

# Coarsening at random: characterizations, conjectures, counter-examples

Richard D. Gill  
Mark J. van der Laan  
Jamie 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 it is still true that the coarsening mechanism can be essentially ignored when making likelihood-based inference about the parameters of the distribution of the variable of interest. This paper explores a number of the consequences of the CAR assumption. First in discrete sample spaces, we give some simple equivalent characterizations of CAR. We show that grouped data *always* fit a CAR model, or in a slogan: **CAR is everything**. More precisely, a nonparametric model for the variable of interest together with the assumption of an arbitrary CAR mechanism actually puts no restriction at all on the distribution of the data. Next we describe what intuitively would seem to be the most general possible way to generate CAR mechanisms, a sequential procedure called *randomized monotone coarsening*. Counter-examples however show that CAR mechanisms exist which cannot be represented in this way. This negative result shows that the CAR assumption, though highly convenient, in itself is difficult to motivate on physical grounds: CAR mechanisms exist which intrinsically depend on information in the 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**. Next we give a new definition of CAR in general sample spaces and establish parallel results to the discrete case. We criticize Jacobsen and Keiding's (1994) proposal, contrasting our absolute definition with their relative one. We extend our definition to the case when more data is available than just the coarsening itself; i.e., we also catch a glimpse of random variables involved in the coarsening mechanism. Though we cannot show that 'CAR is everything' in general sample spaces, we establish a local version of such a result, and also a general identifiability result. Finally we discuss the many open problems still remaining to be solved, and pose some conjectures.

## 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  $X$ . (We give the definition in the next section).

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 (1996) study statistical consequences in non- and semi-parametric models (with positive probability of complete observations). In survival analysis, restricting attention to right-censored observations, coarsening at random, is intimately connected to the central notion of ‘independent censoring’.

‘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 of coarsening at random is still poorly understood and this has dangers in uncritical application. This paper attempts to clarify ‘coarsening at random’ by investigating it from a variety of points of view, to find out what it really means, from a modelling point of view, to make the assumption. In making this attempt we quickly encounter a strange fact: CAR has been defined for discrete data only, ‘to avoid measure-theoretic technicalities’, but actually the mere definition of CAR for continuous data—e.g., censored survival times—is not an obvious matter. Keiding and Jacobsen (1995) offer a definition, but though we learnt much from their results, in our opinion their definition is too complex and restrictive, and 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 state the usual (discrete case) definition of CAR in terms of the distribution of the coarsened data given the underlying data (it should only depend on the underlying data through what the observed data tells us about it). We show that this is equivalent to a factorisation 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 (the same as if the coarsening had been fixed in advance, independently of the underlying data). In section 2 we give one of our main new results: 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. In other words, any coarsened data (at least, in the discrete case!) fits exactly to a CAR model. This result shows that the CAR assumption is pretty weak; without adding further (parametric) assumptions either on the complete-data generating mechanism, or on the coarsening mechanism, the assumption is untestable. This result generalises the familiar survival analysis result that the independent competing risks assumption is untestable.

A counterexample shows that extending to continuous models, the result is no longer true. However we argue that it is so close to being true that the data-analyst should still take it as a fact of life.

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.

At the end of section 3 we showed that CAR mechanisms exist which one would not expect to occur on physical grounds. This is the other side of the coin to the fact we established earlier, that the CAR assumption is so weak that in the nonparametric case, assuming CAR one is assuming exactly nothing. However, CAR implies that the coarsening mechanism is supposed to be irrelevant for inference. Knowing more about the coarsening mechanism than that it is CAR should therefore not be of any use. This is true for strict likelihood-based inference, but we argue that there are many situations (and we give in section 5 examples 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 ‘irrelevant’ information and construct ad hoc (‘frequency based’) methods, which are not asymptotically efficient, but do actually work in small samples. We discuss such examples in section 5.

Though the mathematical results so far are all restricted to a finite sam-

ple space, we emphasize that they are also at least in a moral sense generally applicable, in particular, to missing data problems in survival analysis where survival times are generally modelled as continuously distributed. The second main group of sections, 6 to 10, are devoted to the mathematical generalisation of the earlier results, as far as possible. To begin with, we need a definition of CAR in general sample spaces.

CAR was originally stated in terms of the (discrete) density function of the observed and 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. Keiding and Jacobsen (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 factorisation 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 more ‘relative’ definition of Jacobsen and Keiding (1995), and establishes the connections. This can be considered a technical interlude, for the specialists.

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 sense 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 parallel uniqueness result of the decomposition of the law of the data into distribution of underlying data and coarsening mechanism, though somewhat technical.

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. This supports our claim that survival analysis and mathematical statistics continue to enrich one another over the years.

A companion paper in the same volume concentrates on Missing at Random (missing components of a multivariate vector) and investigates there too, the meaning of the MAR assumption. 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  $\mathcal{E} = 2^E$  denote the power set 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:

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

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

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

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 occurring 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  $x \in A$ )

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

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:

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

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

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

Thus the observation of  $\mathcal{X} = A$  tells us no more, in the sense of what is the conditional distribution of  $X$  given this fact, than “the obvious” fact:  $X \in A$ .

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  $\theta$ , while that of the coarsening mechanism (the conditional distribution of  $\mathcal{X}$  given  $X$ ) on a

distinct, variation independent, parameter  $\gamma$ . We suppose CAR holds, for each  $\gamma$ . Write

$$\begin{aligned} p_x^\theta &= \Pr^\theta(X = x); & p_A^\theta &= \Pr^\theta(X \in A); \\ \pi_A^\gamma &= \Pr^\gamma(\mathcal{X} = A | X = x) & (x \in A) \\ &= \Pr^\gamma(\mathcal{X} = A | X \in A). \end{aligned}$$

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

$$\begin{aligned} f_A^{\theta, \gamma} = \Pr^{\theta, \gamma}(\mathcal{X} = A) &= \Pr^{\theta, \gamma}(\mathcal{X} = A \text{ and } X \in A) \\ &= \Pr^\theta(X \in A) \Pr^\gamma(\mathcal{X} = A | X \in A) \\ &= p_A^\theta \pi_A^\gamma. \end{aligned} \quad (4)$$

