Bernoulli **29**(3), 2023, 2084–2100
https://doi.org/10.3150/22-BEJ1533

Local exchangeability

TREVOR CAMPBELL[1,a], SAIFUDDIN SYED[1,b], CHIAO-YU YANG[2,c], MICHAEL I. JORDAN[2,d] and TAMARA BRODERICK[3,e]

[1]*Department of Statistics, University of British Columbia, Vancouver, Canada,* [a]*trevor@stat.ubc.ca,* [b]*saif.syed@stat.ubc.ca*
[2]*Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, USA,* [c]*chiaoyu@berkeley.edu,* [d]*jordan@eecs.berkeley.edu*
[3]*Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, USA,* [e]*tbroderick@mit.edu*

Exchangeability—in which the distribution of an infinite sequence is invariant to reorderings of its elements—implies the existence of a simple conditional independence structure that may be leveraged in the design of statistical models and inference procedures. In this work, we study a relaxation of exchangeability in which this invariance need not hold precisely. We introduce the notion of *local exchangeability*—where swapping data associated with nearby covariates causes a bounded change in the distribution. We prove that locally exchangeable processes correspond to independent observations from an underlying measure-valued stochastic process. Using this main probabilistic result, we show that the *local empirical measure* of a finite collection of observations provides an approximation of the underlying measure-valued process and Bayesian posterior predictive distributions. The paper concludes with applications of the main theoretical results to a model from Bayesian nonparametrics and covariate-dependent permutation tests.

Keywords: Exchangeability; local; representation; de Finetti; Bayesian nonparametrics

1. Introduction

Let $X = X_1, X_2, \ldots$ be an infinite sequence of random elements in a standard Borel space $(\mathcal{X}, \Sigma)$. The sequence is said to be *exchangeable* if for any finite permutation π of $\mathbb{N}$,

$$X_1, X_2, \ldots \overset{d}{=} X_{\pi(1)}, X_{\pi(2)}, \ldots.$$

At first sight this assumption appears innocent; intuitively, it suggests only that the order in which observations appear provides no information about those or future observations. But despite its apparent innocence, exchangeability has a powerful implication. In particular, the well-known *de Finetti's theorem* (e.g. Kallenberg, 2002, Theorem 11.10) states that an infinite sequence is exchangeable if and only if it is mixture of i.i.d. sequences, i.e., there exists a unique random probability measure G on $\mathcal{X}$ such that

$$\mathbb{P}(X \in \cdot \mid G) \overset{a.s.}{=} G^\infty, \tag{1}$$

where G^∞ is the countable infinite product measure constructed from G. Thus, exchangeability provides a strong justification for the Bayesian approach to modeling (Jordan, 2010), and guarantees a latent conditional independence structure of X useful in the design of computationally efficient inference algorithms. Exchangeability is also the basis of well-known nonparametric permutation testing procedures (Pitman, 1937a,b,c; Fisher, 1966, Ch. 3; Ernst, 2004; Lehmann and Romano, 2005, Ch. 15).

However, although exchangeability may be a useful idealization in modeling and analysis, many data come with covariates that preclude an honest belief in its validity. For example, given a corpus of documents tagged by publication date, one might reasonably expect the data to exhibit a time-dependence

that is incompatible with exchangeability. Nevertheless, one might still expect the distribution not to change too much if we permuted documents published only one day apart; i.e., observations with similar covariates are intuitively "nearly exchangeable." In this work, we investigate how to codify this intuition.

One option is to use a kind of *partial exchangeability* (Camerlenghi et al., 2019, de Finetti, 1938, Diaconis and Freedman, 1978, Le Cam, 1964) in which the distribution is invariant to permutations within equivalence classes. Formally, we endow each observation X_n with a covariate t_n from a set $\mathcal{T}$, and assert that the sequence distribution is invariant only to reordering observations with equivalent covariate values. Under this assumption as well as the availability of infinitely many observations at each covariate value, we have a similar representation of X as a mixture of independent sequences given random probability measures $(G_t)_{t\in\mathcal{T}}$,

$$\mathbb{P}\left(X \in \cdot \,|\, (G_t)_{t\in\mathcal{T}}\right) \overset{a.s.}{=} \prod_{n=1}^{\infty} G_{t_n}. \tag{2}$$

The random probability measures $(G_t)_{t\in\mathcal{T}}$ can have an arbitrary dependence on one another; partially exchangeable sequences encompass those that are exchangeable (where the covariate does not matter), decoupled (where subsequences for each different covariate value are mutually independent), and the full range of models in between. In particular, partial exchangeability does not enforce the desideratum that observations with nearby covariates should have a similar law, and is too weak to be useful for restricting the class of underlying mixing measures for the data.

In this work, we introduce a new notion of *local exchangeability*—lying between partial and exact exchangeability—in which swapping data associated with nearby covariates causes a bounded change in total variation distance. We begin by studying probabilistic properties of locally exchangeable processes in Sections 2.1 and 2.2. The main result from this section is in the spirit of de Finetti's theorem: we prove that locally exchangeable processes correspond to independent observations from a unique underlying smooth measure-valued stochastic process. To the best of our knowledge, this representation theorem is the first to arise from an approximate probabilistic symmetry. Further, the existence of such an underlying process not only shows that de Finetti's theorem is robust to perturbations away from exact exchangeability, justifying the Bayesian analysis of real data, but also imposes a useful constraint on the space of models one should consider when dealing with data that one suspects follows a locally exchangeable random process. Next in Section 2.3, we use this result to show that the *local empirical measure* of a finite collection of observations can be used to provide an approximation of the underlying measure-valued process, Bayesian predictive posterior distributions, and the premetric that governs local exchangeability. These results rely heavily on the intuition that locally exchangeable observations from nearby covariates behave essentially like exchangeable observations. Finally, in Section 3, we provide example applications in two statistical models exhibiting local exchangeability—Gaussian processes (Rasmussen and Williams, 2006) and dependent Dirichlet processes (MacEachern, 1999, 2000)—as well as grouped permutation tests in the presence of covariates. The paper concludes with a discussion of directions for future work. Proofs of all results are provided in the supplementary material (Campbell et al., 2023).

1.1. Related work

Beyond de Finetti's original result for infinite binary sequences (de Finetti, 1931) and its extensions to more general range spaces (de Finetti, 1937, Hewitt and Savage, 1955) and finite sequences (Diaconis, 1977, Diaconis and Freedman, 1980b)—see Aldous (1985) for an in-depth introduction—

correspondences between probabilistic invariances and conditional latent structure (known as *representation theorems*) have been studied extensively. Notions of exchangeability and corresponding latent conditional structure now exist for a wide variety of probabilistic models, such as arrays (Aldous, 1981, Austin and Panchenko, 2014, Hoover, 1979, Jung et al., 2021), Markov processes (Diaconis and Freedman, 1980c), networks (Borgs et al., 2017, Cai, Campbell and Broderick, 2016, Caron and Fox, 2017, Crane and Dempsey, 2016, Janson, 2018, Veitch, 2015), combinatorial structures (Broderick, Pitman and Jordan, 2013, Campbell, Cai and Broderick, 2018, Crane and Dempsey, 2019, Kingman, 1978, Pitman, 1995), random measures (Kallenberg, 1990), and more (Diaconis, 1988, Kallenberg, 2005, Orbanz and Roy, 2015). Furthermore, weaker notions of exchangeability such as conditionally identical distributions (Berti, Pratelli and Rigo, 2004, Kallenberg, 1988) have been developed. All past work on probabilistic invariance and its consequences has pertained to exact invariance.

