models (with positive probability of complete observations). In survival analysis, with right-censored observations, coarsening at random is intimately connected to the central notion of 'independent censoring'; see Jacobsen and Keiding (1995).

'Coarsening at random' has clearly become an important topic in survival analysis, in biostatistics; in general, in applied statistics. Practitioners are keen to be able to assume that coarsened data is 'coarsened at random'. Yet in our opinion the notion is still poorly understood and this has dangers in uncritical application. We want to to find out what it really means, from a modelling point of view, to make the CAR assumption. To underline the problem, we note that CAR has only been defined for discrete data but the mere definition of CAR for continuous data—e.g., censored survival times—is not obvious at all. This is not just a question of measure-theoretic technicalities. Jacobsen and Keiding (1995) offer a definition, but though we learnt a great deal from their results, in our opinion their (rather complex and at the same time restrictive) definition does not capture the intended content of the notion.

The paper has a large number of sections, falling into two main parts. Sections 1 to 5 concentrate on the discrete case; 6 to 10 on the general; section 11 concludes. In section 1 we set out the necessary preliminaries. We give the (discrete case) definition of CAR in terms of the conditional distribution of the coarsened data given the underlying data: the conditional probability of a particular outcome depends on the underlying data value only through what the observed data tells us about this value. We show that this is equivalent to a factorization of the likelihood into separate parts corresponding neatly to the underlying data and the coarsening mechanism, and that it is equivalent to a condition concerning the probability law of the underlying data given the observed: it is the same as if the possible coarsenings had been fixed in advance, independently of the underlying data; so-called Coarsening Completely at Random.

In section 2 we give one of our main new results, 'CAR is everything': if we assume nothing about the distribution of the underlying complete data, and nothing about the coarsening mechanism except that it is CAR, then we are in effect assuming nothing at all about the distribution of the data. Any coarsened data from a finite sample space fits exactly to a CAR model. Without adding further assumptions either on the complete-data generating mechanism, or on the coarsening mechanism, the CAR assumption is untestable. This result generalises the well-known result in survival analysis known as the unidentifiabilty of the independent competing risks assumption. That our result is a reasonable conjecture follows by counting parameters; see Rubin, Stern and Vehovar (1995; section 4) for the case of Missing at Random. Our proof combines likelihood and convexity theory; in a sense, we use statistics to prove a theorem of algebra.

A counterexample shows that extending to infinite sample spaces, the result is not always true. However, every distribution of coarsened data can still be arbitrarily well approximated by a CAR model, so no amount of real data will ever be able to rule out a CAR model.

In section 3 we look at the CAR assumption as a modelling assumption, asking ourselves what kind of coarsening mechanisms could arise in nature, satisfying CAR. We argue that the most general realistic (physical) mechanism which produces CAR data is a sequential procedure we call 'sequential randomised coarsening'. A rather natural conjecture is then that *all* CAR mechanisms can be realised in this way. If the conjecture were true, CAR would not just be an attractive assumption to make (because of its data-analytic consequences) but also an assumption with physical or subject-matter content. If the conjecture is false, then CAR may be convenient, but in itself difficult to justify. Put yet another way, if you *can* justify the CAR assumption on subject-matter grounds, you actually know more about the coarsening mechanism than just the fact that it is CAR. It turns out that CAR mechanisms exist which cannot be represented sequentially, so that CAR is indeed much more than it seems.

Section 4 is a kind of interlude, establishing further nice properties of a rather special sequential coarsening mechanism, called 'monotone coarsening'. In this situation nonparametric maximum likelihood estimation can be done explicitly (without iteration), as in the case of the Kaplan-Meier estimator for right-censored survival data.

If the data are CAR, additional knowledge about the coarsening mechanism is irrelevant for likelihood-based inference. However we argue in section 5 that there are many situations, and we give an example in multivariate survival analysis, where restricting oneself to likelihood based methods may leave one with no practically useful methods at all. It can be better to make use of asymptotically irrelevant additional information about the coarsening mechanism and construct ad hoc (frequency based) methods, which are not asymptotically efficient, but do actually work in realistically sized samples.

The mathematical results so far were restricted to a finite sample space though we argue that their broad implications hold generally, for instance in survival analysis. In section 5 we already drew such implications. However we also want to develop the mathematical theory of CAR in this direction. The second main group of sections, 6 to 10, is devoted to the generalisation of the earlier results to arbitrary sample spaces. To begin with, we need a general definition of CAR.

CAR is usually stated as a condition on the (discrete) density function of the observed given the underlying data. One then is led to define CAR in general also in terms of densities. This leads immediately into technical problems, since coarsened data is a random set, and it is not clear how to introduce density functions into the picture. Jacobsen and Keiding (1995) took this route, defining CAR in terms of the densities of the variables in the model under consideration, relative to a 'reference model'. Our philosophy is different. We remark that in discrete models, a discrete density is just a probability. We read the definition of CAR as a statement about certain conditional *distributions*, not conditional densities. Now it is more or less immediate how one should generalise such a statement to a general case. Our general definition of CAR, in section 6, is that certain conditional distributions should coincide on certain parts of the sample space: 'the distribution of the observed data given the underlying variable of interest only depends on that variable through the information given to us about it, in the data'. There are minor measure-theoretic issues in making this definition mathematically rigorous, since conditional distributions can be changed at will on conditioning events of probability zero. For the more practically motivated reader it is enough to know that it is possible to make the definition precise in such a way that it both can be applied to the cases of interest in practice, and that it has the expected statistical consequences. The section shows that from our general definition of CAR do indeed follow the expected factorization of the likelihood, and the expected property of the conditional distribution of the underlying data given the observed. However these properties though implied by CAR are no longer equivalent to CAR, so here an important difference with the discrete case emerges.

Section 7 contrasts our 'absolute' definition of CAR (for general sample spaces) with the 'relative' definition of Jacobsen and Keiding (1995), and establishes the connections.

In section 8 we consider an important general issue: suppose we also observe to some extent some aspects of the coarsening mechanism. Is there still a natural definition of coarsening at random? For example, often in survival analysis one observes part or all of the censoring variables, even for uncensored observations. The original notion of CAR is only applicable when the actual data is strictly a coarsening of the underlying survival time. No further information is supposed to be available. We show that there are no problems in extending our definition and results to this more general case still. Here again our philosophy of thinking in terms of (conditional) distributions, not densities, pays off.

In sections 9 and 10 we attempt to extend our results on existence and uniqueness of a CAR model for arbitrarily coarsened data to the general case. Recall that in section 2 we show that discrete coarsened data always fits exactly to a CAR model, and that the underlying distribution and the CAR mechanism can be essentially uniquely reconstructed from the law of the data. However this breaks down in general sample spaces (even in countable sample spaces), though there is a good sense in which it is true for practical purposes: CAR is 'almost everything'. Any coarsened data whatsoever can be fit arbitrarily well by a CAR model, if not exactly. Anyway, in section 9 we obtain another 'next best' result stating, in the language of semiparametric models and information bounds, that the CAR model places no restrictions on the distribution of the data. In the neighbourhood of a CAR model one has so much freedom (although subject to the CAR assumption) that the set of possible score-functions is everything, and estimation is as difficult as in a completely non-parametric model. Section 10 gives a uniqueness result on the factorization of the likelihood in a CAR model, generalising the uniqueness of the CAR decomposition established in the discrete case.

In section 11 we conclude and in particular survey the many open problems which remain. We see that Coarsening at Random is not only a topic full of importance and interest from an applied point of view, but also full of challenges to theoreticians, opening a view to a rich and delicate theory. Survival analysis and mathematical statistics continue to enrich one another over the years.

A companion paper in this volume, Gill and Robins (1997), studies coarsening mechanisms with a sequential structure. Following Robins (1996b) we consider a generalisation of CAR in which coarsening occurs in a number of phases, each of which separately is CAR while the overall result is not. We also further investigate sequential mechanisms for producing MAR data. How could one 'physically' realise a general MAR mechanism? Is the MAR assumption an assumption which on its own can be supported by subject matter knowledge, or is it the case that if one can argue for MAR, one actually knows more (and therefore, outside of likelihood-based inference, has more options in data-analysis)?

1. Preliminaries

Suppose X is a random variable taking values x in a finite set E. Let $\mathcal{E}$ denote the set of all subsets of E, and let $\mathcal{X}$ denote a random nonempty subset of E: so $\mathcal{X}$ takes values A in $\mathcal{E} \setminus \{\emptyset\}$. We say that $\mathcal{X}$ is a coarsening of X if, with probability 1, $X \in \mathcal{X}$. The observed data, the random set $\mathcal{X}$, is usually denoted by Y in the literature on CAR. However later we will make a distinction between the random set $\mathcal{X}$ and its representation in the data Y as a list of coordinates, coefficients, or types.

If $\mathcal{X}$ is a coarsening of X, and one observes $\mathcal{X}$ but not X itself, one may ask if the observation '$\mathcal{X} = A$' can be treated for statistical purposes as the observation '$X \in A$'; i.e., as if the value of $\mathcal{X}$ instead of being random, had been provided in advance. Heitjan and Rubin (1991) show that this is the case if the conditional distribution of $\mathcal{X}$ given $X = x$ satisfies the following *coarsened at random* (CAR) assumption:

$$\text{for all } A \in \mathcal{E}, \quad \Pr(\mathcal{X} = A | X = x) \quad \text{is constant in } x \in A. \tag{1}$$

Obviously $\Pr(\mathcal{X} = A|X = x) = 0$ if $x \notin A$, if $\mathcal{X}$ is a coarsening of X. In a moment we derive their main result on ignorability of the coarsening mechanisms under CAR, but first we note that the CAR assumption intuitively seems to say that the observation of $\mathcal{X} = A$ is not influenced by the specific value of X in A which was taken, only by the fact that X *did* take a value in A. In fact CAR is obviously equivalent to

$$\Pr(\mathcal{X} = A|X = x) \quad = \quad \Pr(\mathcal{X} = A|X \in A) \quad \forall A, x \in A. \tag{2}$$

The CAR assumption is an assumption on the coarsening mechanism leading from X to $\mathcal{X}$, by which we emphasize that coarsening is seen as occuring in two stages: firstly the random variable X of interest is realised; secondly, a conceptually different process (usually associated with features of measurement or observational restrictions, rather than the scientific phenomenon under study itself), given the value x taken by X, replaces this value by a set $\mathcal{X} = A \ni x$.

However, having observed $\mathcal{X}$, we are free to consider the conditional distribution of X given $\mathcal{X} = A$, even though this compounds two quite different processes. Since (2) can be rewritten as (for all $x \in A$)

$$\Pr(\mathcal{X} = A|X = x \text{ and } X \in A) \quad = \quad \Pr(\mathcal{X} = A|X \in A)$$

we can recognise it as a conditional independence assumption: given $X \in A$, the events $X = x$ and $\mathcal{X} = A$ are independent. By symmetry of (conditional) independence, we therefore equivalently have:

$$\Pr(X = x|\mathcal{X} = A \text{ and } X \in A) \quad = \quad \Pr(X = x|X \in A)$$

But since the former is equal to $\Pr(X = x|\mathcal{X} = A)$ we have that CAR is equivalent to:

$$\Pr(X = x|\mathcal{X} = A) \quad = \quad \Pr(X = x|X \in A) \quad \text{for all } x \in A. \tag{3}$$

Thus the observation of $\mathcal{X} = A$ tells us no more, in the sense of what is now the conditional distribution of X, than the obvious '$X \in A$'.

So far we have only discussed the (probabilistic) interpretation of the CAR assumption. Now we give Heitjan and Rubin's statistical consequence. Suppose the distribution of X depends on a parameter θ, while the coarsening mechanism, supposed CAR, depends on a distinct, variation independent, parameter γ. We suppose CAR holds, for each γ. Write

$$p_x^\theta \quad = \quad \Pr{}^\theta(X = x); \qquad p_A^\theta = \Pr{}^\theta(X \in A);$$

$$\begin{aligned}\pi^\gamma_A &= \Pr{}^\gamma(\mathcal{X}=A|X=x) \qquad (x\in A)\\ &= \Pr{}^\gamma(\mathcal{X}=A|X\in A).\end{aligned}$$

The marginal distribution of $\mathcal{X}$ is

$$\begin{aligned}f^{\theta,\gamma}_A=\Pr{}^{\theta,\gamma}(\mathcal{X}=A) &= \Pr{}^{\theta,\gamma}(\mathcal{X}=A \text{ and } X\in A)\\ &= \Pr{}^\theta(X\in A)\Pr{}^\gamma(\mathcal{X}=A|X\in A)\\ &= p^\theta_A\pi^\gamma_A. \end{aligned}\tag{4}$$

So under CAR, the joint likelihood for θ and γ *factors* and the θ part can be written down *without knowledge* of *the coarsening mechanism*: as far as θ is concerned, the observation '$\mathcal{X}=A$' can be treated like an observation '$X\in A$'; i.e., as if the coarsening mechanism had been Coarsening Completely at Random: the sample space is partitioned in advance, independently of X, and one just looks to see which element of the partition X falls into. At the same time, the likelihood for γ can be written down without knowing the distribution of X, and moreover the likelihood for γ based on the data $\mathcal{X}$ is the same as the likelihood for γ based on the conditional distribution of $\mathcal{X}$ given X (even though X itself cannot be observed).

2. CAR is everything

Suppose $\mathcal{X}$ is a coarsening of X (in the discrete set-up of the previous section). We observe $\mathcal{X}$ only. If we assume nothing about the distribution of X, but we do assume CAR, does this imply anything about the distribution of the observable $\mathcal{X}$? Put another way, given a random non-empty set $\mathcal{X}$, can we construct a random variable X such that $\mathcal{X}$ is a coarsening of X and CAR holds?

