Exchangeability in Gaussian Process Regression

Gian Carlo Di-Luvi

December 10, 2019

1   Background

2 Exchangeability

Throughout this section we assume that all random variables of interest take values in a standard measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, unless otherwise stated. We start by defining exchangeability for sequences of random variables in such a setting.

Definition 2.1 (Exchangeability). *The random variables $X = X_1, X_2, ..., X_n$ are said to be finitely exchangeable if*

$$(X_1, ..., X_n) \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, ..., X_{\pi(n)})$$

for any permutation π of $\mathbb{N}_n := \{1, ..., n\}$.[1] A countable sequence $(X_n)_{n=1}^{\infty}$ of random variables is said to be exchangeable if every finite subsequence of it is finitely exchangeable.

The notion of exchangeability captures a sense of homogeneity in the population, but it is a weaker assumption than independence and identical distribution, as we show in the following example.

> **Example 2.2.** *Let $X_1, X_2, ...$ be i.i.d. random variables with common distribution μ. For $n \in \mathbb{N}$ consider (without loss of generality) the subsequence $X = (X_1, ..., X_n)$. Then the joint distribution μ_n of X is, due to independence, the n-product of μ: $\mu_n = \mu \times \mu \times \cdots \times \mu$. For any permutation π of $\mathbb{N}_n$ it is possible to trivially rearrange the products, and so clearly $X \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, ..., X_{\pi(n)})$. Hence, any i.i.d. sequence is exchangeable.*

The converse of Example 2.2 is not true. (See the suplementary exercises.)

The main result that follows from exchangeability is de Finetti's representation theorem, which has been proved in increasing generality [de 30; HS55; DF78].

Theorem 2.3 (de Finneti). *$X = (X_n)_{n=1}^{\infty}$ is an exchangeable sequence of random variables if and only if there exists a unique probability measure μ on $\mathcal{P}$—the set of all probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$—such that*

$$\mathbb{P}\{X_1 \in A_1, ..., X_n \in A_n\} = \int_{\mathcal{P}} \prod_{i=1}^{n} F(A_i) \, \mu(dF) \tag{1}$$

for every $n \in \mathbb{N}$ and $A_1, ..., A_n \in \mathcal{B}(\mathcal{X})$.

Observe that the integral in Equation (1) is over a set of numerical functions—namely, the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

> **Example 2.4.** *Let $X = X_1, X_2, ...$ be an exchangeable sequence of binary random variables, that is, $\mathcal{X} = \{0, 1\}$. In this case [see Dia88, p.111], de Finetti's theorem asserts the existence of a unique measure μ such that, for every $n \in \mathbb{N}$,*
>
> $$\mathbb{P}\{X_1 = x_1, ..., X_n = x_n\} = \int \mu(dp) \, p^s (1-p)^{n-s},$$
>
> *where $s = \sum_{i=1}^{n} x_i$ is the number of 1's in the sequence $x_1, ..., x_n$. In other words, there exists a random variable p in the unit interval with distribution μ such that, given p, $X_1, ..., X_n$ are i.i.d. Bernoulli random variables with parameter p.*

Technicalities aside, Theorem 2.3 intuitively tells us that, conditional on an unknown distribution F (which is always guaranteed to exist), any subsequence of an exchangeable process can be thought of as a random sample with distribution F. This measure plays the role of an infinite-dimensional parameter [BS94, Ch. 4.3]. Furthermore, F has a unique prior distribution μ, which justifies the use of a Bayesian approach in settings

---

[1]For the sake of completeness, a permutation π of $\mathbb{N}_n$ is simply a bijection in $\mathbb{N}_n$.

with exchangeable data.

However, there are very general settings where even exchangeability is too strong an assumption. Consider the case of a sequence (X_n) of random variables endowed with covariates (t_n), and such that the distribution of each X_i depends on t_i; namely, the setting for regression analysis. It is clear that (X_n) is not an exchangeable sequence, and so de Finetti's representation theorem cannot be immediatly used. Some more general notions of exchangeability to overcome this difficulty have been proposed.

2.1 Partial exchangeability

de Finetti [de 38] introduced the concept of *partial exchangeability* precisely for settings in which the variables of interest have covariates. The idea, which we formalize below, is to require exchangeability only for variables with the same covariate values.