2. Local exchangeability

2.1. Definition

Let $X = (X_t)_{t\in\mathcal{T}}$ be a stochastic process on an index (or covariate) set $\mathcal{T}$ taking values in a standard Borel space $(\mathcal{X},\Sigma)$. To encode distance between covariates, we endow the set $\mathcal{T}$ with a *premetric* $d : \mathcal{T}\times\mathcal{T} \to [0,1]$ satisfying $d(t,t') = d(t',t)$ and $d(t,t) = 0$ for $t,t' \in \mathcal{T}$. We will formalize local exchangeability based on the finite dimensional projections of X. For any subset $T \subset \mathcal{T}$ and injection $\pi : T \to \mathcal{T}$, let X_T and $X_{\pi,T}$ denote stochastic processes on index set T such that

$$\forall t \in T, \qquad (X_T)_t := X_t \qquad (X_{\pi,T})_t := X_{\pi(t)}. \tag{3}$$

In other words, X_T is the restriction of X to index set T, while $X_{\pi,T}$ is the restriction to T under the mapping π. Definition 1 captures the notion that observations with similar covariates should be close to exchangeable, i.e., the total variation between X_T and $X_{\pi,T}$ is small as long as the distances between t and $\pi(t)$ are small for all $t \in T$.

Definition 1. The process X is *locally exchangeable* with respect to a premetric d if for any finite subset $T \subset \mathcal{T}$ and injection $\pi : T \to \mathcal{T}$,

$$d_{\mathrm{TV}}(X_T, X_{\pi,T}) \leq \sum_{t\in T} d(t,\pi(t)). \tag{4}$$

Definition 1 generalizes both exchangeability and partial exchangeability among equivalence classes. In particular, the zero premetric where $d(t,t') = 0$ identically yields classical exchangeability, while the premetric $d(t,t') = 1 - \mathbb{1}[t \sim t']$ for equivalence relation $\sim$ yields partial exchangeability. Further, any process is locally exchangeable with respect to the discrete premetric $d(t,t') = 1 - \mathbb{1}[t = t']$; in order to say something of value about a process X, it must satisfy Eq. (4) for a tighter premetric.

To quantify differences in distributions, Definition 1 employs the total variation distance, which for random elements Y, Z in a measurable space $(\mathcal{Y},\Xi)$ is defined as

$$d_{\mathrm{TV}}(Y,Z) := \sup_{A\in\Xi} |\mathbb{P}(Y \in A) - \mathbb{P}(Z \in A)| .$$

The choice of total variation distance (as opposed to other metrics and divergences, see e.g. (Gibbs and Su, 2002)) is motivated by its symmetry and generality. We make d a premetric—as opposed to a (pseudo)metric, say—as the triangle inequality and positive definiteness are unused in the theory

below. Further we use a premetric with range $[0,1]$ because total variation always lies in this range, and so any valid bound in Eq. (4) for a premetric $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ can be improved by replacing d with $\min(d,1)$. And although Definition 1 imposes a total variation bound only for all finite sets of covariates, it is equivalent to do so for all countable sets of covariates, as shown in Proposition 2.

Proposition 2. *If X is locally exchangeable with respect to d, then for any countable subset $T \subset \mathcal{T}$ and injection $\pi : T \to \mathcal{T}$,*

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

Example 3. A simple example of local exchangeability that we will return to throughout the paper is the process of observable measurements X from a Bayesian linear regression model on $\mathcal{T} = \mathbb{R}$ with a quadratic trend,

$$\theta \sim \mathcal{N}(0,1), \qquad \forall t \in \mathbb{R}, \quad X_t \overset{\text{indep}}{\sim} \mathcal{N}(\theta t^2, 1). \tag{5}$$

By Lemma 14 in the supplementary material (Campbell et al., 2023), since the X_t are independent conditioned on θ,

$$d_{\mathrm{TV}}(X_T, X_{\pi,T}) \leq \sum_{t \in T} \mathbb{E}\left[d_{\mathrm{TV}}(\mathcal{N}(\theta t^2, 1), \mathcal{N}(\theta \pi(t)^2, 1)) \right].$$

We bound the terms in the sum using the Lipschitz continuity of the standard normal CDF Φ,

$$\begin{aligned}\mathbb{E}\left[d_{\mathrm{TV}}(\mathcal{N}(\theta t^2, 1), \mathcal{N}(\theta \pi(t)^2, 1)) \right] &= \mathbb{E}\left[\Phi\left(\frac{|\theta t^2 - \theta \pi(t)^2|}{2} \right) - \Phi\left(-\frac{|\theta t^2 - \theta \pi(t)^2|}{2} \right) \right] \\ &\leq \frac{\mathbb{E}|\theta||t^2 - \pi(t)^2|}{\sqrt{2\pi}} \leq \frac{|t^2 - \pi(t)^2|}{\sqrt{2\pi}}.\end{aligned}$$

Therefore the process X in the Bayesian linear regression model Eq. (5) is locally exchangeable with respect to the premetric $d(t,t') = \min(|t^2 - t'^2|/\sqrt{2\pi}, 1)$. Note that we are free to take $\min(\cdot, 1)$ because the total variation is bounded above by 1. This example illustrates why we opt for the generality of a premetric; here, observations at points t and $-t$ are exactly exchangeable since $d(t,-t) = 0$, which does not generally hold for a metric, and $|t^2 - t'^2|$ does not satisfy the triangle inequality. Also note that the marginal distribution of X_T is a multivariate Gaussian with off-diagonal covariance terms $\mathbb{E}[X_t X_{t'}] \propto t^2 t'^2$, which varies with t, t'; multivariate Gaussians with exchangeable components must have constant off-diagonal covariance terms. Therefore this example also shows that there exist processes that are locally exchangeable but not exchangeable.

2.2. de Finetti representation

In the previous example, we used the fact that the variables X_t were conditionally independent given a latent random variable θ to demonstrate their local exchangeability. A natural question to ask is whether *all* locally exchangeable processes exhibit a similar structure. Theorem 5 answers this question in the affirmative, by providing a de Finetti-like representation of locally exchangeable processes similar to Eq. (1) and Eq. (2). This representation guarantees the existence of a simple conditional structure that can be leveraged in the design of statistical inference procedures, and justifies a Bayesian approach when dealing with covariate-dependent data. We first require a weak assumption on the space $\mathcal{T}$.

Definition 4 (Infinitely-separable space). A premetric space $(d,\mathcal{T})$ is *infinitely separable* if there exists a countable subset $\mathfrak{T} \subseteq \mathcal{T}$ such that for all $t \in \mathcal{T}$, there exists a Cauchy sequence $(t_n)_{n\in\mathbb{N}}$ in $\mathfrak{T}$ such that $t_n \to t$ and $|\{t_n : n \in \mathbb{N}\}| = \infty$.

When d is a metric, infinite separability is equivalent to $\mathcal{T}$ being separable with no isolated points. When d is a pseudometric, it is equivalent to the existence of a countable dense subset $\mathfrak{T} \subseteq \mathcal{T}$ such that for all $t \in \mathcal{T}$ and $\epsilon > 0$, $|\{t' \in \mathfrak{T} : d(t,t') < \epsilon\}| = \infty$. In general, infinite separability ensures that there are infinitely many elements to swap "nearby" each covariate value of interest $t \in \mathcal{T}$. This assumption precludes the situation where observations satisfy finite exchangeability (Diaconis, 1977, Diaconis and Freedman, 1980b) but not infinite exchangeability.

Theorem 5 shows that under infinite separability, the desired de Finetti-like representation indeed does exist. In particular, we show that there is a unique probability measure-valued process G that renders X conditionally independent, and that G satisfies a continuity property with the same "smoothness" as the observed process. For the precise statement of the result in Theorem 5, recall that a *modification* of a stochastic process G on $\mathcal{T}$ is any other process G' on $\mathcal{T}$ such that $\forall t \in \mathcal{T}, \mathbb{P}\left(G_t = G'_t\right) = 1$.

