EXISTENCE OF PROPER REGULAR CONDITIONAL DISTRIBUTIONS

Pietro Rigo
University of Pavia

Bologna, may 16, 2018

Classical (Kolmogorovian) conditional probabilities

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathcal{G} \subset \mathcal{A}$ a sub-σ-field.

A **regular conditional distribution (rcd)** is a map Q on $\Omega \times \mathcal{A}$ such that

(i) $Q(\omega, \cdot)$ is a probability on $\mathcal{A}$ for $\omega \in \Omega$

(ii) $Q(\cdot, A)$ is $\mathcal{G}$-measurable for $A \in \mathcal{A}$

(iii) $P(A \cap B) = \int_B Q(\omega, A)\, P(d\omega)$ for $A \in \mathcal{A}$ and $B \in \mathcal{G}$

An rcd can fail to exist. However, it exists under mild conditions and is a.s. unique if $\mathcal{A}$ is countably generated.

In the standard framework, thus, conditioning is with respect to a σ-field $\mathcal{G}$ and not with respect to an event H.

What does it mean ?

According to the usual interpretation, it means: For each $B \in \mathcal{G}$, we now whether B is true or false. This naive interpretation is very dangerous.

Example 1 Let $X = \{X_t : t \geq 0\}$ be a process adapted to a filtration $\mathcal{F} = \{\mathcal{F}_t : t \geq 0\}$. Suppose $P(X = x) = 0$ for each path x and

$$\{A \in \mathcal{A} : P(A) = 0\} \subset \mathcal{F}_0.$$

In this case,

$$\{X = x\} \in \mathcal{F}_0 \text{ for each path } x.$$

But then we can stop. We already know the X-path at time 0 !

Example 2 (Borel-Kolmogorov paradox) Suppose

$$\{X = x\} = \{Y = y\}$$

for some random variables X and Y. Let Q_X and Q_Y be rcd's given $\sigma(X)$ and $\sigma(Y)$. Then,

$$P(\cdot\,|X = x) = Q_X(\omega, \cdot) \text{ and } P(\cdot\,|Y = y) = Q_Y(\omega, \cdot)$$

where $\omega \in \Omega$ meets $X(\omega) = x$ and $Y(\omega) = y$. Hence it may be that

$$P(\cdot\,|X = x) \neq P(\cdot\,|Y = y) \text{ even if } \{X = x\} = \{Y = y\}.$$

Example 3 For the naive interpretation to make sense, Q should be **proper**, i.e.

$$\boxed{Q(\omega, \cdot) = \delta_\omega \text{ on } \mathcal{G} \text{ for almost all } \omega}.$$

But Q needs not be proper. In fact, properness of Q essentially amounts to $\mathcal{G}$ countably generated.

Conditional 0-1 laws

An rcd Q is 0-1 on $\mathcal{G}$ if

$$\boxed{Q(\omega,\cdot) \in \{0,1\} \text{ on } \mathcal{G} \text{ for almost all } \omega}$$

Why to focus on such a 0-1 law ?

- It is a (natural) consequence of properness

- It is equivalent to

  $$\boxed{\mathcal{A} \text{ independent } \mathcal{G}, \text{ under } Q(\omega,\cdot), \text{ for almost all } \omega}$$

- It is basic for integral representation of invariant measures

- It is not granted. It typically fails if $\{A \in \mathcal{A} : P(A) = 0\} \subset \mathcal{G}$

Theorem 1

Let $\mathcal{G}_n \subset \mathcal{A}$ be a sub-σ-field and Q_n an rcd given $\mathcal{G}_n$.

The rcd Q is 0-1 on $\mathcal{G}$ if

- The "big" σ-field $\mathcal{A}$ is countably generated
- Q_n is 0-1 on $\mathcal{G}_n$ for each n and $\mathcal{G} \subset \limsup_n \mathcal{G}_n$
- $E(1_A|\mathcal{G}_n) \to E(1_A|\mathcal{G})$ a.s. for each $A \in \mathcal{A}$

Note that, by martingale convergence, the last condition is automatically true if the sequence $\mathcal{G}_n$ is monotonic

Examples

Let S be a Polish space and $\Omega = S^\infty$. Theorem 1 applies to

Tail σ-field: $\mathcal{G} = \cap_n \sigma(X_n, X_{n+1}, \ldots)$

where X_n is a sequence of real random variables

Symmetric σ-field:

$\mathcal{G} = \{B \in \mathcal{A} : B = f^{-1}(B)$ for each finite permutation $f\}$

Thus,

Theorem 1 $\Rightarrow$ de Finetti's theorem

Open problem: Theorem 1 does not apply to the **shift-invariant σ-field:**

$\mathcal{G} = \{B \in \mathcal{A} : B = s^{-1}(B)\}$

where $s(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$ is the shift

Disintegrability

Let $\Pi \subset \mathcal{A}$ be a partition of Ω. P is disintegrable on Π if

$$\boxed{P(A) = \int_{\Pi} P(A|H)\, P^*(dH)}$$

for each $A \in \mathcal{A}$, where

- $P(\cdot|H)$ is a probability on $\mathcal{A}$ such that

  $$\boxed{P(H|H) = 1}$$

- P^* is a probability on a suitable σ-field of subsets of Π

Theorem 2

Given a partition Π of Ω, let

$G = \{(x, y) \in \Omega \times \Omega : x \sim y\}$.

