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*

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

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 $X \overset{\mathrm{d}}{=} (X_{\pi(1)}, ..., X_{\pi(n)})$. 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 dependence 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. 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*

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

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*,*

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

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 measures, and provide some examples of locally-exchangeable processes from the Bayesian nonparametrics literature.

We now state the main result for local exchangeability. The idea behind it is that, so long as the covariate space $\mathcal{T}$ is "nice," a de Finetti-like representation of f-locally exchangeable sequences exists in terms of a stochastic process G, conditional on which the sequence exhibits independence (Equation 4). Furthermore, the function f controls the "smoothness" behaviour of G (Equation 5).

Theorem 2.10 (Campbell et al. (2019))**.** *Let* $X = (X_t)_{t \in \mathcal{T}}$ *be a stochastic process on a separable space* $\mathcal{T}$*, which furthermore has no isolated points under the pseudometric* d*. Then* X *is* f*-locally exchangeable if and only if there exists a random measure-valued stochastic process* $G = (G_t)_{t \in \mathcal{T}}$ *such that, for any finite subset of covariates* $T \in \mathcal{T}$ *and permutation* π *of* T*,*

$$\mathbb{P}\{X_T \in \cdot \,|\, G\} \overset{\text{a.s.}}{=} \prod_{t \in \mathcal{T}} G_t := G_T \tag{4}$$

and

$$\sup_{A} \mathbb{E}\left|G_T(A) - G_{\pi T}(A)\right| \leq \sum_{t \in \mathcal{T}} f(d(t, \pi(t))). \tag{5}$$

Furthermore, G *is unique up to modification, that is, if* G' *also satisfies Equations (4) and (5) then* $\mathbb{P}\left\{G_t = G_t^{'}\right\} = 1$ *for all* $t \in \mathcal{T}$*.*

Local exchangeability manages to relax the requirements of exchangeability and partial exchangeability while still preserving a representation result. However, in doing so, the cost it pays is an increased complexity in the calculations involved. Where most processes can be easily determined to be either (partially) exchangeable or not—sometimes even by construction—actually proving a sequence to be locally exchangeable is not an easy feat. [Cam+19b, Proposition 3] provides necessary conditions which make this task easier, but only marginally so.

2.3 Regression exchangeability

McCullagh [McC05] proposed yet another notion of exchangeability. Unlike the previous ideas so far discussed, McCullagh aims not for generality but for a definition that works well specifically in a regression setting, appealing to the idea that exchangeability should capture a sense of homogeneity.

Definition 2.11 (Regression exchangeability)**.** *Let $X = (X_n)_{n=1}^{\infty}$ be a sequence of random variables, each one of them endowed with a covariate $t_n \in \mathcal{T}$. The sequence X is said to be regression exchangeable (modulo $T = (t_n)$) if given two arbitrary subsets $T_1, T_2 \subset T$ of the covariate space the following two conditions hold:*

1. *If $T_1 \subset T_2$ then the distribution of X_{T_1} must be the marginal distribution of X_{T_2} under co-ordinate deletion.*
2. *If $T_1 = T_2$ then $X_{T_1} \overset{\mathrm{d}}{=} X_{T_2}$.*

Condition 1 in Definition 2.11 simply ensures compatibility with respect to subsampling from X, and in the context of Gaussian process regression it is known as *marginalization* property [RW06, p. 13]. Condition 2 may seem trivial, but observe that X_{T_1} and X_{T_2} may very well be different. However, so long as their covariates are the same, any distinction between the actual values within X_{T_1} and X_{T_2} has no effect on their distribution. We showcase this with an example.

Example 2.12. *Let $X = X_1, X_2, ...$ be independent random variables such that $X_i \sim \mathcal{N}(\eta + \tau_{t_i}, 1)$, where $T = (t_n)_{n=1}^{\infty} = (1, 2, 3, 1, 2, 3, ...)$. X, which can be thought of as the response of an experiment with one factor and three levels, has independent components, but is nonetheless not exchangeable. However, X is clearly regression exchangeable (modulo T): given $T_1, T_2 \subset T$, X_{T_1} is the sample of such an experiment and follows a multivariate Normal distribution with covariance matrix $\sigma^2 I_{|T_1|}$ and mean vector $(\eta + \tau_{t_i})_{t_i \in T_1}$, and similarly with T_2. Clearly if $T_1 \subset T_2$ then the distribution of T_1 is obtained by "removing" the covariates in $T_2 \setminus T_1$. Furhermore, if $T_1 = T_2$ then (even if the actual X_i's selected are different) $X_{T_1} \overset{\mathrm{d}}{=} X_{T_2}$.*

Example 2.12 works well due to the availability of replicates. However, unlike partial exchangeability (Example 2.4), a process may not have replicates at all and still be regression exchangeable, whereas it would not be partially so. However, it is worth noting that Condition 2 in Definition 2.11 does reduce to a triviality in such a setting: the only way $T_1 = T_2$ would be if $X_{T_1} = X_{T_2}$ exactly.