Theorem 5. *Suppose $(d,\mathcal{T})$ is infinitely separable. Then the process X is locally exchangeable with respect to d if and only if there exists a random measure-valued stochastic process $G = (G_t)_{t\in\mathcal{T}}$ (unique up to modification) such that for any finite subset of covariates $T \subset \mathcal{T}$ and $t,t' \in \mathcal{T}$,*

$$\mathbb{P}(X_T \in \cdot \,|\, G) \overset{a.s.}{=} \prod_{t\in T} G_t, \qquad \sup_A \mathbb{E}\,|G_t(A) - G_{t'}(A)| \le d(t,t'). \tag{6}$$

For example, given $\mathcal{T} = \mathbb{N}$ and the zero premetric $d(t,t') = 0$, one recovers the de Finetti representation of exchangeable sequences; the smoothness condition asserts that G_t must be constant for all $t \in \mathcal{T}$ as expected. Similarly, suppose we are given an equivalence relation $\sim$ on $\mathbb{N}$ where each equivalence class has infinite cardinality. Then setting $\mathcal{T} = \mathbb{N}$ and $d(t,t') = 1 - \mathbb{1}[t \sim t']$ recovers the de Finetti representation of partially exchangeable sequences under permutation within equivalence classes; here the smoothness condition asserts that G_t must be constant within each equivalence class, but allows for general dependence between G_t across the equivalence classes. Thus, in the same way that Definition 1 generalizes (partial) exchangeability, Theorem 5 generalizes the de Finetti representation theorem.

Note that we still obtain the "if" direction of Theorem 5 without imposing the infinite separability assumption on $(d,\mathcal{T})$. In particular, if we are given a process G satisfying Eq. (6), then the process X is locally exchangeable with respect to both

$$d_c(t,t') := \sup_A \mathbb{E}\,|G_t(A) - G_{t'}(A)|, \qquad \text{and} \qquad d_{sc}(t,t') := \mathbb{E}\left[d_{\mathrm{TV}}(G_t,G_{t'})\right].$$

We refer to d_c as the *canonical premetric* and d_{sc} as the *strong canonical premetric*. Note that X is locally exchangeable with respect any premetric d satisfying $d \ge d_c$, and in particular, $d_{sc} \ge d_c$. Given a particular G, one can use Lemma 14 in the supplementary material (Campbell et al., 2023) to derive an upper bound on these two premetrics (as demonstrated in Example 3), which then provides insight into the extent to which data X generated from G are exchangeable. Note that $(d_c,\mathcal{T})$ and $(d_{sc},\mathcal{T})$ may or may not be infinitely separable, depending on the characteristics of the process G.

Example (continued). In the linear regression example, the underlying measure-valued process is the collection of normal distributions

$$G_t = \mathcal{N}(\theta t^2, 1), \qquad t \in \mathcal{T}.$$

Theorem 5 guarantees that this process is unique up to modification. In this case, the randomness in G is entirely due to the latent variable $\theta \sim \mathcal{N}(0,1)$; in general G need not be determined by a finite-dimensional quantity. We can also verify that G satisfies the required smoothness condition with respect to d, although it is not surprising in this case given that we originally derived the premetric using the same technique:

$$\sup_A \mathbb{E}\,|G_t(A) - G_{t'}(A)| \le \mathbb{E}\, d_{\mathrm{TV}}(G_t, G_{t'}) \le \min\left(\frac{1}{\sqrt{2\pi}}|t^2 - t'^2|, 1\right) = d(t,t').$$

2.3. Local empirical measure process

The de Finetti result in Theorem 5 guarantees the existence of a unique underlying measure-valued process G, but does not provide any direct insight into the distribution of G or whether it is identifiable given only (countably many) measurements of the process X. In the classical setting of an exchangeable sequence $X_1, X_2, \ldots$, the empirical measure $\widehat{G}_N = \frac{1}{N}\sum_{n=1}^N \delta_{X_n}$ of a finite collection of observations $(X_n)_{n=1}^N$ serves this purpose, as it converges weakly to G almost surely (Varadarajan, 1958), i.e.,

$$d_{\mathrm{P}}(\widehat{G}_N, G) \overset{a.s.}{\to} 0, \qquad N \to \infty, \tag{7}$$

where d_{P} denotes the Lévy-Prokhorov metric. In the setting of local exchangeability more generally, however, the usual empirical measure does not provide a result similar to Eq. (7). If we are interested in understanding the distribution of G_τ for some $\tau \in \mathcal{T}$, and we collect measurements $(X_t)_{t\in T}$ of X at a finite set of covariates $T \subset \mathcal{T}$, the presence of far-away covariates in T from τ can result in a non-vanishing bias in the empirical measure. To address this issue, for each $\tau \in \mathcal{T}$, let $t_i(\tau)$, $i = 1, \ldots, |T|$ be an ordering of the set T such that the values $d_i(\tau) = d(t_i(\tau), \tau)$ are ordered from smallest to largest. Then define

$$M_\tau = \max\left\{M \in [|T|] : \frac{1}{M}\left(1 + \sum_{m=1}^{M} 2d_m(\tau)\right) > 2d_M(\tau)\right\}, \qquad \mu_\tau = \frac{1}{M_\tau}\sum_{m=1}^{M_\tau} d_m(\tau).$$

We construct the *local empirical measure process* $(\widehat{G}_\tau)_{\tau\in\mathcal{T}}$ via

$$\widehat{G}_\tau = \sum_{t\in T} \xi_t(\tau)\delta_{X_t}, \qquad \xi_t(\tau) = \max\left\{0, \frac{1}{M_\tau} + 2(\mu_\tau - d(t,\tau))\right\}.$$

The local empirical measure process $\widehat{G}$ serves as an approximation of the measure-valued process G underlying the locally exchangeable process X. Note that $\sum_{t\in T}\max\{0, \frac{1}{M_\tau} + 2(\mu_\tau - d(t,\tau))\} = 1$, so $\widehat{G}_\tau$ is a probability measure for each $\tau \in \mathcal{T}$. Further note that $(\widehat{G})_{\tau\in\mathcal{T}}$ is measurable with respect to $(X_t)_{t\in T}$. Intuitively, $\widehat{G}$ includes only those observations at covariates sufficiently close to the point of interest $\tau \in \mathcal{T}$ such that the decrease in variance associated with adding another observation outweighs the potential increase in bias. The value M_τ represents how many observations are included in the local empirical measure at that location, and μ_τ represents the average distance of their covariates to τ.

Our goal now is to provide a weak convergence result for the local empirical measure process $\widehat{G}$ in the limit of many observations, similar to that of Eq. (7). As a key step towards that goal, Theorem 6 provides bounds on both the expected squared estimation error (Eq. (8)) as well as error tail probabilities (Eq. (9)) when using the local empirical measure process $\widehat{G}_\tau$ in place of G_τ or $\mathbb{P}(X_\tau \in \cdot \mid X_T)$, for

all $\tau \in \mathcal{T}$. Each bound in Theorem 6 has two terms: the first is related to the variance incurred by estimation via independent sampling, and the second is related to the bias incurred by using observations from $t \neq \tau$. Note that Theorem 6 quantifies the approximation error using the metric

$$\|\nu - \eta\|_{\mathcal{A}} = \sum_{i=1}^{\infty} c_i \, |\nu(A_i) - \eta(A_i)|, \quad \nu, \eta \text{ probability measures},$$

where $\mathcal{A} = \{c_i, A_i\}_{i=1}^{\infty}$, A_i are measurable subsets of $\mathcal{X}$, $c_i \geq 0$, and $\sum_i c_i = 1$. We work with $\|\cdot\|_{\mathcal{A}}$ rather than standard metrics because it simplifies the analysis substantially. Although the properties of $\|\cdot\|_{\mathcal{A}}$ depend on the choice of $\mathcal{A}$ in general, there exists a choice such that $\|\cdot\|_{\mathcal{A}} \to 0$ implies weak convergence (see Lemma 16 in the supplementary material (Campbell et al., 2023)), and the bounds below in Theorem 6 are valid for any choice of $\mathcal{A}$, as indicated by the supremum. We will use the metric $\|\cdot\|_{\mathcal{A}}$ and the results in Theorem 6 as a stepping stone to obtain weak convergence in Corollary 7 below.