Then, P is disintegrable on Π whenever

- $(\Omega, \mathcal{A})$ is nice (e.g. a standard space)
- G is a Borel subset of $\Omega \times \Omega$

Remark: G is actually a Borel set if Π is the partition in the atoms of the tail, or the symmetric, or the shift invariant σ-fields

Remark: The condition on G can be relaxed (e.g., G coanalytic)

Coherent (de Finettian) conditional probabilities

A different notion, introduced by de Finetti, is as follows.

Let

$P(\cdot|\cdot) : \mathcal{A} \times \mathcal{G} \to R$.

For all $n \geq 1$, $c_1, \ldots, c_n \in R$, $A_1, \ldots, A_n \in \mathcal{A}$ and $B_1, \ldots, B_n \in \mathcal{G} \setminus \emptyset$, define

$G(\omega) = \sum_{i=1}^{n} c_i \, 1_{B_i}(\omega) \, \{1_{A_i}(\omega) - P(A_i|B_i)\}$.

Then, $P(\cdot|\cdot)$ is coherent if

$\sup_{\omega \in B} G(\omega) \geq 0 \quad$ where $\quad B = \cup_{i=1}^{n} B_i$.

Such a definition has both merits and drawbacks. In particular, contrary to the classical case:

- The conditioning is now with respect to events,
- $P(B|B) = 1$,
- For fixed B, $P(\cdot|B)$ is "only" a finitely additive probability,
- Disintegrability on Π is not granted, where Π is the partition of Ω in the atoms of $\mathcal{G}$

Bayesian inference

$(\mathcal{X}, \mathcal{E})$ sample space, $(\Theta, \mathcal{F})$ parameter space,

$\{P_\theta : \theta \in \Theta\}$ statistical model,

A **prior** is a probability π on $\mathcal{F}$. A **posterior** for π is any collection $Q = \{Q_x : x \in \mathcal{X}\}$ such that

- Q_x is a probability on $\mathcal{F}$ for each $x \in \mathcal{X}$

- $\int_A Q_x(B)\, m(dx) = \int_B P_\theta(A)\, \pi(d\theta)$

  for all $A \in \mathcal{E}$ $B \in \mathcal{F}$ and for some (possibly finitely additive) probability m on subsets of $\mathcal{X}$

Theorem 3

Fix a measurable function T on $\mathcal{X}$ (a statistic) such that

$P_\theta(T = t) = 0$ for all θ and t.

Under mild conditions, for any prior π, there is a posterior Q for π such that

$$T(x) = T(y) \;\Rightarrow\; Q_x = Q_y$$

Interpretation:

The above condition means that T is **sufficient** for Q. Suppose you start with a prior π, describing your feelings on θ, and a statistic T, describing how different samples affect your inference on θ. Theorem 3 states that, whatever π and T (with $P_\theta(T = t) = 0$) there is a posterior Q for π which makes T sufficient.

Point estimation

The ideas underlying Theorem 3 yield further results. Suppose $\Theta \subset R$ and $d : \mathcal{X} \to \Theta$ is an estimate of θ.

Theorem 4

Under mild conditions, if the prior π is null on compacta, there is a posterior Q for π such that $\int \theta^2 Q_x(d\theta) < \infty$ and

$$\boxed{E_Q(\theta|x) = \int \theta \, Q_x(d\theta) = d(x)}$$

Interpretation:

The above condition means that d is optimal under square error loss. Suppose you start with a measurable map $d : \mathcal{X} \to \Theta$, to be regarded as your estimate of θ. Theorem 4 states that, if the prior π vanishes on compacta, there is a posterior Q for π which makes d optimal

Compatibility

Let $X = (X_1, \ldots, X_k)$ be a k-dimensional random vector and

$X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$

To assess the distribution of X, assign the kernels $Q_1, \ldots, Q_k$, where each Q_i is only requested to satisfy

$Q_i(x, \cdot)$ is a probability for fixed x and

the map $x \mapsto Q_i(x, A)$ is measurable for fixed A

The kernels $Q_1, \ldots, Q_k$ are **compatible** if there is a Borel probability μ on R^k such that

$$\boxed{P_\mu(X_i \in \cdot \,|X_{-i} = x) = Q_i(x, \cdot)}$$

for all i and μ-almost all x.

