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 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*,*

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

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. Proposition 3.4 provides sufficient 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 the *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

Consider a sequence of random variables $X = (X_t)_{t\in\mathcal{T}}$ endowed with covariates in the space $\mathcal{T}$ and consider a finite subset X_T, $T \subset \mathcal{T}$ of them. Suppose we are interested in studying X_{t^*} for $t^* \notin T$, given X_T. A natural estimator would be $\mathbb{E}[X_{t^*} \,|\, \sigma X_T]$, which of course depends on the distributional and dependence structures of the process X. If X is a Gaussian process, which we define below, this expression can be easily computed and the methodology just discusses is called Gaussian process regression.

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 $\kappa : (t,t') \mapsto \mathrm{Cov}(X_t, X_{t'})$ and we write $X \sim \mathrm{GP}(m,\kappa)$. 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]. Unless otherwise stated, we use a squared exponential covariance function,

$$\kappa(t,t') = \kappa(|t-t'|) = \exp\left\{-\frac{1}{2\ell^2}|t-t'|^2\right\}. \tag{6}$$

In practice, GP regression commonly arises when studying functions which are computationally expensive to evaluate. The general idea is to assume that the function of interest f is a GP and use a sample of values of f to estimate the function in unobserved values of the domain. Formally, consider a function $f : \mathcal{T} \to \mathbb{R}$ and denote $X_t := f(t)$ for all $t \in \mathcal{T}$. We assume that $X = (X_t) \sim \mathrm{GP}(m,\kappa)$. Usually, $\mathcal{T}$ will be a subset of $\mathbb{R}^n$, $m = 0$ and we will have access to observations $(X_t, t)_{t\in T}$, where T is a finite subset of $\mathcal{T}$. If this is the case,

$$X_T \overset{\mathrm{d}}{=} \mathcal{N}(0, K(T,T)), \tag{7}$$

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

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

from where, using basic properties of the Normal distribution,

$$\mathbb{P}[X_{t^*} \,|\, \sigma X_T] \overset{\mathrm{d}}{=} \mathcal{N}\left(K(t^*,T)K(T,T)^{-1}X_T,\ \kappa(t^*,t^*)K(t^*,T)K(T,T)^{-1}K(T,t^*)\right). \tag{8}$$

Equation (8) gives us not only the conditional expectation discussed earlier, but the whole conditional distribution: $\mathbb{P}[X_{t^*} \,|\, \sigma X_T]$ is a Normally-distributed random variable. Commonly, the conditional mean in Equation (8) is used as a point estimate $\tilde{f}$ of f:

$$\tilde{f}(t^*) = \mathbb{E}[X_{t^*} \,|\, \sigma X_T] = K(t^*,T)K(T,T)^{-1}X_T. \tag{9}$$

The estimated variance can be used to e.g. compute (conditional) 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 (9) 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 σ^2 is commonly added to the covariance function κ for observations with the same covariates: $\kappa_\sigma(t,t') = \kappa(t,t') + \sigma^2\,\mathbb{1}(t=t')$, where κ is a regular noise-free covariance function. For any finite subset $T \subset \mathcal{T}$, this can be represented as

$$K_\sigma(X_T, X_T) = K(X_T, X_T) + \sigma^2\, I_{|T|},$$

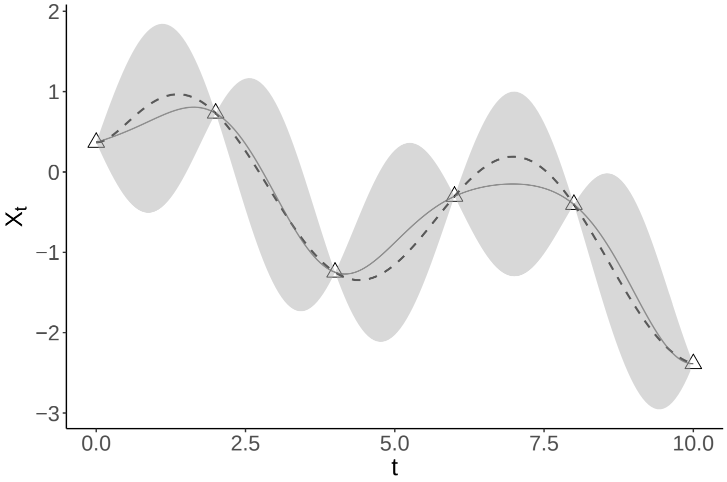

Figure 1: Gaussian process regression used to estimate a function $f = X_t$ (shown in dotted grey lines). The prediction is based on the conditional mean of a GP with squared exponential function with $\ell = 1$, which is shown in a line and was fitted based on 6 observations of f (the triangle shapes). 95% confidence bands computed using the conditional variance are also shown.