Theorem 6. *Let $(d, \mathcal{T})$ be infinitely separable and X be locally exchangeable with respect to d. Then*

$$\forall \tau \in \mathcal{T}, \quad \sup_{\mathcal{A}} \mathbb{E}\left[\|\widehat{G}_\tau - G_\tau\|_{\mathcal{A}}^2\right] \leq \frac{1}{4M_\tau} + \mu_\tau, \tag{8}$$

and for all $\delta > 0$, $\tau \in \mathcal{T}$,

$$\sup_{\mathcal{A}} \mathbb{P}\left(\|\widehat{G}_\tau - G_\tau\|_{\mathcal{A}} > \delta + \sqrt{2\mu_\tau + 1/M_\tau}\right) \leq \exp\left(\frac{-\delta^2}{2\,(2\mu_\tau + 1/M_\tau)}\right) + \frac{2\mu_\tau}{\delta + \sqrt{1/M_\tau}}. \tag{9}$$

Furthermore, the same bounds in Eqs. (8) and (9) apply when G_τ is replaced with $\mathbb{P}(X_\tau \in \cdot \mid X_T)$.

When all of the covariates in the observed set T are close to τ, the bounds in Theorem 6 provide essentially the same guarantees as one would expect for exchangeable random variables. In particular, suppose for all $t \in T$, $d(t, \tau) \lesssim \exp(-|T|)$, and so $\xi_t(\tau) \approx 1/|T|$. In this situation the bounds above reduce to

$$\sup_{\mathcal{A}} \mathbb{E}\left[\|\widehat{G}_\tau - G_\tau\|_{\mathcal{A}}^2\right] = O(|T|^{-1}), \quad \sup_{\mathcal{A}} \mathbb{P}\left(\|\widehat{G}_\tau - G_\tau\|_{\mathcal{A}} > \delta + |T|^{-1/2}\right) = O\left(e^{-|T|\delta^2}\right).$$

Corollary 7 uses the results in Theorem 6 to obtain a weak convergence result for $\widehat{G}_\tau$ similar to Eq. (7). In particular, if we collect measurements of X from a sequence of sets that concentrate around τ—for example, $T_n = \{t_i\}_{i=1}^n$ such that there exists a subsequence $t_{i_k} \to \tau$—then the local empirical measure $\widehat{G}_\tau$ converges weakly to both G_τ and the Bayesian posterior predictive distribution in probability. Recall that d_{P} denotes the Lévy-Prokhorov metric.

Corollary 7. *Fix $\tau \in \mathcal{T}$. Suppose we make observations at a sequence of finite sets $T_n \subset \mathcal{T}$, $n \in \mathbb{N}$ of covariates such that for all $\epsilon > 0$, $|\{t \in T_n : d(t, \tau) \leq \epsilon\}| \to \infty$. Then*

$$d_{\mathrm{P}}(\widehat{G}_\tau, G_\tau) \xrightarrow{p} 0 \quad \text{and} \quad d_{\mathrm{P}}(\widehat{G}_\tau, \mathbb{P}\left(X_\tau \in \cdot \mid X_{T_n}\right)) \xrightarrow{p} 0, \qquad n \to \infty.$$

A byproduct of Corollary 7 is that one can characterize the distribution of G_τ by analyzing the distribution of X_τ conditioned on X_{T_n} for a sequence of sets of covariates T_n that concentrate around τ, i.e., $|T_n| \to \infty$ and $\max\{d(t, \tau) : t \in T_n\} \to 0$ as $n \to \infty$. Note that it is not required to know the premetric

d governing local exchangeability in order to identify G using this technique; one can instead construct the set of covariates T_n such that $\max\{\ell(t,\tau) : t \in T_n\} \to 0$ for any premetric $\ell : \mathcal{T} \times \mathcal{T} \to [0,1]$ that *dominates* d in the sense that for any two sequences of covariates t_n, t_n', $n \in \mathbb{N}$,

$$\ell(t_n, t_n') \to 0 \implies d(t_n, t_n') \to 0, \quad n \to \infty. \tag{10}$$

The requirement in Eq. (10) is typically not stringent; it states only that when covariates get close under ℓ, they must also get close under d, with no other stipulation about relative rates, bounds, etc. In the following linear regression example, we will use the usual metric $\ell(t,t') = |t - t'|$ on $\mathbb{R}$.

Example (continued). We return to the linear regression example to show how the distribution of G_τ can be recovered from the process X via Corollary 7. The joint density of X_T, X_τ is

$$p(x_\tau, x_T) \propto \exp\left(-\frac{1}{2}x_\tau^2 - \frac{1}{2}\sum_{t\in T} x_t^2 + \frac{1}{2}\frac{\left(x_\tau \tau^2 + \sum_{t\in T} x_t t^2\right)^2}{1 + \tau^4 + \sum_{t\in T} t^4} \right).$$

Therefore the conditional distribution of X_τ given X_T is given by

$$X_\tau \sim \mathcal{N}\left(\frac{\tau^2 \sum_{t\in T} X_t t^2}{1 + \sum_{t\in T} t^4}, \frac{1 + \tau^4 + \sum_{t\in T} t^4}{1 + \sum_{t\in T} t^4} \right),$$

If we then consider a sequence of sets T_n of covariates that grows in size and concentrates quickly around τ—e.g., $T_n = \{\tau + i\exp(-n) : i = 1, \ldots, n\}$—we find that the conditional distribution of X_τ given X_T converges to

$$X_\tau \sim \mathcal{N}(Y, 1), \text{ where } Y \sim \mathcal{N}\left(0, \tau^4\right).$$

By setting $\theta = Y\tau^{-2}$, we recover the fact that X_τ is generated from $G_\tau = \mathcal{N}(\theta\tau^2, 1)$, $\theta \sim \mathcal{N}(0,1)$, i.e., the marginal of the original Bayesian linear regression model. Note that one can repeat essentially the same analysis for multiple covariates $\tau_1, \ldots, \tau_K$ to recover finite marginal distributions. For example, if we consider the bivariate distribution of G_{τ_1}, G_{τ_2}, we find that X_{τ_1}, X_{τ_2} are generated independently from

$$G_{\tau_1} = \mathcal{N}(\theta\tau_1^2, 1) \quad G_{\tau_2} = \mathcal{N}(\theta\tau_2^2, 1), \quad \theta \sim \mathcal{N}(0,1).$$

The analysis from the example in Section 2.1 can then be used to bound the strong canonical premetric $d_{sc}(t,t') = d_{\mathrm{TV}}(G_t, G_{t'}) \leq \min\left(|t - t'|/\sqrt{2\pi}, 1\right)$. Thus, given only the process X, we have identified a premetric d under which X is locally exchangeable as well as the measure-valued process G.