So under CAR, the joint likelihood for  $\theta$  and  $\gamma$  factors and the  $\theta$  part can be written down *without knowledge of the coarsening mechanism*: as far as  $\theta$  is concerned, the observation “ $\mathcal{X} = A$ ” can be treated like an observation “ $X \in A$ ”. At the same time, the likelihood for  $\gamma$  can be written down without knowing the distribution of  $X$ , and moreover the likelihood for  $\gamma$  based on the data  $\mathcal{X}$  is the same as the likelihood for  $\gamma$  based on the conditional distribution of  $\mathcal{X}$  given  $X$  (and can be written down 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*

$$f_A = p_A \pi_A \quad (5)$$

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

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

For if the distribution of  $\mathcal{X}$  factors as in (5), construct a random variable  $X$  by letting

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

Under (5), if  $f_A > 0$  then  $p_A > 0$  too so the construction is possible. The construction forces  $X \in \mathcal{X}$  to hold with probability 1, so  $\mathcal{X}$  is a coarsening of  $X$ . 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

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

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

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

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

This argument, together with the conditional independence arguments of the introduction, shows that CAR can equivalently be described in terms of:

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

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

FACTOR( $\mathcal{X}$ ): factorization of the marginal distribution of  $\mathcal{X}$ , (5).

The answer to our Question is *yes*, and moreover the factorization (5) is unique. The distribution  $(p_x)$  is unique too if it is determined uniquely by the  $p_A$  for  $A$  with  $f_A > 0$ .

**Theorem.** *Let  $\mathcal{X}$  be a random non-empty set with distribution  $(f_A : A \in \mathcal{E} \setminus \{\emptyset\})$ . Then there exist CAR probabilities  $(\pi_A)$  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  $f_A > 0$ ,  $\pi_A$  and  $p_A$  are unique.*

**Proof.** Consider the problem of maximization of

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

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

Considered as a function of the  $p_A$  and  $\pi_A$ , varying in  $[0, \infty)$ , for  $A$  with  $f_A > 0$ , to  $[-\infty, \infty)$ ,  $\sum f_A \log p_A + \sum f_A \log \pi_A$  is continuous and strictly concave. The set of  $p_A$  and  $\pi_A$  satisfying the constraints is convex and compact. So the supremum is attained uniquely.

We study the solution for the  $p_A$  separately, in more detail. 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 the 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  $\lambda$  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

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

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

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

So  $\lambda = 1$  and we have

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

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

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

and  $f_A = p_A \pi_A$  for all  $A$  (also if  $p_A = 0$ , and whether or not  $A$  is a singleton  $\{x\}$ ).

Thus a factorization  $f_A = \pi_A p_A$  exists. Since  $\sum f_A \log(p_A \pi_A) \leq \sum f_A \log f_A$  for all  $(\pi_A)$  and  $(p_x)$  satisfying the constraints, the factorization we have

found must be a solution of the maximization problem: maximize  $\sum f_A \log(p_A \pi_A)$ . As we remarked before, the  $p_A$  and  $\pi_A$  for  $A$  with  $f_A > 0$  are uniquely determined in this problem.  $\square$

It is difficult to give necessary and sufficient conditions for uniqueness of *all*  $p_x$  and  $\pi_A$  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  $f_A > 0$ , or  $A = E$ , 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  $(p_A : f_A > 0 \text{ or } A = E)$ .

### 3. Sequential representations of CAR

So far a CAR mechanism is described in an algebraic way: just a collection of probabilities  $\pi_A$  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?

One 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*), which we singled out at the end of the previous section for another attractive property. 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  $1 = \sum_{A'' \ni x} \pi_{A''} < \pi_{A'} + \sum_{A'' \ni x} \pi_{A''} \leq \sum_{A'' \ni x'} \pi_{A''} = 1$ , which is impossible.

Now we describe how the random set  $\mathcal{X}$  can be generated by a random walk on the tree, starting at the root  $E$ , and stopping somewhere,  $\mathcal{X} = A$ , in the tree (i.e., not necessarily at a leaf: a terminal node). Suppose at some stage we have just moved into the node  $A$ . Conditionally on this, we stop in  $A$  with probability  $\pi_A/(1 - \sum_{A' \supset A} \pi_{A'})$ . Since for any  $x \in A$  we have  $1 = \sum_{A' \ni x} \pi_{A'} \geq \pi_A + \sum_{A' \supset A} \pi_{A'}$ , the probability of stopping in  $A$  is indeed a probability. 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.

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 terminal node containing  $x$ . Call this path  $E = A_0 \supset A_1 \supset \dots \supset A_k \ni x$ . The probabilities  $\pi_{A_i}$  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 (or *randomized sequential coarsening*). Initially,  $n = 1$  and  $\mathcal{A}_0 = E$ . By step  $n$  we have generated a sequence of nonempty subsets  $\mathcal{A}_0 \supset \mathcal{A}_1 \supset \dots \supset \mathcal{A}_{n-1}$  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  $\mathcal{A}_{n-1} \setminus B_n$ , and set  $\mathcal{A}_n$  equal to  $B_n$  or  $\mathcal{A}_{n-1} \setminus B_n$  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  $\mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_{n-1}$ . If these probabilities only depend on the current data  $\mathcal{A}_{n-1}$  (and on  $n$ ) then we call the scheme a *Markov coarsening*.

It seems a reasonable conjecture that: *any set of CAR probabilities can be represented by a randomized monotone coarsening scheme*. Note that in a simulation of randomized monotone coarsening, we do not use more information about  $X$  than that which is finally revealed in the value of  $\mathcal{X}$ . Put the other way round: to simulate a CAR mechanism which is *not* randomized monotone, the computer program requires 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  $\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$ . Any chain of decisions terminating in  $\mathcal{X} = \{2, 3\}$  if  $X = 2$  or  $X = 3$  would terminate in  $\{1\}$  if  $X = 1$ . However the probability of such a chain is never larger in the  $\{2, 3\}$  case than in the  $\{1\}$  case since in the  $\{2, 3\}$  case one also has to decide, with probability possibly smaller than 1, to stop at this stage and not to continue. In the  $\{1\}$  case this decision to stop is forced. So we must have  $\pi_{\{2,3\}} \leq \pi_{\{1\}}$ . However for instance

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

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  $E = \{1, 2, \dots, n\}$  for any  $n \geq 3$ , comparing  $\pi_{\{1,2,\dots,n-1\}}$  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 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?)