so that

$$X_T \overset{\mathrm{d}}{=} \mathcal{N}\left(0,\ K(T,T) + \sigma^2\, I_{|T|}\right). \tag{10}$$

The resulting conditional distribution for a new covariate value t^* given T is

$$\mathbb{P}[X_{t^*} \,|\, \sigma X_T] \overset{\mathrm{d}}{=} \mathcal{N}\left(K(t^*,T)\left(K(T,T) + \sigma^2\, I_{|T|}\right)^{-1} X_T,\ \kappa(t^*,t^*)K(t^*,T)\left(K(T,T) + \sigma^2\, I_{|T|}\right)^{-1} K(T,t^*)\right).$$

3.1 Exchangeability of Gaussian process regression

We now study exchangeability in the context of Gaussian process regression. Let $X = (X_t)_{t\in\mathcal{T}} \sim \mathrm{GP}(0,\kappa)$. The process X is said to be noise-free if Equation (7) holds and σ-noisy if Equation (10) holds for any finite subset $T \subset \mathcal{T}$. We assume that κ is given by Equation (6) with $\ell = 1$, possibly with a modification to account for noisy observations if necessary.

Proposition 3.2. *Let $X = (X_t)_{t\in\mathcal{T}} \sim \mathrm{GP}(0,\kappa)$. Then X is not exchangeable.*

Proof. Conceptually this is trivially true, and is due to the fact that the covariance function κ—which completely determines the process—is sensible to the distance between covariates. The order of the covariates matters, and so the sequence cannot be exchangeable. Formally, let $T = \{t_1, t_2, t_3\}$ with $t_2 = t_1+1, t_3 = t_1+2$ and π be a permutation of T such that $\pi(t_1) = t_2$, $\pi(t_2) = t_1$, and $\pi(t_3) = t_3$. Then $X_T = \mathcal{N}(0, K(T,T))$ and $X_{\pi T} = \mathcal{N}(0, K(\pi T, \pi T)$. If X were exchangeble, then $X_T \overset{\mathrm{d}}{=} X_{\pi T}$, and particularly $\kappa(t_1,t_3) = \kappa(t_2,t_3)$, which is clearly not the case because $\kappa(t_1,t_3) = e^{-1}$ but $\kappa(t_2,t_3) = e^{-1/2}$, regardless of whether the process is noisy or not.

□

One may argue that this result is more due to the specific covariance function we chose rather than some underlying principle, but this is not so: the covariance function *has* to be sensible to some notion of distance betweem covariates because we assume that close covariates will yield somewhat similar values of the function. The prediction of the process at t^* includes information from all observations in T, but naturally those values in T closer to t^* will have a greater weight on the estimate $\mathbb{E}[X_{t^*} \,|\, \sigma X_T]$.

The exchangeability requirement in Proposition 3.2 can be weakened to yield a positive result, but first we follow [Cam+19b] and introduce the following notation. Let $X = (X_t)_{t\in\mathcal{T}}$ be a noisy Gaussian process with

$\mathcal{T} = \mathbb{R}$. If it is possible to obtain multiple observations at a single value, then it makes sense to modify $\mathcal{T}$ by $\tilde{\mathcal{T}} = \mathbb{R} \times \mathbb{N}$, where the first entry would denote the location of the observation and the second one the number of observation. So $\tilde{t} = (0.56, 2)$ would refer to the second observation at the value $t = 0.56$. In such a scenario we use both $\mathcal{T}$ and $\tilde{\mathcal{T}}$ interchangeably depending on the context; particularly when dealing with local exchangeability, using $\tilde{\mathcal{T}}$ is more convenient.

Proposition 3.3. *Let $X = (X_t)_{t \in \mathcal{T}} \sim \mathrm{GP}(0, \kappa)$ be a σ-noisy Gaussian process. Then X is partially exchangeable.*

Proof. The idea is to group the process into sequences with the same covariate value and prove that each of those is exchangeable. Formally, let $X_{t_0} = (X_{t_0,\,n})_{n=1}^{\infty}$ be the t_0-class of X for each (distinct) $t_0 \in \mathcal{T}$. Let $N = \{n_1, ..., n_r\}$ and π a permutation of N. The key observation here is that $\kappa(t_0, t_0) = 0$, and so

$$X_{t_0\,N} \overset{\mathrm{d}}{=} \mathcal{N}(0, \sigma^2 I_r) \overset{\mathrm{d}}{=} X_{t_0\,\pi N}. \tag{11}$$

This proves that each t_0-class is exchangeable, and thus the whole process is partially exchangeable.