Such a μ, if exists, should be regarded as the distribution of X

Example: Let $k = 2$ and

$$Q_1(x, \cdot) = Q_2(x, \cdot) = \mathcal{N}(x, 1)$$

This looks reasonable in a number of problems. Nevertheless, Q_1 and Q_2 are not compatible, i.e., no Borel probability on R^2 admits Q_1 and Q_2 as conditional distributions

Compatibility issues arise in: **spatial statistics, statistical mechanics, Bayesian image analysis, multiple data imputation and Gibbs sampling**

Another example are **improper priors**. Given the statistical model $\{P_\theta : \theta \in \Theta\}$, let $Q = \{Q_x : x \in \mathcal{X}\}$ be the "formal posterior" of an improper prior γ (i.e., $\gamma(\Theta) = \infty$). Strictly speaking, Q makes sense only if compatible with the statistical model. In that case, Q agrees with the posterior of some (proper) prior

For $x \in R^k$ and $f \in C_b(R^k)$, let

$$E(f \,|X_{-i} = x_{-i}) = \int f(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_k)\, Q_i(x_{-i}, dt)$$

Theorem 5

Suppose there is a compact set A_i such that

$Q_i(x, A_i) = 1$ for all $x \in R^{k-1}$.

Letting $A = A_1 \times \ldots \times A_k$, suppose also that

$x \mapsto E(f \,|X_{-i} = x_{-i})$ is continuous on A for each $f \in C(A)$

Then, $Q_1, \ldots, Q_k$ are compatible if and only if

$$\boxed{\sup_{x \in A} \sum_{i=1}^{k-1} \{E(f_i \,|X_{-i} = x_{-i}) - E(f_i \,|X_{-k} = x_{-k})\} \geq 0}$$

for all $f_1, \ldots, f_{k-1} \in C(A)$

For each i, fix a (σ-finite) measure λ_i and suppose that

$$\boxed{Q_i(x, dy) = f_i(x, y)\,\lambda_i(dy)} \text{ for all } x \in R^{k-1}$$

Let $\lambda = \lambda_1 \times \ldots \times \lambda_k$ be the product measure

Theorem 6

Suppose $f_i > 0$ for all i. Then, $Q_1, \ldots, Q_k$ are compatible if and only if there are positive Borel functions $u_1, \ldots, u_k$ on R^{k-1} such that

$$\boxed{f_i(x_i\,|x_{-i}) = f_k(x_k\,|x_{-k})\,u_i(x_{-i})\,u_k(x_{-k}),}$$

for all $i < k$ and λ-almost all $x \in R^k$, and

$\int u_k\,d\lambda_{-k} = 1$

Remark: The assumption $f_i > 0$ can be dropped at the price of a more involved statement

An asymptotic result

Let S be a Polish space, (X_n) an **exchangeable** sequence of S-valued random variables, and

$\boxed{\mu_n = (1/n) \sum_{i=1}^n \delta_{X_i}}$ empirical measure

$\boxed{a_n(\cdot) = P(X_{n+1} \in \cdot \,|X_1, \ldots, X_n)}$ predictive measure

Often, a_n can not be evaluated in closed form and μ_n is a reasonable "estimate" of a_n. Here, we focus on the error

$$d(\mu_n, a_n)$$

where d is a distance between probability measures. For instance, if

$$d(\mu_n, a_n) \to 0 \text{ in some sense}$$

then μ_n is a **consistent** estimate of a_n

Fix a class $\mathcal{D}$ of Borel subsets of S and define d as

$$d(\alpha, \beta) = ||\alpha - \beta|| = \sup_{A \in \mathcal{D}} |\alpha(A) - \beta(A)|$$

for all probabilities α and β on the Borel subsets of S

Theorem 7

If $\mathcal{D}$ is a (countably determined) VC-class,

$$\boxed{\limsup_n \sqrt{\frac{n}{\log \log n}}\, ||\mu_n - a_n|| \leq 1/\sqrt{2}} \text{ a.s.}$$

Hence, for any constants r_n,

$$r_n \, ||\mu_n - a_n|| \to 0 \text{ a.s. provided } r_n \sqrt{\frac{\log \log n}{n}} \to 0$$

Remark: If $S = R^k$,

$\mathcal{D} = \{$closed balls$\}$, $\mathcal{D} = \{$half spaces$\}$, and

$\mathcal{D} = \{(-\infty, t] : t \in R^k\}$

are (countably determined) VC-classes

Remark: It is possible to give conditions for

$\sqrt{n}\,||\mu_n - a_n|| \to 0$ in probability

or even for

$n\,||\mu_n - a_n||$ converges a.s. to a finite limit

Example: Let $S = \{0, 1\}$. Then, $\sqrt{n}\,||\mu_n - a_n|| \to 0$ in probability if the prior (i.e, the de Finetti's measure) is absolutely continuous and $n\,||\mu_n - a_n||$ converges a.s. if the prior is absolutely continuous with an almost Lipschitz density