This unsatisfactory state of affairs leaves many questions open. We cannot conceive of a more general mechanism than a randomized monotone coarsening scheme for constructing CAR mechanisms in an honest way, but is this just a lack of imagination? Can one easily recognise if a given CAR

mechanism has a randomized monotone representation? 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 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 overlap non-trivially.

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  $A$  form a partition of  $A$ .

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.

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. Conditional on  $X = x$  the distribution of  $\mathcal{X}$  is obtained by normalising the probabilities  $f_A$  along this main path to add up to 1. But these are exactly the non-zero probabilities  $\pi_A$  for  $A \ni x$ . Having calculated these  $\pi_A$  we may calculate the corresponding  $p_A$  by division; we are really only interested in the smallest  $A \ni x$  with  $f_A > 0$ .

An alternative calculation follows by considering the steps up the main path of the tree from  $E$  to  $\{x\}$  as time-steps. 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 (i.e., 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  $\tilde{p}_{\{x\}}$  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  $\infty$  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  $\pi_A$  also.

We return to monotone coarsening in the next section, where we show how the associated coarsening mechanism can be realised by a random walk on the tree, starting at  $E$ , and at each node  $A$  either stopping at  $A$  (returning ‘ $\mathcal{X} = A$ ’) with conditional probability  $\pi_A / \sum_{A' \subseteq A} \pi_{A'}$ , or continuing with the complementary probability and choosing the branch  $A'$  from  $A$  with conditional probability  $p_{A'}/p_A$ ; in other words, choosing the branch which contains the true value of  $X$ . Thus the CAR probabilities  $\pi_A$  determine (and are determined by) the stopping mechanism; the underlying distribution of  $X$  determines (and is determined by) the branching mechanism.

## 5. Examples in survival analysis

Ordinary censored survival times  $(\tilde{T}, \Delta)$  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  $(\tilde{T}, \infty)$  if  $\Delta = 0$ , and the singleton  $\{\tilde{T}\}$  if  $\Delta = 1$ . Of course this example needs a continuous sample space to be treated properly, but for the time being we ignore this complication; let us suppose that all time points, denoted  $t$  or  $\tilde{t}$ , are integers between say 0 and  $N$ . The coarsened data is now represented by half intervals  $\{\tilde{t} + 1, \dots, N\}$  in the case of censored observations, and of singletons  $\{t\}$  in the case of uncensored. Any two of such sets are either disjoint or one is contained in the other.

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  $1/n$ . Censored survival data is monotone and the maximization can be done explicitly, 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

$\{C+1, \dots, N\}$  if  $T > C$ . By CAR we have that for each  $\tilde{t}$ ,  $\Pr\{\tilde{T} = \tilde{t}|T = t\}$  is the same for each  $t > \tilde{t}$ . These probabilities, for  $\tilde{t} = 0, 1, \dots$ , 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  $\tilde{T}, \Delta$  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  $T, C$ .

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. This situation turns out to be intimately connected with CAR ideas. Let us represent one observation again as  $(\tilde{T}, \Delta)$  where  $\tilde{T}$  is now a *vector* of censored survival times and  $\Delta$  a vector of censoring indicators. If the observation takes the value  $(\tilde{t}, \delta)$  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  $(\tilde{t}_i, \delta_i)$ .

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, the 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 ( $p_t$ ), ( $\pi_A$ ) 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  $n \rightarrow \infty$ , 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) 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  $(\tilde{T}, \Delta)$  is formed componentwise from the components of  $T$  and  $C$  as in the univariate case. Now, although any distribution of  $(\tilde{T}, \Delta)$  can be represented by a CAR model (CAR is everything: now CAR stands for Censoring at Random, which is more general than Random Censoring) it is no longer the case that any distribution can be represented by a Random Censorship model. The nonparametric multivariate random censorship model is a model, it is identified, it is testable. Here is a simple example of bivariate censoring which is Censoring at Random but not random censoring: let  $T = (T_1, T_2)$  be a bivariate survival time. Suppose  $T_1$  and  $T_2$  were actually consecutive durations between events in the lifetime of one individual, starting at time 0. Let  $C$  be an independent censoring time in the ‘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  $(\tilde{T}, \Delta) = ((\min(T_1, C), \min(T_2, \max(0, C - T_1))), (1\{T_1 \leq C\}, 1\{T_1 + T_2 \leq C\}))$ .

Assuming only CAR (a correct assumption), 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 negligible 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. However still its computation is very 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 knew that the data was not just 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 (19..) and of Prentice and Cai (19..)—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.

From a likelihood point of view this is a surprising result. The likelihood has factored; information about nuisance parameters should be irrelevant; yet we have made use of such information to generate alternative and practically valuable estimators. Our explanation is that (as frequentists) we use likelihood methods because of their good large sample properties. The curse of dimensionality may prohibit practical use of likelihood methods, and one may be forced to use irrelevant information to construct well-behaved though asymptotically inefficient statistical procedures; see Robins and Ritov (19..) for an in depth study of this phenomenon.

Our little 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. However 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. Binning of the observations according to values of  $t_1$  solves that problem. One could use the Dabrowska or the Prentice-Cai estimators: that will not require any artificial grouping or smoothing of the data, but they will be asymptotically inefficient.

A less sensible statistician will treat the data precisely as bivariate censored observations of the durations  $(T_1, T_2)$ . If she sticks to NPMLE (or modifications thereof) nothing will go wrong; the data is CAR. However the Dabrowska or Prentice-Cai estimators will now be inconsistent since the Random Censoring model is not true.

#### 4. 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  $X$ ,  $\text{CAR}(\mathcal{X}|X)$ . We showed in section 2 that the definition was equivalent to a condition on the conditional distribution of  $X$  given  $\mathcal{X}$ ,  $\text{CAR}(X|\mathcal{X})$ , and to a factorization of the marginal distribution of  $\mathcal{X}$ ,  $\text{FACTOR}(\mathcal{X})$ .

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 some possible general definitions we must set up enough measure-theoretic background in order that we can indeed talk about all these conditional distributions.