□

A good question is whether a noise-free GP is partially exchangeable. We argue to the contrary, asserting—as in the previous section—that partial exchangeability is defined with replicates in mind. If a function can be precisely estimated (even if expensively so) at any given value t, it makes no sense to obtain replicates: there is no within-covariate variability. Indeed, if one attempts to replicate the previous proof then Equation (11) still holds, albeit with degenerate Normals that have zero variance.

In preparation for the main local-exchangeability result for GP regression, we state a sufficiency condition for determining local exchangeability of a sequence. For a proof, see [Cam+19b, Proposition 3].

Proposition 3.4. *Consider a process $(X_t)_{t \in \mathcal{T}}$ and suppose there exists a σ-algebra $\mathcal{G}$ such that, for any finite subset $T \subset \mathcal{T}$ of covariates, the components of X_T are conditionally independent given $\mathcal{G}$. Let $G_t = \mathbb{P}[X_t \,|\, \mathcal{G}]$. Then X is f-locally exchangeable if, for all pairs $t, t' \in \mathcal{T}$,*

$$\mathbb{E}[d_{\mathrm{TV}}(G_t, G_{t'})] \leq f(d(t, t')). \tag{12}$$

Now we assert that σ-noisy GPs are f-locally exchangeable, where f is inversely proportional to the noise factor σ.

Proposition 3.5. *Let $f(x) = x/\pi\sigma$ and $X = (X_t)_{t \in \mathbb{R}} \sim \mathrm{GP}(0, \kappa)$ be a σ-noisy Gaussian process in $\mathbb{R}$, which we furthermore endow with the metric $d(t, t') = |t - t'|$. Then X is f-locally exchangeable.*

Proof. We prove that the requirements of Proposition 3.4 are met. First observe that

$$\kappa(t, t') = \exp\left\{-\frac{1}{2}|t - t'|^2\right\} \geq 1 - \frac{1}{2}|t - t'|^2 = 1 - \frac{1}{2}d(t, t')^2,$$

which can be proved via Taylor's expansion of $\exp\left\{-\frac{1}{2}y^2\right\}$. This can also be rewritten as

$$1 - \kappa(t, t') \leq \frac{1}{2}d(t, t')^2. \tag{13}$$

Now let $T \subset \mathbb{R}$ be a finite subset of covariates of size n. Denoting $K(T, T) = K$, by Equation (10) we have that $X_T \overset{\mathrm{d}}{=} \mathcal{N}(0, K + \sigma^2 I_n)$. Now let $Y_T \overset{\mathrm{d}}{=} \mathcal{N}(0, K)$ and observe that

$$\mathbb{P}[X_T \,|\, \sigma Y_T] \overset{\mathrm{d}}{=} N(Y_T, \sigma^2 I_n). \tag{14}$$

Thus, for every $t \in T$ the conditional distribution of X_t given Y_t is $G_t \overset{\mathrm{d}}{=} \mathcal{N}(Y_t, \sigma^2)$. We need only prove that $\mathbb{E}[d_{\mathrm{TV}}(G_t, G_{t'})] \leq f(d(t, t'))$. But, as per [Cam+19b, p. 15], the total variation between two Normal distributions can be expressed in terms of the cumulative distribution function Φ of a standard Normal distribution, and so

$$\begin{aligned}
d_{\mathrm{TV}}(G_t, G_{t'}) &= d_{\mathrm{TV}}(\mathcal{N}(Y_t, \sigma^2), \mathcal{N}(Y_{t'}, \sigma^2)) \\
&= \Phi\left(\frac{|Y_t - Y_{t'}|}{2\sigma}\right) - \Phi\left(-\frac{|Y_t - Y_{t'}|}{2\sigma}\right) \\
&\leq \frac{|Y_t - Y_{t'}|/2\sigma + |Y_t - Y_{t'}|/2\sigma}{\sqrt{2\pi}} && \text{(by Lemma B.1)} \\
&= \frac{|Y_t - Y_{t'}|}{\sqrt{2\pi\sigma^2}}. && (15)
\end{aligned}$$

Now, observe that $\kappa(t,t) = 1$ for all $t \in \mathbb{R}$, and so $Y_t, Y_{t'} \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$. Furthermore, $\operatorname{cov}(Y_t, Y_{t'}) = \kappa(t,t')$, and so by Lemma B.2

$$\mathbb{E}|Y_t - Y_{t'}| = \sqrt{\frac{4}{\pi}(1 - \kappa(t,t'))}.$$

Substituting in Equation (15) we deduce

$$\mathbb{E}[d_{\mathrm{TV}}(G_t, G_{t'})] \leq \sqrt{\frac{2}{\pi^2\sigma^2}(1 - \kappa(t,t'))}. \tag{16}$$

Using κ's lower bound (Equation 13) we finally get that, for all t, t',

$$\mathbb{E}[d_{\mathrm{TV}}(G_t, G_{t'})] \leq \sqrt{\frac{1}{\pi^2\sigma^2}d(t,t')^2} = \frac{d(t,t')}{\pi\sigma}. \tag{17}$$

The result follows from Proposition 3.4 with $f(x) = \frac{x}{\pi\sigma}$.