To the best of our knowledge, there is no de Finetti-like representation theorem available for regression exchangeability.

3 Gaussian process regression

Gaussian process regression arises when studying functions which are computationally expensive to evaluate. The general idea is to assume that the function of interest f is a stochastic process and use a sample of values of f to estimate the function in unobserved values of the domain. We start by defining Gaussian processes, which are the building blocks of this idea.

Definition 3.1 (Gaussian process)**.** *A stochastic process $X = (X_t)_{t\in\mathcal{T}}$ is said to be a Gaussian process (GP) if, for any finite subset $T \subset \mathcal{T}$, X_T follows a Normal distribution.*

Remark. A Gaussian process is entirely determined by its mean m and covariance κ functions. Formally, $m : t \mapsto \mathbb{E}[X_t]$ and $\kappa : (t, t') \mapsto \operatorname{Cov}(X_t, X_{t'})$ and we write $X \sim \operatorname{GP}(m, \kappa)$. Commonly, m is assumed to be zero and κ is chosen from some parametric family of functions, many of which have been thoroughly studied in the literature [see RW06, Ch. 4].

Now we formalize GP regression. Consider a function $f : \mathcal{T} \to \mathbb{R}$ and denote $X_t := f(t)$ for all $t \in \mathcal{T}$. We assume that $X = (X_t) \sim \operatorname{GP}(m, \kappa)$. Usually, $\mathcal{T}$ will be a subset of $\mathbb{R}^n$. Assuming that $m = 0$ and that we have access to observations $(X_t, t)_{t\in T}$, where T is a finite subset of $\mathcal{T}$, then

$$X_T \sim \mathcal{N}(0, K(T, T)),$$

where K is a $|T| \times |T|$ matrix with entries $K_{ij} = \kappa(t_i, t_j)$. If we want to predict the value of the function $f(t^*)$, we again use the fact that

$$\begin{pmatrix} X_T \\ X_{t^*} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(T,T) & K(T,t^*) \\ K(t^*,T) & \kappa(t^*,t^*) \end{pmatrix}\right),$$

from where, using basic properties of the Normal distribution,

$$X_{t^*} \mid X_T \sim \mathcal{N}\left(K(t^*,T)K(T,T)^{-1}X_T,\ \kappa(t^*,t^*)K(t^*,T)K(T,T)^{-1}K(T,t^*)\right). \qquad (6)$$

Commonly, the conditional mean in Equation (6) is used as a point estimate of $f(t^*)$; the estimated variance can be used to e.g. compute confidence bands. Also, observe that we could have well chosen t^* to have more than one component if we were interested in values of f only at specific points in the covariate space. The advantage of Equation (6) is that it provides point estimates for *any* point in $\mathcal{T}$. Figure 1 showcases this process.

It is easy to extend this idea for cases in which, rather than having access to exact values of the function f, only estimates with some noise are available. This is the case, for example, when f is simply impossible to evaluate, but can be reliably estimated via Monte Carlo methods—e.g. when f is an expected loss over a high-dimensional parameter space. In these situations an additive noise term ξ is commonly added to the covariance function κ.

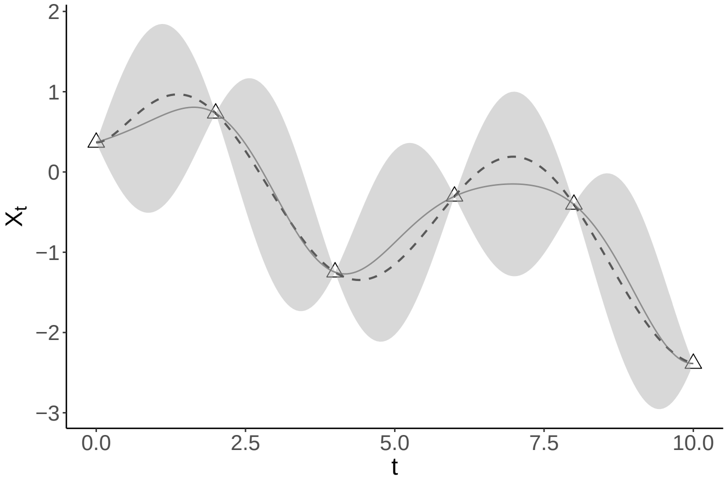

Figure 1: Gaussian process regression used to estimate a function f (shown in dotted grey lines). The prediction is based on the conditional mean of the GP, which is shown in a line and was fitted based on 8 observations of f (the triangle shapes), along with 95% confidence bands.

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.