If  $X$  is, say, a real vector, our random set  $\mathcal{X}$  takes values in the power set of  $\mathbb{R}^k$  ( $k$  the dimension of  $X$ ). There is no natural topology on this very large space, hence no natural Borel  $\sigma$ -algebra. Moreover the space is so large that conditional distributions of  $X$  given  $\mathcal{X} = A$  are not guaranteed in general 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, 1993). 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  $\sigma$ -algebra corresponding to the topologies on the spaces where  $X$  and  $Y$  lie.

From now on, we assume this bare minimum of regularity without com-

ment 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

$\text{CAR}(Y|X)$ . *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  $P_{Y|X=x}(dy) = \Pr(Y \in dy|X = x)$  can be chosen for  $P_X$ -almost all  $x$ , such that for  $x, x'$  not in the exceptional set,*

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

One might hope that if (7) holds except on a set of measure zero, then versions of  $P_{Y|X=x}$  can be chosen making it hold *everywhere*. The following recipe (also used in the proof of the theorem 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.

Anyway, the reason for this particular definition is that it gives us, in the next lemma, an important consequence for Radon-Nikodym derivatives between two *distinct* coarsening mechanisms  $P, P'$  each satisfying CAR separately, namely:

$$\frac{dP'_{Y|X=x}(y)}{dP_{Y|X=x}(y)} \text{ does not depend on } x \in y, \quad (8)$$

it only depends on  $y$  itself. We call this derived property  $\text{CAR}(\text{REL})$ , REL standing for relative, in contrast with (7) which can be called  $\text{CAR}(\text{ABS})$ . The lemma says that  $\text{CAR}(\text{ABS})$  for  $P$  and  $\text{CAR}(\text{ABS})$  for  $P'$  implies  $\text{CAR}(\text{REL})$  for  $P'$  with respect to  $P$ . From the lemma follow a factorization and a result of the type  $\text{CAR}(X|Y)$  concerning conditional distributions in the reverse direction. Our specification of exceptional points  $x$  is as wide as possible subject to allowing these results in the hope of establishing, later, not just sufficient but also necessary conditions for CAR. But these are also open problems still!

**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  $dP'_{Y|X=x}/dP_{Y|X=x}(y)$ ,  $y \ni x$ , 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

$$\frac{dP'_{Y|X=x}}{dP_{Y|X=x}}(y) \neq \frac{dP'_{Y|X=x'}}{dP_{Y|X=x'}}(y)$$

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 (watch out!) 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 of points, with

$$\frac{dP'_{Y|X=x}}{dP_{Y|X=x}}(y) < \frac{dP'_{Y|X=x'}}{dP_{Y|X=x'}}(y) \quad (9)$$

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)

$$P'_{Y|X=x} < P'_{Y|X=x'} \quad (10)$$

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  $P'$ .  $\square$

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

$$P_{X|Y=y}(dx) \stackrel{(\text{def.})}{=} \Pr(X \in dx|Y = y) = \Pr(X \in dx|X \in y) \quad (11)$$

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  $\Pr(X \in dx|X \in y)$  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  $\theta$  and  $\gamma$  respectively;

$$P_{X,Y}^{\theta,\gamma}(dx, dy) = P_X^\theta(dx) P_{Y|X=x}^\gamma(dy). \quad (12)$$

We assume  $P_{Y|X=x}^\gamma(dy)$  satisfies CAR, (7), for each value of  $\gamma$ . Moreover we assume that  $P_X^\theta \ll P_X^{\theta_0}$ ; and for  $P_X^{\theta_0}$  almost all  $x$ ,  $P_{Y|X=x}^\gamma \ll P_{Y|X=x}^{\gamma_0}$ .

**Theorem.** *Under CAR, the likelihood for  $\theta, \gamma$  based on observation of  $Y$  factors:*

$$\frac{dP_Y^{\theta,\gamma}}{dP_Y^{\theta_0,\gamma_0}}(y) = \frac{dP_{Y|X=x}^\gamma}{dP_{Y|X=x}^{\gamma_0}}(y) \cdot E_{\theta_0,\gamma_0} \left( \frac{dP_X^\theta}{dP_X^{\theta_0}}(X) \middle| Y = y \right) \quad (13)$$

(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,  $P_{X|Y=y}^{\theta,\gamma}(dx)$  does not depend on  $\gamma$ .

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

$$\begin{aligned} \frac{dP_{X,Y}^{\theta,\gamma'}}{dP_{X,Y}^{\theta,\gamma}}(x, y) &= \frac{dP_X^\theta}{dP_X^\theta}(x) \frac{dP_{Y|X=x}^{\gamma'}}{dP_{Y|X=x}^\gamma}(y) \\ &= 1 \cdot k(y; \gamma', \gamma) \end{aligned} \quad (14)$$

for some function  $k$ , by the Lemma.

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

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

Thus

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

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

We also have

$$\frac{dP_{X,Y}^{\theta,\gamma'}}{dP_{X,Y}^{\theta,\gamma}}(x, y) = \frac{dP_Y^{\theta,\gamma'}}{dP_Y^{\theta,\gamma}}(y) \frac{dP_{X|Y=y}^{\theta,\gamma'}}{dP_{X|Y=y}^{\theta,\gamma}}(x)$$

hence

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

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

$$\frac{dP_{X|Y=y}^{\theta, \gamma}}{dP_{X|Y=y}^{\theta, \gamma_0}}(x) = 1$$

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

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

The  $\gamma$ -part (CAR mechanism) is the same as the likelihood for  $\gamma$  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  $\theta$ -part (underlying variable of interest) can be written down without knowing the CAR parameter  $\gamma$ : just pick any value, say  $\gamma_0$ , 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  $P_{Y|X=x}^\gamma$ —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  $P_{Y|X=x}^\gamma \ll P_{Y|X=x}^*$ , in other words: the “reference” coarsening mechanism  $P_{Y|X \neq x}^*$  used in the calculations can generate all the sets  $y$  which can occur “in reality” under  $P_{Y|X=x}^\gamma$ .

Put another way: the likelihood for  $\theta$ , under CAR, is the same as under any other specific CAR mechanism which generates the same random sets. We illustrate this and further points with a succession of examples. Examples 2 and 3 concern censored data (univariate and multivariate respectively). Example 1 is a classical paradox from the theory of conditional distributions.