Mathematically we have the following

Question. *Given a probability distribution* $(f_A : A\in\mathcal{E}, A\neq\emptyset)$ *of a random non-empty set* $\mathcal{X}$*, can we write*

$$f_A=p_A\pi_A \tag{5}$$

where $(p_x : x\in E)$ *is a probability distribution on the finite set* E*,* p_A *is defined by* $p_A=\sum_{x\in A}p_x$*, and* $(\pi_A : A\in\mathcal{E}\backslash\{\emptyset\})$ *is a set of probabilities such that, for each* $x\in E$*,*

$$\sum_{A\ni x}\pi_A=1 \quad ?$$

For given that the distribution of $\mathcal{X}$ factors as in (5), construct a joint distribution of X and $\mathcal{X}$ by letting

$$\begin{aligned}&\Pr(X=x|\mathcal{X}=A)=p_x/p_A \quad && x\in A,\ f_A>0,\\ &\Pr(X=x|\mathcal{X}=A)=0 && x\notin A.\end{aligned}$$

Under (5), if $f_A > 0$ then $p_A > 0$ too so the construction is well-defined. The construction forces $X \in \mathcal{X}$ to hold with probability 1. Moreover

$$\Pr(X = x \text{ and } \mathcal{X} = A) = \frac{p_x}{p_A} p_A \pi_A = p_x \pi_A \text{ for } A \neq \emptyset,\ x \in A,\ f_A > 0$$

and trivially

$$\Pr(X = x \text{ and } \mathcal{X} = A) = p_x \pi_A \qquad A \neq \emptyset,\ x \in A$$

if $f_A = 0$ and hence $p_A = 0$ or $\pi_A = 0$. Adding over $A \ni x$ shows that the marginal distribution of X is (p_x). Dividing by p_x shows

$$\Pr(\mathcal{X} = A | X = x) = \pi_A \qquad A \ni x,\ p_x > 0$$

which doesn't depend on x, so CAR holds.

This argument, together with the conditional independence arguments of Section 1, shows that CAR can equivalentally be formulated as:

CAR($\mathcal{X}|X$): the conditional law of $\mathcal{X}$ given X satisfies (1) or (2);

CAR($X|\mathcal{X}$): the conditional law of X given $\mathcal{X}$ satisfies (3); and

FACTOR($\mathcal{X}$): the marginal law of $\mathcal{X}$ factors as in (5).

Now we return to the question posed at the beginning of this section. Just counting equations and unknowns suggests, but does not prove, that the answer is *yes*. The equations are non-linear and the unknowns have to be probabilities. We will prove that the answer is yes and moreover that the factorization (5) is unique for those A with $f_A > 0$. Our proof uses statistical ideas; we consider the expected log likelihood for a CAR model for the given observed data distribution. First we state the result formally:

Theorem. *Let $\mathcal{X}$ be a random non-empty set with distribution $(f_A : A \in \mathcal{E}\backslash\{\emptyset\})$. Then there exist CAR probabilities (π_A) and a distribution (p_x) on E such that $f_A = \pi_A p_A$ for all A, where $p_A = \sum_{x \in A} p_x$. For each A with $f_A > 0$, π_A and p_A are unique.*

Proof. Consider the problem of maximization of

$$\sum f_A \log(p_A \pi_A) = \sum f_A \log p_A + \sum f_A \log \pi_A$$

$$\text{over} \quad p_A = \sum_{x \in A} p_x,\ p_x \geq 0,\ p_E = 1,\ \sum_{A \ni x} \pi_A = 1\ \forall x,\ \pi_A \geq 0.$$

Considered as a function of the p_A and π_A for A with $f_A > 0$, with all these variables varying unrestrictedly in $[0, \infty)$, $\sum f_A \log p_A + \sum f_A \log \pi_A$ is continuous and strictly concave, and takes values in $[-\infty, \infty)$. The subset of p_A and π_A satisfying all the further listed constraints (to be interpreted as 'there exist p_x, π_A such that ...', as far as these variables were not yet involved) is convex and compact. So the supremum is attained uniquely as far as the original restricted set of variables are concerned, but may not be unique when we take account of the supplementary variables as well. We want to show that at least one of the solutions not only solves the maximization problem but also provides a factorization $f_A = \pi_A p_A$.

We study a solution for the p_A separately, in more detail. The analysis would be much simplified if we knew in advance that each $p_x > 0$ (which is for instance the case if each $f_{\{x\}} > 0$) and the reader could first restrict attention to this case. Consider the maximization of $\sum_A f_A \log p_A$, now over variables $p_x \geq 0$, subject to the constraint $p_E = 1$, where $p_A = \sum_{x\in A} p_x$. There exists a solution, and by concavity of $\sum f_A \log(\sum_{x\in A} p_x)$ we know, see for instance Whittle (1961), that there exists a Lagrange multiplier λ such that any solution is also solution of the problem: maximize $\sum_A f_A \log p_A - \lambda p_E$ over $p_x \geq 0$. At a given solution, for those x satisfying $p_x > 0$ differentiating with respect to p_x shows

$$\sum_{A\ni x} \frac{f_A}{p_A} - \lambda = 0. \tag{6}$$

For other x such that $p_x = 0$ we only have

$$\sum_{A\ni x} \frac{f_A}{p_A} - \lambda \leq 0.$$

If at this solution $p_A = 0$, then we must have $f_A = 0$ (otherwise $f_A \log p_A = -\infty$) and $p_x = 0$ for all $x \in A$. Multiplying (6) by p_x and adding over x such that $p_x > 0$ gives

$$\begin{aligned}
0 &= \sum_{x:p_x>0} p_x \sum_{A\ni x} \frac{f_A}{p_A} - \lambda \sum_{x:p_x>0} p_x \\
&= \sum_x \sum_{A\ni x} p_x \frac{f_A}{p_A} - \lambda \qquad\qquad \text{where } 0/0 = 0 \\
&= \sum_A \sum_{x\in A} p_x \frac{f_A}{p_A} - \lambda \\
&= \sum_A p_A \frac{f_A}{p_A} - \lambda = \sum_A f_A - \lambda \qquad \text{since } p_A = 0 \Rightarrow f_A = 0 \\
&= 1 - \lambda.
\end{aligned}$$

So $\lambda = 1$ and we have

$$\begin{aligned}
\sum_{A\ni x} \frac{f_A}{p_A} &= 1 \quad \text{if} \quad p_x > 0 \\
\sum_{A\ni x} \frac{f_A}{p_A} &\leq 1 \quad \text{if} \quad p_x = 0.
\end{aligned}$$

Define now $\pi_A = f_A/p_A$ *except* that if $p_x = 0$, so also $f_{\{x\}} = 0$, define $\pi_{\{x\}} = 1 - \sum_{A \ni x} f_A/p_A$ (with $0/0 = 0$ throughout). We then have:

$$\sum_{A \ni x} \pi_A = 1 \quad \text{for all } x;$$

and $f_A = p_A \pi_A$ for all A (also A with $p_A = 0$, also singletons).

Thus a factorization $f_A = \pi_A p_A$ exists. Since $\sum f_A \log \widehat{f}_A \leq \sum f_A \log f_A$ for all $\widehat{f}_A = \pi_A p_A$, the factorization we have found must also be a solution of the maximization problem we considered at the outset. As we remarked, this uniquely determines the p_A and π_A for A with $f_A > 0$. □

It is difficult to give necessary and sufficient conditions for uniqueness of *all* p_x and π_A in the factorization $f_A = p_A \pi_A$. If $f_{\{x\}} > 0$ for all x, then $p_x > 0$ and is uniquely determined for all x, hence $p_A > 0$ for all A and $\pi_A = f_A/p_A$ is uniquely determined for all A. Consider the incidence matrix with rows corresponding to A with $f_A > 0$ and augmented if necessary with a row for $A = E$; columns corresponding to $x \in E$; and the (A, x) element equal to the indicator of $x \in A$. The vector of p_A's with $f_A > 0$ augmented with p_E, equals this matrix times the vector of (p_x); so if the matrix has rank equal to the number of elements of E, (p_x) is uniquely determined. This rank condition is however not necessary, since the inequalities $p_x \geq 0$ might also help to uniquely determine (p_x) from $(p_A : f_A > 0 \text{ or } A = E)$.

One would hope that the result 'car is everything' remains true in general sample spaces but the following counter-example, due to Ya'akov Ritov, shows that this hope fails already in a countable sample space. The example may seem artificial, but actually it corresponds to a survival analysis situation where every observation is censored and arbitrarily large censored observations can occur. Let E be the natural numbers $\{0, 1, \ldots\}$ and suppose the only subsets of E which get positive probability are $\{n, n+1, \ldots\}$ for $n = 1, 2, \ldots$. If we try to factor $f_A = p_A \pi_A$ by maximizing the log likelihood $\sum f_A \log p_A$ we see that the likelihood is always increased by moving probability from the left to the right. The maximizer would like to put all the probability mass at $+\infty$ but there is no such point in E so the maximizer does not exist. Hence there can be no factorization, since if there were one, it would maximize the log likelihood by fitting the f_A exactly. One could try to save the situation by adding a point $+\infty$ to E but this only helps if one also adds the same point to all the sets $A = \{n, n+1, \ldots\}$. In other words, this example can be repaired by compactifying both the sample space and all the observed random sets in a careful way. That might be true in general.

Even if CAR is not everything in the strict sense, we do argue that it is 'almost everything'. At least from a practical point of view, every coarsened data model can be arbitrarily well approximated by a model for discrete coarsened data, and a CAR model fits that exactly. Combining these two facts, a CAR model fits arbitrarily well, even if not exactly. Similarly, the empirical distribution of any finite sample of coarsened data can be exactly fit to a CAR model: computation of the non-parametric maximum likelihood estimator does exactly this job, where without loss of generality one can assume that the support of the underlying variable X is finite.

3. Sequential representations of CAR

So far a CAR mechanism is described in an algebraic way: just a collection of probabilities π_A satisfying

$$\sum_{A\ni x} \pi_A = 1$$

for each $x \in E$. Is there a more appealing way to describe all CAR mechanisms? Is there a convenient way to simulate any CAR mechanism?

The direct way to simulate the random set $\mathcal{X}$ is first to generate X according to the law $(p_x : x \in E)$, then $\mathcal{X}$ according to the conditional law $(\pi_A : A \ni x)$. This makes no use of the fact that the coarsening mechanism is actually CAR. Moreover in the course of the simulation we have to look at the specific value taken by X, even though this value is not later revealed by $\mathcal{X}$. Another way is to directly generate $\mathcal{X}$ from its marginal distribution $(f_A = p_A \pi_A : A \subseteq E)$. Again, once the probabilities f_A have been calculated, no use is made of the fact that coarsening is CAR.

A rather special kind of CAR does allow an appealing simulation construction: so-called *monotone coarsening* (or *monotone missingness*). Consider the collection of subsets A with $\pi_A > 0$. Suppose no two of these subsets overlap non-trivially. Consider the directed graph on $\{A : \pi_A > 0\} \cup \{E\}$ where there is an edge from A to A' if and only if $A' \subset A$ and no A'' exists with $A' \subset A'' \subset A$ (and $\pi_{A''} > 0$). This graph forms a tree with root at E.

The leaves of the tree form a partition of E; and in fact the branches leading from any node A form a partition of A. For suppose the contrary were true: there exists A, $x \in A$, $A' \subset A$ with $\pi_{A'} > 0$, and $x \notin A'$. Moreover $x \notin A''$ for any $A'' \subset A$ with $\pi_{A''} > 0$. Choose $x' \in A'$. We have $1 = \sum_{A''\ni x} \pi_{A''} < \pi_{A'} + \sum_{A''\ni x} \pi_{A''} \le \sum_{A''\ni x'} \pi_{A''} = 1$, which is impossible.

The random set $\mathcal{X}$ can be generated by a random walk up the tree, starting at the root E, and stopping somewhere, $\mathcal{X} = A$, on the way up. Suppose at some stage we have just moved into the node A. We then stop in A with probability $\pi_A/(1 - \sum_{A'\supset A} \pi_{A'})$. (This expression is a probability, since for any $x \in A$ we have $1 = \sum_{A'\ni x} \pi_{A'} \ge \pi_A + \sum_{A'\supset A} \pi_{A'}$.) Conditionally on *not* stopping in A, we choose a branch A' with probability $p_{A'}/p_A$ and move into A'. Since the branches A' from A form a partition of A, the branching probabilities add to 1. An equivalent description of this step is that, knowing now that $X \in A$ and that we do not stop here, we look to see which element of the partition of A contains X, and move to that element. Note that we use the CAR probabilities π_A to decide whether to stop or not; we use the underlying probabilities p_A to choose a branch, given that we did not stop.

A direct calculation shows that this procedure generates $\mathcal{X}$ with probability distribution $p_A\pi_A$. To see what is going on more intuitively, consider the pair X, $\mathcal{X}$. If the value $X = x$ were known in advance, only one path through the tree would be relevant, the path starting at E and ending at the leaf containing x. Call this path $E = A_0 \supset A_1 \supset \ldots \supset A_k \ni x$. The probabilities π_{A_i} along this path form the distribution of $\mathcal{X}$ given $X = x$, and the 'stopping probability' $\pi_A/(1 - \sum_{A' \supset A} \pi'_A)$, for $A = A_j$, equals $\pi_{A_j}/(1-\sum_{i<j} \pi_{A_i})$. In fact in our simulation we do not generate X in advance but at a given step, when we know already that $X \in A$ and that we do not stop here, we decide which branch A' from A contains X, according to the conditional distribution of X given $X \in A$.

Of course CAR probabilities such that the sets A with $\pi_A > 0$ lie on a tree are rather special. However the idea of generating $\mathcal{X}$ by successively partitioning a set in which X is known to lie, and observing in which element of the partition X lies, seems to us the most general way conceivable to physically realize a CAR mechanism. In monotone coarsening the partitions are given in advance. Now we will allow the partitions to be chosen at random; in principle the choice could depend on what happened at previous stages. However it should not depend on future choices since that would require foreknowledge of X, in conflict with the required CAR property.