2.4. Regularity

The smoothness property of G in Eq. (6) may seem unsatisfying at a first glance; it bounds the absolute difference in the underlying mixing measure process at nearby locations only *in expectation*, leaving room for the possibility of sample discontinuities in G_t as a function of t. However, there are many probabilistic models that, intuitively, generate observations that should be considered locally exchangeable but which have discontinuous latent mixing measures. For example, some dynamic nonparametric mixture models (Chen et al., 2013, Lin and Fisher, 2010) have components that are created

and destroyed over time, causing discrete jumps in the mixing measure. As long as the jumps happen at diffuse random times, the probability of a jump occurring between two times decreases as the difference in time decreases, and the observations may still be locally exchangeable. However, intuitively, if there is a fixed location t_0 with a nonzero probability of a discrete jump in the mixing measure process, the observations X cannot be locally exchangeable. Corollary 8 provides the precise statement.

Corollary 8. *Suppose $(d, \mathcal{T})$ is infinitely separable and X is locally exchangeable with respect to d. Then for all $A \in \Sigma$, $t_0 \in \mathcal{T}$, and $\epsilon > 0$,*

$$\lim_{\eta \to 0} \sup_{t\,:\, d(t,t_0) \leq \eta} \mathbb{P}\left(|G_t(A) - G_{t_0}(A)| > \epsilon\right) = 0.$$

That being said, it is worth examining whether different guarantees on properties of the underlying measure process G result as a consequence of different properties of the premetric d. Theorem 9 answers this question in the affirmative for processes on $\mathcal{T} = \mathbb{R}$; in particular, the faster the decay of $d(t,t')$ relative to $|t - t'|$ as $t \to t'$, the stronger the guarantees on the behavior of the mixing measure G. Note that while this result is presented for covariate space $\mathbb{R}$, the result can be extended to processes on $\mathbb{R} \times \mathbb{N}$ and more general separable spaces (Potthoff, 2009, Theorems 2.8, 2.9, 4.5).

Theorem 9. *Let $\mathcal{T} = \mathbb{R}$, $\gamma \geq 0$, and X be locally exchangeable with respect to a premetric d satisfying $d(t,t') = O(|t - t'|^{1+\gamma})$ as $|t - t'| \to 0$. Then:*

1. *($\gamma > 1$): X is exchangeable and G is a constant process.*
2. *($0 < \gamma \leq 1$): X is stationary and for any $A \in \Sigma$ and $\alpha \in (0, \gamma)$, $(G_t(A))_{t \in \mathbb{R}}$ is weak-sense stationary with an α-Hölder continuous modification.*
3. *($\gamma = 0$): G may have no continuous modification.*

Remark. A rough converse of the first point holds: X exchangeable implies constant G, and $d(t,t') = 0$ is trivially $O(|t - t'|^{1+\gamma})$ for $\gamma > 1$. But a similar claim for the second point is not true in general: X stationary and locally exchangeable does not necessarily imply that $d(t,t') = O(|t - t'|^{1+\gamma})$ for $0 < \gamma \leq 1$. For a counterexample, consider a square wave shifted by a uniform random variable, i.e., the process $X_t = \text{sign}\,(\sin(2\pi(t - U)))$ for $U \sim \mathsf{Unif}[0,1]$. Here X_t is stationary and locally exchangeable with $d(t,t') = \min(|t - t'|, 1)$, but $|t - t'| \neq O(|t - t'|^{1+\gamma})$ for any $\gamma > 0$ as $|t - t'| \to 0$.

2.5. Approximate conditional independence

In the classical setting of exchangeable sequences $X_1, X_2, \ldots$, the empirical measure $\widehat{G} = \frac{1}{N} \sum_{n=1}^{N} \delta_{X_n}$ satisfies the following property: for all bounded measurable functions $h : \mathcal{X}^N \to \mathbb{R}$,

$$\mathbb{E}\left[h(X_1, \ldots, X_N) | \widehat{G}, G\right] = \mathbb{E}\left[h(X_1, \ldots, X_N) | \widehat{G}\right]. \tag{11}$$

Thus G and $(X_1, \ldots, X_N)$ are conditionally independent given $\widehat{G}$. In other words, the fact that $(X_1, \ldots, X_N)$ corresponds to covariate values $(1, \ldots, N)$ provides no additional information about G beyond $\widehat{G}$ itself.

In the setting of local exchangeability, the question of how important the covariate values are in inferring the measure-valued process G is relevant in practice: we do not often get to observe the true covariate values $\{t_1, \ldots, t_N\} = T \subset \mathcal{T}$, but rather we observe discretized versions that are grouped into "bins." For example, if X_T corresponds to observed document data with timestamps T, we may

know those timestamps up to only a certain precision (e.g. days, months, years). This section shows that a "binned" version of the empirical measure $\widehat{G}$ provides an approximate conditional independence similar to Eq. (11), where the error of approximation decays smoothly by an amount corresponding to the uncertainty in covariate values.

Formally, suppose we partition our covariate space $\mathcal{T}$ into disjoint bins $\{\mathcal{T}_k\}_{k=1}^{\infty}$, where each bin has observations $T_k = \mathcal{T}_k \cap T$. We may use a finite partition by setting all but finitely many $\mathcal{T}_k$ to the empty set. Although we know the number of points in each bin (i.e., the cardinality of T_k), we will encode our lack of knowledge of their positions as randomness: $T_k \sim \mu_k$, where μ_k is a probability distribution capturing our belief of how the unobserved covariates are generated within each bin. Following the intuition from the classical de Finetti's theorem, we define the binned empirical measures $\widetilde{G}_k = \sum_{t \in T_k} \delta_{X_t}$, $\widetilde{G} := (\widetilde{G}_1, \widetilde{G}_2, \dots)$, and let $\mathcal{G}$ denote the subgroup of permutations $\pi : T \to T$ that permute observations only within each bin, i.e., such that $\forall k \in \mathbb{N}, \pi(T_k) = T_k$. Note that $|\mathcal{G}| = \prod_{k=1}^{\infty} |T_k|! < \infty$ since there are only finitely many observations in total. Unlike classical exchangeability, $\widetilde{G}$ does not provide exact conditional independence of X_T and G; but Theorem 10 guarantees that it provides a form of approximate conditional independence, with error that depends on $(\mu_k)_{k=1}^{\infty}$.

Theorem 10. *Suppose $(d, \mathcal{T})$ is infinitely separable. If X is locally exchangeable with respect to d, and $h : \mathcal{X}^T \to \mathbb{R}$ is a bounded measurable function,*

$$\mathbb{E}\left|\mathbb{E}\left[h(X_T) \mid \widetilde{G}, G\right] - \mathbb{E}\left[h(X_T) \mid \widetilde{G}\right]\right| \leq 4\|h\|_{\infty}\mathbb{E}\left[\sum_{t \in T} d(t, \pi(t))\right],$$

where $\pi \sim \mathsf{Unif}(\mathcal{G})$ and $T_k \overset{indep}{\sim} \mu_k$.

Remark. Note that the expectation on the right hand side averages over the randomness both in the uncertain covariates T and the permutation π.

If X is exchangeable within each bin $\mathcal{T}_k$, Theorem 10 states that X_T and G are conditionally independent given $\widetilde{G}$, as desired. Further, the deviance from independence is controlled by the deviance from exchangeability within each bin. In particular,

$$\mathbb{E}\left[\sum_{t \in T} d(t, \pi(t))\right] \leq \sum_{k=1}^{\infty} |T_k| \operatorname{diam} \mathcal{T}_k \leq |T| \sup_k \{\operatorname{diam} \mathcal{T}_k\}, \tag{12}$$

where $\operatorname{diam} \mathcal{T}_k := \sup_{t,t' \in \mathcal{T}_k} d(t, t')$. Both bounds in Eq. (12) are independent of μ_k; thus the result holds even if we are unwilling to express our uncertainty in the binned covariates via a distribution.