**Example 1. Borel’s paradox.**

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

One easily computes that  $\Theta$  and  $\Phi$  are independent,  $\Theta$  is uniformly distributed while  $\Phi$  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  $\theta$  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  $\phi$ .

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 place 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).

Now suppose we change the coordinate system, placing (new) North and South poles at 0 and 180 degrees longitude on the equator and letting the (old) equator be the new 0 and 180 degree longitude line. Call the new coordinate system  $\Theta'$ ,  $\Phi'$ . Conditional on  $\Theta' = 0$  the point  $X$  lies very non-uniformly distributed on the equator. Conditional on  $\Phi = 0$  the point  $X$  lies uniformly distributed on the equator. But the coarsened data is the same!

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.  $\square$

**Example 2.** *Univariate right censoring.*

TO BE WRITTEN!  $\square$

**Example 3.** *Multivariate right censoring.*

Suppose  $X = (X_1, \dots, X_k)$  and the possible realisations of  $\mathcal{X} = \alpha(Y)$  are Cartesian products of singletons  $\{x_i\}$  and half-lines  $(x_i, \infty)$ . Such sets are generated by the *random-censoring* model:  $C = (C_1, \dots, C_k)$  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  $\tilde{X}_i = X_i \wedge C_i$ ,  $\Delta_i = 1\{X_i \leq C_i\}$ .

It is easily checked that this specific model is CAR. Moreover, for a point  $(\tilde{x}, \delta)$  and vector  $x$  let  $x_\delta = (x_i : \delta_i = 1)$ ,  $x_{\bar{\delta}} = (x_i : \delta_i = 0)$ . Then likelihoods can be computed as

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

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

$$P_{X_\delta}(d\tilde{x}_\delta)P_{X_{\bar{\delta}}|X_\delta=\tilde{x}_\delta}((\tilde{x}_{\bar{\delta}}, \infty_{\bar{\delta}})).$$

This example allows sets  $y$ , having probability zero, but which are not singletons.  $\square$

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  $\theta$ . 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 special model,  $P_{X|Y=y} = P_{X|X \in y}$  so (13) leads to the ‘right’ answer  $P_X^\theta(y)$ .

Our theorem satisfactorily shows that our general notion of  $\text{CAR}(Y|X)$  has the required consequences  $\text{CAR}(X|Y)$  and  $\text{FACTOR}(Y)$ . It is an open question as to whether (and how) these can be made actually *equivalent* to  $\text{CAR}(Y|X)$ . 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,  $\sigma$ -finite, measure. One may check that it then conversely holds that the distributions of  $X$  given  $Y$  are also dominated; in fact,  $\text{DOM}(Y|X) \iff \text{DOM}(X|Y)$ , and under this condition,  $\text{CAR}(Y|X) \iff \text{CAR}(X|Y)$ . However this special case hardly has interesting applications beyond the discrete case.

Our definition of  $\text{CAR}(Y|X)$  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  $\theta$  fixed in  $\gamma$ , 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  $\text{CAR}(\text{REL})$  and not  $\text{CAR}(\text{ABS})$  was needed to prove the theorem. It is easy to show that if a statistical model satisfies  $\text{CAR}(\text{REL})$  (with respect to  $\gamma$ , for each  $\theta$ ), and one point in it satisfies  $\text{CAR}(\text{ABS})$  (for each  $\theta$ ), then so do the rest. One may say: if  $\text{CAR}(\text{REL})$  holds, then  $\text{CAR}(\text{ABS})$  either holds at all points in the model or none. It is possible that interesting statistical models can be found which satisfy  $\text{CAR}(\text{REL})$  without  $\text{CAR}(\text{ABS})$  holding anywhere. In fact  $\text{CAR}(\text{REL})$  is simply a classical sufficiency condition: assuming domination, it is the factorization criterion, in the model when *both*  $X$  and  $Y$  are observed, for  $Y$  to be sufficient for  $\gamma$  for each fixed  $\theta$ . Consequently (and equivalently) we have sufficiency according to the definition in terms of conditional distributions: for each  $\theta$ , the distribution of  $X$  given  $Y = y$  does not depend on  $\gamma$ . This is just the second part of our theorem. (See Chang and Pollard (1994) 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.

### 5. CAR according to Jacobsen and Keiding (1994)

Jacobsen and Keiding (1994) 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.

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 known partition of the sample space, and we observe the element of the partition  $Y$  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  $Y = Y(X, G)$ ; because of the partitioning structure we have, for every  $y = Y(x, g)$  for some  $x, g$ , that  $x \in y$  and, if also  $x' \in y$ , then  $Y(x', g) = y$  too. In fact  $\{(x, g) : Y(x, g) = y\} = y \times \{g : Y(x, g) = y \text{ for some } x\}$ .

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  $\mu$ ,  $G$  with distribution  $\nu$ . In fact they give  $P^*$  the name  $\rho$ . 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  $P_{G|X=x}^\gamma \ll \nu$ . 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  $\gamma$ . They prove that under CAR, the likelihood based on  $Y$ , for  $\theta$  and  $\gamma$ , factors, and the  $\theta$  part is  $E^*(f(X; \theta)|Y = y)$ . When  $y$  is a singleton, or a set of positive  $\mu$ -probability, the likelihood becomes as one would hope  $f(X; \theta)$  or  $P_X^\theta(y)$  respectively.

It can now 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 and stronger conclusions. In fact, in the terminology of the end of the last section, their 'reference model' satisfies CAR(ABS), while the rest of the model satisfies CAR(REL). We do not know if (under the appropriate regularity conditions) some kind of

converse is true. In particular, given a joint distribution of  $X, Y$  it is not clear how one can set about constructing a random variable  $G$  and mapping  $x, g \mapsto Y(x, g)$  such that  $Y = Y(X, G)$ .

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  $\theta$  part of the likelihood factorization should be, and how the distribution of  $X$  given  $X \in y$  should be defined.

From our point of view their set-up is too restrictive. In particular, it is not a helpful starting point for investigating the non-parametric nature of CAR given the distribution of a random non-empty set  $Y$ : does there exist (or can one construct) a variable  $X$  so that  $Y$  is a CAR coarsening of  $X$ ?