Partitioning a set into say k subsets and observing in which X lies can also be carried out by a series of partitions in 2 subsets. Thus we arrive at the following definition of a *randomized monotone coarsening* scheme. Initially, $n = 1$ and $\mathcal{A}_0 = E$. By step n we have generated a sequence of nonempty subsets $\mathcal{A}_0 \supset \mathcal{A}_1 \supset \ldots \supset \mathcal{A}_{n-1}$ and we know $X \in \mathcal{A}_{n-1}$. We may now decide to terminate, and set $\mathcal{X} = \mathcal{A}_{n-1}$, or we decide to continue. In the latter case we choose at random a subset B_n of $\mathcal{A}_{n-1}$. We observe whether X lies in B_n or in $\mathcal{A}_{n-1}\backslash B_n$, and set $\mathcal{A}_n$ equal to B_n or $\mathcal{A}_{n-1}\backslash B_n$ accordingly. Increment n by one, and repeat. The probability of stopping and the probability distribution of B_n, given we do not stop, may depend in an arbitrary way on the past sequence $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{n-1}$. If these probabilities only depend on the current data $\mathcal{A}_{n-1}$ (and on n) then we call the scheme a *Markov coarsening.*

A very natural conjecture is: *any set of CAR probabilities can be represented by a randomized monotone coarsening scheme.* In a simulation of randomized monotone coarsening, as a separate procedure taking place after the generation of X, we do not use more information about X than that which is finally revealed in the value of $\mathcal{X}$. Conversely, given X, to simulate a CAR mechanism which is *not* randomized monotone, the computer program will require in the course of the procedure information about X—perhaps even its precise value—which is ultimately not revealed in the value of $\mathcal{X}$ output by the computer in its final `print` statement. But the fact that the computer has had to hide information from us does not affect our face-value inference, '$\mathcal{X} = A$ tells us no more than that $X \in A$'.

The conjecture is easily found to be true when $\#E = 2$. However already when $\#E = 3$ there are counter-examples. Let $E = \{1, 2, 3\}$; the list of possible A is $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1, 2, 3\}$. The only way the procedure can stop with the set $\{2, 3\}$ is when at the very first stage, the partition $\{1\}, \{2, 3\}$ is chosen (and $X = 2$ or 3). Note that the very first partition is chosen independently of X. So if $X = 1$, we would also have stopped immediately if this had been the first partition chosen. However one can also arrive, when $X = 1$, at $\mathcal{X} = \{1\}$ after various other first stage choices. So the probability to terminate with $\mathcal{X} = \{1\}$, when $X = 1$, is at least as large as the probability to terminate with $\mathcal{X} = \{2, 3\}$ when $X = 2$, or when $X = 3$; i.e., $\pi_{\{1\}} \geq \pi_{\{2,3\}}$. However for instance

$$\begin{gathered}\pi_{\{1\}} = \pi_{\{2\}} = \pi_{\{3\}} = 0.1 \\ \pi_{\{1,2\}} = \pi_{\{1,3\}} = \pi_{\{2,3\}} = 0.4 \\ \pi_{\{1,2,3\}} = 0.1\end{gathered}$$

satisfies $\sum_{A \ni x} \pi_A = 1 \ \forall x = 1, 2, 3$ but $\pi_{\{2,3\}} > \pi_{\{1\}}$.

Obviously the counter-example can be extended to $E = \{1, 2, \ldots, n\}$ for any $n \geq 3$, comparing $\pi_{\{1,2,\ldots,n-1\}}$ with $\pi_{\{n\}}$.

The probabilities in our counter-example are pretty arbitrary. A more extreme example is obtained by letting the probability of each duplet $\{i, j\}$, $i \neq j$, be equal to 0.5, while letting the probabilities of the singeletons and the triplet be equal to zero. Now it is more clear why these probabilities cannot be realized sequentially: when $X = 1$ the computer ultimately has to choose between reporting '$X \in \{1, 2\}$' and '$X \in \{1, 3\}$'; for the other values of X it has to make similar choices. How can it make its choice without observing X completely? (The reader familiar with the so-called quiz-master problem might like to ponder if that is an example of CAR. A prize car is hidden between one of three doors. You choose one door. The quizmaster, knowing the location of the car, opens another, showing that no car is behind it. He then asks if you would like to revise your choice. Do you? We come back to this example in section 8.)

This unsatisfactory state of affairs leaves many questions open. The authors cannot conceive of a more general mechanism than a randomized monotone coarsening scheme for constructing the CAR mechanisms which one would expect to meet with in practice, but is this just a lack of imagination? Can one easily recognise if a given CAR mechanism has a randomized monotone representation or not? As the set E gets larger, do 'most' CAR mechanisms admit a representation?

4. Computing the monotone CAR model

In general, the problem of computing the decomposition $f_A = p_A \pi_A$ of a given (f_A) does not have an explicit solution. As we saw, it can be phrased as a constrained maximization problem $(\max \sum f_A \log p_A)$, which can be solved by various numerical procedures including the iterative EM or Turnbull algorithm, $p_x^{\text{new}} = \sum_{A \ni x} f_A p_x^{\text{old}} / p_A^{\text{old}}$.

One case (and one case only?) does have an explicit solution. This is when (f_A) is *monotone* by which, similarly to the definition of a monotone coarsening mechanism in the previous section, we mean that for $A \neq A'$ such that $f_A > 0$ and $f_{A'} > 0$, either $A \subset A'$, or $A' \subset A$, or $A' \cap A = \emptyset$. In other words: no two A with positive probability in the data overlap non-trivially. Before, the same statement was supposed to be true conditionally on $X = x$, for any x.

We saw in the last section that a monotone coarsening mechanism could be simulated on the computer as a random walk on a tree. This representation can also be used to compute explicitly the decomposition $f_A = p_A \pi_A$ from given (f_A), using a 'tree'-version of the Kaplan-Meier estimator from ordinary survival analysis—right censored survival times supplying a particularly simple example of monotone data.

Consider the collection of A such that $f_A > 0$, augmented (if they are not already included), with E and all the singletons. This collection also has no non-trivially overlapping members. Define a directed graph on this collection of nodes, with a branch (directed edge) from A to A' if $A \supset A'$ but there is no A'' with $A \supset A'' \supset A$. The symbol $\supset$ means strict inclusion. The graph is easily seen to be a tree with root in E and with the singletons $\{x\}$, $x \in E$, as its leaves. The branches from any node A partition A.

We know that a decomposition $f_A = p_A \pi_A$ exists. Moreover it may be chosen so that $f_A = 0 \Rightarrow \pi_A = 0$ except possibly if A is a singleton. So all A with $\pi_A > 0$ are on our tree: the coarsening mechanism is also monotone.

Consider a point x and in imagination hold the tree at the root E and leaf $\{x\}$, pulling it tight, and so forming a straight path from the ground to the sky with side branches on the way up. Consider the steps up the main path of the tree from E to $\{x\}$ as discrete time-steps, with the leaf at time infinity. Consider all the f_A as empirical (relative) frequencies of a large sample of data. Each observation A, together with the underlying true value of $X = x' \in A$, represents a path on the tree, starting from E, going through A, and ending at $\{x'\}$. At some point this path must branch off from the main route up to $\{x\}$ (unless $x' = x$); either before, at, or after A. Consider the branching time as an underlying survival time (you die when you leave the straight and narrow path; you live for ever if $x' = x$). The underlying survival time is observed exactly if A is off the main path or if $A = \{x\}$; but is unknown if A is on the main path, before $\{x\}$. If A lies on the main path the observation is censored just *before* this time point. The underlying survival function between consecutive time steps $A \supset A'$

going up the main path is $p_{A'}$. Hence $p_{\{x\}}$ is the estimated probability to be still alive just after the last branching before $\{x\}$, or as one could say, the probability of eternal life. This Kaplan-Meier estimated survival function just before time ∞ is as usual undefined if the 'last observation is censored'. In this context that occurs when the last node before $\{x\}$, say A, which necessarily satisfies $f_A > 0$, is such that $f_{A'} = 0$ for all $A' \subset A$. Then we can only calculate p_A itself, the probability of surviving till just before A. The estimated survival curve actually tells us all the p_A for A on the main route to $\{x\}$ and by division we can recover the π_A also.

5. Examples in survival analysis

Ordinary censored survival times $(\widetilde{T}, \Delta)$ provide a classic example of monotone data. The underlying variable of interest will be denoted by T instead of X. A coarsened observation is the interval $(\widetilde{T}, \infty)$ if $\Delta = 0$, and the singleton $\{\widetilde{T}\}$ if $\Delta = 1$. Of course this example needs a continuous sample space to be treated properly, and we will do that in the next section. However our main points on CAR and survival analysis can be made while pretending the survival times are discrete. So let us suppose that all time points, denoted t or $\widetilde{t}$, are integers between 0 and N say. In ordinary survival analysis (one-dimensional right-censored survival times) the coarsened data is represented by half intervals $\{\widetilde{t}+1, \ldots, N\}$ in the case of censored observations, and by singletons $\{t\}$ in the case of uncensored. Two outcomes are either disjoint or one contains the other: the data is monotone.

Suppose we assume nothing about the distribution of the underlying survival time, and nothing about the coarsening mechanism except that it is CAR. Since CAR is everything we are assuming nothing at all about the distribution of the observations. Computing the non-parametric maximum likelihood estimator comes down to maximising $\sum f_A \log p_A$ over underlying distributions (p_t), where (f_A) is the empirical distribution of observed sets A. If each of n observations is different, then each observation yields one set A with f_A equal to $1/n$. The maximization can be done explicitly (because of monotonicity), yielding the Kaplan-Meier estimator for (p_t). The CAR probabilities can be computed explicitly too. The CAR mechanism is actually a random (independent) censoring model: in other words, one can generate the observed coarsening by choosing a C independent of $T \sim (p_t)$, and then reporting the set $\{T\}$ if $T \leq C$, and $\{C+1, \ldots, N\}$ if $T > C$. By CAR we have that for each $\widetilde{t}$, $\Pr\{\widetilde{T} = \widetilde{t} | T = t\}$ is the same for each $t > \widetilde{t}$. These probabilities, for $\widetilde{t} = 0, 1, \ldots$, supply the claimed distribution of C. The fact 'CAR is everything' is well known, for censored survival data, as *the unidentifiability of the independent competing risks assumption*: any pair $\widetilde{T}, \Delta$ whose first element is a random time and whose second element is a zero/one variable can be written, in distribution, as $\min(T, C), 1\{T \leq C\}$, for an *independent* couple T, C (each possibly taking the value $+\infty$ with positive probability). Note that the possibility of mass at infinity is crucial to the nonidentifiability of dependent competing risks. Specifically it follows from Ritov's counterexample to 'CAR is everything' (end of section 2) that we could rule out independence of competing risks (i.e., independence of censoring and failure time) if we knew that with probability one every observation will be censored and we knew the failure time distribution has no mass at infinity.

Turning to multivariate survival analysis, the situation becomes much more complex. The problem of nonparametric estimation for multivariate censored survival times has remained open for a long time and only recently was a lot of striking progress made. The difficulty is directly connected with CAR. Let us represent one observation again as $(\widetilde{T}, \Delta)$ where $\widetilde{T}$ is now a *vector* of censored survival times and Δ a vector of censoring indicators. If the observation takes the value $(\widetilde{t}, \delta)$ then we know that the underlying survival vector T lies in the set A formed by taking the Cartesian product of singletons or half intervals defined precisely as in univariate censoring from each pair of components $(\widetilde{t}_i, \delta_i)$.

Let us assume nothing about the distribution of T. Our aim is to estimate its (multivariate) survival function, let us call it S, based on n censored observations. Ignoring again the fact that the sample space should be continuous, an assumption of CAR together with no assumption on the distribution of T means that we are not assuming anything about the distribution of the data at all. Computing the non-parametric maximum likelihood estimator of S by maximising the sum (over the n observations) of logs of probabilities of observed sets is no more than computing the reparametrisation from observed data probabilities (f_A) to underlying (p_t), (π_A) where we plug in as (f_A) the empirical distribution of the observed data: probability $1/n$ for each observed set A. Since our model is completely nonparametric there are no other reasonable estimators. In fact, as $n \to \infty$, in a situation where root n consistent estimation is possible at all—roughly speaking, when the CAR probabilities to get an exact observation are all positive–the nonparametric maximum likelihood estimator for S should be asymptotically efficient and moreover any other asymptotically regular estimator of S will be asymptotically equivalent with it. (Sometimes the NPMLE itself may fail to have good asymptotic behaviour, but still all asymptotically efficient estimators will be asymptotically equivalent to one another. Typically a simple modification of the NPMLE turns it into one of these good estimators. See van der Laan (1993, 1995) for general theory and many applications).

In the univariate case these facts are true and nowadays quite well-known. Assuming CAR (with otherwise completely unknown coarsening mechanism) is in fact equivalent to assuming Random Censorship with unknown censoring distribution. If both survival and censoring distribution are completely unknown, the model for the data is completely nonparametric. There is no essential alternative for nonparametric estimation of the survival function to the Kaplan-Meier estimator. Apparent alternatives such as nonparametric Bayes estimators, or the negative exponential of the

Nelson estimator of cumulative hazard, are asymptotically equivalent to the Kaplan-Meier estimator. Only if one assumes some knowledge of the censoring distribution (or is also able to observe censoring times of uncensored observations) do inefficient, strictly different estimators become available such as the reduced sample estimator (see Kaplan and Meier, 1958), or the reweighted (according to the censoring survival function) empirical distribution of the uncensored observations.

With multivariate censored data one can also consider the natural analogue of the univariate Random Censoring model. This says that there exists a vector of censoring times C, independent of T, and with completely unknown distribution, such that $(\widetilde{T}, \Delta)$ is formed componentwise from the components of T and C as in the univariate case. This does define a CAR mechanism. But there are many CAR mechanisms producing multivariate censored observations which cannot be represented in this way. The nonparametric multivariate random censoring model really is a model.

Here is a simple example of bivariate censoring which is coarsened at random but not random censored. Let $T = (T_1, T_2)$ be a bivariate survival time. Suppose T_1 and T_2 are actually consecutive durations between events in the lifetime of one individual, starting at time 0. Let C be an independent censoring time in the one-dimensional 'calendar time' time scale at which there are two events of two different types at times T_1 and $T_1 + T_2$. Thus $(T_1, T_1 + T_2)$ is randomly censored by (C, C). The data may still be represented, in 'two-dimensional duration time', as $(\widetilde{T}, \Delta) = (\,(\min(T_1, C), \min(T_2, \max(0, C - T_1)))\,,\,(1\{T_1 \leq C\}, 1\{T_1 + T_2 \leq C\})\,)$.