3. Examples

In this section, we provide example applications of the theory in Section 2. First, we use a case study of Gaussian processes to show how one can use posterior predictive distributions to analyze the local exchangeability of a process. In particular, we show how to derive the underlying measure process G, as well as an appropriate premetric d governing local exchangeability, using only finite marginals of the process X. Second, we use a case study of dependent Dirichlet processes to show that one can use local empirical measures as a surrogate for otherwise intractable posterior predictive distributions in discrete Bayesian nonparametric models. See the supplementary material (Campbell et al., 2023) for

other examples of Bayesian nonparametric models exhibiting local exchangeability—e.g., kernel beta process feature models (Hjort, 1990, Ren et al., 2011) and dynamic topic models (Blei and Lafferty, 2006, Wang, Blei and Heckerman, 2008), among others. Finally, we demonstrate a usage of local exchangeability as a tool to analyze the inflation of type-I error in matched permutation tests involving covariates.

3.1. Obtaining the underlying measure-valued process and premetric

We will first provide an example of how one can use the Bayesian posterior predictive distributions of a locally exchangeable process X to derive the distribution of the underlying measure-valued process G as well as the premetric of local exchangeability d. This example applies the same strategy as in the running example from Section 2.3, albeit in a more sophisticated nonparametric model.

Consider a Gaussian process $X \sim \mathrm{GP}(m, \kappa)$ on $\mathcal{T} = \mathbb{R}^d$ with continuous mean function $m : \mathbb{R}^d \to \mathbb{R}$, and covariance function $\kappa(x, y) = \sigma^2(x)\mathbb{1}[x = y] + k(x, y)$ for continuous nonnegative $\sigma^2 : \mathbb{R}^d \to \mathbb{R}_+$ and continuous symmetric positive-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. Define a set of k unique covariate values $\tau_1, \ldots, \tau_k \in \mathcal{T}$, and consider the Euclidean metric on $\mathcal{T}$. For each $n \in \mathbb{N}$ and $i = 1, \ldots, k$, let T_{in} be a finite subset of covariates such that $|T_{in}| = n$ and $\max\{\|\tau_i - t\| : t \in T_{in}\} = o(1/n)$. Direct analysis of the conditional density yields that as $n \to \infty$, the conditional distribution of $X_{\tau_1}, \ldots, X_{\tau_k}$ given $X_{T_{1n}}, \ldots, X_{T_{kn}}$ converges to

$$(X_{\tau_1}, \ldots, X_{\tau_k}) \sim \mathcal{N}\left((Y_1, \ldots, Y_k), \mathrm{diag}\left(\sigma^2(\tau_1), \ldots, \sigma^2(\tau_k)\right)\right), \tag{13}$$

where

$$(Y_1, \ldots, Y_k) \sim \mathcal{N}\left((m(\tau_1), \ldots, m(\tau_k)), K\right), \qquad K_{ij} = k(\tau_i, \tau_j). \tag{14}$$

Eqs. (13) and (14) demonstrate that X is conditionally independently drawn from the process G where

$$\forall \tau \in \mathcal{T}, \quad G_\tau = \mathcal{N}(Y_\tau, \sigma^2(\tau)) \qquad Y \sim \mathrm{GP}(m, k).$$

We now derive the strong canonical premetric of local exchangeability. In this setting,

$$d_{sc}(t, t') = \mathbb{E}\left[d_{\mathrm{TV}}(G_t, G_{t'})\right] = \mathbb{E}\left[d_{\mathrm{TV}}(\mathcal{N}(Y_t, \sigma^2(t)), \mathcal{N}(Y_{t'}, \sigma^2(t')))\right].$$

By Devroye, Mehrabian and Reddad (2020, Theorem 1.3),

$$d_{\mathrm{TV}}(\mathcal{N}(Y_t, \sigma^2(t)), \mathcal{N}(Y_{t'}, \sigma^2(t'))) \leq \frac{3|\sigma^2(t) - \sigma^2(t')|}{2\max\{\sigma^2(t), \sigma^2(t')\}} + \frac{|Y_t - Y_{t'}|}{2\max\{\sigma(t), \sigma(t')\}}.$$

Applying Jensen's inequality $\mathbb{E}|Y_t - Y_{t'}| \leq \sqrt{\mathbb{E}(Y_t - Y_{t'})^2}$, then evaluating the expectation and using the bounds $|\sigma^2(t) - \sigma^2(t')| \leq 2\max\{\sigma(t), \sigma(t')\}|\sigma(t) - \sigma(t')|$, and $\sqrt{x^2 + y^2} \leq x + y$ yields

$$d_{sc}(t, t') \leq \min\left(1, \frac{6|\sigma(t) - \sigma(t')| + |m(t) - m(t')| + \sqrt{k(t,t) + k(t',t') - 2k(t,t')}}{2\max\{\sigma(t), \sigma(t')\}}\right). \tag{15}$$

In the usual setting with zero mean $m(t) = 0$, constant noise variance $\sigma(t) = \sigma$ for some $\sigma > 0$, and stationary kernel $k(t,t') = r(\|t - t'\|)$ for some $r : \mathbb{R}_+ \to \mathbb{R}_+$, Eq. (15) reduces to

$$d_{sc}(t,t') \leq \min\left(1, \frac{\sqrt{r(0) - r(\|t - t'\|)}}{\sigma}\right).$$

This example demonstrates that Gaussian processes are locally exchangeable in the presence of measurement noise, i.e. where $\sigma(t) > 0$. However, note that $\sigma(t) > 0$ is not strictly necessary for local exchangeability; to obtain a necessary and sufficient characterization of local exchangeability in Gaussian processes, we could instead analyze the canonical metric d_c per Theorem 5.

3.2. Approximate predictive distributions in discrete Bayesian nonparametrics

Next, we demonstrate that the local empirical measure can serve as a useful surrogate for otherwise intractable posterior predictive distributions in discrete Bayesian nonparametric models. The Dirichlet process (Ferguson, 1973) is a popular prior for the weights and component parameters in nonparametric mixture models. Draws from a Dirichlet process are discrete probability measures,

$$G = \sum_{k=1}^{\infty} w_k \delta_{\theta_k},$$

where $(w_k)_{k=1}^{\infty}$ are weights satisfying $w_k \geq 0$, $\sum_k w_k = 1$, and $(\theta_k)_{k=1}^{\infty}$ are component parameters, each with distribution given by (Sethuraman, 1994)

$$\theta_k \overset{\text{i.i.d.}}{\sim} H, \qquad v_k \overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(1,\alpha), \qquad w_k = v_k \prod_{i=1}^{k-1}(1 - v_i), \qquad k \in \mathbb{N},$$

for some distribution H and concentration parameter $\alpha > 0$. Given draws $X_n \overset{\text{i.i.d.}}{\sim} G$, the posterior predictive distribution of X_{N+1} given the first N draws $X_1, \ldots, X_N$ is

$$X_{N+1} \sim \frac{\alpha}{\alpha + N} H + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{X_n} = \frac{\alpha}{\alpha + N} H + \frac{N}{\alpha + N} \widehat{G}. \tag{16}$$

The fact that one can marginalize the (infinitely many) weights and parameters to arrive at Eq. (16) is critical in tractable computational inference for models involving the Dirichlet process (Neal, 2000).