## 6. 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  $\alpha$  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  $\phi$ , and  $\alpha(y) = \{x : \exists g \text{ with } \phi(x, g) = y\}$ .

The only difference with the set-up of Section 7 is that we do not suppose the function  $\alpha$  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  $\alpha$ . We will define a notion of coarsening at random in terms of these two ingredients. If actually  $Y = \phi(X, G)$  for some  $\phi, G$ , and corresponding  $\alpha$ , the CAR property can also be rephrased as a condition on the laws of  $G$  given  $X = x$  and  $\phi$ . However it turns out that the latter is much more clumsy.

Now in Section 7,  $Y$  was introduced as essentially just a synonym for  $\mathcal{X} = \alpha(Y)$ . However the implication of this that  $\alpha$  is one-to-one was not used anywhere. Therefore *we maintain the definition* (21) of  $\text{CAR}(Y|X)$  in

our new context, and *all the results of Section 7 remain valid*. Everywhere, the statement ‘ $x \in y$ ’ should just be read as shorthand for ‘ $x \in \alpha(y)$ ’.

Also the discrete-case results of Section 1 can 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:

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

Consequently  $\Pr\{Y = y | X = x\} = \Pr\{Y = y | X \in \alpha(y)\}$  for  $x \in \alpha(y)$ ; of course  $\Pr\{Y = y | X = x\} = 0$  for  $x \notin \alpha(y)$ .

We compute the marginal distribution of  $Y$ : it is

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

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

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

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

so  $\text{CAR}(X|Y)$  holds.

Suppose actually  $Y = \phi(X, G)$  for some grouping or censoring variable  $G$ . Take

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

The  $\text{CAR}(Y|X)$  assumption (in the discrete case) is immediately rewritten as

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

This is an assumption on the conditional distribution of  $G$  given  $X = x$ , but 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  $\phi$  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 thus equivalent to simultaneous coarsenings of both  $X$  and  $G$ . In particular,  $\alpha(y) = \phi_X^{-1}(y)$ . Also

$$\Pr\{\phi(x, G) = y | X = x\} = \Pr\{G \in \phi_G^{-1}(y) | X = x\}.$$

Under Cartesian coarsening,  $\text{CAR}$  is *implied* by the assumption:  $\Pr\{G = g | X = x\}$  does not depend on  $x \in \phi_X^{-1}(y)$ , for each  $g \in \phi_G^{-1}(y)$ . This sufficient condition can be reformulated in general sample spaces analogously to (21).

A final remark is that under Cartesian coarsening, one can always pretend that  $G$  is actually completely observed: define  $G^*$  as a function of  $Y$  by letting  $g^*$  be a measurably selected element of  $\phi_G^{-1}(y)$ . Now  $Y^* = (Y, G^*) = (\phi(X, G), G^*) = (\phi(X, G^*), G^*) = \phi^*(X, G^*)$ , showing that the augmented data  $Y^*$  is a Cartesian coarsening, with coarsening variable  $G^*$ , and such that  $\phi_{G^*}^{-1}(y^*)$  is the singleton  $\{g^*\}$  in  $y^* = (y, g^*)$ .

### 7. Locally, CAR is everything

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  $E^0 h(X) = 0$ ,  $P^{\theta, h}$  defined by  $P^{\theta, h}(dx) = (1 + \theta h(x))P^0(dx)$  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 (w.r.t.  $\mathcal{L}^2(P^0)$ ) 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  $\mathcal{L}_0^2(P^0)$ , the space of all square-integrable, mean-zero functions of  $X \sim P^0$ . We also write  $\mathcal{L}_0^2(X)$  for the same space when the distribution of  $X$  under which we work is clear from the context.

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

$$P^{\theta, h}(dx) = \left(1 + \frac{1}{2}\theta h(x)\right)^2 P^0(dx) / \left(1 + \frac{1}{4}\theta^2 E^0(h(X)^2)\right).$$

Now 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:  $\mathcal{L}_0^2(X)$ .

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  $\mathcal{L}_0^2(Y)$* . Locally, we are not assuming anything about the distribution of  $Y$ .

This result would be a corollary to a ‘global’ result: each distribution of  $Y$  admits representation as a CAR model.

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

$$\begin{aligned} P_X^\theta(dx) &= (1 + \theta h(x))P_X(dx) \\ P_{Y|X=x}^\gamma(dy) &= (1 + \gamma k(y; x))P_{Y|X=x}(dy). \end{aligned}$$

If  $h$  is bounded and  $E(h(X)) = 0$  this defines a one-dimensional parametric submodel for the distribution of  $X$  with parameter  $\theta$  (sufficiently close to zero). Similarly if  $k$  is bounded and  $E(k(Y; x) | X = x) = 0$  we have a model for the distribution of  $Y$  given  $X$  with parameter  $\gamma$ . 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  $\theta$  and  $\gamma$  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:

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

Write  $\|\cdot\|_X$ ,  $\langle \cdot, \cdot \rangle_X$  etc. for the corresponding norms and inner products. Note that  $E^X$  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:

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

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  $\theta$  and  $\gamma$  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):

$$\frac{dP_Y^{\theta, \gamma}}{dP_Y^{\theta_0, \gamma_0}}(y) = \frac{dP_{Y|X=x}^{\gamma}}{dP_{Y|X=x}^{\gamma_0}}(y) \cdot E_{\theta_0, \gamma_0} \left( \frac{dP_X^{\theta}}{dP_X^{\theta_0}}(x) \mid Y = y \right),$$