Assuming only CAR (a correct assumption in this example), and assuming nothing about S, we have no option than to compute the NPMLE of S, or an asymptotically equivalent version of it. Such estimators have been studied by van der Laan (1996). In fact, in general the NPMLE of bivariate censored survival data does not work correctly as it stands; one has to modify it slightly by an asymptotically negligeable further coarsening of the data. In the two dimensional case the possible observations are points, half-lines, and quadrants. The half-lines cause problems because we have to put probability mass in these lines, but have no information about how to do that since there will typically be no point-observations within the lines. The half-lines should be slightly expanded to thin strips, containing a few uncensored observations, and then the NPMLE makes sense and can be made asymptotically efficient. Its computation is time-consuming and its mathematical analysis very delicate. These problems are associated with the curse of dimensionality: under the completely nonparametric model we are forced to use the NPMLE or a modification thereof, and that forces us into binning or smoothing high-dimensional data in order to estimate conditional densities of some components given others. This only makes sense with huge data sets.

Suppose however we know that multivariate censored data is not just

Coarsened (in this case, censored) at Random but actually Randomly Censored. Although assuming nothing about survival or censoring distributions, we are now making identifiable assumptions; we have a real (restrictive) model. Going for full asymptotic efficiency gives us again no options: the same, delicate, NPMLE. However we can use our information on the censoring mechanism to generate a multitude of inefficient estimators. Some of these—the beautiful Kaplan-Meier generalisations of Dabrowska (1988) and of Prentice and Cai (1992)—do not lose much efficiency, are easy to calculate, and work very well already with quite small sample sizes. The other side of the coin is that there are CAR mechanisms which are not Random Censoring under which those estimators are inconsistent. They truly need the 'nuisance assumption' to work.

The likelihood factored; information about nuisance parameters should be irrelevant; yet we have made use of such information to generate alternative and practically valuable estimators. But a major justification for likelihood methods is their good large sample properties. The curse of dimensionality may prohibit their practical use, and one can be better off using asymptotically irrelevant information to construct well-behaved though (asymptotically) inefficient statistical procedures; see Robins and Ritov (1996) for an in depth study of this phenomenon.

Our example with calendar and duration time mix-up illustrates again the pitfalls. A sensible statistician would represent the data as censored times of events $(T_1, T_1 + T_2)$, knowing that from the joint distribution of these two times one can easily compute the joint distribution of (T_1, T_2). The data is actually monotone. The NPMLE can be computed explicitly. It is based on combining the marginal Kaplan-Meier estimator of the distribution of T_1 with conditional Kaplan-Meier estimators of the distribution of $T_1 + T_2$ given T_1, for each observation for which T_1 is uncensored. Actually, because we will be using one observation to estimate each conditional survival function for each observed value t_1, one can expect this estimator to make nonsense. But binning of the observations according to values of t_1 solves that problem if the sample size is large enough, and yields an asymptotically efficient estimator. One could alternatively use the Dabrowska or the Prentice-Cai estimators: they do not require any artifical grouping or smoothing of the data, but are asymptotically inefficient.

A less sensible statistician will treat the data precisely as bivariate censored observations of the durations (T_1, T_2). If he or she sticks to NPMLE (or modifications thereof) nothing will go wrong; the data is CAR and likelihood based methods do not need further information about the coarsening mechanism. However the Dabrowska or Prentice-Cai estimators will now be inconsistent since the Random Censoring model is not true.

Section 3 considered sequential mechanisms for generating CAR data. We showed that not all CAR mechanisms can be represented in this way. An important direction for future reseach is to define and study similar mechanisms in the context of multivariate censored data. This will lead to a class of censoring models intermediate between Random Censoring and Censoring (Coarsening) at Random.

6. CAR in general sample spaces

In a discrete sample space, equivalent definitions of CAR and important consequences of it were easy to obtain. In a general sample space, the various possible definitions may not be easy to formulate any more; moreover, even if they can be formulated in a natural way, they may no longer be equivalent. In that case, which definition one takes as primary should be influenced by which desirable results can be obtained from it.

In the first section we defined CAR in terms of the conditional distribution of a coarsening $\mathcal{X}$ given the coarsened variable X, CAR($\mathcal{X}|X$). We showed in section 2 that the definition was equivalent to a condition on the conditional distribution of X given $\mathcal{X}$, CAR($X|\mathcal{X}$), and to a specific factorization of the marginal distribution of $\mathcal{X}$, FACTOR($\mathcal{X}$).

The original definition—in terms of the distribution of $\mathcal{X}$ given X—respects the idea that *after* the random variable X has been generated, it is coarsened to the observation $\mathcal{X}$ by a conceptually distinct process. The condition on the conditional distribution of X given $\mathcal{X}$ describes in an appealing way that under CAR, knowing $\mathcal{X} = A$ tells us no more about X than the obvious fact $X \in A$. It is moreover useful in statistical inference—e.g., in the E step of the EM algorithm, in running the Gibbs' sampler, in calculating score functions; in all cases using exactly this conditional distribution. Finally factorization of the marginal distribution allows likelihood based inference on the distribution of X to be carried out completely ignoring the coarsening mechanism.

Before giving a general definition we must set up the measure-theoretic background to be able to talk about all these conditional distributions. If X is, say, a k-dimensional vector, our random set $\mathcal{X}$ takes values in the set of *all* subsets of $\mathbb{R}^k$. There is no natural topology on this very large space, no natural Borel σ-algebra. The space is so large that conditional distributions of X given $\mathcal{X} = A$ are not guaranteed to exist. In practice however, the range space of $\mathcal{X}$ can be taken to be quite small (e.g., rectangles only). Each possible value can typically be described by a short list of types, coefficients, coordinates or whatever. So we suppose that $\mathcal{X}$ can be described in a 1–1 way as function of some, say real vector, Y; $\mathcal{X} = \alpha(Y)$. In fact if we just suppose that X and Y take values in *Polish* spaces (separable, metric spaces) then sets of regular conditional distributions of X given Y and of Y given X both exist (see, e.g., Chang and Pollard, 1997). We also want the values of $\mathcal{X}$ to be measurable sets for X, and the set of values of Y consistent with a given value of X, to be measurable too. This is taken care of by assuming that the mapping $(x, y) \mapsto 1\{x \in \alpha(y)\}$ is jointly measurable in x and y, where the domain of the mapping is given the Borel σ-algebra corresponding to the topologies on the spaces where X and Y lie.

From now on, we assume this bare minimum of regularity without comment and also, when it is not relevant in the present context, drop the distinction between the set $\mathcal{X}$ and its description Y. Suppose then Y is a coarsening of X so, abusing our notation as announced already, $X \in Y$ with probability 1. The natural generalisation of CAR is

CAR($Y|X$). *The conditional distributions of Y given X do not depend on the values x taken by X, except for the restriction implied by Y being a coarsening of X, namely that given $X = x$, the random set Y takes values in $\{y : y \ni x\}$. More precisely, taking account of the fact that conditional distributions are not uniquely defined on sets of probability zero, we suppose that versions of $P_{Y|X=x}(\mathrm{d}y) = \Pr(Y \in \mathrm{d}y|X = x)$ can be chosen for P_X-almost all x, such that for x, x' not in the exceptional set,*

$$P_{Y|X=x}(\mathrm{d}y) = P_{Y|X=x'}(\mathrm{d}y) \text{ on } \{y : y \ni x\} \cap \{y : y \ni x'\}. \tag{7}$$

One might hope that if CAR is true according to this definition, then versions of $P_{Y|X=x}$ can be chosen making (7) hold *everywhere.* The following recipe (also used in the proof of 'car is everything' in section 2) might work: for the bad x, redefine $P_{Y|X=x}$ on $\{y \ni x\}$ to be equal, for each good x', to $P_{Y|X=x'}$ on $\{y \ni x'\}$. If probability mass still remains to be assigned, put all the remainder as an atom on the singleton $\{x\}$. One must check that this pasting together of bits of many other probability distributions does not entail using *more* than total probability 1. This problem is open. Alternatively, one could simply delete all bad x from the original sample space, and merge corresponding y (with and without bad x), arriving at a new coarsening model in which (7) holds without exception, and only differing from the original in indistinguishable events.

The precise formulation we have taken of CAR allows us to establish the following property of the Radon-Nikodym derivative (likelihood ratio) between *two* distinct coarsening mechanisms P, P' *each* satisfying CAR separately:

$$\frac{\mathrm{d}P'_{Y|X=x}}{\mathrm{d}P_{Y|X=x}}(y) \quad \text{does not depend on } x \in y, \tag{8}$$

it only depends on y itself. We call this derived property CAR(REL), REL standing for relative, in contrast with (7) which can be called CAR(ABS). In other words, we show that CAR(ABS) for P and CAR(ABS) for P' implies CAR(REL) for P' with respect to P. From CAR(REL) will follow a factorization and a result of the type CAR($X|Y$) concerning conditional distributions in the reverse direction.

Lemma. *Suppose that* (7) *holds for each of two coarsening mechanisms P, P', with the* same *marginal distribution for X, thus $P_X = P'_X$, and* (*without loss of generality*) *with the same exceptional set. Then* (8) *holds: i.e., versions of $\mathrm{d}P'_{Y|X=x}/\mathrm{d}P_{Y|X=x}(y)$, $y \ni x$, can be chosen which only depend on y, for P_X-almost all x.*

Note that we do not assume any dominatedness, so Radon-Nikodym derivatives may be zero or infinite on non-null sets. With the natural conventions $1/0 = \infty$, $1/\infty = 0$, (8) is symmetric with respect to P and P': it does not make any difference which is placed in numerator and which in denominator.

Proof. We will prove the lemma by establishing that 'not (8)' implies '(7) cannot hold for both P and P''. This is equivalent to showing that if (8) is not true while P does satisfy (7), then P' does not satisfy (7).

Now the negation of (8) implies that for each of a P_X-positive set of points x one may find at least one, and possibly many, points x' with

$$\frac{\mathrm{d}P'_{Y|X=x}}{\mathrm{d}P_{Y|X=x}}(y) \neq \frac{\mathrm{d}P'_{Y|X=x'}}{\mathrm{d}P_{Y|X=x'}}(y)$$

on a $P_{Y|X=x}$- or a $P'_{Y|X=x}$-positive set of points y in $\{y \ni x\} \cap \{y \ni x'\}$. Also all the points x' so involved must together have positive P_X probability, for otherwise we could also simply put them in the exceptional set.

For each such pair (x, x'), either we must have '$<$' or '$>$' on either a $P_{Y|X=x}$- or a $P'_{Y|X=x}$-positive set of points y. The resulting four combinations define four (possibly overlapping) sets of pairs (x, x'). At least one of these four sets must involve both a P_X-positive set of points x and a P_X-positive set of points x': otherwise (8) is saved by simply augmenting the exceptional (but still null) set. Almost without loss of generality we suppose the surviving combination has '$<$'; more obviously, we must consider the cases of a $P_{Y|X=x}$-positive or a $P'_{Y|X=x}$-positive set of y separately.

Suppose first that for each of a P_X-positive set of points x one can find one or more points x', altogether also making up a P_X-positive set, with

$$\frac{\mathrm{d}P'_{Y|X=x}}{\mathrm{d}P_{Y|X=x}}(y) < \frac{\mathrm{d}P'_{Y|X=x'}}{\mathrm{d}P_{Y|X=x'}}(y) \tag{9}$$

on a $P_{Y|X=x}$-positive set of points in $\{y \ni x\} \cap \{y \ni x'\}$. Since we have strict inequality, the left-hand side is finite *everywhere* on this set. Integrating over the set with respect to $P_{Y|X=x}$, by (7) equal to $P_{Y|X=x'}$, gives (both integrals of course finite)

$$P'_{Y|X=x} < P'_{Y|X=x'} \tag{10}$$

on some set of points y, for a collection of pairs (x, x'), each coordinate covering a P_X-non null set. Thus (7) fails for P'. Though integrating '$+\infty$' over a null set may give a positive result, this can only happen on the right-hand side, not harming the inequality. Note that the argument does not depend on the direction of the inequality in (9) and (10), since by our assumption (7) for P one can interchange the roles of x and x' if desired.

Suppose on the other hand (9) holds, now with positive $P'_{Y|X=x}$ probability, for each of the usual collection of pairs (x, x'). Again integrate with respect to $P_{Y|X=x} = P_{Y|X=x'}$ over the indicated set. The integrand on the left-hand side must have been finite *everywhere* (since we had *strict* inequality) and the result is strictly positive since it is just the $P'_{Y|X=x}$ probability of the set over which we integrate. The set must therefore also have had $P_{Y|X=x} = P_{Y|X=x'}$ positive probability and hence the result on the right-hand side is strictly larger giving us (10) again; thus again (7) fails for P'.

This last argument relied on our having '<' rather than '>' in the inequality (9). With the reverse inequality, now the right-hand side is finite everywhere. If the $P_{Y|X=x} = P_{Y|X=x'}$ measure of our set of points y is zero, the right-hand side integrates to zero, which must also be its $P'_{Y|X=x'}$ measure, and consequently less than its a priori known positive $P'_{Y|X=x}$ measure. But if we start with positive $P_{Y|X=x}$ measure, then finiteness of the integrand on the right-hand side everywhere preserves the strict inequality on integration. In either case we get the desired (reversal) of (10); and again (7) fails for P'. □

We now want to establish a factorization of the marginal distribution of Y, FACTOR(Y), and a property of the conditional distribution of X given Y, CAR($X|Y$). For the latter, we would like to derive:

$$P_{X|Y=y}(\mathrm{d}x) \overset{\text{(def.)}}{=} \Pr(X \in \mathrm{d}x | Y = y) \ = \ \Pr(X \in \mathrm{d}x | X \in y) \qquad (11)$$

However, the last expression here is *not* well-defined in general. If y is a singleton, or y is a set of positive probability for X, then we know how $\Pr(X \in \mathrm{d}x | X \in y)$ should be interpreted; in the other specific examples we also may be able to guess a reasonable definition. But the general set-up so far does not permit a unique interpretation (more on that later).

