

# On the testability of the CAR assumption

By Eric Cator

*Delft University of Technology*

In recent years a popular non-parametric model for coarsened data is an assumption on the coarsening mechanism called Coarsening At Random (CAR). We will give a short definition here; in the talk we will introduce a more general version. Let  $\mathcal{Y}$  be the space of the parameter of interest  $Y$ . On  $\mathcal{Y}$  we have a dominating measure  $Q_0$ . Let  $\mathcal{C}$  be a space with a dominating measure  $P_0$  such that the data contains the *coarsening variable*  $C \in \mathcal{C}$  and a function of  $Y$  and  $C$ , so

$$X = (C, \phi(C, Y)).$$

The coarsening mechanism is a joint distribution  $\mu(dcdy) = f(c, y)P_0(dc)Q_0(dy)$  on  $\mathcal{C} \times \mathcal{Y}$ . We can write this as

$$\mu(dcdy) = (f(c|y)P_0(dc))h(y)Q_0(dy),$$

where  $f(c|y)$  is the conditional density of  $C$  given  $Y = y$ .

**Definition 1 (the CAR assumption)** *In the notation given above, the CAR assumption states that  $\mu \ll P_0 \otimes Q_0$  is a possible (or admitted) distribution of  $(C, Y)$  (also called a “coarsening”) if and only if there exists a function  $g$  with*

$$f(c|y) = g(c, \phi(c, y)).$$

This loosely means that we assume that given  $Y$ , the *unknown* part by which the coarsening mechanism chooses  $C$  (note that  $P_0$  is known!) may only be a function of the data. The reason why this is useful, is because it enables a factorization of the likelihood of the data in such a way, that maximizing the likelihood can be done separately for  $g$  (the nuisance parameter) and  $h$ , the density of the parameter of interest. Hence the maximum likelihood estimator for the distribution of  $Y$  can be calculated directly.

To illustrate this, we will use the well known example of current status data. Here we have a variable of interest  $T$  (for example the time of onset of a disease), but we cannot observe this  $T$  directly. Instead, we observe

a censoring time  $C$  (for example the time of a visit to the doctor) and the indicator  $\Delta = 1_{\{T \leq C\}}$ , so the data consists of  $X = (C, \Delta)$ . One can prove that the CAR assumption in this case just states that  $T$  and  $C$  have to be independent.

It has been conjectured in several papers that this assumption cannot be tested by the data, i.e., the assumption does not restrict the possible distributions of the data. In this talk we will show that this conjecture is not always true; an example will actually be current status data, for which we will show that there exists a set of distributions of the data which can be separated from the possible distributions you get assuming CAR by two linear tests. We will also give exact conditions when the conjecture is true, and in doing so, we will introduce a generalized version of the CAR assumption. As an easy consequence we will show that in right-censored data, the CAR assumption cannot be tested.