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  $\kappa(P_X) = \int \kappa(x) P_X(dx)$  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  $\tilde{\kappa} = \kappa - E_X(\kappa)$ . Then by BKRW or by van der Vaart (1991), we have: if  $I = A^\top A$  has (at  $\tilde{\kappa}$ ) an inverse  $I^{-1}$ , then the asymptotic information bound for estimation of  $\kappa$  is

$$\|A(A^\top A)^{-1} \tilde{\kappa}\|^2 < \infty.$$

In fact  $g(X) = (A^\top A)^{-1} \tilde{\kappa}(X)$  generates a ‘hardest’ one-dimensional submodel for estimating  $\kappa$  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  $\tilde{\kappa}$  lies in the range of  $A^\top$ ; this is obviously implied by  $\tilde{\kappa} = A^\top A g$  for some  $g$ . We will later argue that a finite information bound means in some sense local identifiability. But first we give 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  $I = A^\top A : \mathcal{L}_0^2(X) \rightarrow \mathcal{L}_0^2(X)$  is onto and has a bounded inverse; in fact  $\|I^{-1}h\| \leq \delta^{-1/2}\|h\|$  for all  $h$ . Consequently  $\|A(A^\top A)^{-1} \tilde{\kappa}\|^2 \leq \delta^{-1} \|\tilde{\kappa}\|^2$  or: the asymptotic information bound for estimating  $\kappa$  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

$$\begin{aligned} \|Ah\|^2 &= E\left(E(h(X) | Y)^2\right) \\ &\geq E\left(h(X)^2 1\{Y = \{X\}\}\right) \\ &= E\left(h(X)^2 \Pr(Y = \{X\} | X)\right) \\ &\geq \delta \|h\|^2. \end{aligned}$$

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

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

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),

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

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  $(\sum_{n=0}^{\infty} (1 - \delta)^n) = \delta^{-1}$ .  $\square$

**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  $n$  i.i.d. 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  $\|\tilde{\kappa}\|^2/(n\delta)$ . This shows directly that the information bound for estimation of  $\kappa$  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  $P_{Y^\gamma}^{\theta, \gamma}$  constructed from  $P_X^\theta$  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  $\text{IC}_{\text{opt}} = A(A^\top A)^{-1}\tilde{\kappa}$  where we work at the fixed point  $\theta = \theta_0$ ,  $\gamma = \gamma_0$ . Then

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

is an estimator of  $\kappa(P_X^{\theta_0})$  based on  $Y_1, \dots, Y_n$  such that  $n^{1/2}(\hat{\kappa} - \kappa(P_X^{\theta_0})) \xrightarrow{\mathcal{D}} \mathcal{N}(\mu, \sigma^2)$  as  $n \rightarrow \infty$ , under  $P_Y^{\theta_0 + n^{-1/2}\eta, \gamma_0 + n^{-1/2}\psi}$ , where  $\mu = \lim_{n \rightarrow \infty} n^{1/2}(\kappa(P_X^{\theta_0 + n^{-1/2}\eta}) - \kappa(P_X^{\theta_0}))$  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.,  $P_X^0 = P_X^{\theta_0}$ ,  $P_{Y|X=x}^0 = P_{Y|X=x}^{\gamma_0}$ . Since even under CAR the tangent space at  $P^0$  is everything, any  $P_Y$  close to  $P_Y^0$  lies to a close approximation on one of these submodels. Thus in a local asymptotic sense, for  $P_Y$  close to a given model  $P_Y^0$  determined by  $P_X^0, P_{Y|X=x}^0$ , one can recover  $P_X$  from  $P_Y$ .

## 8. Global identifiability of CAR

Suppose the triple  $X, Y, \alpha$  is such that  $Y, \alpha$  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  $(\pi_A)$  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  $\alpha$  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  $Q_X = \frac{1}{2}(P_X + P'_X)$  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  $Q$ . Consider also  $P^{\theta, \gamma}$  defined by

$$P_X^\theta = (1 - \theta)P_X + \theta P'_X,$$

$$P_{Y|X=x}^\gamma = (1 - \gamma)P_{Y|X=x} + \gamma P'_{Y|X=x}.$$

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

$$\frac{dP_Y^{\theta,\gamma}}{dQ_Y}(y) = E_Q\left(\frac{dP_X^\theta}{dQ_X}(X) \mid Y = y\right) \cdot \frac{dP_{Y|X=x}^\gamma}{dQ_{Y|X=x}}(y), \quad x \in \alpha(y).$$

Thus

$$\begin{aligned} E_P \log \frac{dP_Y^{\theta,\gamma}}{dQ_Y}(Y) &= E_P \log \left( E_Q \left( \frac{dP_X}{dQ_X}(X) \mid Y \right) + \theta \left( E_Q \left( \frac{dP'_X}{dQ_X}(X) - \frac{dP_X}{dQ_X}(X) \mid Y \right) \right) \right) \\ &\quad + E_P \log \left( \frac{dP_{Y|X}}{dQ_{Y|X}}(Y) + \gamma \left( \frac{dP'_{Y|X}}{dQ_{Y|X}}(Y) - \frac{dP_{Y|X}}{dQ_{Y|X}}(Y) \right) \right) \end{aligned}$$

where  $(dP_{Y|X}/dQ_{Y|X})(Y)$  is defined, for  $Y = y$ , as  $(dP_{Y|X=x}/dQ_{Y|X=x})(y)$  for any  $x \in \alpha(y)$ .

Now the above function of  $\theta$  and  $\gamma$  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  $\theta$  and  $\gamma$ , or:

$$E_Q \left( \frac{dP'_X}{dQ_X}(X) \mid Y = y \right) = E_Q \left( \frac{dP_X}{dQ_X}(X) \mid Y = y \right)$$

for  $P$  almost all  $y$ ,

$$\frac{dP'_{Y|X=x}}{dQ_{Y|X=x}}(y) = \frac{dP_{Y|X=x}}{dQ_{Y|X=x}}(y)$$

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

$$\frac{dP_Y}{dQ_Y}(y) = E_Q \left( \frac{dP_X}{dQ_X}(X) \mid Y = y \right) \cdot \frac{dP_{Y|X=x}}{dQ_{Y|X=x}}(y), \quad x \in \alpha(y)$$

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

## 9. 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. We do not know if equivalence holds in general.

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  $\alpha_k$  and  $\beta_k$  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.