Definition 2.5 (Partial exchangeability). *Let $X = (X_n)_{n=1}^{\infty}$ be a sequence of random variables, each one of them endowed with a covariate $t_n \in \mathcal{T}$. For every $t \in \mathcal{T}$ define the t-class X_t to be the subsequence of X such that all the covariates of X_t are t, that is,*

$$X_t = \{X_n \ : \ t_n = t\}.$$

Then the sequence X is said to be partially exchangeable if all of its classes are exchangeable.

A partially exchangeable sequence can be naturally grouped in an array-like fashion according to the different values that the covariates take. Specifically, suppose that $\mathcal{T} = \{t_1, ..., t_d\}$, so that there are d classes. Then the sequence $\{(X_{t_i, n})_{n=1}^{\infty} \ : \ i = 1, ..., d\}$ is a partially exchangeable sequence if $(X_{t_i, n})_{n=1}^{\infty}$ is exchangeable for each $i = 1, ...d$.

Under the assumption that every class is countably infinite, a representation theorem analogous to 2.3 exists [Cam+19b]. Before giving a more formal statement, we showcase an example.

Example 2.6. *Consider an experiment in which two types of binary observations, $(X_{0, n})_{n=1}^{\infty}$ and $(X_{1, n})_{n=1}^{\infty}$, are under study. For example, $X_{0, i}$ could measure if a male patient recovered (1) or not (0) after taking a certain medication, while $X_{1, i}$ would measure the same outcome but for a female patient. In this case the covariate t_n is also binary, and only indicates whether the patient is male (0) or female (1). Hence, there are two classes: X_0 and X_1, which correspond to male and female patients, respectively.*

If it is deemed that the value of the covariate does not affect the outcome of the variables, then exchangeability of the sequence $X = (X_0, X_1)$ might be a feasible assumption. If this is not so, it might still be feasible to assume exchangeability within male patients and whithin female patients. In that case, X would be partially exchangeable.

Furthermore, for such a partially exchangeable process there exists a measure μ such that, for every $l, r \in \mathbb{N}$,

$$\mathbb{P}\{X_{0, 1} = m_1, ..., X_{0, l} = m_l, X_{1, 1} = f_1, ..., X_{1, r} = f_r\} = \iint \mu(dp_0, dp_1) \, p_0^{s_0} (1 - p_0)^{l - s_0} \, p_1^{s_1} (1 - p_1)^{r - s_1},$$

where $s_0 = \sum_{i=1}^{l} m_i$ and $s_1 = \sum_{i=1}^{r} f_i$ are the number of 1's in the male and female groups, respectively [Dia88, p. 112-113].

We now state the representation theorem for general partially exchangeable sequences. (Adapted from [Cam+19a, p. 69].)

Theorem 2.7. $\{(X_{t_i,\,n})_{n=1}^{\infty} \,:\, i = 1, ..., d\}$ *is a partially exchangeable sequence if and only if there exists a unique probability measure* μ *on* $\mathcal{P}^d$ *such that, for all* $n_1, ..., n_d \in \mathbb{N}$ *and* $A_1, ..., A_d \in \mathcal{B}(\mathcal{X})^{n_i}$,

$$\mathbb{P}\left\{(X_{t_i,\,k})_{k=1}^{n_i} \in A_i \,:\, i = 1, ..., d\right\} = \int_{\mathcal{P}^d} \prod_{i=1}^{d} F_i(A_i)\, \mu(dF_1, ..., dF_d). \tag{2}$$

As per Theorem 2.3, every distribution F_i in Equation (2) is itself a product measure on $\mathcal{X}^{n_i}$.

Finally, observe that partial exchangeability requires access to so-called replicates: at each value of the covariate, an infinite number of response variables can be, at least in theory, obtained (or considered). Although this is true for cases as the one in Example 2.4, it does not always hold.

2.2 Local exchangeability

Campbell et al. [Cam+19b] propose the notion of *local exchangeability* for data with covariates. Both exchangeability and partial exchangeability require some sort of strict invariance under permutations. In this case this is relaxed to allow for some distributional variation under permutations, so long as this variation is bounded and proportional to how close the corresponding covariates are. Interestingly, a representation theorem can still be obtained in such a scenario. Before formalizing these ideas, we define a way to measure variation between distributions.