When the observations come with additional covariate information, the *dependent* Dirichlet process mixture model (MacEachern, 1999, 2000) may be used instead. There are many instantiations of the dependent Dirichlet process; for simplicity we consider a model where the weights are a function of a covariate but the component parameters are constant across covariate values, i.e.,

$$X_{x,n} \overset{\text{indep}}{\sim} \sum_{k=1}^{\infty} w_{x,k} \delta_{\theta_k}, \qquad n \in \mathbb{N},\ x \in \mathbb{R},$$

where $w_{x,k} = v_{x,k} \prod_{i=1}^{k-1}(1 - v_{x,i})$, and the stick variables $v_{x,k}$ are now i.i.d. stochastic processes on $\mathbb{R}$. The marginal distributions of $v_{x,k}$ at $x \in \mathbb{R}$ are designed to be $\mathsf{Beta}(1,\alpha)$ so that the dependent Dirichlet

process is marginally a Dirichlet process for each covariate value. But even for simple stochastic processes $v_{x,k}$, the posterior predictive distribution is not tractable to obtain in closed-form. However, we can note that the process X is locally exchangeable with strong canonical premetric

$$d_{sc}(t,t') = \mathbb{E}\left[d_{\mathrm{TV}}\left(\sum_{k=1}^{\infty} w_{x,k}\delta_{\theta_k}, \sum_{k=1}^{\infty} w_{x',k}\delta_{\theta_k}\right)\right] = \frac{1}{2}\sum_{k=1}^{\infty}\mathbb{E}\left|w_{x,k} - w_{x',k}\right|,$$

where $t = (x,n)$ and $t' = (x',n')$. Since $w_{x,k}$ is a product of independent variables, Lemma 13 in the supplementary material (Campbell et al., 2023) yields

$$d_{sc}(t,t') \leq \frac{1}{2}\mathbb{E}\left[\left|v_{x,1} - v_{x',1}\right|\right]\sum_{k=1}^{\infty}\left(\left(\frac{\alpha}{\alpha+1}\right)^{k-1} + \frac{k-1}{1+\alpha}\left(\frac{\alpha}{\alpha+1}\right)^{k-2}\right).$$

The infinite sum converges to some $0 < C < \infty$, and so

$$d_{sc}(t,t') \leq \min\left(1, C\mathbb{E}\left|v_{x,1} - v_{x',1}\right|\right).$$

Therefore, as long as the stochastic process $v_{x,1}$ is smooth enough, and we condition on X_T, where T concentrates closely around $\tau \in \mathcal{T}$, the posterior predictive distribution of X_τ given X_T is approximately equal to the local empirical measure $\widehat{G}_\tau$, by Theorem 6; the latter has a tractable closed-form expression.

3.3. Type-I error inflation in grouped permutation tests

One of the key applications of exchangeability in statistical data analysis is in the design of nonparametric *permutation tests* with exact type-I error bounds (Pitman, 1937a,b,c; Fisher, 1966, Ch. 3). In the notation of this work, we are given observations of a stochastic process X at a finite set of covariates $T \subset \mathcal{T}$, a subgroup of $\mathcal{G}$ permutations $\pi : T \to T$, and a test statistic $S : \mathcal{X}^T \to \mathbb{R}$. The null hypothesis is that X_T is exchangeable; so we set a desired threshold $\alpha \in [0,1]$, and reject the null with type-I error at most α if

$$\frac{1}{|\mathcal{G}|}\sum_{\pi\in\mathcal{G}}\mathbb{1}\left[S(X_T) \leq S(X_{\pi,T})\right] \leq \alpha,$$

where $X_{\pi,T}$ is defined as in Eq. (3). This setup is commonly used in observational studies with a control group and treatment group, where $\mathcal{G}$ consists of permutations that swap *matched pairs* of elements in the control and treatment groups. However, a typical problem is that elements in the two groups are not exactly comparable due to the presence of covariates. In this case, a standard approach is to construct $\mathcal{G}$ to permute only those elements with similar covariates from the control and treatment groups, under some metric d (Baiocchi et al., 2010, Cochran, 1965, Greevy et al., 2004, Hansen, 2004, Hansen and Klopfer, 2006, Lu and Rosenbaum, 2004, Lu et al., 2011, Rosenbaum, 1989, 2002, Rubin, 1973b,a). Local exchangeability provides a general way to analyze the type-I error of these methods; Proposition 11 shows that for a locally exchangeable process, the type-I error α may potentially be increased by the average distance between pairs of covariates permuted by $\pi \in \mathcal{G}$. Eq. (17) also incidentally provides a rigorous justification for past work that formulates the construction of $\mathcal{G}$ as the minimization of this penalty (e.g., Rosenbaum (1989)).

Proposition 11. *Let X be locally exchangeable with respect to d. For $\alpha \in [0,1]$,*

$$\mathbb{P}\left(\frac{1}{|\mathcal{G}|} \sum_{\pi \in \mathcal{G}} \mathbb{1}\left[S(X_T) \leq S(X_{\pi,T}) \right] \leq \alpha \right) \leq \alpha + \frac{1}{|\mathcal{G}|} \sum_{\pi \in \mathcal{G}} \sum_{t \in T} d(t, \pi(t)). \tag{17}$$

4. Discussion

The major question posed in this paper is what we can do with data when we do not believe that they are exchangeable, but are willing to believe that they are *nearly* exchangeable. This paper answers the question with a relaxed notion of *local* exchangeability in which swapping data associated with nearby covariates causes a bounded change in total variation distance. We have demonstrated that classical results for exchangeable processes are "robust to the real world;" indeed, locally exchangeable processes have a de Finetti representation that may be leveraged in the design of statistical models and inference procedures. Finally, many popular covariate-dependent statistical models—which violate the assumptions of exchangeability—satisfy local exchangeability, extending the reach of exchangeability-based analyses to these models.

One limitation of local exchangeability is the infinite separability assumption. There are applications in which the covariate space $\mathcal{T}$ has isolated points that violate this condition, e.g., discrete time series where the covariate space is $\mathcal{T} = \mathbb{N}$ endowed with the Euclidean metric. However, if X can be extended to a process on $\mathcal{S} \supseteq \mathcal{T}$ such that $(d, \mathcal{S})$ is infinitely separable and $(X_s)_{s \in \mathcal{S}}$ is locally exchangeable with respect to d, then the theoretical results from this work hold for the marginal process $(X_t)_{t \in \mathcal{T}}$. Another limitation is that the total variation bound in the definition of local exchangeability is quite weak, which has downstream consequences for the tightness of the error bounds in Section 2.3. Further study on alternate definitions of local exchangeability is warranted to strengthen these guarantees.

As a final note, it is also possible that an analogue of the theory of finite exchangeability (Diaconis and Freedman, 1980b) holds in the local setting; but it is not yet clear whether this is indeed true or what form it would take. It would also be of interest to investigate more general notions of local exchangeability under group actions, e.g., permutations that preserve some statistic of the data, which have been used in past work on randomization testing in the presence of covariates (Rosenbaum, 1984).

Acknowledgements

The authors thank Jonathan Huggins for illuminating discussions.

Funding

T. Campbell is supported by a National Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and Discovery Launch Supplement. T. Broderick is supported in part by an NSF CAREER Award, an ARO YIP Award, ONR, and a Sloan Research Fellowship.

Supplementary Material

Supplement to "Local exchangeability" (DOI: 10.3150/22-BEJ1533SUPP; .pdf). Proofs for all results developed in this paper, with some additional examples from Bayesian nonparametrics.