Many coarsened data applications actually involve observation of more than just a random set in which  $X$  lies; for instance, in random censoring, one is sometimes able to observe the censoring variable  $C$  even when  $X < C$  so  $X$  is observed. Our theory needs to be extended to cover this kind of application. The data  $Y$  is then typically supposed to be a known function  $\phi$  of both  $X$  and another variable  $G$  carrying the random part of the coarsening. The difference with Jacobsen and Keiding’s set-up is that  $Y$  does not necessarily take values in the power set of the range of  $X$ . However let us suppose that we still have the product structure that  $\phi^{-1}(y)$  is a rectangle, written say as  $\phi_X^{-1}(y) \times \phi_G^{-1}(y)$ , for each  $y$ . Observation of  $Y = y$  therefore ‘looks like’ observation that  $X$  and  $G$  each lie in certain sets  $\phi_X^{-1}(y)$  and  $\phi_G^{-1}(y)$ . The proper definition of CAR, generalising (21), should now be

$$P_{G|X=x}(dg) = P_{G|X=x'}(dg) \quad \text{on } \{g : \phi(x, g) = \phi(x', g)\}.$$

Actually it is not immediately clear that our first set-up is contained in this, since an auxiliary variable  $G$  and function  $\phi$  are now assumed given ingredients in the definition, while we were only given a coarsening: a random set  $\mathcal{X} = \alpha(Y) \ni X$ , represented by coordinates  $Y$ . However it seems that one may then *define*  $G \equiv Y$  and  $\phi(x, g) = \phi(x, y) \equiv y$ ; we have the product structure and the new definition of CAR contains the old one. (How does this work in our counterexamples to sequential representations?) We expect that all our results carry over to the wider definition of CAR. An intermediate step could be to show that  $G$  may be replaced by a new variable  $G^*$  drawn, given  $Y = y$ , from the conditional distribution of  $G$  given  $G \in \phi_G^{-1}(y)$ . The data  $Y$  can now be replaced by  $(Y, G^*)$ ; in other words, without loss of generality, one may assume that  $G$  is also observed, and is conditionally independent of  $X$  given  $Y$ .

It remains to investigate whether, given a random non-empty set  $\mathcal{X}$  (and subject to a minimum of structural conditions),  $\mathcal{X}$  can be considered as a

coarsening of some  $X$ , satisfying CAR; in our slogan, ‘**CAR is everything**’. Furthermore, is the resulting distribution of  $X$  unique? In other words, is the general CAR assumption completely nonparametric: does assuming nothing about the distribution of  $X$  and nothing about the CAR mechanism, imply *nothing* is assumed about the distribution of the data  $\mathcal{X}$ ?

Here is a possible approach based on the EM algorithm. Writing again  $Y$  for the data, suppose one can pick a random point  $X_0 \in Y$  in a measurable way. Suppose  $P_{X|X \in y}$  is well defined for each possible  $y$  and all  $P_X$ . Now generate a sequence of distributions  $P_{X_0}, P_{X_1}, \dots, P_{X_n}, \dots$  as follows:  $P_{X_{n+1}}$  is the distribution obtained by drawing  $Y = y$  from  $P_Y$  and then  $X_{n+1}$  from  $P_{X_n|X_n \in y}$ . We conjecture that the sequence  $P_{X_n}$  converges weakly to a limit  $P_X$  which is a fixed point of the just described iteration and such that the resulting joint distribution of  $X$  and  $Y$  is a CAR model with  $X \sim P_X$ .

Our negative results on sequential representations of CAR and MAR need further study. **CAR is more than it seems!** The problem is reminiscent of the need in quantum probability, see e.g., Maassen and Kümmerer (1994), to construct random variables with ‘impossible’ correlations. We have shown that 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 also obtained a positive result on MAR mechanisms with sequential representation, namely they are mixtures of deterministic monotone procedures. Does this hold more generally for CAR?

We proved in section 2 that in finite sample spaces, CAR is everything: any  $f_A$  factors. One would hope that this 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. Let  $E$  be the natural numbers  $\{0, 1, \dots\}$  and suppose the only subsets of  $E$  which get positive probability are  $\{n, n+1, \dots\}$  for  $n = 1, 2, \dots$ . If we try to factor  $f_A = p_A \pi_A$  by maximising 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 maximiser would like to put all the probability mass at  $+\infty$  but there is no such point in  $E$  so the maximiser does not exist. Hence there can be no factorisation, since if there were one, it would maximise 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  $A = \{n, n+1, \dots\}$ . In other words, this example can only be repaired by compactifying both the sample space and all the observed random sets in a careful way. In general it might be the case that the theorem can be

made true after a suitable compactification of sample space and observed points. Important examples like classical right censored data have observed sets which are open intervals  $(c, \infty)$ , so this does not look very attractive.

If there are only finitely many sets  $A$  which get positive probability then the factorisation does hold: one can restrict attention to the finite sets formed by picking one point in each non-empty intersection of all  $A$ 's and their complements. In particular, an empirical distribution of  $n$  observed sets  $A$  is always exactly matched by some CAR model. For practical purposes one could say that all distributions, also those of coarsened data, can be arbitrarily well approximated by discrete distributions, for which the theorem is true; therefore even if CAR is not everything, it is almost everything.

We showed that in finite sample spaces, monotone coarsened data could be modelled by monotone coarsening rules, and that the factorisation could be explicitly recovered from the observed data distribution. It is also a challenge to extend this to the general case. Some kind of product-integration technique should be possible to mimic the Kaplan-Meier method we used.

Whatever else CAR may be, we think it may be said 'CAR is fun'.

## 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 (1993), *Conditioning as disintegration*, preprint, Dept. Math., Yale Univ.
- R.D. Gill (1989), Non- and semi-parametric maximum likelihood estimators and the von Mises method, Part 1, *Scand. J. Statist.* **16**, 97–128.
- 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 (1994), 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.
- B. Kümmerer and H. Maassen (1996), Elements of quantum probability, in: *Quantum Probability Communications X*, ed. R.L. Hudson and J.M. Lindsay, World Scientific.
- M.J. van der Laan (1993), *Efficient and Inefficient Estimation in Semiparametric Models*, Ph.D. Thesis, Dept. Mathematics, University Utrecht; re-

- vised, reprinted (1995) as CWI tract, Centre for Mathematics and Computer Science, Amsterdam.
- R.J.A. Little and D.B. Rubin (1987), *Statistical Analysis with Missing Data*, Wiley, New York.
- J.M. Robins (1996), 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 ... (eds), Kluwer, Dordrecht.
- 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.
- A.W. van der Vaart (1991), On differentiable functionals, *Ann. Statist.* **19**, 178–204.
- P. Whittle (1971), *Optimization under Constraints*, Wiley, New York