Definition 2.8 (Total variation distance)**.** *Let* X, Y *be random variables taking values in a measurable space* $(E, \mathcal{E})$. *Then the total variation between* X *and* Y *is*

$$d_{\mathrm{TV}}(X, Y) = \sup_{A \in \mathcal{E}} |\mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\}|.$$

Definition 2.9 (Local exchangeability)**.** *Let* $X = (X_t)_{t \in \mathcal{T}}$ *be a sequence of random variables with covariate space* $\mathcal{T}$ *and* $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ *a pseudometric (distances between different points need not be positive). The process* X *is said to be* f-*locally exchangeable if there exists a function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *continuous at zero and with* $f(0) = 0$ *such that, for every finite subset* $T \subset \mathcal{T}$ *and permutation* π *of* T,

$$d_{\mathrm{TV}}(X_T, X_{\pi T}) \le \sum_{t \in T} f(d(t, \pi(t))). \tag{3}$$

For a very rich discussion of this idea see [Cam+19b], where the authors discuss e.g. the usage of the total variation distance instead of other divergence, and provide some examples of locally-exchangeable processes from the Bayesian nonparametrics literature. In the interest of brevity, we give a very simple example and state the main result, namely, the de Finetti-like representation theorem.

Example 2.10. *Consider a*

3   Gaussian process regression

4   Open questions and research directions

A Exercises

References

[Ald10]   D. J. Aldous. "More uses of exchangeability: representations of complex random structures". In: *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*. Ed. by N. H. Bingham and C. M. Goldie. London Mathematical Society Lecture Note Series. Cambridge University Press, 2010, pp. 35–63.

[Ber96]   J. M. Bernardo. "The concept of exchangeability and its applications". In: *Far East Journal of Mathematical Sciences* 4 (1996), pp. 111–122.

[BS94]   J. M. Bernardo and A. F. Smith. *Bayesian Theory*. 1st ed. Wiley, 1994.

[Cam+19a] F. Camerlenghi et al. "Distribution Theory for Hierarchical Processes". In: *The Annals of Statistics* 47.1 (2019), pp. 67–92.

[Cam+19b] T. Campbell et al. "Local Exchangeability". In: *arXiv e-prints*, arXiv:1906.09507 (2019), arXiv:1906.09507. arXiv: 1906.09507 [math.ST].

[de 30]   B. de Finetti. "Funzione caratteristica di un fenomeno aleatorio". In: *R. Academia Nazionale dei Lince* 4.1 (1930), pp. 86–133.

[de 38]   B. de Finetti. "Sur la condition d'equivalence partielle". In: *Actualites Scientifiques et Industrielles* 739 (1938). In French; translated as "On the condition of partial exchangeability," P. Benacerraf and R. Jeffrey (eds) in *Studies in Inductive Logic and Probability II*, 193-205, Berkeley, University of California Press, 1980.

[Dia88]   P. Diaconis. "Recent Progress on de Finetti's Notions of Exchangeability". In: *Bayesian Statistics* 3 (1988), pp. 111–125.

[DF78]   P. Diaconis and D. Freedman. *de Finetti's Generalizations of Exchangeability*. Tech. rep. 109. Stanford University, 1978.

[Fra18]   P. I. Frazier. "A Tutorial on Bayesian Optimization". In: *arXiv e-prints*, arXiv:1807.02811 (2018), arXiv:1807.02811. arXiv: 1807.02811 [stat.ML].

[HS55]   E. Hewitt and L. J. Savage. "Symmetric Measures on Cartesian Products". In: *Trans. Amer. Math. Soc.* 80 (1955), pp. 470–501.

[Kal02]   O. Kallenberg. *Foundations of Modern Probability*. 2nd. Springer-Verlag New York, 2002.

[Kin78]   J. F. C. Kingman. "Uses of Exchangeability". In: *The Annals of Probability* 6.2 (1978), pp. 183–197. URL: http://dx.doi.org/10.1214/aop/1176995566.

[McC05]   P. McCullah. "Exchangeability and regression models". In: *Celebrating Statistics: Papers in honour of Sir David Cox on his 80th birthday*. Oxford": Oxford University Press, 2005. Chap. 4, pp. 89–114.

[OR15]   P. Orbanz and D. M. Roy. "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pp. 437–461.

[PD16]   G. Pokharel and R. Deardon. "Gaussian process emulators for spatial individual-level models of infectious disease". In: *The Canadian Journal of Statistics* 44.4 (2016), pp. 480–501.

[RW06]   C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. 1st ed. The MIT Press, 2006.

[Woo+17]   D. C. Woods et al. "Bayesian design of experiments for generalized linear models and dimensional analysis with industrial and scientific application". In: *Quality Engineering* 29.1 (2017), pp. 91–103.