References

Aldous, D.J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

Aldous, D.J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—1983*. *Lecture Notes in Math.* **1117** 1–198. Berlin: Springer. MR0883646 https://doi.org/10.1007/BFb0099421

Austin, T. and Panchenko, D. (2014). A hierarchical version of the de Finetti and Aldous-Hoover representations. *Probab. Theory Related Fields* **159** 809–823. MR3230009 https://doi.org/10.1007/s00440-013-0521-0

Baiocchi, M., Small, D.S., Lorch, S. and Rosenbaum, P.R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550 https://doi.org/10.1198/jasa.2010.ap09490

Berti, P., Pratelli, L. and Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Probab.* **32** 2029–2052. MR2073184 https://doi.org/10.1214/009117904000000676

Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *International Conference on Machine Learning*.

Borgs, C., Chayes, J.T., Cohn, H. and Holden, N. (2017). Sparse exchangeable graphs and their limits via graphon processes. *J. Mach. Learn. Res.* **18** Paper No. 210. MR3827098

Broderick, T., Pitman, J. and Jordan, M.I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Anal.* **8** 801–836. MR3150470 https://doi.org/10.1214/13-BA823

Cai, D., Campbell, T. and Broderick, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*.

Camerlenghi, F., Lijoi, A., Orbanz, P. and Prünster, I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92. MR3909927 https://doi.org/10.1214/17-AOS1678

Campbell, T., Cai, D. and Broderick, T. (2018). Exchangeable trait allocations. *Electron. J. Stat.* **12** 2290–2322. MR3832093 https://doi.org/10.1214/18-EJS1455

Campbell, T., Syed, S., Yang, C.-Y., Jordan, M.I., Broderick, T. (2023). Supplement to "Local Exchangeability." https://doi.org/10.3150/22-BEJ1533SUPP

Caron, F. and Fox, E.B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

Chen, C., Rao, V., Buntine, W. and Teh, Y. (2013). Dependent normalized random measures. In *International Conference on Machine Learning*.

Cochran, W.G. (1965). The planning of observational studies of human populations. *J. R. Stat. Soc., A* **128** 234–266.

Crane, H. and Dempsey, W. (2016). Edge exchangeable models for network data. Available at arXiv:1603.04571v3.

Crane, H. and Dempsey, W. (2019). Relational exchangeability. *J. Appl. Probab.* **56** 192–208. MR3981153 https://doi.org/10.1017/jpr.2019.13

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6.* **4** 251–299. In Italian.

de Finetti, B. (1937). La prévision : Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. MR1508036

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualites Scientifiques et Industrielles* **739**. 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.

Devroye, L., Mehrabian, A. and Reddad, T. (2020). The total variation distance between high-dimensional Gaussians. Available at arXiv:1810.08693.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability: Foundations of probability and statistics, II. *Synthese* **36** 271–281. MR0517222 https://doi.org/10.1007/BF00486116

Diaconis, P. (1988). Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics, 3 (Valencia, 1987)*. *Oxford Sci. Publ.* 111–125. New York: Oxford Univ. Press. MR1008047

Diaconis, P. and Freedman, D. (1978). de Finetti's generalizations of exchangeability. Technical Report No. 109, Univ. California, Berkeley.

Diaconis, P. and Freedman, D. (1980b). Finite exchangeable sequences. *Ann. Probab.* **8** 745–764. MR0577313

Diaconis, P. and Freedman, D. (1980c). de Finetti's theorem for Markov chains. *Ann. Probab.* **8** 115–130. MR0556418

Ernst, M.D. (2004). Permutation methods: A basis for exact inference. *Statist. Sci.* **19** 676–685. MR2185589 https://doi.org/10.1214/088342304000000396

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

Fisher, R. (1966). *The Design of Experiments*, 8th ed. Edinburgh: Oliver & Boyd.

Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.

Greevy, R., Lu, B., Silber, J. and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5** 263–275.

Hansen, B.B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. MR2086387 https://doi.org/10.1198/016214504000000647

Hansen, B.B. and Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

Hewitt, E. and Savage, L.J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80** 470–501. MR0076206 https://doi.org/10.2307/1992999

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294. MR1062708 https://doi.org/10.1214/aos/1176347749

Hoover, D. (1979). Relations on probability spaces and arrays of random variables. Technical Report, Institute for Advanced Study, Princeton Univ.

Janson, S. (2018). On edge exchangeable random graphs. *J. Stat. Phys.* **173** 448–484. MR3876897 https://doi.org/10.1007/s10955-017-1832-9

Jordan, M. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications.

Jung, P., Lee, J., Staton, S. and Yang, H. (2021). A generalization of hierarchical exchangeability on trees to directed acyclic graphs. *Ann. Henri Lebesgue* **4** 325–368. MR4213163 https://doi.org/10.5802/ahl.74

Kallenberg, O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *Ann. Probab.* **16** 508–534. MR0929061

Kallenberg, O. (1990). Exchangeable random measures in the plane. *J. Theoret. Probab.* **3** 81–136. MR1031426 https://doi.org/10.1007/BF01063330

Kallenberg, O. (2002). *Foundations of Modern Probability*, 2nd ed. *Probability and Its Applications (New York)*. New York: Springer. MR1876169 https://doi.org/10.1007/978-1-4757-4015-8

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. *Probability and Its Applications (New York)*. New York: Springer. MR2161313

Kingman, J.F.C. (1978). The representation of partition structures. *J. Lond. Math. Soc. (2)* **18** 374–380. MR0509954 https://doi.org/10.1112/jlms/s2-18.2.374

Lauritzen, S. (1974). On the interrelationships among sufficiency, total sufficiency, and some related concepts. Technical Report, Institute of Mathematical Statistics, Univ. Copenhagen.

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. New York: Springer. MR2135927

Lin, D. and Fisher, J. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems*.

Lu, B. and Rosenbaum, P.R. (2004). Optimal pair matching with two control groups. *J. Comput. Graph. Statist.* **13** 422–434. MR2063993 https://doi.org/10.1198/1061860043470

Lu, B., Greevy, R., Xu, X. and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist.* **65** 21–30. MR2899649 https://doi.org/10.1198/tast.2011.08294

MacEachern, S. (1999). Dependent nonparametric processes. Technical Report, The Ohio State Univ.

MacEachern, S. (2000). Dependent Dirichlet processes. Technical Report, The Ohio State Univ.

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

Orbanz, P. and Roy, D. (2015). Bayesian models of graphs, arrays, and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.

Pitman, E. (1937a). Significance tests which may be applied to samples from any populations. *Suppl. J. R. Stat. Soc.* **4** 119–130.

Pitman, E. (1937b). Significance tests which may be applied to samples from any populations II: The correlation coefficient test. *Suppl. J. R. Stat. Soc.* **4** 225–232.
Pitman, E. (1937c). Significance tests which may be applied to samples from any populations III: The analysis of variance test. *Biometrika* **29** 322–335.
Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158. MR1337249 https://doi.org/10.1007/BF01213386
Potthoff, J. (2009). Sample properties of random fields. II. Continuity. *Commun. Stoch. Anal.* **3** 331–348. MR2604006 https://doi.org/10.31390/cosa.3.3.02
Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. MR2514435
Ren, L., Wang, Y., Dunson, D. and Carin, L. (2011). The kernel beta process. In *Advances in Neural Information Processing Systems*.
Rosenbaum, P.R. (1984). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574. MR0763575
Rosenbaum, P. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.
Rosenbaum, P.R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487 https://doi.org/10.1214/ss/1042727942
Rubin, D. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.
Rubin, D. (1973a). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.
Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433
Varadarajan, V.S. (1958). On the convergence of sample probability distributions. *Sankhyā* **19** 23–26. MR0094839
Veitch, V. and Roy, D. (2015). The class of random graphs arising from exchangeable random measures. Available at arXiv:1512.03099.
Wang, C., Blei, D. and Heckerman, D. (2008). Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence*.

Received May 2021 and revised July 2022