However, an important feature of (11) is that the very right hand side should be computable from the marginal distribution of X, without knowledge of the conditional distribution of Y given X (the coarsening mechanism); in fact, it should be the *same* whatever the distribution of Y given X. So just as when discussing factorization of a likelihood function, we consider a family of joint distributions of X and Y, each satisfying the probabilistic CAR assumption (7), and the property we derive is a property of the resulting *statistical model.*

To respect the distinction between the underlying variable of interest X and the coarsening mechanism leading to Y we suppose their joint distribution depends on variation independent parameters belonging separately to these two aspects, say θ and γ respectively;

$$P^{\theta,\gamma}_{X,Y}(\mathrm{d}x, \mathrm{d}y) = P^{\theta}_{X}(\mathrm{d}x) P^{\gamma}_{Y|X=x}(\mathrm{d}y). \qquad (12)$$

We assume $P^{\gamma}_{Y|X=x}(\mathrm{d}y)$ satisfies CAR, (7), for each value of γ. Moreover we assume that $P^{\theta}_{X} \ll P^{\theta_0}_{X}$; and for $P^{\theta_0}_{X}$ almost all x, $P^{\gamma}_{Y|X=x} \ll P^{\gamma_0}_{Y|X=x}$.

Theorem. *Under CAR, the likelihood for θ, γ based on observation of Y factors:*

$$\frac{\mathrm{d}P_Y^{\theta,\gamma}}{\mathrm{d}P_Y^{\theta_0,\gamma_0}}(y) = \frac{\mathrm{d}P^{\gamma}_{Y|X=x}}{\mathrm{d}P^{\gamma_0}_{Y|X=x}}(y) \cdot E_{\theta_0,\gamma_0}\left(\frac{\mathrm{d}P_X^{\theta}}{\mathrm{d}P_X^{\theta_0}}(X) \,\middle|\, Y = y \right) \tag{13}$$

(for arbitrary $x \in y$ not in the null set for x, this being possible for $P_Y^{\theta_0,\gamma_0}$-almost all y). Moreover, $P^{\theta,\gamma}_{X|Y=y}(\mathrm{d}x)$ does not depend on γ.

Proof. Consider, for $x \in y$,

$$\begin{aligned} \frac{\mathrm{d}P^{\theta,\gamma'}_{X,Y}}{\mathrm{d}P^{\theta,\gamma}_{X,Y}}(x,y) &= \frac{\mathrm{d}P_X^{\theta}}{\mathrm{d}P_X^{\theta}}(x)\, \frac{\mathrm{d}P^{\gamma'}_{Y|X=x}}{\mathrm{d}P^{\gamma}_{Y|X=x}}(y) \\ &= 1 \cdot k(y;\gamma',\gamma) \end{aligned} \tag{14}$$

for some function k, by the Lemma.

Since the right hand side of (14) does not depend on x, we must have

$$\frac{\mathrm{d}P_Y^{\theta,\gamma'}}{\mathrm{d}P_Y^{\theta,\gamma}}(y) = k(y;\gamma',\gamma).$$

Thus

$$\begin{aligned} \frac{\mathrm{d}P_Y^{\theta,\gamma}}{\mathrm{d}P_Y^{\theta_0,\gamma_0}}(y) &= \frac{\mathrm{d}P_Y^{\theta,\gamma}}{\mathrm{d}P_Y^{\theta,\gamma_0}}(y)\, \frac{\mathrm{d}P_Y^{\theta,\gamma_0}}{\mathrm{d}P_Y^{\theta_0,\gamma_0}}(y) \\ &= k(y;\gamma,\gamma_0)\, E_{\theta_0,\gamma_0}\left(\frac{dP^{\theta,\gamma_0}_{X,Y}}{dP^{\theta_0,\gamma_0}_{X,Y}}(X,Y) \,\middle|\, Y = y \right) \\ &= k(y;\gamma,\gamma_0)\, E_{\theta_0,\gamma_0}\left(\frac{dP_X^{\theta}}{dP_X^{\theta_0}}(X) \,\middle|\, Y = y \right), \end{aligned}$$

which by (14) is the claimed factorization (13).

We also have

$$\frac{\mathrm{d}P^{\theta,\gamma'}_{X,Y}}{\mathrm{d}P^{\theta,\gamma}_{X,Y}}(x,y) = \frac{\mathrm{d}P_Y^{\theta,\gamma'}}{\mathrm{d}P_Y^{\theta,\gamma}}(y)\, \frac{\mathrm{d}P^{\theta,\gamma'}_{X|Y=y}}{\mathrm{d}P^{\theta,\gamma}_{X|Y=y}}(x)$$

hence

$$k(y;\gamma',\gamma) = k(y;\gamma',\gamma)\, \frac{\mathrm{d}P^{\theta,\gamma'}_{X|Y=y}}{\mathrm{d}P^{\theta,\gamma}_{X|Y=y}}(x).$$

Since $k(y;\gamma,\gamma_0) = (\mathrm{d}P_Y^{\theta,\gamma}/\mathrm{d}P_Y^{\theta,\gamma_0})(y)$, it is positive for $P_Y^{\theta,\gamma}$ almost all y. So for such y, and all $x \in y$,

$$\frac{\mathrm{d}P^{\theta,\gamma}_{X|Y=y}}{\mathrm{d}P^{\theta,\gamma_0}_{X|Y=y}}(x) = 1$$

or $P^{\theta,\gamma}_{X|Y=y} = P^{\theta,\gamma_0}_{X|Y=y}$ for $P^{\theta,\gamma}_{Y}$ almost all y. □

The theorem not only shows there is a factorization in the likelihood for θ, γ but also fairly explicitly tells what the two factors are.

The γ-part (CAR mechanism) is the same as the likelihood for γ based on the conditional distribution of Y given $X = x$, even though X itself is not observed, only Y. In other words, inference about the coarsening mechanism can be done as if X had also been observed.

The θ-part (underlying variable of interest) can be written down without knowing the CAR parameter γ: just pick any value, say γ_0, and compute the second factor of the right hand side of (13). It does seem that we do need to know the *structure* of the coarsening mechanism. However, even if we believe in a particularly complex mechanism—so a particular family $P^{\gamma}_{Y|X=x}$—we can calculate the likelihood from the second factor of (13) using a set of conditional distributions $P^{*}_{Y|X=x}$ *outside* this family, and perhaps of much simpler structure. We only need to have $P^{\gamma}_{Y|X=x} \ll P^{*}_{Y|X=x}$, in other words: the 'reference' coarsening mechanism $P^{*}_{Y|X=x}$ used in the calculations can generate all the sets y which can occur 'in reality'.

Put another way: the likelihood for θ, under CAR, is the same as under any specific CAR mechanism which, given $X = x$, can generate the same (or more) random sets. We illustrate this and further points with a succession of examples. Examples 1 and 2 concern censored data (univariate and multivariate respectively). Example 3 is a classical paradox from the theory of conditional distributions. We concentrate on the main features of the examples, leaving more technical details to the interested reader.

Example 1. *Univariate right censoring.* Suppose $X \in [0, \infty)$ and $\mathcal{X} = \alpha(Y)$ is either a singleton $\{x\}$ or a half-line $(\widetilde{x}, \infty)$. The data Y is traditionally the pair $(\widetilde{X}, \Delta)$ where $\widetilde{X}$ takes values in $[0, \infty)$, Δ in $\{0, 1\}$, and $\Delta = 1$ if $\mathcal{X} = \{\widetilde{X}\} = \{X\}$, $\Delta = 0$ if $\mathcal{X} = (\widetilde{X}, \infty) \ni X$.

Suppose such data arose from a CAR coarsening of X. The data is monotone and we expect to be able to retrieve the coarsening mechanism and the underlying law of X fairly explicitly from the law of the data. We also expect that such data in general fits a CAR model. We will see that for practical purposes, that is true. Every outcome is either a singleton or a set of positive probability for X, so we can guess in advance the likelihood for the interest part of the model.

The CAR property is that for any $x < x'$,

$$\text{law}((\widetilde{X}, \Delta)|X = x) \;=\; \text{law}((\widetilde{X}, \Delta)|X = x')$$

restricted to values of $(\widetilde{X}, \Delta)$ such that $\mathcal{X}$ contains both x and x'. This means $\Delta = 0$ and $\widetilde{X} < x$. So CAR is equivalently the assumption $\Pr\{\widetilde{X} \in \mathrm{d}\widetilde{x}, \Delta = 0 | X = x\}$ does not depend on x, for $\widetilde{x} < x$. We may therefore define $G(t) = \Pr\{\widetilde{X} \le t, \Delta = 0 | X = x\}$ for arbitrary $x > t$ and G is a (sub)-distribution function defined at least up to the right hand end-point of the support of $\widetilde{X}$. Letting t increase up to x from below, we obtain $G(x-) = \Pr\{\Delta = 0|X = x\}$ and hence $1 - G(x-) = \Pr\{\Delta = 1|X = x\}$. We now have described the law of the coarsened data given X completely in terms of a distribution G. It could be generated on a computer by drawing a random variable C from the distribution G, independently of X, and reporting $\Delta = 1, \widetilde{X} = X$ if $C \geq X$, $\Delta = 0, \widetilde{X} = C$ if $C < X$.

Suppose X has distribution function F and survival function S. We compute the marginal distribution of the data as $\Pr\{\widetilde{X} \in \mathrm{d}\widetilde{x}, \Delta = 0\} = G(\mathrm{d}\widetilde{x})(1 - F(\widetilde{x}))$, $\Pr\{\widetilde{X} \in \mathrm{d}\widetilde{x}, \Delta = 1\} = F(\mathrm{d}\widetilde{x})(1 - G(\widetilde{x}-))$. Moreover $\Pr\{\widetilde{X} \geq \widetilde{x}\} = (1 - F(\widetilde{x}-))(1 - G(\widetilde{x}-))$ so that the hazard measure of X can be calculated from the distribution of the data as $\Lambda(\mathrm{d}x) = F(\mathrm{d}x)/(1 - F(x-)) = \Pr\{\widetilde{X} \in \mathrm{d}x, \Delta = 1\}/\Pr\{\widetilde{X} \geq x\}$, at least, for x such that $\Pr\{\widetilde{X} \geq x\} > 0$. Product-integration of the hazard yields the survival function $S = \prod(1 - \mathrm{d}\Lambda)$, for the values of x just mentioned. Once S has been recovered we can also recover G from the distribution of the data.

Even if we did not know that $\mathcal{X}$ was CAR, we could use these calculations to produce a CAR model exactly fitting the data. This would succeed unless the calculations implied a defective distribution of X, suggesting that $X = \infty$ has positive probability. For instance, if in reality the data was generated by a censoring time C identically equal to one minus the failure time X, and arbitrary large values of X can occur, then knowing that X is actually finite we could conclude from the distribution of the data that the coarsening was not CAR. □

Example 2. *Multivariate right censoring.* Suppose $X = (X_1, \ldots, X_k)$ and the possible realisations of $\mathcal{X} = \alpha(Y)$ are Cartesian products of singletons $\{x_i\}$ and half-lines (x_i, ∞). Such sets are generated by the *random-censoring* model: $C = (C_1, \ldots, C_k)$ is independent of X and for each i we observe X_i if $X_i \leq C_i$, C_i if $X_i > C_i$. Write $\widetilde{X}_i = X_i \wedge C_i$, $\Delta_i = 1\{X_i \leq C_i\}$.

It is easily checked that this specific model is CAR. For a point $(\widetilde{x}, \delta)$ and vector x let $x_\delta = (x_i : \delta_i = 1)$, $x_{\overline{\delta}} = (x_i : \delta_i = 0)$. The distribution of the data $\Pr(\widetilde{X} \in \mathrm{d}\widetilde{x}, \Delta = \delta)$ can be expressed as

$$P_{X_\delta}(\mathrm{d}\widetilde{x}_\delta) P_{X_{\overline{\delta}}|X_\delta = \widetilde{x}_\delta}((\widetilde{x}_{\overline{\delta}}, \infty_{\overline{\delta}})) \cdot P_{C_{\overline{\delta}}}(\mathrm{d}\widetilde{x}_{\overline{\delta}}) P_{C_\delta|C_{\overline{\delta}} = \widetilde{x}_{\overline{\delta}}}([\widetilde{x}_\delta, \infty_\delta)).$$

Intuitively, for a particular value of the vector of censoring indicators δ, we first write down the probability for the exactly observed variables to take values in tiny intervals around those values, and multiply by the conditional probability for the other variables to exceed the relevant values. We use the independence of X and C to further split each term in two. Finally we regroup to exhibit the CAR factorization.

Therefore under *any* CAR model producing the same sets the likelihood for parameters of the distribution of X is

$$P_{X_\delta}(\mathrm{d}\widetilde{x}_\delta) P_{X_{\overline{\delta}}|X_\delta = \widetilde{x}_\delta}((\widetilde{x}_{\overline{\delta}}, \infty_{\overline{\delta}})).$$

This example allows sets y having probability zero which are not singletons. One must check that any CAR mechanism which produces right-censored data is dominated by some random censorship model, as required to apply the theorem. □

Example 3. *Borel's paradox.* This is a famous example of two different ways in which one could observe that a point, uniformly distributed on the surface of a sphere, actually lies on a particular great circle. The two data generating mechanisms are such that the conditional distribution of the point, given it lies on the circle, is uniform in the one case, and non-uniform in the other case. The example can be 'fixed' by changing conditional distributions on an event of probability zero, but it is also a genuine example of the fact that the law of X given $X \in y$ cannot be talked about without reference to a 'reference model' generating the same random sets y as occur in the CAR model under consideration.

Suppose X is uniformly distributed on the surface of the unit sphere, and let Θ be its longitude $\in [0, 2\pi)$ and Φ its latitude $\in [-\pi/2, \pi/2]$. If $\Phi = -\pi/2$ or $+\pi/2$ (South or North pole) then Θ can be defined arbitrarily: this case has probability zero anyway.

One easily computes that Θ and Φ are independent, Θ is uniformly distributed while Φ has density $\frac{1}{2}\cos\phi$. Consequently, given $\Theta \bmod \pi = \theta$, the point X is distributed on the great circle through the poles on longitudes θ and $\theta + \pi$ with probability $\frac{1}{2}$ to be on each side of the globe, and its latitude having density $\frac{1}{2}\cos\phi$. On the other hand, given $\Phi = \phi$, the point X is uniformly distributed on the circle of constant latitude ϕ.