□

It is clear from this proof

4 Open questions and research directions

A Exercises

Exercise A.1 (Adapted from [Kin78]). *Let $X = (X_n)_{n=1}^{\infty}$ be an exchangeable sequence of random variables taking values in a measurable space $(\mathcal{X}, \mathcal{F})$ with common expectation. A random variable $Y = g(X)$ is said to be n-symmetric if it is bounded and invariant under permutations of its first n components, that is, if*

$$g(X_1, X_2, ...) = g(X_{\pi(1)}, ..., X_{\pi(n)}, X_{n+1}, ...) \tag{18}$$

for every permutation π of $\mathbb{N}_n$. Define $\mathcal{F}_n$ to be the smallest σ-algebra with respect to which all n-symmetric random variables are measurable. Let f be an integrable numerical $\mathcal{F}$-measurable function, that is, $\mathbb{E}|f(X_n)| < \infty$ for all n. Prove that $\frac{1}{n}\sum_{j=1}^{n} f(X_j)$ is a version of $\mathbb{E}[f(X_1) \,|\, \mathcal{F}_n]$ for every n.

Solution A.1. Fix $n \in \mathbb{N}$. Because the components of X have common expectation, clearly $\mathbb{E}[f(X_1)] = \mathbb{E}[f(X_j)]$ for every $1 \leq j \leq n$ and, moreover, $\mathbb{E}[f(X_1)g(X)] = \mathbb{E}[f(X_j)g(X)]$ for every $g \in \mathcal{F}_n$. Thus,

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} f(X_j)g(X)\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[f(X_j)g(X)] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[f(X_1)g(X)] = \mathbb{E}[f(X_1)g(X)],$$

and so $\frac{1}{n}\sum_{j=1}^{n} f(X_j)$ satisfies the projection property. Furthermore, it is clearly n-symmetric as a function of X, and hence is in $\mathcal{F}_n$, which completes the proof.

□

Observe also that $\mathcal{F}_{n+1} \subset \mathcal{F}_n$: if a function is $n+1$-symmetric it is also n-symmetric. A reverse-martingale argument [see Theorem B.124 (Lévy's theorem: part II) in Sch95, p. 650] can be used to show that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} f(X_j) = \mathbb{E}[f(X_1) \,|\, \mathcal{F}_\infty] \quad \text{a.s.}, \tag{19}$$

where

$$\mathcal{F}_\infty = \bigcap_{n=1}^{\infty} \mathcal{F}_n.$$

Equation (19) is a version of the Strong Law of Large Numbers for exchangeable sequences.

B Additional technical results

Lemma B.1. *Let Φ and ϕ denote the cdf and pdf of a standard Normal distribution, respectively. Then, for $x_1 \leq x_2$,*

$$\Phi(x_2) - \Phi(x_1) \leq \frac{x_2 - x_1}{\sqrt{2\pi}}.$$

In other words, $1/\sqrt{2\pi}$ is a Lipschitz constant of Φ.[2]

Proof. Recall that $\phi(x) \leq \phi(0) = 1/\frac{1}{2\pi}$, from where trivially

$$\Phi(x_2) - \Phi(x_1) = \int_{x_1}^{x_2} du\, \phi(u) \leq \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} du = \frac{x_2 - x_1}{\sqrt{2\pi}}.$$

□

Lemma B.2. *Let $X_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(0, \sigma_2^2)$ be Normal random variables with covariance σ_{12}. Then $\mathbb{E}|X_1 - X_2| = \sqrt{\frac{2}{\pi}(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})}$.*

Proof. We have that $X_1 - X_2 \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$, and so $W := |X_1 - X_2| \sim \mathrm{HN}(\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}})$, where HF refers to the Half-Normal distribution. Indeed, it is well known that if $U \sim \mathrm{HN}(\xi)$ then $\mathbb{E}[U] = \sqrt{2/\pi}\xi$, from where result follows.

□

[2] Actually, it is *the* Lipschitz constant of Φ, although we only prove the weaker result here stated for the sake of brevity.

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.

[Sch95] M. Schervish. *Theory of Statistics*. 1st ed. Springer, 1995.

[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.