A short note on coarsening at random

Yen-Chi Chen
University of Washington
April 1, 2020

The coarsening at random (CAR) is a more general concept than the usual missing at random (MAR). Here is a short note about it.

Let $X \in \mathbb{R}^k$ be the variable of interest and $C \in \mathcal{C}$ be a coarsening variable. C may or may not be observed. Our observation is the random vector

$$Y = \Phi(X,C) \in \mathbb{R}^q,$$

where $\phi(X,C)$ is a many to one mapping. Let

$$\mathcal{X}(y) = \{x : \Phi(x,c) = y, c \in \mathcal{C}\}$$

be the collection of X such that under some coarsening case, we observe the same $Y = y$.

The CAR assumes that

$$P(Y = y \mid X = x) = P(Y = y \mid X = x') \quad \text{for all } x,x' \in \mathcal{X}(y). \tag{1}$$

Namely, the conditional probability of Y given X will not change as long as $X \in \mathcal{X}(y)$. Equation (1) further implies

$$P(Y = y \mid X = x) = P(Y = y \mid X \in \mathcal{X}(y)) = h(y) \tag{2}$$

for some function h.

It is possible to generalize the above notations to probability densities of Y. The generalization allow equations (1) an d (2) to be re-written as

$$p_{Y|X}(y \mid x) = p_{Y|X}(y \mid x') \quad \text{for all } x,x' \in \mathcal{X}(y)$$

and

$$p_{Y|X}(y \mid x) = h(y) \tag{3}$$

for some function h. Note that, we may relax the assumption by only requiring (1) holds almost surely.

Relation to MAR. To see how CAR and MAR are related, note that if Y contains C, the coarsening variable, then (CAR) is equivalent to

$$P(C = c \mid X = x) = P(C = c \mid X = x') \quad \text{for all } x,x' \in \mathcal{X}(y). \tag{4}$$

which, together with equation (2), implies

$$P(C = c \mid X = x) = h(y) \quad \text{for all } x \in \mathcal{X}(y). \tag{5}$$

If one view C as the missing indicator, equation (5) implies that the conditional probability of a missing pattern given the variable of interest X only depends on the observable Y, which is how the MAR assumption is formulated.

Relation to MLE. It is well-known that the MAR has the ignorability property that when doing a likelihood inference, there is no need to model the missing part (there is one small additional assumption to achieve this call separation of parameters). A similar pattern occurs here at the CAR. Because all we observed is Y, we can connect the marginal density of Y to the density of X using

$$\begin{aligned}
p_Y(y) &= \int_{\mathcal{X}(y)} p_{Y,X}(y,x)\mu(dx) \\
&= \int_{\mathcal{X}(y)} p_{Y|X}(y|x)p_X(x)\mu(dx) \\
&\overset{(3)}{=} \int_{\mathcal{X}(y)} h(y)p_X(x)\mu(dx) \\
&= h(y)\int_{\mathcal{X}(y)} p_X(x)\mu(dx).
\end{aligned}$$

Assume that we assign a parametric model of X that

$$p_X(x) = p_X(x;\theta).$$

This implies that

$$p_Y(y;\theta) = h(y)\int_{\mathcal{X}(y)} p_X(x;\theta)\mu(dx) = h(y)p_X(\mathcal{X}(y);\theta),$$

where $p_X(\mathcal{X}(y);\theta) = \int_{\mathcal{X}(y)} p_X(x;\theta)\mu(dx)$. Thus, the log-likelihood of θ using Y is

$$\ell(\theta|Y) = \log p_Y(Y;\theta) = \log h(Y) + \log p_X(\mathcal{X}(Y);\theta)$$

so the total log-likelihood with n observation is

$$\ell(\theta|Y_1,\cdots,Y_n) = \Omega(Y_1,\cdots,Y_n) + \sum_{i=1}^n \log p_X(\mathcal{X}(Y_i);\theta)$$

Thus, maximizing the log-likelihood using Y is equivalent to maximizing the log-likelihood constructed from $\sum_{i=1}^n \log p_X(\mathcal{X}(y);\theta)$. Again, we obtain the same result as the ignorability of MAR! Note that the maximization of $\sum_{i=1}^n \log p_X(\mathcal{X}(y);\theta)$ is similar to the case of a mixture model or a latent variable model. The MLE is often obtained by applying an EM algorithm.