Taking $Y = \Theta \bmod \pi$ or $Y = \Phi$ is in both cases a coarsening at random of X. In the second case it is possible that $\Phi = 0$, conditional on which X is uniformly distributed on the equator. In the first case it is possible that $\Theta \bmod \pi = 0$ and then X is *non*-uniformly distributed on the great circle through North pole and Greenwich (England).

We took the coordinate system as being the same in the two cases. But we could have used different coordinate systems, so that the equator in one case was the same as the Greenwich meridian in the other. Then we have two CAR coarsenings which can produce exactly the same set y, but such that the distribution of X given $X \in y$ depends on which coarsening was involved.

There is no conflict with our main theorem. The theorem tells us that two coarsening at random mechanisms *which produce the same sets* have the same conditional distributions of X given the set. In the present example the sets produced by the two coarsening mechanisms are completely disjoint except for the single case of the equator which has zero probability under both mechanisms. □

If y is a singleton, then the distribution of X given $Y = y$ is degenerate at this point, so (13) gives the usual 'complete data' likelihood for θ. Suppose on the other hand sets y can occur with positive P_X probability. Construct a reference CAR coarsening model by choosing one of the sets at random, independently of X, and observing whether or not X lies in the set. For this model, $P_{X|Y=y} = P_{X|X \in y}$ and (13) gives the 'right' answer $P_X^\theta(y)$.

Our theorem satisfactorily shows that our general notion of CAR($Y|X$) has the required consequences CAR($X|Y$) and FACTOR(Y). It is an open question as to whether (and how) these can be made actually *equivalent* to CAR($Y|X$). A special case in which that can be shown, generalizing the discrete case with which we started, is when the distributions of Y given $X = x$ are dominated (over x) by a single, σ-finite, measure. One may check that it then conversely holds that the distributions of X given Y are also dominated; in fact, DOM($Y|X$) $\iff$ DOM($X|Y$), and under this condition, CAR($Y|X$) $\iff$ CAR($X|Y$). However this special case hardly has interesting applications beyond the discrete case.

Our definition of CAR($Y|X$) was an absolute or probabilistic definition for each coarsening mechanism in the model separately. In the lemma we derived a relative or statistical consequence concerning the likelihood ratios between different coarsening mechanisms, and that was all we used in our theorem. The theorem did not show *what* the factorisation was, nor *what* is the, for each θ fixed in γ, distribution of X given $Y = y$. In order to formulate necessary and sufficient conditions for CAR we must further specify these ingredients.

As we saw, only CAR(REL) and not CAR(ABS) was needed to prove the theorem. It is easy to show that if a statistical model satisfies CAR(REL) (with respect to γ, for each θ), and one point in it satisfies CAR(ABS) (for each θ), then so do the rest. One may say: if CAR(REL) holds, then CAR(ABS) either holds at all points in the model or none. It is possible that interesting statistical models can be found which satisfy CAR(REL) without CAR(ABS) holding anywhere. In fact CAR(REL) is simply a classical sufficency condition: assuming domination, it is the factorization criterion, in the model when *both* X and Y are observed, for Y to be sufficient for γ for each fixed θ. Consequently (and equivalently) we have sufficiency according to the definition in terms of conditional distributions: for each θ, the distribution of X given $Y = y$ does not depend on γ. This is just the second part of our theorem. (See Chang and Pollard (1997) for a modern proof of this equivalence). The first part of our theorem is the classical result that the likelihood function based on the sufficient statistic is the same as the likelihood function based on the original data.

7. CAR according to Jacobsen and Keiding

Jacobsen and Keiding (1995) have a somewhat different definition of CAR in general sample spaces. Their definition assumes much more structure on X and Y, which allows rather concrete representations of the various conditional distributions of interest, without making our regularity conditions. Their definition of CAR becomes less transparent since it is stated in terms of a density with respect to a particular reference experiment. Their conclusions are partly more strong, since more explicit, but on the other hand do not reveal so explicitly as ours the practical interpretation and calculation of the factors in the likelihood. We summarize their results below, first describing the main features of their set-up. Further analysis of the difference between our and their approach is given by Nielsen (1996).

Suppose besides X there is a non-observable random variable G generating the coarsening of X. Thus for each value g of G there is a partition of the sample space, and we observe the element Y of the partition in which X lies. G may be dependent on X. The coarsening mechanism is described by the distribution of G given $X = x$, for each x, and the partition generated by $G = g$, for each g. Write $Y = \Phi(X, G)$; because of the partitioning structure we have, for every $y = \Phi(x, g)$ for some x, g, that $x \in y$ and, if also $x' \in y$, then $\Phi(x', g) = y$ too. In fact $\{(x, g) : \Phi(x, g) = y\} = y \times \{g : \Phi(x, g) = y \text{ for some } x\}$.

Jacobsen and Keiding assume that there exists a reference model, which we shall call P^*, under which: X and G are *independent*, X with distribution μ, G with distribution ν. (They give P^* the name ρ.) Then they consider families of distributions $P^{\theta,\gamma}$ such that X has marginal distribution $P^\theta \ll \mu$, and, for each x, $G|X = x$ has distribution $P^\gamma_{G|X=x} \ll \nu$. Call the corresponding densities $f(x; \theta)$ and $h(g; x, \gamma)$. Define

$$k(y; x, \gamma) = E^*(h(G; x, \gamma)|Y(x, G) = y);$$

i.e., write the conditional expectation $E^*(h(G; x, \gamma)|Y(x, G))$ as a function of $Y(x, G)$. They note that k is the density of the conditional $P^{\theta,\gamma}$ distribution of Y given $X = x$, with respect to its distribution under P^*. The P^*-independence of X and G plays a crucial role in these calculations. Finally they define CAR as: $k(y; x, \gamma)$ does not depend on $x \in y$, for each γ. They prove that under CAR, the likelihood based on Y, for θ and γ, factors, and the θ part is $E^*(f(X; \theta)|Y = y)$. When y is a singleton, or a set of positive μ-probability, the likelihood becomes as one would hope $f(X; \theta)$ or $P^\theta_X(y)$ respectively.

It can be checked that if both Jacobsen and Keiding's structure, and our regularity conditions, are present, then CAR according to Jacobsen and Keiding implies CAR according to us, thereby giving more interpretability to their condition and giving further conclusions. In the terminology of the last section, their 'reference model' satistifies CAR(ABS), while the rest of the model satisfies CAR(REL). We do not know if their set-up is essentially more restrictive than ours. Starting from a given law of (X, Y) as we do, one might take for G a collection of uniform random variables used in a computer simulation of the law of Y given X, and let Φ be the deterministic part of the algorithm. However then different CAR mechanisms will correspond to different functions Φ, not to different distributions of G.

Jacobsen and Keiding's definition of CAR works for a given statistical model and is relative to a specific 'reference model'. In fact, the reference model P^* itself satisfies CAR according to their own definitions (take the densities f and h identically equal to 1, then also k is identically equal to 1). Their reference model is also CAR according to our, absolute (non-relative) definition.

One could say that Jacobsen and Keiding define CAR precisely through *assuming* the factorization holds of likelihoods, *with respect* to a specific reference model, which itself in our broader sense is CAR. Their results therefore do give a context in which it is true that CAR (absolutely—in terms of $P_{Y|X=x}$) and the factorization are *equivalent*, certainly a very nice theorem to have. This works by having a reference model through which it is defined what the θ part of the likelihood factorization should be, and how the distribution of X given $X \in y$ should be defined.

8. More data

In many applications the coarsening mechanism depends on an underlying random variable G which may be observed, or partially observed, along with the coarsening of X. For instance in survival analysis, potential censoring times are sometimes known even for uncensored observations.

Let us represent the data by some random variable Y. Suppose there is a function α from the sample space for Y to the non-empty subsets of E such that $\mathcal{X} = \alpha(Y)$ is a coarsening of X: thus $X \in \alpha(Y)$ with probability one. Typically we will have $Y = \Phi(X, G)$ for some known function Φ, and $\alpha(y) = \{x : \exists g \text{ with } \Phi(x, g) = y\}$.

The only difference with the set-up of Section 6 is that we do not suppose the function α is one-to-one: two different points y, y' could give the same functional information about X, namely $X \in \alpha(y) = \alpha(y')$. So Y is not just a parametrisation of $\mathcal{X}$.

The coarsening mechanism is fixed by describing the conditional laws of Y given $X = x$ together with the function α. We will define Coarsening at Random in terms of these two ingredients. If actually $Y = \Phi(X, G)$ for some Φ, G, and corresponding α, the CAR property can be rephrased as a more clumsy looking condition on the laws of G given $X = x$ and the function Φ.

In Section 6, Y was thought of as a convenient encoding of the observed coarsening of X, $\mathcal{X} = \alpha(Y)$. But the implicit fact that α is one-to-one was not used anywhere at all! Therefore we may *maintain the definition* (7) of CAR($Y|X$) in our new context, and *all the results of Section 6 remain valid.* Everywhere, the statement '$x \in y$' should just be read as shorthand for '$x \in \alpha(y)$'.

The discrete-case results of Section 1 can also be copied. Let X, Y be discrete random variables and $\mathcal{X} = \alpha(Y)$ a coarsening of X. Our definition of coarsening at random becomes:

$$\text{CAR}(Y|X): \quad \Pr\{Y = y \mid X = x\} \quad \text{does not depend on } x \in \alpha(y).$$

Consequently $\Pr\{Y = y \mid X = x\} = \Pr\{Y = y \mid X \in \alpha(y)\}$ for $x \in \alpha(y)$;

of course $\Pr\{Y = y \mid X = x\} = 0$ for $x \notin \alpha(y)$.

We compute the marginal distribution of Y: it is

$$\begin{aligned} \Pr\{Y = y\} &= \sum_{x \in \alpha(y)} \Pr\{Y = y \mid X = x\} \Pr\{X = x\} \\ &= \Pr\{Y = y \mid X = x\} \Pr\{X \in \alpha(y)\} \end{aligned}$$

for arbitrary $x \in \alpha(y)$; this is the factorisation property FACTOR(Y).

Finally, for $x \in \alpha(y)$,

$$\begin{aligned} \Pr\{X = x \mid Y = y\} &= \frac{\Pr\{X = x\} \Pr\{Y = y \mid X = x\}}{\Pr\{X \in \alpha(y)\} \Pr\{Y = y \mid X = x\}} \\ &= \Pr\{X = x \mid X \in \alpha(y)\}, \end{aligned}$$

so CAR($X|Y$) holds.

Sometimes one has in mind a model $Y = \Phi(X, G)$ involving a specific grouping or censoring variable G. Take

$$\alpha(y) = \{x : \exists g, \Phi(x, g) = y\}.$$

CAR($Y|X$) (in the discrete case) is immediately rewritten as

$$\Pr\{\Phi(x, G) = y \mid X = x\} \quad \text{does not depend on } x \in \alpha(y).$$

This is hard to rephrase in a more attractive way (without reverting to Y) and hard as it stands to generalise to arbitrary sample spaces. A little progress can be made in an important special case when Φ is *Cartesian* by which we mean that $\Phi^{-1}(y) = \{(x, g) : \Phi(x, g) = y\}$ is a Cartesian product, say $\Phi_X^{-1}(y) \times \Phi_G^{-1}(y)$. The data Y is equivalent to simultaneous coarsenings $\mathcal{X}$ and $\mathcal{G}$ of X and G. In particular, $\mathcal{X} = \alpha(y) = \Phi_X^{-1}(y)$. Then

$$\Pr\{\Phi(x, G) = y \mid X = x\} = \Pr\{G \in \Phi_G^{-1}(y) \mid X = x\}.$$

We see from this that under Cartesian coarsening, CAR is *implied* by the assumption that $\Pr\{G = g \mid X = x\}$ does not depend on $x \in \Phi_X^{-1}(y)$, for each $g \in \Phi_G^{-1}(y)$. This more pleasant looking sufficient condition, which one could call CAR($G|X$), could be reformulated in general sample spaces analogously to (7). If the coarsening is both Cartesian and CAR and moreover G is observed exactly—i.e., $\Phi_G^{-1}(y)$ is always a singleton—then the condition is also necessary.

On the other hand, it is easy to construct toy examples where CAR($Y|X$) is true but the coarsening is not Cartesian or CAR($G|X$) does not hold.

Under Cartesian coarsening, one can pretend that G is actually completely observed, as follows. The data Y is equivalent to coarsenings $\mathcal{X}, \mathcal{G}$ of X and G. Define G^* to be a function of Y which picks out in a deterministic, measurable, way an arbitrary element g^* of $\mathcal{G}$. Now $Y^* = (\mathcal{X}, G^*)$ is not only a function of Y, but actually equivalent to the data Y since one can recompute $Y = \Phi(X, G) = \Phi(x, G^*)$ by picking any x in $\mathcal{X}$. This trick is sometimes useful in calculations, because CAR($G^*|X$) does now hold, even if CAR($G|X$) didn't; see Robins and Rotnitzky (1992), Robins (1996a).

Nielsen (1996), following Jacobsen and Keiding (1995) has interesting further results in this area.

We conclude this section with a discussion of the quizmaster problem, mentioned at the end of section 3. Let us avoid modelling the psychology of the quizmaster and the player (us) by an initial randomization by which the three doors get secretly relabelled by us with the numbers $1, 2, 3$ uniformly and independently of the location of the car behind one of the doors. We choose one of the numbers at random and indicate that door. The quizmaster opens one of the other doors and shows that there is no car hidden behind it. (It is an important model assumption that he *always* does this; of course, he always *can*). He asks us to reconsider our choice.

Because of our initial randomisation, the location of the car X is uniformly distributed over $\{1, 2, 3\}$. Our choice of door is also independently and uniformly distributed. Given our choice of door and the location of the car, the quizmaster either has chosen one of the two possible doors with equal probability if we are right, and has made the only possible choice if we are wrong. Represent our data Y with the ordered pair whose first component is the number of the door we chose and whose second is the number of the remaining closed door. The coarsening $\alpha(Y)$ of X is the set whose unordered elements are the components of the ordered pair Y. We know that the car lies behind one of these two doors. It is easy to calculate that conditional on the data, the car is twice as likely to lie behind the other closed door than the door we chose. The law of X given $Y = y$ is therefore different from the law of X given $X \in \alpha(y)$, so this data cannot have been coarsened at random. If however we reduce the data to the coarsened value of X—for example $(1, 2)$ and $(2, 1)$ are reduced to $\{1, 2\}$—then we do have CAR, as is also easy to verify!

This perhaps explains why most people on first being told this story, say that the information given by the quizmaster is no use and that one might just as well keep to the same door. The coarsening of X on its own was coarsening at random, so not informative of the value of X. But the complete data is is not CAR, and tells us something nontrivial about X.

9. Locally, CAR is everything

We saw that 'car is everything' breaks down, strictly speaking, in general sample spaces. In this section we show that something very close is true, in a sense connected to asymptotic efficiency theory. Starting from a given CAR model, and then allowing the CAR mechanism to vary arbitrarily and the underlying distribution to vary arbitrarily, together allow the distribution of the data to vary away from the starting point in any direction whatsover, just as if we had assumed nothing at all from the start. This has important consequences in asymptotic optimality theory. If we assume just that the coarsening is CAR, and assume nothing about the distribution of the variable of interest, all regular estimators of the underlying distribution of interest are asymptotically equivalent. Esentially there is only one estimator available: the NPMLE. Other (more technical) consequences are given below.

Suppose we have one observation of a random vector X and assume nothing whatsoever about its distribution. If X actually has distribution P^0 then for every bounded function h of X, such that $E^0h(X)=0$, $P^{\theta,h}$ defined by $P^{\theta,h}(\mathrm{d}x)=(1+\theta h(x))P^0(\mathrm{d}x)$ is for small enough $|\theta|$ also a probability distribution of X. In fact $(P^{\theta,h}:|\theta|\leq\varepsilon)$ is a one-dimensional parametric submodel for X with score function, at $\theta=0$, equal to $h(X)$. The *tangent space* is by definition the closure (in the $\mathcal{L}^2(P^0)$ sense) of the linear span of all score-functions, at P^0, of regular one-dimensional parametric submodels for the distribution of X, passing through P^0. We see that if we assume nothing about X, then the tangent-space at P_0 is $\mathcal{L}^2_0(P^0)$, the space of all square-integrable, mean-zero functions of $X\sim P^0$. We also write $\mathcal{L}^2_0(X)$ for the same space when the distribution of X under which we work is clear from the context.

In fact, any element of $\mathcal{L}^2_0(X)$, not just the bounded ones, are score-functions of submodels: define alternatively

$$P^{\theta,h}(\mathrm{d}x)=\Bigl(1+\tfrac{1}{2}\theta h(x)\Bigr)^2P^0(\mathrm{d}x)\,/\,\Bigl(1+\tfrac{1}{4}\theta^2E^0(h(X)^2)\Bigr).$$

The tangent-space plays a central role in the theory of semi-parametric models. In particular, the asymptotic Cramèr-Rao lower bound for estimation of functionals of the distribution of X based on i.i.d. replicates is calculated via a calculation of the tangent space. The larger the tangent space, the harder is estimation and the larger is the Cramèr-Rao bound. As we have just seen, assuming nothing about the distribution of X leads to the largest possible tangent space: $\mathcal{L}^2_0(X)$.

Suppose now Y is a coarsening of X satisfying the CAR assumption. Our model for Y is built up of a model for P_X, the distribution of X, and for $P_{Y|X=x}$, the family of distributions of Y given X. We show here that: *if nothing is assumed about P_X, and nothing is assumed about $P_{Y|X=x}$ beyond the CAR assumption, then the tangent space at a particular point P_Y in the resulting model for the distribution of Y is $\mathcal{L}^{\,2}_0(Y)$*. Locally, we are not assuming anything about the distribution of Y.

Let P_X, $P_{Y|X=x}$ be given, the latter satisfying CAR, and define for given functions $h(x)$ and $k(y;x)$

$$\begin{array}{rcl} P^\theta_X(\mathrm{d}x) & = & \bigl(1+\theta h(x)\bigr)P_X(\mathrm{d}x)\\ P^\gamma_{Y|X=x}(\mathrm{d}y) & = & \bigl(1+\gamma k(y;x)\bigr)P_{Y|X=x}(\mathrm{d}y). \end{array}$$

If h is bounded and $E(h(X))=0$ this defines a one-dimensional parametric submodel for the distribution of X with parameter θ (sufficiently close to zero). Similarly if k is bounded and $E\big(k(Y;x) \mid X=x\big)=0$ we have a model for the distribution of Y given X with parameter γ. In order that the CAR assumption holds under $P^{\theta,\gamma}$ we require that $k(y;x)$, $y \ni x$, does not depend on x; so in fact $k(y;x)=k(y)$.

If we had observed X and Y the score functions (at $\theta=0$, $\gamma=0$) for θ and γ would have been $h(X)$ and $k(Y)$ respectively. When we observe only Y, the score functions are transformed to their conditional expectations given $Y: E(h(X)|Y)$ and $k(Y)$ respectively; see Gill(1989, §3, Ex. 2) for a heuristic derivation of this result and Bickel, Klaassen, Ritov and Wellner (1993, Prop. A5.5) for a rigorous one. (We refer in the sequel to this work as BKRW.)

Write E^X, E^Y for conditional expectation operators given X and Y respectively, considered as mappings on the following Hilbert spaces:

$$\begin{aligned} E^X &: \mathcal{L}_0^2(Y) \to \mathcal{L}_0^2(X) \\ E^Y &: \mathcal{L}_0^2(X) \to \mathcal{L}_0^2(Y). \end{aligned}$$

Write $\|\cdot\|_X$, $\langle\cdot,\cdot\rangle_X$ etc. for the corresponding norms and inner products. E^X and E^Y are one-another's adjoint: writing $A=E^Y$ and defining $A^\top$ by $\langle g, Ah\rangle_Y = \langle A^\top g, h\rangle_X$ for all $g \in \mathcal{L}_0^2(Y)$ and $h \in \mathcal{L}_0^2(X)$ we have:

$$\begin{array}{rcccccc} \langle g, Ah\rangle_Y &=& \langle g, E^Y h\rangle_Y &=& E(g(Y)E(h(X)|Y)) &=& E(g(Y)h(X)) \\ &&&=& E(E(g(Y)|X)h(X)) &=& \langle E^X g, h\rangle_X \end{array}$$

proving that if $A=E^Y$, then $A^\top = E^X$.

We have shown that for each bounded function h of X, with mean zero, $E^Y h$ is a score function of a one-dimensional parametric submodel for the distribution of X. Similarly, for each bounded function $k(Y)$, with conditional mean given X zero, k is a score function of a one-dimensional parametric CAR submodel for the distribution of Y given X. Since taking $\theta \equiv \gamma$ gives a score function equal to the sum of the scores for θ and γ separately, we find by taking closures that our tangent space based on observation of Y contains $\overline{\mathcal{R}(A)} + \mathcal{N}(A^\top)$ where $\mathcal{R}$ and $\mathcal{N}$ denote range and null-space respectively. However, it is a well-known (and easily proved) fact from the theory of Hilbert spaces that for any bounded linear operator A from one Hilbert space H to another H', $\overline{\mathcal{R}(A)} + \mathcal{N}(A^\top)$ is a decomposition of the range space H' into two orthogonal components: for suppose g is orthogonal to $\overline{\mathcal{R}(A)}$. Then $\langle g, Ah\rangle_{H'}=0$ for all h, thus $\langle A^\top g, h\rangle_H = 0$ for all h, thus $A^\top g = 0$ or $g \in \mathcal{N}(A^\top)$. So the tangent space is $\overline{\mathcal{R}(A)} + \mathcal{N}(A^\top) = \mathcal{L}_0^2(Y)$, the largest possible tangent space, corresponding globally to making no assumptions whatever on the distribution of Y.

We next prove, under an assumption concerning the probability to get a *complete observation, $y=\{x\}$, that the distribution of X is locally identified* under the completely nonparametric CAR model described above. The result was already given in van der Laan (1993, Lemma 3.3) and Robins and Rotnitzky (1992).

Under CAR, we obtained the factorization (13):

$$\frac{\mathrm{d}P_Y^{\theta,\gamma}}{\mathrm{d}P_Y^{\theta_0,\gamma_0}}(y) = \frac{\mathrm{d}P_{Y|X=x}^{\gamma}}{\mathrm{d}P_{Y|X=x}^{\gamma_0}}(y)\cdot E_{\theta_0,\gamma_0}\left(\frac{\mathrm{d}P_X^{\theta}}{\mathrm{d}P_X^{\theta_0}}(x)\,\middle|\, Y=y\right),$$

where the first factor depends only on y (not on $x \in y$). Fixing $\theta = \theta_0$ we see that the space of score functions of one-dimensional parametric submodels for the coarsening mechanism $P_{Y|X=x}$ not only contains but is actually exactly equal to $\mathcal{N}(A^\top)$, the space of zero-mean, square integrable functions of Y with conditional mean given X identically zero. Similarly, the space of score-functions of one-dimensional parametric submodels for the distribution of interest P_X is exactly equal to $\overline{\mathcal{R}(A)}$. Since $\overline{\mathcal{R}(A)}$ and $\mathcal{N}(A^\top)$ are orthogonal, we find from the theory of semiparametric models that the asymptotic Cramèr-Rao lower bound for estimation of functionals of P_X is the same when $P_{Y|X=x}$ is known and fixed, as when it is completely unknown (subject in both cases to CAR). Suppose we want to estimate $\kappa(P_X) = \int \kappa(x)P_X(\mathrm{d}x)$ for some bounded function $\kappa(x)$, e.g. $\kappa(x) = 1_A(x)$ corresponding to estimation of $\Pr(X \in A)$ for a given set A. Define $\widetilde{\kappa} = \kappa - E_X(\kappa)$. Then by BKRW or by van der Vaart (1991), we have: if $I = A^\top A$ has (at $\widetilde{\kappa}$) an inverse I^{-1}, then the asymptotic information bound for estimation of κ is

$$\|A(A^\top A)^{-}\widetilde{\kappa}\|^2 < \infty.$$

In fact $g(X) = (A^\top A)^{-}\widetilde{\kappa}(X)$ generates a 'hardest' one-dimensional submodel for estimating κ at P_X (maximizes the Cramèr-Rao bound over all parametric submodels). A slightly weaker condition for a finite asymptotic information bound is just that $\widetilde{\kappa}$ lies in the range of $A^\top$; this is obviously implied by $\widetilde{\kappa} = A^\top A g$ for some g. We will later argue that a finite information bound means in some sense local identifiability. But first the result:

Theorem. *(van der Laan, 1993). Suppose for each x, $P_{Y|X=x}(\{x\}) \geq \delta > 0$; i.e., the conditional probability of a complete observation is bounded away from zero. Then $I = A^\top A : \mathcal{L}_0^2(X) \to \mathcal{L}_0^2(X)$ is onto and has a bounded inverse; in fact $\|I^{-1}h\| \leq \delta^{-1/2}\|h\|$ for all h. Consequently $\|A(A^\top A)^{-1}\widetilde{\kappa}\|^2 \leq \delta^{-1}\|\widetilde{\kappa}\|$ or: the information bound for estimating κ based on Y is not more than $1/\delta$ times its bound based on observing X.*

Proof. The argument is based on van der Laan (1993; Lemma 2.2 and Lemma 3.3) with a minor supplement. To start with (cf. Lemma 3.3, van der Laan 1993), consider

$$\begin{aligned}
\|Ah\|^2 &= E\Big(E\big(h(X)\mid Y\big)^2\Big)\\
&\geq E\Big(h(X)^2\,1\{Y=\{X\}\}\Big)\\
&= E\Big(h(X)^2\,\Pr\big(Y=\{X\}\mid X\big)\Big) \quad\geq\quad \delta\|h\|^2.
\end{aligned}$$

So $0 < \delta \leq \|A\|^2 \leq 1$ and, if $\|h\| = 1$,

$$\begin{array}{rcll} \|A^\top Ah\| & = & \|A^\top Ah\|\,\|h\| & \\ & \geq & \langle A^\top Ah, h\rangle & \text{(by Cauchy-Schwartz)} \\ & = & \|Ah\|^2 \ \geq \ \delta. & \end{array}$$

Thus for any h, $\|A^\top Ah\| \geq \delta\|h\|$. This shows in particular that $A^\top A$ is 1–1 since, if $A^\top Ah = A^\top Ah'$, then $A^\top A(h - h') = 0$ and $\delta\|h - h'\| \leq \|A^\top A(h-h')\| = 0$, implying $h = h'$. Now (following van der Laan's Lemma 2.2) let us consider the operator $1 - A^\top A$, where 1 is the identity. This operator is self-adjoint. It is also bounded, since A, $A^\top$ and 1 are bounded. Therefore from Hilbert space theory (see, e.g., Kress, 1989, Theorem 15.9),

$$\begin{array}{rcl} \|1 - A^\top A\| & = & \sup\limits_{h:\|h\|=1} |\langle h, (1 - A^\top A)h\rangle| \\ & = & \sup\limits_{h:\|h\|=1} |1 - \|Ah\|^2| \ \leq \ 1 - \delta \ < \ 1. \end{array}$$

Consequently we have that $(A^\top A)^{-1} = (1 - (1 - A^\top A))^{-1}$ exists and is in fact given by $\sum_{n=0}^{\infty}(1 - A^\top A)^n$. The squared norm of the inverse is bounded by $(\sum_{n=0}^{\infty}(1 - \delta)^n) = \delta^{-1}$. □

Remark. If we know $P_{Y|X=x}$ and moreover $P_{Y|X=x}(\{x\}) \geq \delta > 0$ for all x, one could estimate $\kappa(P_X) = \int \kappa dP_X$ based on n observations of Y by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1\{Y_i = \{X_i\}\}}{P_{Y|X=x}(\{x\})|_{x=X_i}}\kappa(X_i).$$

This estimator is unbiased and its variance is easily seen not to exceed $\|\widetilde{\kappa}\|^2/(n\delta)$. This shows directly that the information bound for estimation of κ is finite and not more than $(1/\delta)$ times the bound when X is observed, when $P_{Y|X=x}$ is known. By orthogonality the same bound applies even when $P_{Y|X=x}$ is unknown.