Example: missing data. Consider a simple missing data problem where we have two variables per individual: W, the response variable, and Z, the covariate. However, not all response variable W is observed. Some individuals we only observe the covariate. Let R be the observed pattern where $R = 1$ means that we observe both W and Z while $R = 0$ is the case we only see Z. In this case, the variable of interest is $X = (W,Z)$ and the coarsening variable is $C = R$. Our observation can be written as $Y = (R,Z,WR+\star(1-R))$, where $\star$ denotes the missing value. Since the coarsen variable R is inside Y, the CAR is equivalent to

$$P(R=r\mid W=w,Z=z) = P(R=r\mid W=w',Z=z') \quad \text{for all } (w,z),(w',z')\in\mathcal{X}(y).$$

When $r = 1$, this does not tell us much information but when $r = 0$, $y = (0,z,\star)$ so this implies that

$$P(R=0\mid W=w,Z=z) = P(R=0\mid W=w',Z=z)$$

for all w, w'. This implies that

$$P(R=0 \mid W=w, Z=z) = P(R=0 \mid Z=z),$$

which is the MAR assumption.

Example: censoring data. Consider the censoring problem where we have a time-to-event variable T of interest and a censoring variable S. Our observations are $Y = (I(T \le S), \min\{T, S\})$. In this case, the censoring variable $S = C$ is our coarsening variable and the time-to-event variable $T = X$ is the variable of interest. Because the coarsening variable is not directly observed in Y, we use the original form of the CAR:

$$P(Y=y \mid X=x) = P(Y=y \mid X=x') \quad \text{for all } x, x' \in \mathcal{X}(y).$$

When $y = (\delta, \omega)$ where $\delta \in \{0,1\}$ is binary and $\omega \in \mathbb{R}$, $\mathcal{X}(0,\omega) = \{x : x > \omega\}$ and $\mathcal{X}(1,\omega) = \{x : x = \omega\}$. The case where $y = (1,\omega)$ does not give us much information so we focus on the case $y = (0,\omega)$. The CAR implies

$$P(Y=(0,\omega)|T=t) = P(Y=(0,\omega)|T=t') \quad t, t' > \omega = S.$$

This implies that $p_{S|T}(S=\omega|T)$ does not depend on T if $T > S$. Namely, CAR implies

$$S \perp T \mid T > S,$$

the censoring time is independent of the time-to-event of interest when $T > S$, which is the common assumption assumed in handling the censoring data.

Example: causal inference (counterfactual model). Consider the counterfactual model that the binary variable A denotes the reception of treatment or not (1 is treated) and Z is the observed outcome. Under the counterfactual model, there are two potential outcomes $Z(0)$ and $Z(1)$. The observed data is (A, Z), where $Z = A \cdot Z(1) + (1-A) \cdot Z(0)$. In this case, the variable of interest are $Z(0)$ and $Z(1)$ and the coarsening variable is A, which is directly observable in this case. Using equation (4), the CAR assumption is

$$P(A=1|Z(0),Z(1)) = P(A=1|Z(1)), \quad P(A=0|Z(0),Z(1)) = P(A=0|Z(0)).$$

Both equality holds for any pairs z_0, z_1 such that $Z(0) = z_0$ and $Z(1) = z_1$, which implies

$$\begin{aligned}
P(A=1|Z(0)=z_0, Z(1)=z_1) &= P(A=1|Z(1)=z_1) \\
&= 1 - P(A=0|Z(0)=z_0, Z(1)=z_1) \\
&= 1 - P(A=0|Z(0)=z_0)
\end{aligned}$$

for all (z_0, z_1). This is equivalent to $A \perp Z(0), Z(1)$, which is the common assumption on the independence of treatment from the potential outcome.

Reference: *Unified Methods for Censored Longitudinal Data and Causality* by van der Laan & Robins