Now we discuss the interpretation of this result as a kind of local identifiability. Suppose we have n i.i.d. observations Y_i of Y and consider any parametric model $P_Y^{\theta,\gamma}$ constructed from P_X^θ and $P_{Y|X=x}^\gamma$. Consider the local models $\theta = \theta_0 + n^{-\frac{1}{2}}\eta$, $\gamma = \gamma_0 + n^{-\frac{1}{2}}\psi$. Define the optimal influence curve $\mathrm{IC}_{\mathrm{opt}} = A(A^\top A)^{-1}\widetilde{\kappa}$, working at the point $\theta = \theta_0$, $\gamma = \gamma_0$. Then

$$\widehat{\kappa} = \kappa(P_X^{\theta_0}) + \frac{1}{n}\sum_{i=1}^{n}\mathrm{IC}_{\mathrm{opt}}(Y_i)$$

is an estimator of $\kappa(P_X^\theta)$ based on $Y_1, \ldots, Y_n$ such that $n^{1/2}(\widehat{\kappa} - \kappa(P_X^{\theta_0})) \xrightarrow{\mathcal{D}} \mathcal{N}(\mu, \sigma^2)$ as $n \to \infty$, under $P_Y^{\theta_0 + n^{-1/2}\eta, \gamma_0 + n^{-1/2}\psi}$, where the limiting mean

$\mu = \lim_{n\to\infty} n^{1/2}(\kappa(P_X^{\theta_0+n^{-1/2}\eta}) - \kappa(P_X^{\theta_0}))$, and $\sigma^2 < \infty$. Thus asymptotically we can recover $\kappa(P_X^{\theta_0+n^{-1/2}\eta})$ from $(P_Y^{\theta_0+n^{-1/2}\eta,\gamma_0+n^{-1/2}\psi})^n$.

This holds separately for every parametric model passing through the same given point P^0, i.e., $P^0_X = P_X^{\theta_0}$, $P^0_{Y|X=x} = P^{\gamma_0}_{Y|X=x}$. Since even under CAR the tangent space at P^0 is everything, any P_Y close to P^0_Y lies to a close approximation on one of these submodels. Thus in a local asymptotic sense, for P_Y close to a given model P^0_Y determined by $P^0_X, P^0_{Y|X=x}$, one can recover P_X from P_Y.

10. Global identifiability of CAR

Suppose the triple X, Y, α is such that Y, α is a coarsening at random of X. The question we study here is: given the distribution of the data Y, and the coarsening $\mathcal{X} = \alpha(Y)$, are the marginal distribution of X and the conditionals of Y given $X = x$ uniquely determined? In other words, if a factorization of the distribution of Y exists, is it unique?

In Section 2 we saw that in the discrete case the factorisation $f_A = p_A\pi_A$ (which was always possible) was uniquely determined for A with $f_A > 0$. The (p_A) and (π_A) of the factorisation might not be hereby completely determined for A with $f_A = 0$. There might be some free choice between having $p_A = 0$ or $\pi_A = 0$, and consequently some free choice in the value given to the non-zero member of the pair.

In general sample spaces there is a similar non-uniqueness (if a factorisation exists at all). Let the function α and the marginal law of Y be fixed. Let P and P' denote two CAR models, such that the possible P-null exceptions x for the CAR property of $P_{Y|X=x}$ also form a P'-null set and vice-versa. We assume $P_Y = P'_Y$.

Define $Q_X = \frac{1}{2}(\mathrm{P}_X + P'_X)$ and $Q_{Y|X=x} = \frac{1}{2}(P_{Y|X=x} + P'_{Y|X=x})$. Then Q is also CAR, and P and P' are dominated by Q. Let $P^{\theta,\gamma}$ be defined by

$$\begin{aligned} P^\theta_X &= (1-\theta)P_X + \theta P'_X, \\ P^\gamma_{Y|X=x} &= (1-\gamma)P_{Y|X=x} + \gamma P'_{Y|X=x}. \end{aligned}$$

Thus $Q = P^{0.5,0.5}$. By the theorem of Section 7,

$$\frac{\mathrm{d}P_Y^{\theta,\gamma}}{\mathrm{d}Q_Y}(y) = E_Q\Big(\frac{\mathrm{d}P^\theta_X}{\mathrm{d}Q_X}(X)\,\Big|\, Y = y\Big)\cdot\frac{\mathrm{d}P^\gamma_{Y|X=x}}{\mathrm{d}Q_{Y|X=x}}(y), \quad x \in \alpha(y).$$

Thus

$$\begin{aligned} E_P\ \log\frac{\mathrm{d}P_Y^{\theta,\gamma}}{\mathrm{d}Q_Y}\quad (Y) = {} & E_P\log\Bigg(E_Q\Big(\tfrac{\mathrm{d}P_X}{\mathrm{d}Q_X}(X)\,\Big|\,Y\Big) \\ & +\theta\Big(E_Q\Big(\tfrac{\mathrm{d}P'_X}{\mathrm{d}Q_X}(X) - \tfrac{\mathrm{d}P_X}{\mathrm{d}Q_X}(X)\,\Big|\,Y\Big)\Big)\Bigg) \\ & +\ E_P\log\Bigg(\tfrac{\mathrm{d}P_{Y|X}}{\mathrm{d}Q_{Y|X}}(Y) + \gamma\Big(\tfrac{\mathrm{d}P'_{Y|X}}{\mathrm{d}Q_{Y|X}}(Y) - \tfrac{\mathrm{d}P_{Y|X}}{\mathrm{d}Q_{Y|X}}(Y)\Big)\Bigg) \end{aligned}$$

where $(\mathrm{d}P_{Y|X}/\mathrm{d}Q_{Y|X})(Y)$ is defined on $Y=y$ as $(\mathrm{d}P_{Y|X=x}/\mathrm{d}Q_{Y|X=x})(y)$ for any $x\in\alpha(y)$.

Now the above function of θ and γ is concave in both arguments, and maximal both at $\theta=0$, $\gamma=0$, and at $\theta=1$, $\gamma=1$. Therefore it must be constant in θ and γ, or:

$$E_Q\Big(\frac{\mathrm{d}P'_X}{\mathrm{d}Q_X}(X)\;\Big|\;Y=y\Big) \;=\; E_Q\Big(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}(X)\;\Big|\;Y=y\Big)$$

for P almost all y,

$$\frac{\mathrm{d}P'_{Y|X=x}}{\mathrm{d}Q_{Y|X=x}}(y) \;=\; \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}Q_{Y|X=x}}(y)$$

for P almost all y. Now the particular choice of Q dominating P was not important so we have that all CAR models reproducing P_Y have the same decomposition

$$\frac{\mathrm{d}P_Y}{\mathrm{d}Q_Y}(y) \;=\; E_Q\Big(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}(X)\;\Big|\;Y=y\Big).\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}Q_{Y|X=x}}(y),\qquad x\in\alpha(y)$$

provided the same exceptional points x are involved; in particular, if actually CAR$(Y|X)$ holds without any exceptional points.

11. Open questions

We have shown that, in fairly general sample spaces, a certain definition of CAR in terms of $P_{Y|X=x}$ has desired consequences for $P_{X|Y=y}$ and for factorization of P_Y. In discrete sample spaces, these three properties are actually equivalent. The equivalence does not hold in general. But one can look for supplementary conditions to make this hold.

Part of this problem is the wish to be able to have from CAR (or even equivalent to CAR): $P_{X|Y=y}=P_{X|X\in y}$. In a general set-up however, there is not a unique way to interpret $P_{X|X\in y}$. Perhaps one should restrict attention to cases where Y has further special structure. The following covers all specific examples of which we are aware: it has features both from missing observations in a multivariate vector and from grouped (including censored) observations. Suppose observation of $Y=y$ is equivalent to observation of a discrete 'type' K, and, when $K=k$, observation that $\alpha_k(X)=a_k$ and $\beta_k(X)\in B_k$ for certain measurable functions α_k and β_k where furthermore $\Pr(\beta_k(X)\in B_k|\alpha_k(X)=a_k)>0$ for all possible values a_k and sets B_k. Conditional on $Y=y$, we would now want $P_{X|Y=y}$ to coincide with the conditional distribution of X given $\alpha_k(X)=a_k$ and $\beta_k(X)\in B_k$ which, for each k, a_k and B_k is unambiguously defined, and which we may justly call $P_{X|X\in y}$. Now one could try to construct a CAR mechanism which produces observations of this form only, and which is generated by an underlying independent 'typing and grouping' variable G as in Keiding and

Jacobsen's reference experiment. In this reference model one should be able to compute $P_{X|Y=y}$ and show that it equals $P_{X|X\in y}$. Then by our result that $P_{X|Y=y}$ does not depend on the specific CAR mechanism at hand, it remains equal to $P_{X|X\in y}$ for all CAR mechanisms.

We showed that in discrete sample spaces, `CAR is everything`, but that this result breaks down in general. Is there a repair-job which will make it generally true? We conjecture that even if in general CAR is not everything, that MAR is: in other words, any model for the distribution of a random subset of the components of a multivariate vector can be reproduced exactly with a Coarsening (Missing) at random model.

Our negative results on sequential representations of CAR in this paper and on MAR in our companion paper need further study. `CAR is more than it seems`! CAR and MAR mechanisms exist, whose computer implementation has the following property: the computer needs to know more about X than it's willing to output in its final `print` statement, yet this fact does not affect our face-value inference.

We cannot conceive of more general mechanisms for generating CAR and MAR in an honest way, but is this just a lack of imagination? Can one easily recognise if a given CAR or MAR mechanism has a sequential representation? In 'large' spaces, do 'most' CAR and MAR mechanisms admit a representation?

We showed that in finite sample spaces, monotone coarsened data could be modelled by monotone coarsening rules, and that the CAR factorisation could be explicitly recovered from the observed data distribution. Robins and Rotnitzky (1992) and Robins (1996a) have obtained analogous results when monotone coarsened data results from right censoring by a continuously distributed censoring variable. It is a challenge to extend this to the general monotone case. Some kind of product-integration technique should be possible to mimic the Kaplan-Meier method we used. The branching structure of a tree should be replaced by the continuous branching of histories of a continuous time stochastic process, where the total history of the process 'up to and including time infinity' tells us the complete data. We conjecture that all monotone CAR data can be represented as a stopped stochastic process with the unstopped process corresponding to the underlying data X and the stopping time corresponding to a generalised censoring time.

Bibliography

P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and Adaptive Inference in Semi-parametric Models*, John Hopkins University Press, Baltimore.

J.T. Chang and D. Pollard (1997), Conditioning as disintegration, *Statistica Neerlandica* **51** (to appear).

D.M. Dabrowska (1988), Kaplan-Meier estimation on the plane, *Ann. Statist.* **16**, 1475–1489.

R.D. Gill (1989), Non- and semi-parametric maximum likelihood estimators and the von Mises method, Part 1, *Scand. J. Statist.* **16**, 97–128.

R.D. Gill and J.M. Robins (1997), Sequential models for coarsening and missingness, *Proc. First Seattle Symposium on Biostatistics: Survival Analysis*, ed. D.Y. Lin, Springer-Verlag.

D.F. Heitjan (1993), Ignorability and coarse data: some biomedical examples, *Biometrics* **49**, 1099–1109.

D.F. Heitjan (1994), Ignorability in general incomplete-data models, *Biometrika* **81**, 701–708.

D.F. Heitjan and D.B. Rubin (1991), Ignorability and coarse data, *Ann. Statist.* **19**, 2244–2253.

M. Jacobsen and N. Keiding (1995), Coarsening at random in general sample spaces and random censoring in continuous time, *Ann. Statist.* **23**, 774–786.

R. Kress (1989), *Linear Integral Equations*, Springer-Verlag, Berlin.

M.J. van der Laan (1993), *Efficient and Inefficient Estimation in Semiparametric Models*, Ph.D. Thesis, Dept. Mathematics, University Utrecht; reprinted (1995) as CWI tract 114, Centre for Mathematics and Computer Science, Amsterdam.

M.J. van der Laan (1996), Efficient estimation in the bivariate censoring model and repairing NPMLE, *Ann. Statist.* **24**, 596–627.

R.J.A. Little and D.B. Rubin (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

S.F. Nielsen (1996), *Incomplete Observations and Coarsening at Random*, preprint, Institute of Mathematical Statistics, Univ. of Copenhagen.

R.L. Prentice and J. Cai (1992), Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika* **79**, 495–512.

J.M. Robins (1996a), Locally efficient median regression with random censoring and surrogate markers, pp. 263–274 in: *Lifetime Data: Models in Reliability and Survival Analysis*, N.P. Jewell, A.C. Kimber, M.L. Ting Lee, G.A. Whitmore (eds), Kluwer, Dordrecht.

J.M. Robins (1996b), Non-response models for the analysis of non-monotone non-ignorable missing data, *Statististics in Medicine*, Special Issue, to appear.

J.M. Robins and R.D. Gill (1996), Non-response models for the analysis of non-monotone ignorable missing data, *Statistics in Medicine*, to appear.

J.M. Robins and Y. Ritov (1996), Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models, *Statistics in Medicine*, to appear.

J.M. Robins and A. Rotnitzky (1992), Recovery of information and adjustment for dependent censoring using surrogate markers, pp. 297–331 in: *AIDS Epidemiology—Methodological Issues*, N. Jewell, K. Dietz, V. Farewell (eds), Birkhäuser, Boston.

J.M. Robins, A. Rotnitzky and L.P. Zhao (1994), Estimation of regression coefficients when some regressors are not always observed, *J. Amer. Statist. Assoc.* **89**, 846–866.

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

D.B. Rubin, H.S. Stern and V. Vehovar (1995), Handling "Don't Know" survey responses: the case of the Slovenian plebiscite, *J. Amer. Statist. Assoc.* **90**, 822–828.

A.W. van der Vaart (1991), On differentiable functionals, *Ann. Statist.* **19**, 178-204.

P. Whittle (1971), *Optimization under Constraints*, Wiley, New York.

Authors:

R.D.Gill, Mathematical Institute, University Utrecht, PO Box 80 010, 3508 TA Utrecht, Netherlands; email: gill@math.ruu.nl

M.J. van der Laan, Dept. of Biostatistics, University of California, Berkeley, Ca. 94720, USA; email: laan@stat.berkeley.edu

J.M. Robins, Depts of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Road, Boston Ma. 02115, USA; email: robins@hsph.harvard.edu